Proceedings of the
Midwest Instruction and Computing Symposium 2023

(MICS 2023)

March 31 to April 1, 2023

University of Northern Iowa
Cedar Falls, Iowa 50614

MICS 2023 Program

Technical Session 1: 1:30 – 2:30 Friday March 31

Deep Learning: Sabin Hall Room 2		**Session Chair: Elliott Forbes**
1:30	Regenerating Audio Data from Silent Video through Deep Neural Networks	Konrad Rozpadek, Adam Haile, Samir Mahmud and Alexander Neuwirth
2:00	Error-Correcting Music Transformers	Jonathan Keane, Josiah Yoder and Michael Conner

Security: Wright Hall Room 9		**Session Chair: Erich Rice**
1:30	Can Hackers Cash-in On the Sensitive Data Contained in Cache?	Erich Rice, Dennis Guster and Li Dai
2:00	Discovering Vulnerabilities in Web Browser Extensions Contained by Google Chrome	Chapin Johnson, Sharveen Paramiswaran and Akalanka Mailewa

Cloud Computing: Wright Hall Room 10		**Session Chair: Akhtar Hussain**
1:30	Survey on Security and Privacy of Cloud Computing Paradigm: Challenges and Mitigation Methods	Akhtar Hussain, Jun Liu and Eunjin Kim
2:00	Survey on Security and Privacy Issues in Cloud-based Big Data Applications	Vedant Kharche and Jun Liu

CS Education: Wright Hall Room 105		**Session Chair: Tim Krause**
1:30	Interactive Mood Boards to Teach User Experience (UX) Principles as Part of an Agile Methodology	Tim Krause
2:00	Tutorial on TensorFlow Spark for BCI Augmented Robotics	Adriano Cavalcanti

Technical Session 2: 3:00 – 4:00 Friday March 31

Deep Learning: Sabin Hall Room 2		**Session Chair: Joshua Grant**
3:00	Transforming MoonBoard Climbing Route Classification and Generation	Joshua Grant, Michael Kirkton, Aiden Miller, Aydin Ruppe, Benjamin Weber and Ryan Kruk
3:30	Separating Spaces in Relative Attention for Music Generation	Michael Conner, Josiah Yoder and Jonathan Keane

Security: Wright Hall Room 9		**Session Chair: Anushka Hewarathna**
3:00	Encryption Methods and Key Management Services for Secure Cloud Computing: Review	Tristan Moore, Samuel Conlon, Anushka Hewarathna, Thivanka Dissanayaka M and Akalanka Mailewa
3:30	Darknet Traffic Classification using Deep Learning	Quinn Sullivan and Muhammad Abusaqer

Image Processing: Wright Hall Room 10		**Session Chair: Brendan Betterman**
3:00	Imaging Using 2.4GHz	Brendan Betterman, Richard Anderson and Baozhong Tian
3:30	Monocular Vision and Sensor Coupling for Indoor Localization	Houlin Chen, Lu Liang and Lei Wang

CS Education: Wright Hall Room 105		**Session Chair: Mark Fienup**
3:00	Computing for Data Science Course	Mark Fienup
3:30	Catapult Launch for Python Data Science Libraries	Leon Tabak

Robotics Contest and Pizza Party

Sponsored by:

Digi-Key® ELECTRONICS

Technical Session 3: 9:00 – 10:00 Saturday April 1

Deep Learning: Sabin Hall Room 2		**Session Chair: Muhammad Abusaqer**
9:00	Cyberbullying Classification Using Three Deep Learning models: GPT, BERT, and RoBERTa	Muhammad Abusaqer and Charles Fofie Jr
9:30	Automated Categorization of Cybersecurity News Articles through State-of-the-Art Text Transfer Deep Learning Models	Nathan Scott, Jt Snow and Muhammad Abusaqer

Security: Wright Hall Room 9		**Session Chair: Juliana Nkafu**
9:00	Survey of Application of Machine Learning Methods in the Development of Network Intrusion Detection and Prevention Systems.	Juliana Nkafu and Jun Liu

Societal Impact of CS: Wright Hall Room 10		**Session Chair: Roger Massmann**
9:00	Quantum Computing: An Assessment into the Impacts of Post-Quantum Cryptography	Roger Massmann, Nick Grantham and Akalanka Mailewa
9:30	Automation in the Food Service Industry, and It's Wide Reaching Effects	Sieger Canney

CS Education: Wright Hall Room 109		**Session Chair: Jim Seliya**
9:00	Investigating Curiosity in Student Text Data	Paul Meisner, Mitchell Hanson, Naeem Seliya, Benjamin Fine, Rushit Dave and Mounika Vanamala
9:30	Practical studying and conscious lifestyle	Thao Huy Vu and Asaad Saad

Technical Session 4: 10:30 – 11:00 Saturday April 1

Deep Learning: Sabin Hall Room 2		**Session Chair: Autumn Beyer**
10:30	Relative Attention For Video Frame Generation Tasks	Autumn Beyer, Mitchell Johnstone, Sam Keyser, Ryan Kruk, Tillie Pasternak, Tyler Schreiber and Michael Conner

User-Interface Testing: Wright Hall Room 9		**Session Chair: Ariana Beeby**
10:30	Constructing a UX Testing Platform using Embedded Computing Systems	Ariana Beeby and Erik Steinmetz

Image Processing: Wright Hall Room 10		**Session Chair: Sydney Balboni**
10:30	XprospeCT: CT Volume Generation from Paired X-rays	Sydney Balboni, Natalia Bukowski, John Cisler, Andrew Crisler, Joshua Goldshteyn, Julia Kalish, Ben Paulson and Theodore Colwell

Microarchitecture GUI Tool: Wright Hall Room 109		**Session Chair: Adam Grunwald**
10:30	dptv: A new pipetrace viewer for microarchitectural analysis	Adam Grunwald, Phuong Nguyen and Elliott Forbes

Keynote Speaker

Dheryta Jaisinghani is an Assistant Professor in the Department of Computer Science at University of Northern Iowa since August 2020. Her research lab – SyNthesIs (Systems for Next generation of Intelligent networkS) at UNI aims to develop user-friendly and cost-effective systems for smart buildings (offices and classrooms), mobile applications to solve student health challenges at the university, and algorithms to improve the performance of operational WiFi networks.

Regenerating Audio Data from Silent Video through Deep Neural Networks

Adam Haile   Konrad Rozpadek   Samir Mahmud   Alexander Neuwirth

Department of Electrical Engineering and Computer Science
Milwaukee School of Engineering
Milwaukee, WI, 53110
{hailea, rozpadekk, mahmuds, neuwirtha} @msoe.edu

Midwest Instruction and Computing Symposium, 2023

Abstract

Abundant internet video data with high-quality paired audio presents an opportunity to train a model capable of synthesizing new audio for silent video clips. Our work targets a first pass solution to this problem based on deep convolutional neural networks by encoding individual video frames and generating a relevant audio clip. The model is trained using the Youtube-8M dataset. We train a video classification model for frame context and use it to condition a custom WaveNet model trained for video audio generation. Finally, we propose a full audio generation pipeline using these techniques, and discuss the capabilities, limitations and further directions for this approach.

Introduction

The field of audio generation was established relatively recently. The relatively recent development of this research community likely stems from the inherent challenges of working with audio and the wide variety of signals that audio can represent - from speech to music to everyday ambient noises. Some of the more recent breakthroughs come mostly through attempts to generate musical audio. Developing an algorithm capable of writing music has been much more deeply explored and is generally much easier to model as it can be easily represented as a discrete set of musical tones and rhythms (e.g., sheet music or the digital MIDI format) rather than an audio waveform. To enable generation of arbitrary sounds, not just music, our work builds on the smaller body of work around arbitrary waveform generation: producing audio through spectrograms of the output waveform.

Video has become an ever more present form of media in the past few decades; Although not all video has the same form of presentation. Some videos might have a focus on action, change, and vision. On the other hand, some videos have a focus on verbal discussions, text, and other ways of expressing words. The central problem that we addressed has to do with the former, the focus on visuals. Not all videos have audio, or if they do have audio, it might not be very relevant to what is happening on screen. By addressing this problem through the generation of novel, relevant audio it is possible to enhance a video to better convey the ideas expressed to the viewer. Furthermore, it may also enable those with visual impairments to better sense context about the video, increasing accessibility.

Combining these two modalities poses a particularly interesting challenge. Some of the most recent research has come from Vladimir Iashin and others [1], in their research of utilizing transformers to generate a spectrogram from video. For the purpose of our research, we utilize a CNN encoder conditioning a WaveNet style generation approach in order to gain a better initial understanding of how audio can be generated with artificial intelligence, but we do have plans to develop an end-to-end transformer-based encoder/decoder in the future.

Dataset

The dataset that was chosen to train the model was the Youtube-8m dataset. The Youtube-8M dataset consists of video identifiers of YouTube videos that have been assigned labels to fit various common categories and ideas that are expressed in the labeled videos. These videos have been filtered to require at least 1000 views, a length between 120-500 seconds, and a clear display of one of the defined labels in the Youtube-8M label vocabulary. These requirements ensured that the videos that could be used to train the model had a significant amount of accurate audio for the idea that was being expressed.

The Youtube-8M dataset consists of over 6.1 million videos that have been labeled. This resulted in most labels having a large list of videos that can be used to train. In addition to that, Youtube-8M has a vocabulary of 3862 labels which allows the model to be generalized across a wide variety of topics and styles [2].

Video Input Preprocessing

All input video frames are preprocessed before being used as data in the categorizing pipeline. The frames are resized into a 224 by 224 square in order to make sure all input data from any video resolution is normalized. In addition to that, the frames are converted to grayscale to simplify the input. Videos used to train the audio synthesizer pipeline are fetched and have all but their audio tracks removed. The audio is then all converted into the same .wav audio format to be used to train the model.

Processed Image Classification

Each processed frame of video input is categorized via a convolutional neural network, this allows an effective and fast system for scanning our data. We use a 2D convolution to learn filters for our data with a collection of 3x3 kernels, following a flatten in order to reach a densely connected output layer. This is where the convolutional neural network classifies the output into a category to condition the audio generator. As a per-frame operation, this system decides the target audio type for the audio generation process, selecting the dataset the audio generation model pulls from.

WaveNet

For human observers, being able to look at a silent video and make inferences on the sound is often intuitive. By examining the different characteristics of features in a video, one can make grounded guesses as to what sounds they would make from countless examples in learned experience. WaveNet follows a very similar process. WaveNet is a type of generative model, designed by DeepMind, which is able to generate raw audio data. They apply this architecture to realistic text-to-speech generation. However, this is not where the capabilities of WaveNet end. WaveNet has the ability to generate audio on any type of data it is trained on by predicting new samples of audio based upon prior samples. It does this by taking a context window of the previous audio samples at each time step and passing it through a series of dilated convolutional layers. Each layer of convolution has a dilation factor which controls the size of the receptive field. Dilated convolutions were chosen because they have been successfully applied previously in signal processing and image segmentation and provide significant memory size optimizations when applied to very large inputs, such as long audio waveforms [3].

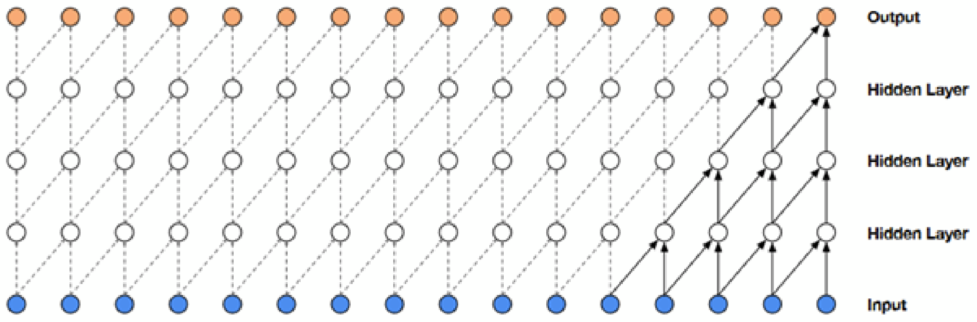


Figure 1: Visualization of a dilated convolutional later [1]

For the scope of our work, we utilized a preexisting implementation of WaveNet for TensorFlow, built by Kotaro Onishi [4], with additional reference to another implementation that was built on a different framework version, built by Igor Babuschkin and others [5].

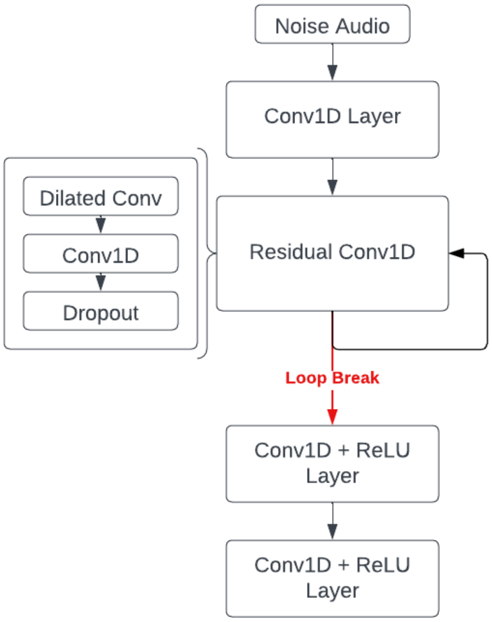


Figure 2: Flowchart of custom WaveNet implementation

Beyond simply extending the style and content found in its training data, WaveNet also enables fine-grained control of its output by operating on two types of conditioning: global and local conditioning. Within the context of our research, apply global conditioning to WaveNet. Global conditioning causes WaveNet to generate with reference to a single, fixed-length vector which possesses all of the style/category data of the audio sample. Local conditioning, which we do not apply in our first-pass implementation, allows for multiple inputs to be added into the sequence which, in our context, would allow for us to add the information of new frames to each time step of the audio. We do not apply local conditioning to this model due to the toolset limitations. Currently, the best public TensorFlow implementation of WaveNet does not include capabilities to generate audio for local conditioning. Implementing such a feature from scratch would require a much deeper analysis into WaveNet's architecture. It is, however, a logical future step for this project, as it would allow for a much closer generation alignment to the source video, rather than focusing the generation with only data from the first frame.

Pipeline

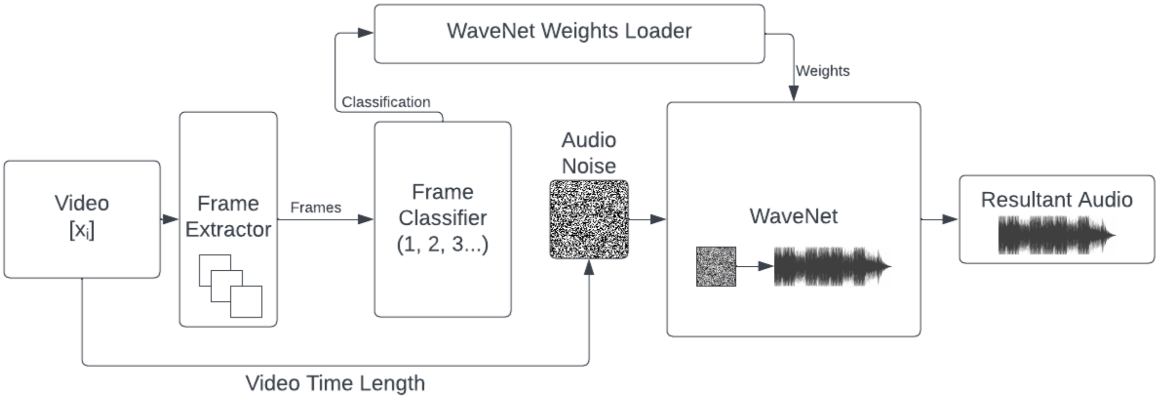


Figure 3: Visualization of implemented pipeline

The pipeline implements two key stages:
1. Categorizing a frame into one of the selected categories that WaveNet was trained on, and
2. Generating audio representative of the video frame’s category.

Categorization

There are 3 steps that go into categorization. The first step is the extraction and preprocessing of frames from a video. These frames are then fed into the frame classifier neural network that outputs a resulting category. Frames from various videos in selected categories are used to train this classifier. This category is then matched with the associated data that should be fed into WaveNet as a global conditioning parameter.

Audio Generation

First, the audio extracted from our training videos is used to train WaveNet along with the relevance weights for the video that was being represented. During the generation of new audio, the output from the categorization step is set as the global conditioning parameter. The length of the video is also fed as a parameter in order to generate an audio track of the same length that consists of random noise. This noise is fed as the input into WaveNet, and the resulting conditioned output is the generated audio output of the pipeline.

Limitation

This pipeline limits how specific the generated audio can be to a video. The implementation is limited to one category, from one frame, to be fed into WaveNet and is unable to synthesize multiple ideas into the generated audio. A more expanded implementation could apply local conditioning in order to have multiple frames feed multiple categories or semantic information unique to specific subjects in the frames as additional conditioning into WaveNet. Furthermore, an expanded implementation could generate multiple segments of audio tracks, then stitch them in order to better reflect rapid changes in visuals.

Challenges

During the process of building this pipeline, we encountered a variety of challenges, many of which we overcame, such as the complex modality-specific issues of working with both video and audio, and also several which we have plans to address in future work. Among these issues is the enhancement from global conditioning to local conditioning.

As described earlier when outlining the differences between global and local conditioning, the WaveNet implementation we have been working with lacks local conditioning capabilities. Local conditioning would prove the capability to factor new information into the generation of different parts of the waveform. A workaround to this is to create short, global conditioning, audio samples. These could then be stitched together to create the full-length sample. This loses out on the major factor of what makes WaveNet powerful though and it is no longer able to use these past audio segments as context for future generation. This is why local conditioning is important, to create a much more detailed generation specific to each frame of the video. Local conditioning should be able to generate much faster than this proposed workaround, as it would be able to continually generate without need to restart generation and be more consistent too.

Translating the meaning of video to audio in general is challenging due to the fact that there is no direct way to express the ideas of a video in an easily transferable way. We attempted this by attaching specific categories that express what is happening in a frame, but this does not express more subtle expressions such as an object moving faster, or an image becoming brighter. These ideas are expressed over multiple frames and cannot be simply determined by examining a single frame.

Future Directions

Much possible improvement remains, but this work demonstrates a full pipeline for extracting category themes from a video and generating relevant audio. We demonstrate that it can be implemented with off-the-shelf tools and components. For a more complete design, we plan to create an end-to-end encoder and decoder to encode an input video and generate the audio corresponding to the frames of the video. This end-to-end encoder/decoder model could be much faster and would condense down much of the bulk of the current model into a format which would be more easily used by others. This encoder would look something similar to what was developed by Vladimir Iashin and others in their transformer design [1].

Prior to the end-to-end encoder and decoder, the implementation of local conditioning into our WaveNet model could provide significant improvements. Local conditioning is necessary to utilize the full capacity of WaveNet.

Conclusion

We propose a deep learning pipeline for audio generation conditioned on paired video frames. Throughout this project, we’ve explored the capabilities and limitations of tuned, pre-existing components. Through the development of a frame classification architecture, and a custom trained WaveNet conditioned to generate audio in each classified category, we have been able to form the beginnings of future audio generation based entirely on off-the-shelf convolutional neural networks. This structure can be further developed and serves as a baseline for our future development of an end-to-end transformer structure. This type of structure should allow for a greatly improved speed of audio generation, and improved accuracy in the results of generated audio as well. Further research into WaveNet is necessary to properly tune and condition it for optimal generation quality, and to generate based on local conditioning rather than global conditioning. As we continue to further our research into this field, we plan to greatly improve our architecture, and create a full-scale model which can be utilized for accessibility, restoration, and further data recognition. This work represents a concrete step forward toward our eventual goal to have a single model take the entirety of any length video and generate the full audio to match.

References

[1] “Papers with Code - Taming Visually Guided Sound Generation,” *paperswithcode.com*. https://paperswithcode.com/paper/taming-visually-guided-sound-generation (accessed Mar. 13, 2023).

[2] “YouTube-8M: A large and diverse labeled video dataset for Video Understanding Research,” *Google*. [Online]. Available: https://research.google.com/youtube8m/index.html. [Accessed: 14-Mar-2023].

[3] A. Van Den Oord *et al.*, “WAVENET: A GENERATIVE MODEL FOR RAW AUDIO.” Available: https://arxiv.org/pdf/1609.03499.pdf

[4] K. Onishi, “WaveNet Tensorflow v2,” *GitHub*, Sep. 15, 2022. https://github.com/kokeshing/WaveNet-tf2/.

[5] I. Babuschkin, “A TensorFlow implementation of DeepMind’s WaveNet paper,” *GitHub*, Jul. 13, 2022. https://github.com/ibab/tensorflow-wavenet

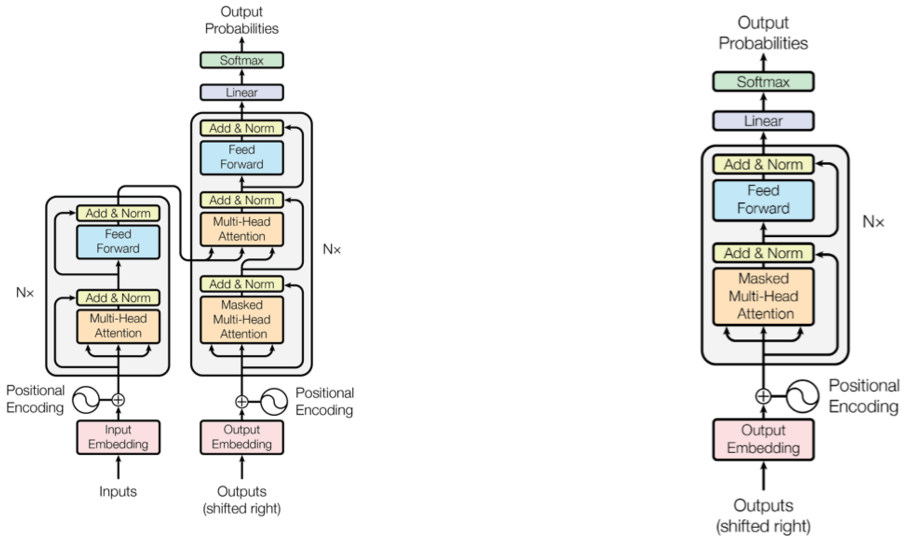
Error-Correcting Music Transformers

Jonathan Keane, Michael Conner, and Josiah Yoder
EECS
Milwaukee School of Engineering
Milwaukee, WI 53202
{keanej, connerm, yoder}@msoe.edu

March 18, 2023

Abstract

In state-of-the-art music Transformers today, a decoder-only Transformer autoregressively produces a musical work, basing its decisions for the next event on the previous sequence of events. With this process, however, there are no means for the model to iterate on itself and correct the original piece when musical inconsistencies occur. Additionally, because decoder-only Transformers only can look backward at the sequence created before it, current music Transformers have no future context that can provide the semantic meaning of future notes, which may be beneficial in informing better decisions about the correct note/event to be used at points earlier in the sequence. Therefore, we propose training a second encoder-decoder Transformer to correct music by training this Transformer to return songs with discretely inserted abnormalities back to its original piece. With this second Transformer, we can then use a generated piece of music from a decoder-only Transformer as the encoder input such that this "error-correction" Transformer can iterate on the original work to improve its quality. This encoder-decoder Transformer can then attend to the context from the whole generated piece and will have learned during training how to correct abnormalities it comes across.



(a) Encoder-Decoder Transformer (reproduced from [3]) (b) Decoder-Only Transformer (as in [1])

Figure 1: Two types of Transformers used for generating sequences.

1 Introduction

When musicians are composing a new piece of music, they do not often come across brilliance upon their first try. They often look at the piece they currently have and make small adjustments, such as making a note more staccato or legato to better fit the piece or changing the note that is played in a piece because it is too sharp or flat. While we see this in the case of human composers, state-of-the-art music Transformers today only perform a single pass while composing a song. Therefore, we hope to show with this research that by allowing a music Transformer to iterate and correct its original piece of music, we can improve the results of the current state-of-the art.

There are two basic architectures of Transformers used for generating sequences. The first kind of Transformer is the original model proposed by Vaswani et al. [3] (Fig. 1a) where there is both an encoder and a decoder, and the goal of the model is to be able to take a text from one "language" and generate a new sequence in a "target language." This encoder-decoder Transformer is commonly used in the NLP field for the task of translating between languages (for example, generating English translations from Spanish source-texts). The second form of the Transformer is a decoder-only model (Fig. 1b) where there is only an encoder block, which uses self-attention to extend an input sequence. This decoder-only Transformer is currently what is used in the state-of-the-art music Transformers (such as [1]), where novel musical works are being generated without a reference work from which they are derived.

While music Transformers [1] use the same general architecture as text-to-text translation Transformers, qualitatively, music generators do not yet achieve musical compositions with

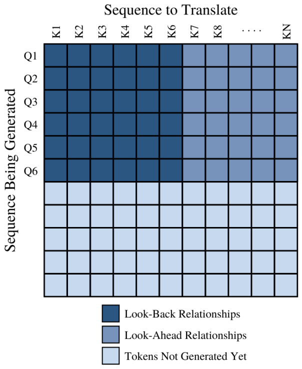


Figure 2: In autoregressively building a sequence with an encoder-decoder Transformer, when you perform cross-attention, the model can take context from information ahead and behind the token being generated.

the same level of coherence seen in text to text translations. One reason for this may be that single note events (such as a "note on" event) convey less information than a word of text. For example, in natural language, when you have the prefix "the dog" in a sentence, because a verb would follow this far more naturally than an adjective. For music, a single note could be followed by almost any other note on the scale, depending on what key the song is in. Thus, a music Transformer must model ideas such as key or phrasing in addition to the larger-scale "grammar" of music.

To address this challenge, we propose to improve the generative quality of music Transformers by training an "error-correction Transformer" that uses the encoder-decoder version of a Transformer (Fig. 1a) to translate music with errors ("bad music") into music without these errors ("good music"). We propose that by discretely augmenting a data set of professional music with errors inserted programmatically and training to correct these errors, this Transformer can learn to correct musical errors. With this, we can take this trained model to iterate on the music from a trained state-of-the-art music Transformer (based on the decoder-only model, Fig. 1b) that generates a sequence and improve upon the initial piece by correcting its "musical errors," improving the overall results of the model as a whole. By having a full song to provide context for sequence generation, when we perform cross-attention in our encoder-decoder Transformer during inference, we will have both look-back as well as look-ahead relationships in our attention matrix as seen in Figure 2. As a result, this model provides additional information to the decisions the model makes that are not possible in an encoder-only Transformer.

2 Prior Work

For the task of music generation, the music Transformer produced by Huang et al. [1] is the foundation that this work builds upon. In [1] Transformer using relative self-attention described by Shaw et. al [2] was used and shown to be capable of producing music that captured long term structure, outperforming previously used techniques for music generation, including both an LSTM and a baseline Transformer with no relative self-attention, when evaluated by humans. This model was not rated as highly as the testing data produced by actual musicians. Our work attempts to improve the performance of the generated music toward the quality of actual musicians.

3 Experiments: Random Addition/Removal of Notes

With the proposed architecture combining both the encoder-decoder and decoder-only forms of the Transformer, we look to have an encoder-decoder Transformer that can learn to correct sequences of music that have randomly inserted/removed notes from a piece of music to the original piece from a professional musician. With inserting random augmentations into the a piece of music, we believe that if there are not enough augmentations, the Transformer will only learn to output the original piece because the model will have lower cost for keeping the errors than attempting to change them. In the case that there are many augmentations, the model may learn to ignore the encoder input, regarding it as random noise, learning only to heed the input prefix and its own previous output. We expect that there is some region that falls in between these two possibilities such that the model learns to correct to the original piece when it comes across these augmentations. We will perform a grid search on the number of insertions and deletions to try and find what balance of augmentations improves results.

To determine if there are improvements, we want to compare the results from the initial decoder-only Transformer to the results that are passed through the initial encoder-decoder Transformer. Users will be presented with the two samples for *n* songs and be asked to determine which they believe is the better music without knowing which Transformer produced which. This will then be used to determine the improvements in an A/B test. From these results, we will use statistical tests to see if our results achieve any form of statistical significance.

3.1 Decoder-Only Transformer & Data Set

Our decoder-only Transformer is the Music Transformer [1], a decoder-only model (Fig. 1b). We use an open-source implementation [1] as the basis for our model. We adjusted some of the hyperparameters and the model that we used for our experiments has the dimensions described in Table 1 above. This baseline model of the music Transformer was trained on the e-Piano competition data set. All pieces in this data set were solo pieces played on the piano. All pieces are of classical music, performed by expert musicians. For training our

[1] https://github.com/jason9693/MusicTransformer-tensorflow2.0

Table 1: Model dimensions of our baseline Transformer used for generating music.

Model Attribute	Dimension
Max Sequence Length	2048
Embedding Dimension	512
Heads per Attention Mechanism	8
Layers	6
Optimization	Adam[a]
Dropout	0.2

[a](learning rate = 0.001,$\beta_1 = 0.9$, $\beta_2 = 0.9$)

baseline model, we took the original piano-e-competition data set and added an augmentations on pitch (shifting the whole song up or down for pitch shifts of up to 3 notes) as well as time shift augmentations where we stretched/compressed the time shifts for an entire piece between notes by a factor of 0.05. At this point, our baseline Transformer makes sequences that are fairly well-formed pieces of music.

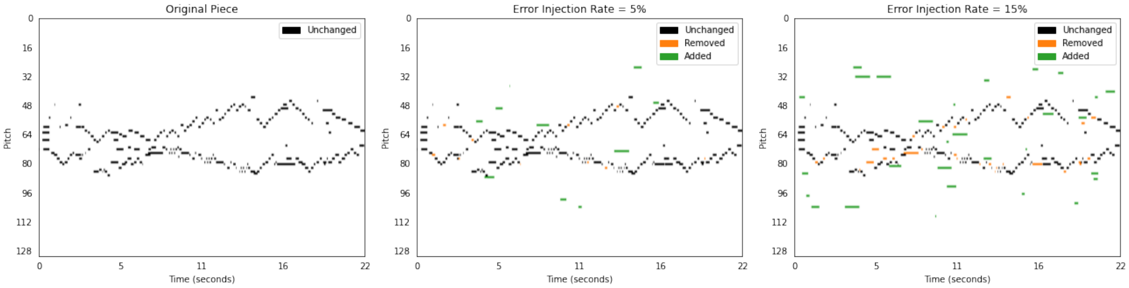
We encode performances as a series of events, using the encoding in Huang et al. [1]. This encoding has four different categories of events that represent events for NOTE_ON, NOTE_OFF, VELOCITY, and TIME_SHIFT. In the encoding scheme, there are 128 NOTE_ON and 128 NOTE_OFF events, which defines events to begin playing or stop playing a given pitch. 32 VELOCITY events are used to represent how hard any keys following this event will be pressed. Finally, to represent the natural timing of a performance, there are 100 TIME_SHIFT events that represent increases in the current timestamp in a piece, taking up all 10 ms intervals between 10 ms and 1 second. When a TIME_SHIFT event is processed, the current timestamp of the piece advances the defined amount and any notes that have been activated by a NOTE_ON event and have not been turned off by their respective NOTE_OFF event will be played for the duration specified by the TIME_SHIFT event.

3.2 Encoder-Decoder Transformer & Note Removal/Insertion

For the second Transformer being trained, we have an encoder-decoder setup that translates from one piece of music to another with differences. The dimensions of this model are defined in Table 1, but there is now both a decoder and an encoder in the model (as in Fig. 1a), which will be trained with encoder input as the augmented music and the expected decoder output as the original piece.

The augmentation algorithm we use randomly inserts and removes a specific number of notes. In our experiments, we took our samples from the beginning of the song, so that the error-filled input sequence and correct output training sequence would be more closely aligned. We do not expect that this would limit the trained model to operate only at the beginning of sequences.

Figure 3: Visualization of when notes are being pressed down during a song, comparing the original song, and the error-injected songs with error insertion rates of 5% and 15%, respectively. In the error injected figures, the notes removed from the original song and the new random notes added to the song are marked.



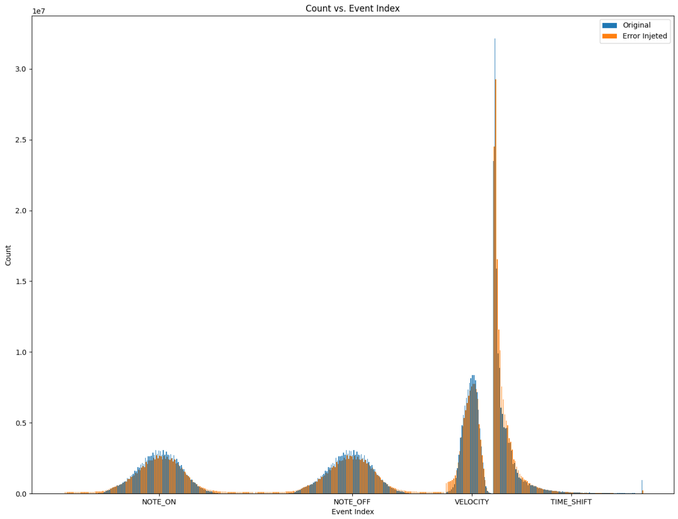
For our error injection procedure of adding/removing notes, in our preprocessing, for each sample in the augmented data set, we convert the event sequence to a list of full notes (encompassing the duration, pitch, and velocity) and use Algorithm 1 to remove and insert notes.

Algorithm 1 Error removal/insertion algorithm.

```
$P =$ number of pitches, $V =$ number of velocites, $T =$ number of time shifts
$R =$ removal rate
$N =$ notes
$U = Uniform([0,1))$
$D_{max} = \max\{\forall\ N_i \in N, Duration(N_i)\}$
$D_{min} = \min\{\forall\ N_i \in N, Duration(N_i)\}$
$L = \max\{\forall\ N_i \in N, StartTime(N_i)\}$
$V = U^{1\times|N|}$
$N_{errors} = \{N_i | V_i > R\}$
while $|N_{errors}| < |N|$ do
    $Start = U * L$
    $Duration = U * (D_{max} - D_{min}) + D_{min}$
    $Velocity = RandomInteger([0,V))$
    $Pitch = RandomInteger([0,P])$
    $N_{errors} = N_{errors} \cup \{Note(Start, Duration, Velocity, Pitch)\}$
end while
return $N_{errors}$
```

When we transform our original data set to uniformly distribute the notes/velocities/time shifts across their respective ranges, we see that the less frequent of each type of event in the original songs become more prominent in the error injected songs and vice versa for more common events in the original songs. This transformation of the distribution of events is visualized in Figure 4, as we see the NOTE_ON and NOTE_OFF events having their distributions become more uniform and flatten out across the entire range of possible notes. When we apply this transformation across the course of the whole song, we see that

Figure 4: Transformation of the distribution of the frequencies of the different note events across the original songs and their error-injected forms. The y-axis represents the number of each event across the training set.



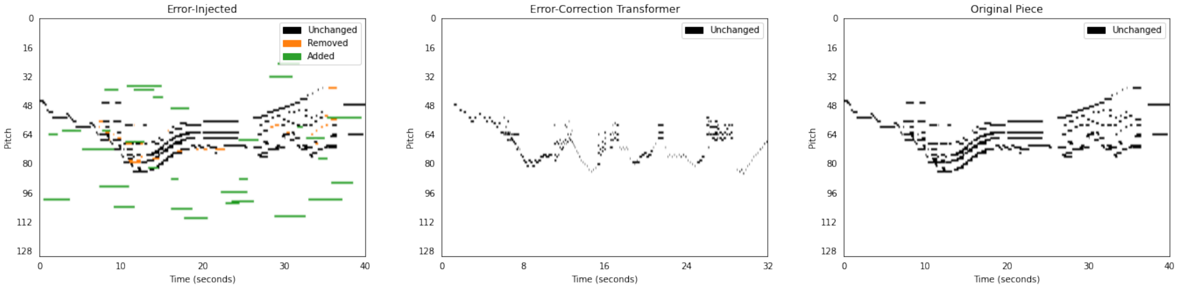
there is typically an increase in the average length of notes, as the correct distribution of the duration of notes is heavily weighted towards shorter notes compared to error-injected sequences across this data set, which is visualized in Figure 3, in the peak on the left end of the fourth rise.

3.3 Model Size & Training

We trained our Transformers on the ROSIE supercomputer at the Milwaukee School of Engineering on a DGX-1 pod using a single NVIDIA V100 Tensor Core GPU for 8 epochs across our augmented data set of 23072 samples, with each sample representing a whole song or an augmentation of a whole song. We use an 80/10/10 training/validation split of our data so we can observe the categorical accuracy measure on validation data during training, selecting the first 80% of the data set for the training set, the next 10% for validation, and the final 10% for testing to keep data consistently separated through different experiments.

3.4 Experiment Set 1: Batch Size = 1

In our first round of experiments, we used a batch size of 1 and a sequence length of 2048, to fit the model within memory during training. We tried training with error-insertion/removal rate from 2 to 30%, but the network continuously fell into producing repeated events or notes.



(a) (b) (c)

Figure 5: Example of the proposed error-correcting Transformer attempting to correct errors in a random error-injected song in the testing data. (a) Input to the encoder-decoder. Green bars show erroneous notes inserted into the song. Orange bars show notes removed (these are not given to the encoder-decoder). Black bars show notes retained. (b) Output of the encoder-decoder, generated with the goal of removing errors. (c) The original song corrupted to produce the input in (a).

3.5 Experiment Set 2: Batch Size = 8

Not finding any improvements with the hyperparameters explored in our first experiment set, we hypothesized that these failures could be due to a batch size that was too small to allow the model to generalize. We rationalized this as potentially being because with each training step, the model would gravitate towards the song it had just seen, but this would mean that the model would never have to try and synthesize results from many samples at once, leading to poor training results. To test this, we switched to a batch size of 8, while also reducing our max sequence length to 1024 to keep the entirety of our model within a single GPU. Performing training on the compressed vocabulary dataset with 15% error-insertion/removal rate, we saw that this model was able to exceed the accuracy score beyond what we saw in our first experiment set. In performing inference, generated sequences created coherent strings of notes.

This model could predict correct tokens in both the training and validation set with approximately 50% accuracy. While this measurement could potentially be misleading because the model could be learning to build sequences that used the most frequent events, a qualitative review of a small subset of the songs, such as the song demonstrated in Figure 5, showed that the pieces generated by the error-correction Transformer seemed to be much more authentically musical compared to the error-injected data it was given as input, while removing many of the errors injected into the piece. However, it seems that in learning to correct errors, the model learned to have a tendency towards shorter notes and less harmony when compared to the original piece. This may be because we inserted notes with random durations from the uniform distribution, our error-injected songs have longer notes compared to our correct songs, as we see the number of TIME_SHIFT events in Figure 4 has many more smaller TIME_SHIFT events compared to longer shifts. Therefore, the model may have learned to always use shorter notes because music with errors tended to have longer notes, which is visible in the error-injected piece in Figure 5.

In a second experiment in this set, we began by training the Transformer to learn the identity transformation (encoder sequence producing a target sequence that was the same encoder sequence advanced one event) before switching to training with errors introduced in the input sequence. Even though we used a batch size of 8, this model trained poorly. We suspect that the pretraining may have hurt the performance by pruning out the more semantic relationships to favor weights that would be most suitable for the identity transformation only.

4 Conclusion

While we have not yet tested the encoder-decoder's ability to correct errors in music generated by a decoder-only music Transformer, in the preliminary experiments we present here, the music Transformer was able to both remove randomly-added notes and add in notes to compensate for randomly-removed notes, creating a coherent melody consistent with the corrupted source melody. With further experiments and tuning, this technique may be able to improve on the state-of-the-art in music Transformers.

5 Future Work

The encoder-decoder Transformer may have a more concrete understanding of what a note is because it must identify and remove spurious notes consisting of several related events: a velocity shift, a series of time shifts for the duration of the piece, and both a NOTE_ON and NOTE_OFF event. This potentially can be visualized by comparing the attention scores that come from the final head of Transformer when generating notes using the encoder-decoder Transformer compared to that of the original Transformer to see the attention to different previous notes. We would expect that in the encoder-decoder network, the weighting of the relationships of the NOTE_OFF event should have stronger weighting towards the NOTE_ON, VELOCITY, and TIME_SHIFT event that were part of the creation of this note, compared to these relationships in the decoder-only music Transformer.

Beyond improving unguided synthesis of music, the strategy proposed in this paper adds the value that we can inject knowledge from the problem domain into the network, making the model capable of being tuned towards any corrections that can be synthesized in the training output data and away from any errors that can be synthesized in the training input data.

There are a variety of strategies for synthesizing errors and improvements to a performance. Are there efficient ways to use errors synthesized by the generator Transformer? Could repetitions be inserted so the encoder-decoder learns to remove them? Can the melody line of a performance be identified automatically and used to synthesize performance harmony from a performance of a novel melody? Error-correcting music Transformers have many possibilities.

References

[1] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, and Douglas Eck. An improved relative self-attention mechanism for transformer with application to music generation. *CoRR*, abs/1809.04281, 2018.

[2] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *CoRR*, abs/1803.02155, 2018.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

CAN HACKERS CASH-IN ON THE SENSITIVE DATA CONTAINED IN CACHE?

Erich Rice, Dennis Guster, Li Dai

Department of Information Systems

St. Cloud State University

St. Cloud, MN 56301

eprice@stcloudstate.edu

Abstract

The threat to IT systems and applications continues to be a concern for organizations. While many of the current attacks are centered at the application layer, these types of attacks are not the only threat that requires protection. While it's easy for semi-sophisticated hackers to procure an organization's email addresses and target them by sending mass phishing emails, more sophisticated hackers may utilize techniques which allow for stealthier extraction of data. These more sophisticated hackers could seek to extract data directly from memory, thereby bypassing security controls placed to try and block or log unusual activity. By extracting the confidential data directly from the real or virtual memory a hacker could get what they desire without raising alerts, which might bring about efforts to try and stop their activities. This type of attack could be especially useful if it was on shared hardware, such as within a public computing cloud.

1 Introduction

Hackers are constantly looking for ways to compromise sensitive data. Certainly, the easiness of the method to compromise a system or application is a major consideration for them (Chng et al., 2022). Also of concern, is the probability of being detected in their hacking attempts. If standard file systems or databases are attacked, those attacks will typically be logged as well (Tayag, De Vigal Capuno 2019).

One potential way for hackers to bypass the logging system is to use a lower-level architecture of the computer system and pull the data they seek directly from memory. Although this method is more complex than the attack methods that are run through the application layer, such as through phishing emails, once they are developed and proven they are well shared and become viable for serious hackers (Koon, 2022). Therefore, these types of attacks should not be taken lightly as indicated by the nearly 70% of Microsoft security issues being memory safety related (Cimpanu, 2019). This type of scenario is especially of concern when it occurs on the backend server side. For example, a Linux host that is acting as a server used to support a company's e-commerce sales system. In this case it is not the data of one client that could be compromised, but the data of all the client devices that connect to the server to use the e-commerce system.

A good summary of how memory is allocated on a Linux host can be obtained by using the "free -h" command (Linuxize, 2020). In the example below taken from a Linux host on the authors' private computing cloud, the key is the first row that delineates the real memory. In that row it is found that the system has 31 GB of total real memory.

```
dennis.guster@eros2:~$ free -h

            total     used     free   shared/buff  cache   available

Mem:        31G       694M     24G       21M       5.9G       30G

Swap:       8.0G      0B       8.0G
```

However, only 694 MB are in use, leaving 24 GB of memory free. Of interest to this paper is the fact that there are 5.9 GB allocated for buffers and cache. In both cases, the use of a buffer and cache, are used to speed up the processing of data. A good example of a buffer would be the process of reading data from a sequential file from the disk. Disk access, even on a solid-state drive (SSD), is significantly slower than reading directly from memory. So, the read request goes directly into memory from disk and the process then pulls from the buffer speeding up the process. However, the process from the disk to the buffer is on-going and the read from the buffer goes on when the process gets the interrupt. The assumption is that the process may have to go through a series of wait states caused by reading other files, waiting for human input, or writing to a log file.

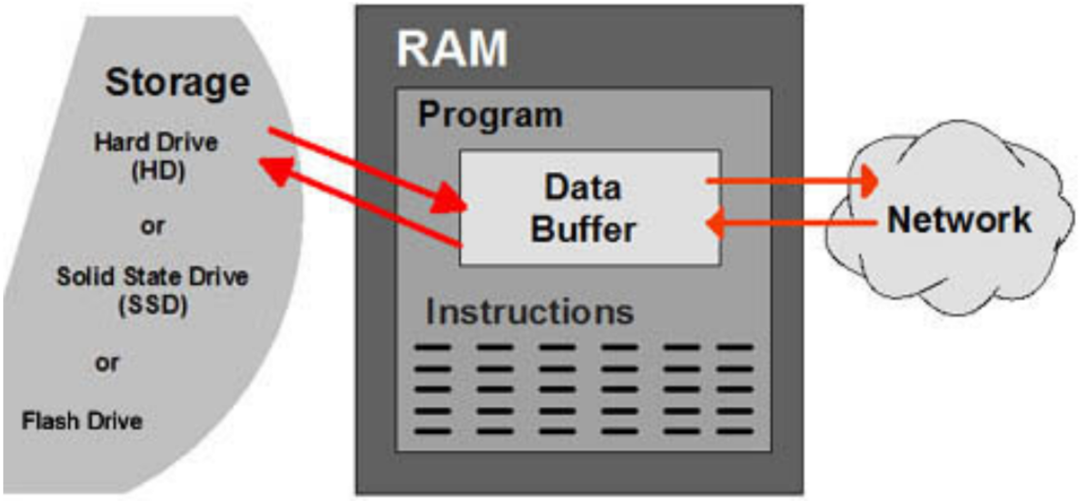


Figure 1: How a Memory Buffer Works (*Definition of buffer,* n.d.*)*

In the case of memory cache, also sometimes called a "CPU cache", information is either read or written quickly to the cache and periodically the cache is cleared (*Definition of cache,* n.d.). After this process has been completed then the cache space can then be reused. It is possible to evaluate some of the characteristics of a file that has been placed in memory on a Linux host. In the example below, a simple java program creates a file called Data.txt and two lines are sent to that file.

```
dennis.guster@eros2:~$ java ReadWriteZ

Type characters to write in File – Press Ctrl+z to end

Line 1: ps

Line 2: this file contains Data
```

However, when one looks in memory using the "lsof" command (Zivanov, 2022) at the open file for process ID 22177 there is no data in the file as indicated by the 0 in italics as seen below.

```
dennis.guster@eros2:~$ lsof -p 22177

java    22177 dennis.guster    6w   REG   0,52        0   27659826 /rhome/dennis.guster/Data.txt
(10.10.3.5:/exports/rhome)
```

A check on the file system level also reveals that no data has been written as well.

```
dennis.guster@eros2:~$ ls -l /rhome/dennis.guster/Data.txt

-rw-r--r-- 1 dennis.guster professors  0  Nov 10 11:46 /rhome/dennis.guster/Data.txt
```

After closing the java program, the data is then written to the file.

```
dennis.guster@eros2:~$ cat Data.txt

Line 1: ps

Line 2: this file contains Data
```

From the examples above it is clear that the data is initially going into a memory buffer. However, that buffer could be contained in real memory. As the free command displayed earlier, it indicated there was 5.9 GB allocated by the operating system for that purpose. It is also possible that the initial location of the data could be in virtual memory. Regardless, it is hoped that the data that is stored in memory will not be compromised by a hacker. For a hacker to accomplish such a feat they would need to know the relative address of the memory storage area where the pertinent data resides. The structure of those memory addresses is a hexadecimal number, which represents the relative bit address of the data. In some cases, those addresses are fixed, while in other cases (typically involving virtual memory) those address can vary (Sterling, Anderson, & Brodowicz 2018). In a 64-bit computer the memory addresses range from 0000000000000000 to ffffffffffffff or 16^16 possible bits (Vostokov, 2023). So even with fixed addresses, guessing where data might be stored is not a trivial matter for a potential hacker. With virtual memory the addresses are likely to change from execution to execution (Sterling, Anderson, & Brodowicz 2018). So, finding a virtual memory address to attack may have a very limited and short value.

However, the processes themselves need to know where the data is stored so it can map to the data needed to exist within the operating system. In the example below, one can see the memory segments related to the process ID 21087. Obviously, this is still a lot of memory to evaluate on a trial-and-error basis to find what you are looking for.

```
dennis.guster@eros2:/proc/21087/map_files$ ls

55f0371b3000-55f0371bb000  7fe994f19000-7fe994f1b000  7fe995137000-7fe995138000  7fe99513f000-
7fe995140000 55f0373ba000-55f0373bb000  7fe994f1f000-7fe994f48000  7fe995138000-7fe995139000
7fe995140000-7fe995147000
```

To investigate the probability of a hacker finding the memory address of potentially sensitive data contained in either real or virtual memory, the example that follows uses an existing sequential file and appends data to it. Of course, the file could be attacked directly, however the probability of an attack like that being detected is quite high. So, the rationale of a hacker reading the sensitive data from memory would be that it would be stealthier, thus resulting in a lower chance of detection and an alert being raised. In the next example seen below, the file "filexyz" is opened and appended to by using the Linux "cat" command.

```
dennis.guster@eros2:~$ ls file*

file2011b  fileEX   filetcp483  filetoremove2  filexyz
```

```
dennis.guster@eros2:~$ cat >> filexyz

sample output

more sample output
```

By looking at the "io" file for process PID 21087 one can see that 36 characters have been appended to the file but note that the write bytes value is 4096 Bytes, which matches the page size currently being used by the Linux based operating system. This makes addressing and allocating memory a little cleaner.

```
dennis.guster@eros2:/proc/21087$ cat io

rchar: 4979

wchar: 36

syscr: 13

syscw: 4

read_bytes: 0

write_bytes: 4096

cancelled_write_bytes: 0
```

It is also possible to see if the file is open in conjunction with the process PID of the application. To do so, the list open files command “lsof” can be used (Zivanov, 2022). The file "filexyz" shows up as a regular file containing 62 Bytes of data at this point. Also of note is the fact that it is not currently on the Linux host, but on another as indicated by the private "10.10.3.5" IP address as the file is housed on an NFS (Network File System) server in the same private cloud environment. This hopefully would make it more difficult for a hacker to attack the file system directly, however it is important to remember that data will be temporarily stored in memory on the host computer prior to being sent across the network to the NFS server for storage.

```
dennis.guster@eros2:~$ lsof -p 19798 | grep filexyz

cat     19798 dennis.guster    1w   REG   0,52        62 27659830 /rhome/dennis.guster/filexyz
(10.10.3.5:/exports/rhome)
```

To determine if the file is in real or virtual memory the "vmtouch" command can be used to ascertain its location (Carrigan, 2020). In the example below it is checked to see if it is currently in the virtual memory, however it can be seen that it is stored in one page of memory, and it is used to store it in the real memory. Then the “filexyz” file is "touched" into virtual memory once again taking up one page of 4KB of space within the virtual memory space. On a side note, it is also possible to remove a file from virtual memory again using the "vmtouch" command (Carrigan, 2020).

```
dennis.guster@eros2:~$ vmtouch -v filexyz

filexyz

           Files: 1

     Directories: 0

  Resident Pages: 1/1  4K/4K  100%

         Elapsed: 0.001655 seconds


dennis.guster@eros2:~$ vmtouch -vt filexyz

filexyz

           Files: 1

     Directories: 0

   Touched Pages: 1 (4K)

         Elapsed: 0.001412 seconds


dennis.guster@eros2:~$ vmtouch -ve filexyz

Evicting filexyz

           Files: 1

     Directories: 0

   Evicted Pages: 1 (4K)

         Elapsed: 0.001375 seconds
```

A logical first step a hacker might take to try and resolve the actual relative address in memory is to take advantage of some of the files in the /proc directory. In the example below for process 30463 the entries for the "cat" command used to open and allow appending to the file "filexyz" are displayed. Note that it provides a beginning and ending relative bit address range in hexadecimal. The first line deals with execution related activities for the file. The second line with reads and the third with writes to the file. In all lines the memory is protected to prevent overwrites as indicated by the "p" flag.

It is really interesting to note that in line 3, if one subtracts the beginning address from the ending address a value of 1000 hex is obtained. If converted to decimal this would be a value of 4096 or 4KB. By coincidence could this be the 4KB page that contains the buffered data for the file "filexyz"? If so, it may be at risk and a hacker could use a program such as “gdb” to debug it and to read its contents (Stallman et al., 2002).

```
dennis.guster@eros2:/proc/30463$ cat smaps | grep /bin/cat

5647a41e3000-5647a41eb000 r-xp 00000000 08:02 1310762       /bin/cat

5647a43ea000-5647a43eb000 r--p 00007000 08:02 1310762       /bin/cat

5647a43eb000-5647a43ec000 rw-p 00008000 08:02 1310762       /bin/cat
```

Below an attempt is made to use the “gdb” debug program to dump the 3rd memory segment range from above (5647a43eb000-5647a43ec000) using the rights profile of the user that owns that process ID. However, this results in a “Cannot access memory” error.

```
dennis.guster@eros2:/proc/30463$ gdb

GNU gdb (Ubuntu 8.1.1-0ubuntu1) 8.1.1

(gdb) dump memory ~/catlog 0x5647a43eb000 0x5647a43ec000

Cannot access memory at address 0xa43eb000
```

From the examples above one can see that finding the address of memory within the Linux operating system is quite possible. Its design is predicated on providing functionality to the system administrator and developers creating applications to be run on it. However, finding an address is one thing, but being able to compromise the data at that address is another. Remember that with virtual addressing the addresses are constantly changing so right off the bat there is a limited window of opportunity for a hacker to take advantage of it. There are also a number of other built in security precautions as well, most notably related to user rights on the profiles.

Therefore, the purpose of this paper is to evaluate the possibility of data being found in a buffer or virtual address space and actually being compromised. The experiments will be discussed from three different rights levels: user/owner, root and kernel module related.

2 Methodology

In the Introduction section an address was found in memory and an attempt was made to try to read that data, though to no avail when using the user level rights privilege. However, that experiment was then rerun and the sudo command was used to provide root level rights. But once again the data could not be read using the “gdb” command. Thus, a user space process, even running as root, is still limited in what it can do as it is running in "user mode" and the kernel is running in "kernel mode" which are actually distinct modes of operation for the CPU itself. In kernel mode a process can access any memory or issue any instruction. In user mode (on x86 CPUs there are actually a number of different protected modes), a process can only access its own memory and can only issue some instructions. Thus, a user space process running as root still only has access to the kernel mode features that the kernel exposes to it.

Thus, even the root user has limitations. Those limitations are imposed by the design of the operating system to differentiate between user space and kernel space. For instance, even though you are a root user, you can't change the speed at which the hard disk rotates if that option isn't provided to you through the driver (you can write a driver that will allow the function, but even then you are not accessing the hardware directly but through the driver), the reason for this is that the actual control of the hardware is all done in kernel space and the way user space accesses it is through system calls. A kernel space is not a place for a user!

Note below that when the “gdb” command was run via “sudo” that the address is slightly different than from the previous example. This is because virtual memory is currently being used. So therefore, a hacker trying to use past addressing of a process would be stymied due to the dynamic nature of the virtual memory addressing.

```
dennis.guster@eros2:~$ sudo gdb

(gdb) dump memory ~/catlog 0x000055c7b524c000 0x000055c7b524c100

Cannot access memory at address 0xb524c000
```

One might find this situation confusing because one is often taught erroneously that the root has all rights everywhere. Further, it is often taught that if the root doesn't have the rights needed that it has the right to give itself those rights. All of this is true except in entities controlled by the kernel module (Wazan et al., 2022).

To illustrate this concept, one might look at a zombie process, a zombie process refers to any process that is essentially removed from the system as “defunct”, yet still resides in the CPU’s memory as a “zombie”. As one might expect a zombie process is one that cannot be killed even by using the root and the “kill” command with the “-9” option (Linuxize, 2019). So, if the root owns an apparent process that is a zombie, and if that is killed the zombie process would usually be killed as well. However, because the zombie process is often left over from some kernel function such as a remote procedure call to install the NFS client on a host the process is then owned by the kernel module and hence the root cannot directly kill it.

Fortunately, from a security perspective this concept of kernel ownership carries over to various memory segments as depicted in the example below in which not even the root can read the selected memory address. By using the program from (Kannan, 2018) it is possible to determine the virtual address of a memory variable and later link that virtual address to the actual physical address. When the C executable file named (vm-addr) is run the process ID of that executable is returned along with a virtual memory address.

```
dennis.guster@eros2:~/OLDHOME$ ./vm-addr
```

```
my pid: 29675

virtual address to work: 0x55dcd9526260
```

Once the program executes it remains in a wait state (as shown below) so that one can evaluate the memory area.

```
dennis.guster@eros2:~$ ps -al

F S      UID          PID    PPID   C  PRI  NI ADDR SZ   WCHAN   TTY    TIME       CMD

0 S 1895401321 29675  28848   0   80     0 - 1129             wait_w    pts/0  00:00:00   vm-addr
```

So, then step one is to ascertain if the virtual memory area can be accessed using the debug program "gdb". However, as is shown in the output below it cannot be found using the debug program as it again provides a "Cannot access memory" error.

```
(gdb) dump memory ~/vmmemlog 0x55dcd9526260 0x55dcd9526360

Cannot access memory at address 0xd9526260
```

Step two then is using the second program found on the East River Village utilities (mem-addr) from (Kannan, 2018), this is to get access to the physical address for the memory variable.

```
dennis.guster@eros2:~/OLDHOME$ ./mem-addr 30004 0x562c0bb28260

getting page number of virtual address 94747174797920 of process 30004

opening pagemap /proc/30004/pagemap

moving to 185053075776

physical frame address is 0x0

physical address is 0x260
```

Step three is then to determine if that memory can be accessed from both a user and root level using the "gdb" debug program. However, once again the memory is protected because it is effectively owned by the operating system kernel.

(For both user and root accounts)

```
(gdb) dump memory ~/phyaddlog 0x260 0x270

Cannot access memory at address 0x260
```

Even though certain ranges of memory addresses are protected, it is sometimes possible to gain access to data via registers and their related addressing. To illustrate this concept a C executable named "add" is run and returns the process PID 30222.

```
dennis.guster@eros2:~/OLDHOME$ ps -al
```

```
F S      UID         PID    PPID  C  PRI  NI ADDR SZ WCHAN  TTY      TIME      CMD

0 S 1895401321 30222  28848  0   80      0 - 1128       wait_w    pts/0    00:00:00   add
```

The utility asks for input via the keyboard and note that the last value entered is an ASCII "5" followed by the enter key. Also, note that the process is still running because only two integers have been entered instead of the three indicated in the first stage. Again, this provides an opportunity to examine memory addressing related to this program.

```
dennis.guster@eros2:~/OLDHOME$ ./add

Enter the number of integers you want to add

3

Enter 3 integers

4

5
```

Again, using the "gdb" debug program it is possible to get a summary of the registers and their associated addresses for this utility. This follows on the logic that was developed in (Farra, Guster, & Rice, 2017), where the source index register (rsi) is shown to contain the buffered data from keyboard entries. Which based on its name, the index contains what you would expect to see, for clarification on how it morphed from its original purpose see (Intel 64, rsi and rdi registers, 2014). The "info registers" command shows the starting relative address of the register related to the add program, which is running as process ID 30222. Note that root access was needed to do this as depicted by the use of the "sudo" command. Next, the first 16 Bytes of that register are dumped to a file called "regmem22".

```
dennis.guster@eros2:~/OLDHOME$ sudo gdb -p 30222

(gdb) info registers

rsi           0x1bd6670        29189744

(gdb) dump memory ~/regmem22 0x1bd6670 0x1bd6680
```

The contents of that file can now be evaluated using the "xxd" (hexadecimal dump) command. The first two characters of the dump are "35" which is an ASCII "5" the next two characters are ASCII "0a" which is a carriage return (or enter) on the keyboard. Off to the right in the interpreted part one can see that the "35" is depicted as a "5". While this only provides a piece of potentially sensitive data a bot could be programmed to record each data chunk as it happens and store them in a file to be read at a later time. There are of course other registers that could be monitored and potentially compromised as well, but this concept provides the basic scenario.

```
dennis.guster@eros2:~/OLDHOME$ xxd ~/regmem22

00000000: 350a 0000 0000 0000 0000 0000 0000 0000  5..............
```

The next question that needs to be addressed then is, will putting an open file that that is in append mode in virtual space, defeat compromising its data via the register attack scenario used above? In this example a file called Data.txt containing 26 Bytes is appended to via the keyboard using the "cat" command.

```
dennis.guster@eros2:~/OLDHOME$ ls -l Data.txt

-rw-r--r-- 1 dennis.guster professors 26 Jan  3  2023 Data.txt

dennis.guster@eros2:~/OLDHOME$ cat >> Data.txt

this is more data.
```

The file then shows up as a regular open file.

```
cat     21120 dennis.guster    1w   REG   0,52       45 27660056 /rhome/dennis.guster/OLDHOME/Data.txt
(10.10.3.5:/exports/rhome)
```

Next, the file is placed in virtual memory using the "vmtouch" command which was used earlier. This provides another level of abstraction in regard to the true memory address of the data contained therein.

```
dennis.guster@eros2:~/OLDHOME$ vmtouch -vt Data.txt

Data.txt

[O] 1/1

    Files: 1

    Directories: 0

    Touched Pages: 1 (4K)

    Elapsed: 0.001294 seconds
```

By using “gdb” debug as before, it is possible to ascertain if all the data is hidden. The same logic as used in the prior example is applied herein and an address for the rsi register (which is being used as a keyboard buffer) is obtained. Note the hex address returned is different than in previous examples seen above. This is a good characteristic of the memory management of the Linux operating system.

```
(gdb) info registers

rsi            0x7f37f9297000   139878380171264
```

The first 48 Bytes of the register are dumped to a file called vmtouchfile (2nd address is 30x higher). When reading this file, it turns out that the register is unaffected by placing the online file in new virtual memory space. By looking at the interpreted portion of the dump one can see that the line "this is more data" is in fact readable. Again, by monitoring the ongoing transactions of the rsi register, data being written to the file Data.txt could potentially be compromised.

```
(gdb) dump memory ~/vmtouchfile 0x7f37f9297000 0x7f37f9297030

dennis.guster@eros2:~$ xxd vmtouchfile

00000000: 7468 6973 2069 7320 6d6f 7265 2064 6174  this is more dat

00000010: 612e 0a00 0000 0000 0000 0000 0000 0000  a...............

00000020: 0000 0000 0000 0000 0000 0000 0000 0000  ................
```

These results illustrate an interesting property about computers. That data is often stored in several places before it reaches its ultimate destination, in this instance via the NSF service across the internal network of a private cloud to a SAN (Storage Area Network), as well as on the Linux host the file was being manipulated on. Further, some of that data is retained along the way. A good example of this is the propensity to create multiple replications of data in cloud computing. Other examples might be data contained in the keyboard or print buffers on a computer.

A computer system is a series of devices that are often operating at different speeds, hence the need to buffer data by various means. In the latest example, the online file's ultimate destination was hidden space in virtual memory. However, the data had to go from the keyboard to a register before it got to that point and hence was potentially vulnerable until it got to the hidden within the virtual memory space.

Therefore, simply using virtualization to obscure the location of data is not to be considered as a viable nor comprehensive security strategy. The literature touts the use of a multilayer security approach and certainly this is an imperative strategy (Koon, 2022). However, it may need to go beyond the rights of users that can access the sensitive data of an organization, as many IAM (Identity Access Management) regimes are currently set up in enterprises. The results herein indicate that monitoring memory usage and low-level access to data also merit consideration in a comprehensive security strategy, especially within a public cloud computing environment, which many organizations have moved towards over the course of the last decade and a half.

3 Discussion and Conclusions

As has been pointed out the threat to IT systems and applications continues to be a huge concern for most organizations, be they public, private or governmental organizations.

While some organizations still host much of their IT infrastructure on-premises, many others have moved to the “Cloud”. The continued push towards cloud computing, especially via large public cloud providers such as Amazon AWS, Microsoft Azure, Google and others means that organizations typically no longer have direct control over the hardware that their systems and applications reside on, although those providers do provide a high degree of physical and network security. However, you don’t know who your “neighbor” is that resides on the physical machines your virtual machines reside on in in the “Cloud”. Thus, the ability to protect against memory mining attacks becomes even more important. If a hacker could get onto an on-premises server how much data could they possibly collect? But if that same server was hosting multiple organizations sensitive confidential data, then the potential repercussions could be even greater, if the hacker was able to mine the memory on the machine. While many of the current attacks are centered at the application layer, the Verizon DBIR (Data Breach Investigations Report) typically lists phishing emails as the cause of roughly 80% of data breaches year-over-year, these types of attacks are not the only threat that organizations need to protect themselves against.

The ability to monitor and place protections against the potential theft of data via memory mining or leakage is also a critical security control that organizations need to be aware of and put in place. More competent hackers or hacking groups might also make use of memory-based fileless methods, such as an Excel spreadsheet file embedded with macros or a website running Adobe Flash to carry out attacks that could be undetectable by conventional defenses, such as the (WAF) web application firewall or some form of endpoint protection, though this would still be delivered typically via an email. Without an adequate amount of cybersecurity in place, malware could be allowed to infect systems without end-users being aware of the fact that it has occurred. Because of the nature of computer memory, organizations need to be aware of the fact that memory-based attacks can be a potential threat vector that they need to be aware of, and place necessary controls and countermeasures in place to monitor and alert if malicious attacks do arise.

References

Carrigan, T. (2020, June 10). *Linux commands: Exploring virtual memory with vmstat*. Enable Sysadmin. Retrieved February 19, 2023, from https://www.redhat.com/sysadmin/linux-commands-vmstat

Chng, S., Lu, H. Y., Kumar, A., & Yau, D. (2022). Hacker types, motivations and strategies: A comprehensive framework. *Computers in Human Behavior Reports*, *5*, 100167. https://doi.org/10.1016/j.chbr.2022.100167

Cimpanu, C. (2019, February 11). *Microsoft: 70 percent of all security bugs are memory safety issues*. ZDNET. Retrieved March 18, 2023, from https://www.zdnet.com/article/microsoft-70-percent-of-all-security-bugs-are-memory-safety-issues/

Definition of buffer. PCMAG. (n.d.). Retrieved February 10, 2023, from https://www.pcmag.com/encyclopedia/term/buffer

Definition of cache. PCMAG. (n.d.). Retrieved February 10, 2023, from https://www.pcmag.com/encyclopedia/term/cache

Farra, H., Guster, D. & Rice, E. (2017). Security Concerns of Registers in Linux Hosts: Using Debug to Find Memory Addresses of Sensitive Data. Proceedings of MICS 2017, https://www.micsymposium.org/mics_2017_proceedings/docs/MICS_2017_paper_2.pdf

Intel 64, rsi and rdi registers. (2014, April 29). Retrieved February 16, 2023, from https://stackoverflow.com/questions/23367624/intel-64-rsi-and-rdi-registers

Kannan, B. (2018, March 2). *Virtual memory to physical memory*. East River Village. Retrieved January 11, 2023, from https://eastrivervillage.com/Virtual-memory-to-Physical-memory/

Kernel Self-protection. (n.d.) Retrieved February 27, 2023, from https://www.kernel.org/doc/html/v4.14/security/self-protection.html

Koon, J. (2022, November 3). *Memory-based cyberattacks become more complex, difficult to detect*. Semiconductor Engineering. Retrieved March 1, 2023, from https://semiengineering.com/memory-based-cyberattacks-become-more-complex-difficult-to-detect/

Linuxize. (2019, December 2). *Kill command in linux*. Linuxize. Retrieved March 18, 2023, from https://linuxize.com/post/kill-command-in-linux/

Linuxize. (2020, July 18). *Free command in linux*. Linuxize. Retrieved February 8, 2023, from https://linuxize.com/post/free-command-in-linux/

Stallman, R., Pesch, R., & Shebs, S., et al. (2002, January). *Debugging with GDB*. Debugging with GDB - Table of Contents. Retrieved March 1, 2023, from https://ftp.gnu.org/old-gnu/Manuals/gdb/html_chapter/gdb_toc.html

Sterling, T., Anderson, M., & Brodowicz, M. (2018, January 5). *Chapter 11 - Operating systems*. High Performance Computing. Retrieved February 24, 2023, from https://www.sciencedirect.com/science/article/pii/B9780124201583000113

Tayag, M. I., & De Vigal Capuno, M. E. (2019). Compromising systems: Implementing hacking phases. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3391093

Wazan, A. S., Chadwick, D. W., Venant, R., Billoir, E., Laborde, R., Ahmad, L., & Kaiiali, M. (2022). Rootasrole: A security module to manage the administrative privileges for linux. *Computers & Security*, 102983. https://doi.org/10.1016/j.cose.2022.102983

Vostokov, D. (2023). Bytes, Halfwords, Words, and Doublewords. In: Foundations of ARM64 Linux Debugging, Disassembling, and Reversing. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-9082-8_5

Zivanov, S. (2022, October 25). *LSOF command in linux {14 practical examples}*. Knowledge Base by phoenixNAP. Retrieved February 12, 2023, from https://phoenixnap.com/kb/lsof-command

Discovering Vulnerabilities in Web Browser Extensions Contained by Google Chrome

Chapin A. Johnson
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
cajohnson5@go.stcloudstate.edu

Sharveen Paramiswaran
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
sharveen.paramiswaran@go.stcloudstate.edu

Akalanka B. Mailewa
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
amailewa@stcloudstate.edu

Abstract

In today's world, web browsers are used by most everyone daily. Many of us take this incredible functionality for granted, and don't recognize the potential risks that are involved with simple internet use. These risks exist in both the browsers and their extensions. Google Chrome has cemented itself as one of the go-to browsers for both commercial and everyday consumer use. In fact, Google Chrome is currently the most used web browser in the world. But with such heavy global use, comes a feeling of false security. Just like many applications and browsers that are widely available and free to use, Google Chrome has its weaknesses and vulnerabilities. These exist within the browser itself, but for the purposes of this research, Chrome's extensions will be the main focus. This research explores the potential risks that are involved in utilizing browser extensions for Google Chrome. We will also look at a variety of attacks that extensions can execute, and exactly how they work. These attacks aren't guaranteed to cause malicious behavior, but we will also discuss ways of increasing user safety when operating a web browser, and specifically browser extensions. The objective of this research is to test a variety of different attacks using browser extensions on Google Chrome. By researching and implementing a variety of different attacks, authors plan to find where Chrome is susceptible to allowing malicious extensions. This research will help to inform browser users of the dangers that exist when using extensions, and how threat actors may be deceiving them to perform malicious activities on their computers. This research also shows the different vulnerabilities that exist within browsers, and demonstrates the privileges that browsers have on a computer. It is expected the research output to show that browser vulnerabilities aren't a one size fits all type of attack and also expected some websites to have lower levels of protection allowing for poor security against malicious extensions.

Keywords: Vulnerabilities; Security; Risk; Google-Chrome; Attacks; DOM; Passwords; Browser-Extensions

1 INTRODUCTION

The problem that we have identified in this research is the significant vulnerabilities that browser extensions can exploit [1][2]. The threats we intend to identify are applicable to any chrome user who utilizes extensions. According to Google's 2020 statistics, most users have at least one extension installed on their browser, and with over 60% of internet users saying they prefer Chrome as their browser of choice, this has massive implications [3]. Most users who download an extension assume that they are safe, though this is far from the truth. Extensions have significant access to the browser itself, along with limited access to the computer they're installed on [4]. In the world of technology, it's important to stay vigilant when browsing online as there is a constant threat affecting users like school students, all the way up to business executives. In order to better protect ourselves when using Google Chrome, we must first understand how Chrome can be vulnerable to attacks, and what steps we can use to better protect ourselves [5][6].

With the security of Chrome's extension coming under fire, several questions need to be answered. Initially, we need to know what kind of privileges Chrome provides to its extensions. If a malicious extension is installed, the privileges given by Chrome dictate how much damage can be done. There's also the possibility that Chrome could further limit the privilege of these extensions to better secure the privacy of its users. Knowing the privileges will also give us better insight as to what attacks could be executed with an extension alone. But security isn't all up to the browser [7]. That duty also falls to the websites that user's access. Therefore, websites can choose options to better secure their webpages, and decrease the chances that users with malicious extensions could be affected. Knowing which popular websites display low levels of security is important to increasing the safety of users. Finally, it's important to understand Google's policy regarding extensions, and discover what more can be done for the benefit of user security.

The objective of this project is to test a variety of different attacks using browser extensions on Google Chrome. By researching and implementing a variety of different attacks, we can find where Chrome is susceptible to allowing malicious extensions [8]. The scope would entail any type of extension that could be installed in Chrome, with malicious intent. These types of extensions can contain a variety of different attack vectors and target many different types of websites or machines. Understanding the weaknesses of Chrome is important to maintaining a safer browsing experience. As the world of technology is always changing, and staying up to date in this field is paramount to security. This research will help to inform browser users of the dangers that exist when using extensions, and how threat actors may be deceiving them to perform malicious activities on their computers. This research also shows the different vulnerabilities that exist within browsers, and demonstrates the privileges that browsers have on a computer [9]. We expect the research to show that browser vulnerabilities aren't a one size fits all type of attack. In order to successfully exploit a machine, a certain set of circumstances must first be met before an exploit is successful [10]. We also expect some websites to showcase lower levels of protection allowing for poor security against malicious extensions.

2 BACKGROUND

Chrome's extension architecture is based on component isolation and privilege separation. When it comes to Chrome extensions, it's a zipped bundle of files that include various formats like HTML, CSS, JavaScript, and many more [11]. When talking about extensions for Google Chrome, it has three types of components. One of them being content scripts that directly interact with web pages. The other being an extension core that interacts with the browser. Lastly, an optional native binary that interacts with the operating system [12]. The extension core becomes active when either the browser starts or after a user logs into their computer provided that the extension has background permission. With an extension, it can inject content scripts into web pages loaded by the browser and each page has its own instance of an extension's content scripts. Each content script runs in the same process as the web page into which it's injected. A content script is a part of the extension that runs in the context of a particular web page. Background scripts can access all the WebExtension JavaScript APIs, but they can't directly access the content of web pages [13]. So, if your extension needs to do that, you need content scripts. Just like the scripts loaded by normal web pages, content scripts can read and modify the content of their pages using the standard DOM APIs [14]. There is only one instance of the extension core per extension and the extension's native binary each run in a separate process. Throughout this research experiment, we will implement code for an extension and edit the variables using Firebase. Firebase is a product from Google that enables developers to build, manage, and grow their apps. No programming is required on the firebase side which makes it easy to use its features more efficiently. This real-time database enables users to sync-up application data in the cloud and make it available across all devices [15].

With a browser like Google Chrome, it provides more than 40 API's for chrome extensions [16]. API is known as an application programming interface and it enables companies to open up their applications' data and functionality to external third-party developers, business partners, and internal departments within their companies. This allows services and products to communicate with each other and leverage each other's data and functionality through a documented interface. Through these APIs, extension cores would be able to get real-time status of the browser [17]. An example of this can be seen as the list of tabs and running extensions or apps. We also would be able to access and modify user's data, update browser components, hijack or modify arbitrary web requests, and send messages to other extensions. Another concept to remember are iframes which would be applied later in our experiments. An iframe or inline frame is a HTML element that loads another HTML page within the document. It basically puts another webpage within the parent page and is usually used for ads, embedded videos, interactive content, and web analytics. When the web browser encounters an iframe element, it creates a new HTML document environment to load the content within. It takes the code from the referenced src or srcdoc and renders it as its own website that is then put entirely within the parent browsing page. It is called an inline frame because to the user it is all one web page. The child iframe is a complete browsing environment within the parent frame. It can load its own JavaScript and CSS separate from the parent. They can also be refreshed and loaded asynchronously from the parent site [18].

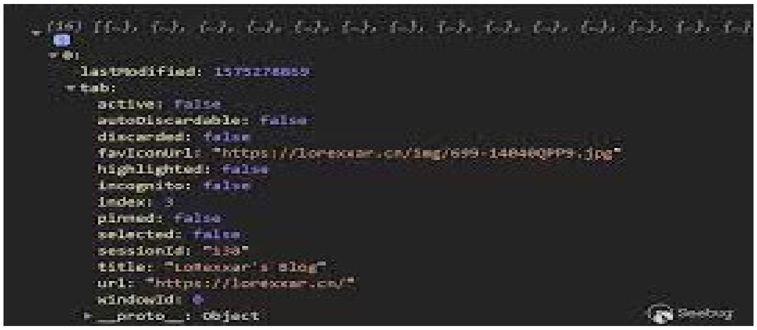
One of the major security features of Google Chrome's extension architecture is that the capabilities of components are limited based on their type and permissions granted to them. One of the high risks associated with extensions are content scripts and how they can be exploited by malicious websites because they directly interact with web pages [19][20]. Because of this reason, content scripts have the lowest privilege and can only use the APIs provided to web pages which are called browser APIs. Browser APIs include JSON, HTML5, and XMLHttpRequest APIs. When talking about content scripts, it is only accessible through the subset of Chrome APIs that support messaging between an extension and its content scripts (chrome.extension API). As previously mentioned the capabilities of the extension is limited by its permissions, so for Chrome APIs, browser APIs, and access to the web pages are guarded by these permissions [21]. The user is notified of these permissions by declaring them in its manifest file. When dealing with permissions, there are two main types which are API permissions and host permissions. Host permissions specify which pages an extension can inject content scripts and are basically a set of URLs. An example of this is when a password manager extension has host permission for one site, then it cannot access any other site. When it comes to API permissions, the extension core can only access it when it is guarded by permissions if it has the corresponding permissions in its manifest [22][23][24]. To add on, gaining access to certain Chrome extension APIs and browser APIs are constrained by host permissions. An example of this is if an extension does not have host permission to a website like http://www.facebook.com or an encompassing permission like "*://*.*", then it can't make an XMLHttpRequest to http://www.facebook.com or even block a web request to www.facebookcom even if it has API permissions webRequest and webRequestBlocking [25].

3 METHODOLOGY

In regard to Google Chrome and the extensions that it uses, theft and forgery of user data can be a highly rated risk associated with the extensions. Users that use Google Chrome have sensitive data like usernames, passwords, social security numbers, and credit card numbers which are normally communicated through different web pages [26][27]. There are different attack vectors associated with stealing user data. As explained before, extensions that are granted permissions to access the pages that contain private and sensitive data can also easily steal the data. One of the flaws in Chrome extensions that bleed into different attack vectors is the abuse of the 'http://*/*' host permission. The most common type of permission that extensions are given are the permissions to inject content scripts into the websites of Google Chrome. This permission enumerates the pages an extension is allowed to access [28][29]. In most cases, the extension's content scripts are allowed to run on any page that is browsed by the user. This permission is denoted with 'http://*/*' and the injected content scripts can read any content on the page, and this includes sensitive data from user input, extensions like password managers, and the browser itself with its built-in auto form filler. An example of this is when a user visits a web page, we can create a malicious extension that would inject scripts into the specified page and since they are running in the page's environment, the scripts would have the ability to read from the DOM the password that the user enters. For this malicious extension to execute, it needs to be installed and active in the browser at the time the user accesses the page. Additionally, finding the specific information to extract in the DOM of the target website is usually page specific [30][31]. Throughout this experiment, we will focus on three main attack vectors which are stealthy attacks using background tabs, stealthy attacks using iframes, and forging user data from specific extensions. Some of these attack vectors are difficult for users to detect and some require fewer permissions which may be difficult to attack as well.

3.1 STEALTHY ATTACKS USING BACKGROUND TABS

In relation to stealthy attacks using background tabs, there are at least two different methods in which to open or redirect a tab to target websites that do not require additional permissions beyond the "http://*/*" host permission. This means that the tab permission is not required. The first method in using background tabs is to redirect an inactive tab to the target web page and from there the extension can steal the sensitive data and later, redirect the tab to the original website. As shown in the figure 1, in detail, by calling 'chrome.tabs.query' which is the queryInfo where the queryInfo's active flag is set to false, an extension can get the list of inactive tabs [32].



```
▾(16) [{…}, {…}, {…}, {…}, {…}, {…}, {…}, {…}, {…}, {…}, {…}, {…}, {…}, {…
  ▾0:
     lastModified: 1575278859
    ▾tab:
       active: false
       autoDiscardable: false
       discarded: false
       favIconUrl: "https://lorexxar.cn/img/699-14040QPP9.jpg"
       highlighted: false
       incognito: false
       index: 3
       pinned: false
       selected: false
       sessionId: "138"
       title: "LoRexxar's Blog"
       url: "https://lorexxar.cn/"
       windowId: 0
     ▸__proto__: Object
```

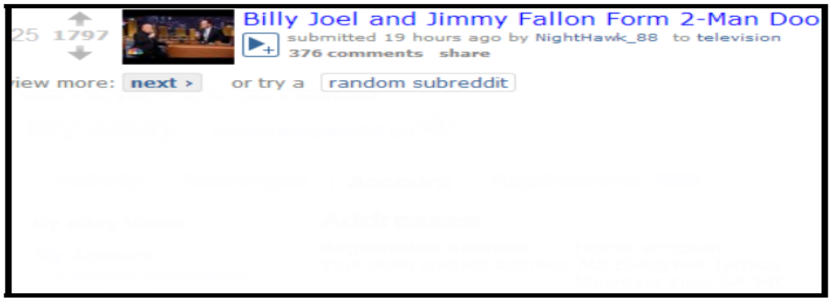
Figure 1: QueryInfo's Active Flag Set to False

From there, the query can then be further restricted to tabs that are open in the background windows by setting the queryInfo's currentWindow field to false. From there, the extension may be able to use 'chrome.tabs.update' to redirect the tab. A user may be able to do this because the tab API methods are not considered sensitive to Google Chrome, and this allows the extension to not claim the tab permission in its manifest. This is considered a stealthy attack because the only way of noticing the attack is when the tab icon redraws when a different page is loaded [33]. The other method in using the 'chrome.tabs.query' to find out whether or not a tab is visible when using the Windows API. The Windows API can determine which browser windows are currently focused on. For example, there could be a browser window open at the foreground or have the pointer hovering over the windows browser which would enable a malicious extension to launch attacks only when the user is using an application other than the browser. This is sort of a workaround in not requiring the extension to have any permissions. The implementation of stealthy attacks in stealing information can be seen later when it comes to forging user data.

3.2 STEALTHY ATTACKS USING I-FRAMES

Another stealthy attack vector when dealing with Chrome extensions is to use iframes. This is done by loading extensions into iframes which could be placed in background tabs or to make them unnoticeable or hidden to users by making the iframes fully transparent or displaying them at a very small size. Stealthily attacking web pages with iframes can be done in two methods [34][35]. The first method deals with modifying the DOM and the autofill feature. For example, when an extension's content script is running on a page, the extension can modify the DOM of that page to create a new iframe by executing "document.write("<iframe src=\"http://target.com\"> </iframe>");". After executing, the page (target.com) is then loaded in the iframe and the autofill feature or a password manager extension would automatically fill in the credentials or content needed for 'target.com'. In order to read the content inside the iframe, an extension needs to have host permission to the iframed page as well as the 'all_frames' option specified in its manifest. With the addition of the 'all_frames' option, it causes no warning to be shown to the user on installation.

The other method that deals with iframes takes an even stealthier approach [36]. In this example, by having host permission to the specific page loaded in the tab, an extension has a content script running in the same tab that doesn't contain the target page. With this, the extension can then create an iframe and load the target page similar to the previous method. This then allows Chrome to use its autofill functionality or a password manager's ability to fill content for the target page. After this, the extension would be able to take a screenshot of the page that is running and it would include the auto filled content. This method is used to steal information in plain text like credit card numbers, usernames, date of births, etc. In order to make the iframe stealthier than the previous attack, we could make the iframe meticulously transparent. Another option instead of making the iframe transparent is to change the size of the iframe in making it really small, so that the user would not be able to notice it. We can make the iframe show a single character at a time and move the field of view character by character until the data has been captured by using 'frame.contentWindow.scrollTo(xcoord,ycoord)'. As shown below in figure 2 and figure 3, we were able to hide an iframe on a target user's reddit page with the help of the malicious extension and it provided us with the login credentials which would help us find more information like addresses.



25 1797

Billy Joel and Jimmy Fallon Form 2-Man Doo

submitted 19 hours ago by NightHawk_88 to television

376 comments share

iew more: next > or try a random subreddit

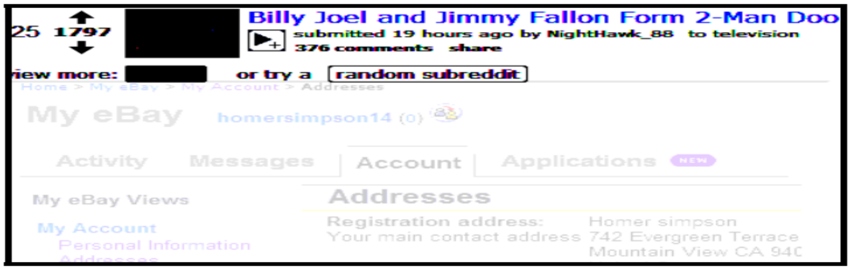
Account

Addresses

Registration address:

Your main contact address 742 Evergreen Terrace

Figure 2: Reddit Page with Transparent iframe on the Bottom



25 1797 [redacted] Billy Joel and Jimmy Fallon Form 2-Man Doo

submitted 19 hours ago by NightHawk_88 to television

376 comments share

iew more: [redacted] or try a random subreddit

Home > My eBay > My Account > Addresses

My eBay homersimpson14 (0)

Addresses

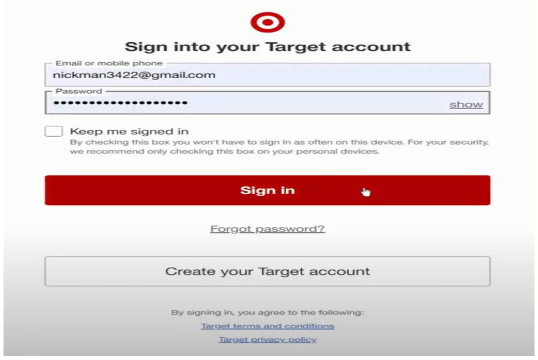
Registration address: Your main contact address: Homer simpson
742 Evergreen Terrace
Mountain View CA 94C

Figure 3: Reddit Page with Visible iframe on the Bottom

Occasionally, sensitive information like login credentials are only available after the user has logged into the target page. In order to get sensitive information, the extension has to mount the attack after the user has logged in but before the user's session expires. One of the ways that a malicious extension can detect the user has recently logged in is in the case where extensions have host permission to the designated page because it can observe the loading of the login page [37][38]. To add on, other permissions given to the extension like viewing the history of the browser and web Requests also lets the extension know that the user had recently logged in. Some of the limitations for these attack methods to work is that the data has to be in plain text. Furthermore, the websites that would be able to implement in an iframe does not include some of the top websites like Facebook, Twitter, and Amazon. Another limitation in these two methods dealing with iframes is that the user must enable or click on the malicious extension's icon in Google Chrome in order for the malicious extension to run or execute. Moreover, the user must already be on the website that the malicious extension plans to extract information.

3.3 FORGING USER DATA

The last attack vector in dealing with Chrome extensions relates to forging user data or web requests from the extensions itself so that it would appear it came from the user [39][40]. From a malicious extension, it can gather information on the user. It could even log into the targeted website and access the information like addresses or transaction history provided that the user is logged in as well. The amount of information or data that can be extracted also depends on the implementation of the website's login process itself. In most cases, it would be difficult if the website has two factor authentication. Figure 4 shows an example of how an extension can take the autofill information and translate it for us.



Sign into your Target account

Email or mobile phone: nickman3422@gmail.com

Password: •••••••••••••••••••• show

□ Keep me signed in

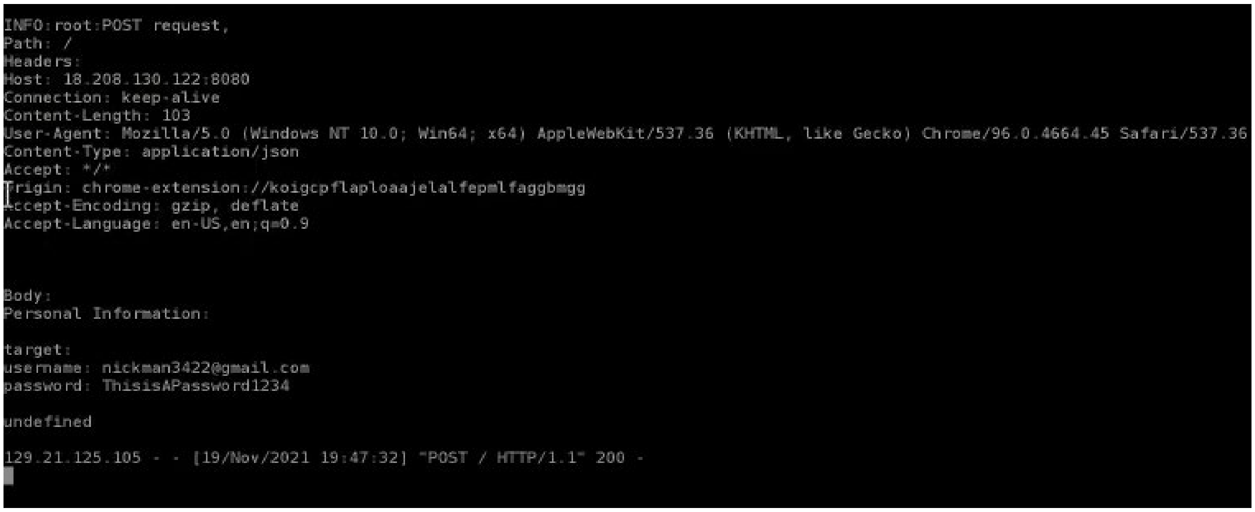
By checking this box you won't have to sign in as often on this device. For your security, we recommend only checking this box on your personal devices.

By signing in, you agree to the following:

Target terms and conditions

Target privacy policy

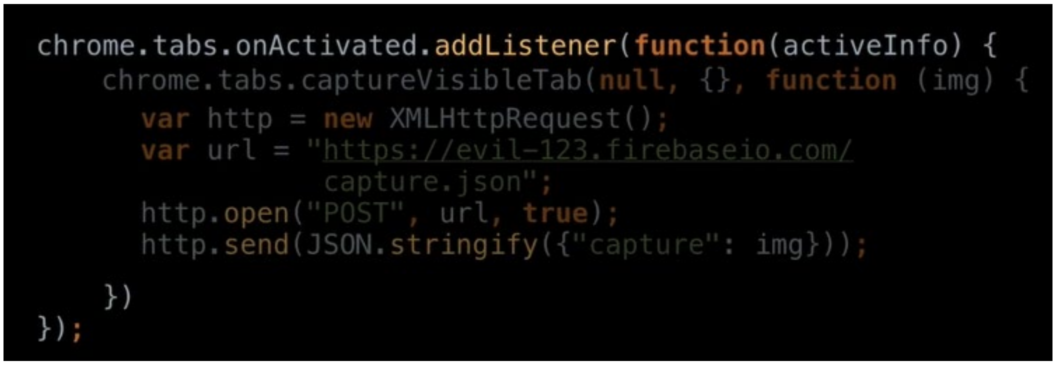
Figure 4: Target Login Page



```
INFO:root:POST request,
Path: /
Headers:
Host: 18.208.130.122:8080
Connection: keep-alive
Content-Length: 103
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/96.0.4664.45 Safari/537.36
Content-Type: application/json
Accept: */*
Origin: chrome-extension://koigcpflaploaajelalfepmlfaggbmgg
Accept-Encoding: gzip, deflate
Accept-Language: en-US,en;q=0.9


Body:
Personal Information:

target:
username: nickman3422@gmail.com
password: ThisisAPassword1234

undefined

129.21.125.105 - - [19/Nov/2021 19:47:32] "POST / HTTP/1.1" 200 -
```

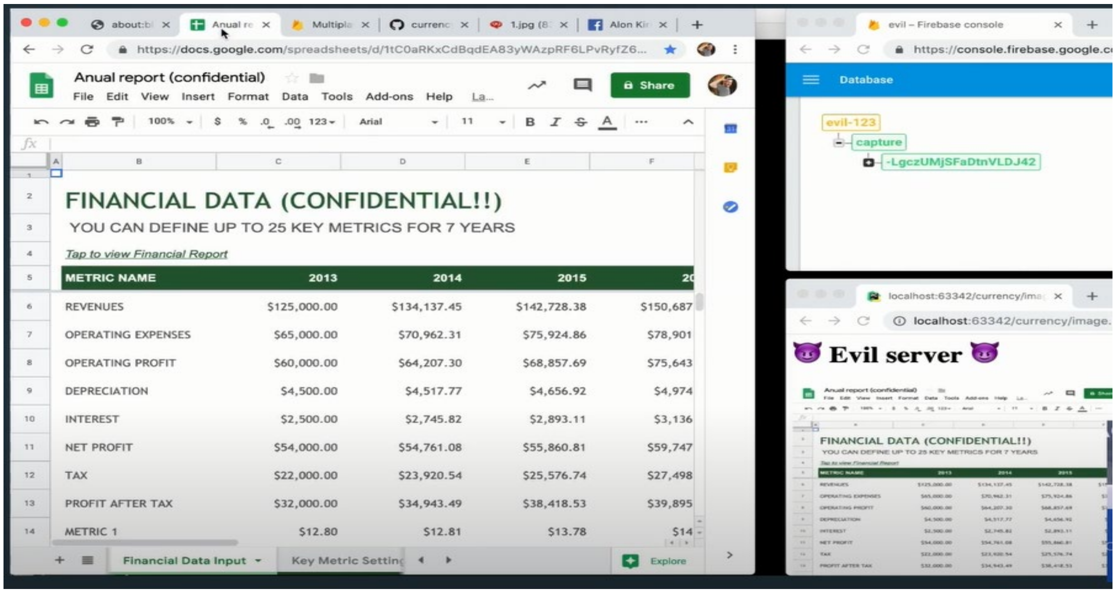
Figure 5: Console View of Captured Credentials

As seen in the figure 5 above, the malicious extension was able to log in to “Target.com” with the identity of the user whose username and password have been auto-filled by the browser or a password-manager extension. With this malicious extension, more damage can be done than just taking the username and password. One of them includes data integrity attacks when changing passwords since we can modify the extension to log in with an incorrect password forcing the user to be locked out. To add on, we can further the damage by making the user reset the password since the old password wouldn’t work and in turn, it would create an opportunity to get the answers for the password reset questions which could be then used for other websites of the user. Lastly, we would also be able to change banking operations such as transferring money with the login credentials. Another implementation of forging user data is through page captures where it would show us captured information of all the sites the user would visit. This is considered a big threat because whatever website the user visits, the attacker would be able to gain information and use it for malicious purposes. Figure 6 illustrations an example of the implemented code and how the extension would capture the webpage and provide the attacker the information.



```
chrome.tabs.onActivated.addListener(function(activeInfo) {
    chrome.tabs.captureVisibleTab(null, {}, function (img) {
      var http = new XMLHttpRequest();
      var url = "https://evil-123.firebaseio.com/
                 capture.json";
      http.open("POST", url, true);
      http.send(JSON.stringify({"capture": img}));

    })
});
```

Figure 6: Extension Code for Screen Capturing



about:bl
Anual re
Multipl
currenc
1.jpg (8
Alon Ki
https://docs.google.com/spreadsheets/d/1tC0aRKxCdBqdEA83yWAzpRF6LPvRyfZ6...
Anual report (confidential)
File Edit View Insert Format Data Tools Add-ons Help La...
Share
100%
Arial
11
Financial Data Input
Key Metric Settin
Explore
evil – Firebase console
https://console.firebase.google.c
Database
evil-123
capture
-LgczUMjSFaDtnVLDJ42
localhost:63342/currency/ima
localhost:63342/currency/image.
Evil server

FINANCIAL DATA (CONFIDENTIAL!!)

YOU CAN DEFINE UP TO 25 KEY METRICS FOR 7 YEARS

Tap to view Financial Report

METRIC NAME	2013	2014	2015	20
REVENUES	$125,000.00	$134,137.45	$142,728.38	$150,687
OPERATING EXPENSES	$65,000.00	$70,962.31	$75,924.86	$78,901
OPERATING PROFIT	$60,000.00	$64,207.30	$68,857.69	$75,643
DEPRECIATION	$4,500.00	$4,517.77	$4,656.92	$4,974
INTEREST	$2,500.00	$2,745.82	$2,893.11	$3,136
NET PROFIT	$54,000.00	$54,761.08	$55,860.81	$59,747
TAX	$22,000.00	$23,920.54	$25,576.74	$27,498
PROFIT AFTER TAX	$32,000.00	$34,943.49	$38,418.53	$39,895
METRIC 1	$12.80	$12.81	$13.78	$14

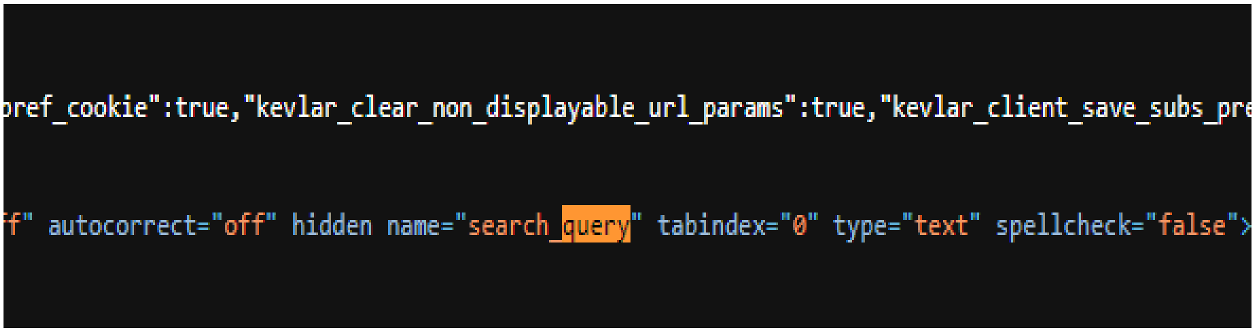
Figure 7: Application of Screen Capture with Malicious Extension

According to the figure 7 above, the page on the right shows the target's page and the websites that we would want to capture, whereas the page on the top right shows the Firebase console and the page on the bottom right shows a small gallery that is built on top of the Firebase console. With this, we can see that whatever website the target user visits, we would be able to capture the information and use it for malicious purposes as mentioned before.

4 RESULTS

From the attacks that we tested, there was a variety of success across the board. This was expected however, as we originally hypothesized that attacks would vary in their success rate and method. For the results, we'll go through each of the three attacks that we previously mentioned to see how successful each one was.

As far as the background tab attack, we expected this to be a go-to attack because of how straight forward it is. The development is simple as extensions using this attack have been around for a long time and aren't anything groundbreaking. Because of this, many of the extensions we wanted to test had already been banned due to being reported for this type of malicious activity. However, several extensions still exist out there that contain one of the attacks we plan to discuss in this project. Regarding our first kind of attack, the attack is pulled off by first checking a series of requirements for performing the attack. The attack takes place by finding tabs that Chrome has running in the background, and then gathering data like login information from those tabs without the user knowing. The malicious extension must first determine if Chrome has the "queryInfo active" flag. If the flag returns false, then the attack can continue. If the flag returns true, then the background tabs won't update, showing the information that the attacker is looking for.



```
pref_cookie":true,"kevlar_clear_non_displayable_url_params":true,"kevlar_client_save_subs_pre

f" autocorrect="off" hidden name="search_query" tabindex="0" type="text" spellcheck="false">
```

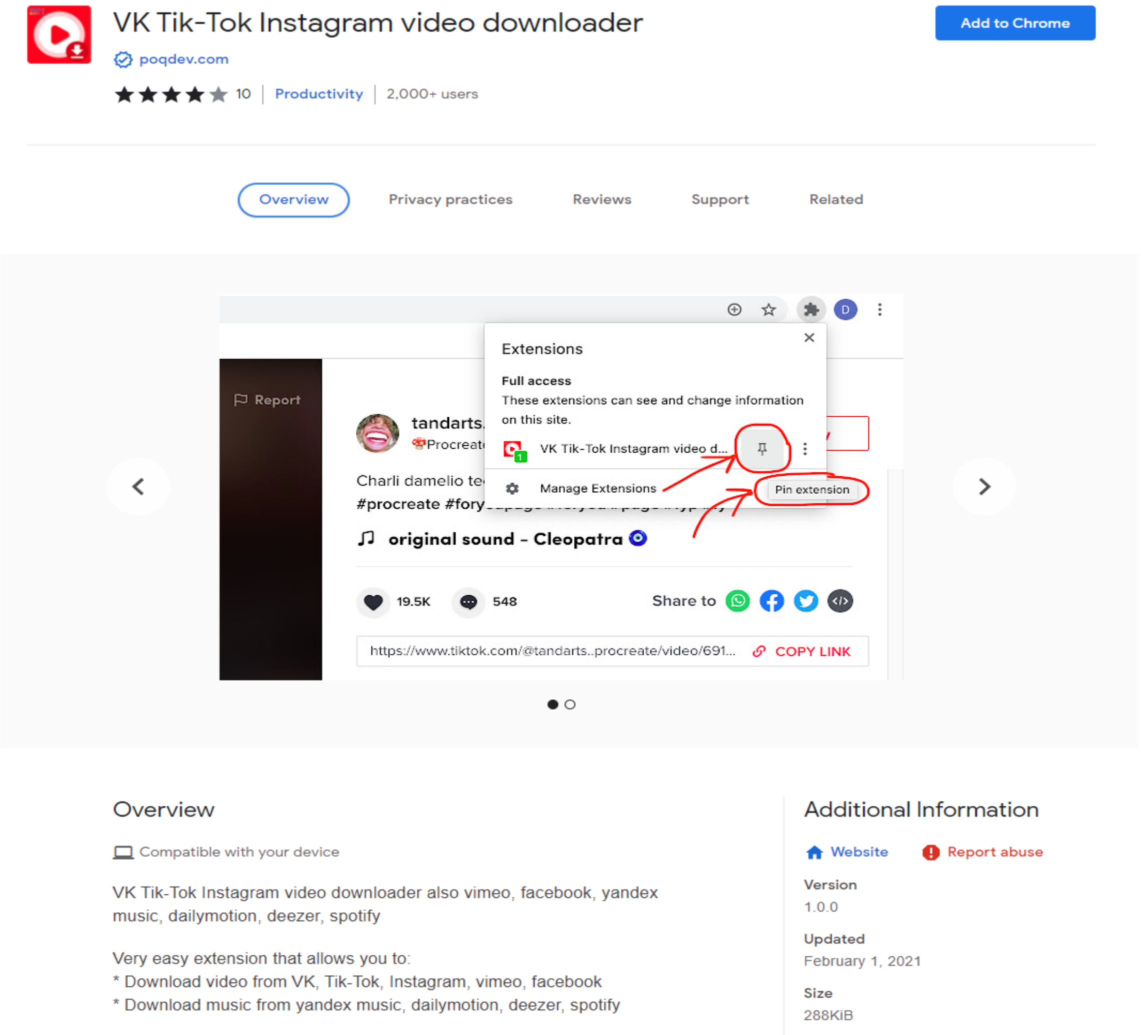
Figure 8: Malicious Extension with “QueryInfo Active” Flag

As shown on the figure 8, just by looking through the source information on a given website, we can see critical information about the options for that site. For most of the testing, we used popular websites, in this case Youtube.com. By checking the query flags, we can see that search queries on this website are hidden. This helps to hide information like search results from other tools like extensions. In this case, an extension without having a direct view of the window, wouldn’t be able to collect search queries on this site.

The downside of attacks using background tabs is the stealth element. The extension must first look to see if the tab is in the background, and then choose to start updating the tab using requests. This method is based on assuming the user can’t see the tab since it’s being covered by something else on the screen. However, obviously this isn’t always the case, and stealth is a major downside for this kind of attack since in some cases it’s easy to get caught. Because of that, we regard this attack as one of the most popular because of its simplicity and widespread use, however far from the most effective.

The next attack we investigated, was a far stealthier attack using iframes. The major advantage here is the ability to access information in a similar fashion with far less chances of the user noticing. This is done using iframes, which resemble small windows that can be opened separate from the original Chrome tab. The advantage here is the flexibility that these windows have. Just like a regular tab, these frames can request web pages and display data. However, these iframes can be altered to be extremely small, so that the user doesn’t see them, or even transparent. Because of this increased functionality, this kind of attack can be much more effective than utilizing background tabs. The methods of how this attack works remain relatively the same. Within this iframe, the attacker can use the extension to take screenshots and record information shown in these frames. Strangely enough, the success rate for this attack was the same for our previous type of attack. In fact, these use similar attack vectors making the attacks almost interchangeable. This would explain why the simple background tab attack is so rare now, since it’s been replaced by something with higher stealth, and the same success rate.

Looking at our final attack utilizing forged user information, we saw different results than expected. This type of attack has a guaranteed success rate if completed correctly, but it must be completed under specific circumstances. For this type of attack, the attacker simply uses stolen data to access restricted websites using login credentials. The extension is only a piece of spyware designed to look at what the user is doing, and any chrome extension you download would have this possible functionality. After seeing the user login, the attack would record the inputs they saw, and mimic those inputs later when they wanted to gain access. After digging through website HTML, and doing research on internet security, we were unable to find any flag options that would prevent this. The downside is the length of time an attack might have to wait before the user decides to login to something personal like social media or a bank account. Finding an existing chrome extension that performs this kind of attack was simple. There is plenty of documentation online giving links to malicious extensions for research purposes. Below was one that we found almost instantly.



VK Tik-Tok Instagram video downloader

poqdev.com

★★★★★ 10 | Productivity | 2,000+ users

Full access

These extensions can see and change information on this site.

Charli damelio te

#procreate #fory

https://www.tiktok.com/@tandarts..procreate/video/691...

Overview

Compatible with your device

VK Tik-Tok Instagram video downloader also vimeo, facebook, yandex music, dailymotion, deezer, spotify

Very easy extension that allows you to:
* Download video from VK, Tik-Tok, Instagram, vimeo, facebook
* Download music from yandex music, dailymotion, deezer, spotify

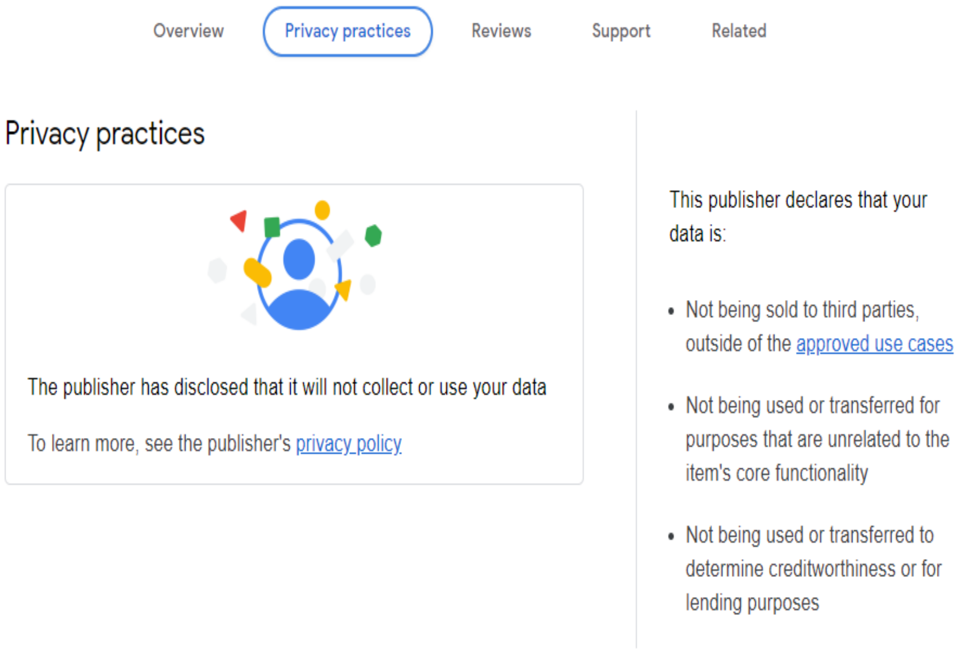
Additional Information

Version
1.0.0

Updated
February 1, 2021

Size
288KiB

Figure 9: Chrome Extension with VK Tik-Tok Instagram Downloader



Privacy practices

The publisher has disclosed that it will not collect or use your data

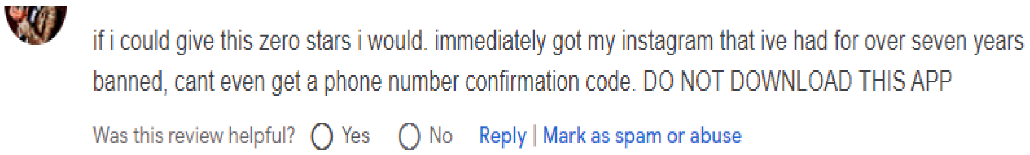
To learn more, see the publisher's privacy policy

This publisher declares that your data is:

- Not being sold to third parties, outside of the approved use cases
- Not being used or transferred for purposes that are unrelated to the item's core functionality
- Not being used or transferred to determine creditworthiness or for lending purposes

Figure 10: Chrome Privacy Practices

As you can see in the figure 9, the extension describes itself as a tool that can download videos from Tik-Toc and Instagram. And upon examining the privacy practices shown in the figure 10, we can see that it claims not to sell or collect any personal data for malicious purposes. However, looking at every other Chrome extension available on the Google store, we found that every single one has this exact same private policy page. This suggests that in order to even publish an extension, the contents of this page must be true. However, extensions are added to Chrome all the time, and Google doesn't have the resources to filter through each one, so developers can simply lie and claim their product to be harmless. Upon looking into the reviews, we found following as shown in the figure 11.


if i could give this zero stars i would. immediately got my instagram that ive had for over seven years banned, cant even get a phone number confirmation code. DO NOT DOWNLOAD THIS APP

Was this review helpful? ○ Yes ○ No Reply | Mark as spam or abuse

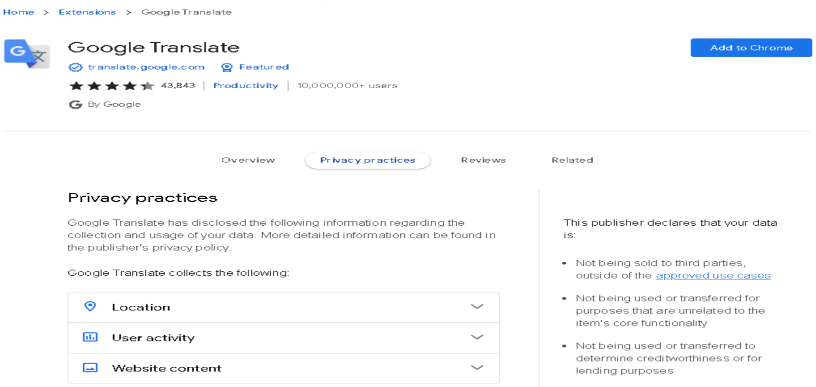
Figure 11: User Review

It's clear that this was a malicious extension, intended to steal password data for websites. The comment warns other users not to download this, as their accounts could get banned. These extensions exist all over the Google store, and when installing extensions, it's important to understand what you're downloading, before you download it.

5 CONCLUSION

Google Chrome's browser extensions can be either useful or malicious depending on the extension downloaded and applied. We have seen so far the capability of malicious Chrome extensions and the extent to what is available in terms of attack vectors in Google Chrome. We had tested out different ways in which a malicious extension can steal users' sensitive data, track their behavior, and forge their input. Even though these malicious extensions are not as popular as other extensions and would not be easily installed by a user willingly, many published extensions still have sufficient privileges to carry out these attacks. A benefit to our approach is that we were able to extract more information than expected from targeted websites. Furthermore, we were able to showcase different attack vectors with simple code to steal user information. Lastly, we were also able to steal this information discreetly without the awareness of the targeted user. The downside to our approach is that we were testing the implementations of these malicious extensions by installing them ourselves and executing them. If it were to apply to practical situations, then we would have to figure out a method in getting the user to install the extension willingly by themselves or we would have to install it on their Chrome browser. To add on, the implications of the extension would be stated on the extension's details and if the user were to read it and have an understanding, then they would be neglected to install the extension.

When installing an extension on Google Chrome or any browser, it is always suggested or required to read the extension details and see the permissions that are allowed for the extension. Moreover, Chrome also verifies and declares what privacy practices the extension would follow, and it would be necessary to read them if one wanted their browsing secure with the installed extensions. Figure 12 shows an example of the privacy practices for the extension Google Translate.



Privacy practices

Google Translate has disclosed the following information regarding the collection and usage of your data. More detailed information can be found in the publisher's privacy policy.

Google Translate collects the following:

- Location
- User activity
- Website content

This publisher declares that your data is:

- Not being sold to third parties, outside of the approved use cases
- Not being used or transferred for purposes that are unrelated to the item's core functionality
- Not being used or transferred to determine creditworthiness or for lending purposes

Figure 12: Privacy Practices of Google Translate Extension

Another method in ensuring that no malicious extension is installed is to enable Google Chrome's Enhanced Safe Browsing feature which is easily enabled by going to the browser's security settings, where one would discover three different degrees of protection being enhanced protection, standard protection, and no protection.

6 FUTURE WORKS

In the future we would attempt to see if the same attack vectors can be used across different web browsers like Mozilla Firefox and Safari. Through research, we found that different browsers have different methods in implementing their extensions. With this, we would like to test out how the attack vectors would work for these specified browsers. We would also delve into more attack vectors for Google Chrome itself. There are several different attacks that we could still attempt in order to steal user information. Some of them include implementing a key-logger or cross-site scripting. In addition to discovering more attack vectors, we would also like to figure out a method in making the user download or install the malicious extension willingly and without full awareness of its capabilities even after reading the details of the extension.

References

[1] Fass, Aurore, Dolière Francis Somé, Michael Backes, and Ben Stock. "Doublex: Statically detecting vulnerable data flows in browser extensions at scale." In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, pp. 1789-1804. 2021.

[2] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Security assurance of MongoDB in singularity LXCs: an elastic and convenient testbed using Linux containers to explore vulnerabilities." Cluster Computing 23 (2020): 1955-1971.

[3] Picazo-Sanchez, Pablo, Lara Ortiz-Martin, Gerardo Schneider, and Andrei Sabelfeld. "Are chrome extensions compliant with the spirit of least privilege?." International Journal of Information Security 21, no. 6 (2022): 1283-1297.

[4] Pantelaios, Nikolaos, Nick Nikiforakis, and Alexandros Kapravelos. "You've changed: Detecting malicious browser extensions through their update deltas." In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 477-491. 2020.

[5] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Vulnerability prioritization, root cause analysis, and mitigation of secure data analytic framework implemented with mongodb on singularity linux containers." In Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis, pp. 58-66. 2020.

[6] Zhang, Mingming, Xiaofeng Zheng, Kaiwen Shen, Ziqiao Kong, Chaoyi Lu, Yu Wang, Haixin Duan, Shuang Hao, Baojun Liu, and Min Yang. "Talking with familiar strangers: An empirical study on https context confusion attacks." In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 1939-1952. 2020.

[7] Kariryaa, Ankit, Gian-Luca Savino, Carolin Stellmacher, and Johannes Schöning. "Understanding users' knowledge about the privacy and security of browser extensions." USENIX, 2021.

[8] Picazo-Sanchez, Pablo, Lara Ortiz-Martin, Gerardo Schneider, and Andrei Sabelfeld. "Are chrome extensions compliant with the spirit of least privilege?." International Journal of Information Security 21, no. 6 (2022): 1283-1297.

[9] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Dynamic & portable vulnerability assessment testbed with Linux containers to ensure the security of MongoDB in Singularity LXCs." In Companion Conference of the Supercomputing-2018 (SC18). 2018.

[10] Sapkota, Bhumika, and Akalanka B. Mailewa. "A Scalable Framework to Detect, Analyze, and Prevent Security Vulnerabilities in Enterprise Software-Defined Networks." Journal homepage: www. ijrpr. com ISSN 2582: 7421. (DOI:10.55248/gengpi.2022.3.2.1)

[11] Agarwal, Shubham, and Ben Stock. "First, Do No Harm: Studying the manipulation of security headers in browser extensions." In Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb). https://doi. org/10.14722/madweb. 2021.

[12] Mailewa, Akalanka, and Jayantha Herath. "Operating Systems Learning Environment with VMware" In The Midwest Instruction and Computing Symposium. Retrieved from http://www.micsymposium.org/mics2014/ProceedingsMICS_2014/mics2014_submission_14.pdf. 2014.

[13] Hiremath, Panchakshari N., Jack Armentrout, Son Vu, Tu N. Nguyen, Quang Tran Minh, and Phu H. Phung. "MyWebGuard: toward a user-oriented tool for security and privacy protection on the web." In Future Data and Security Engineering: 6th International Conference, FDSE 2019, Nha Trang City, Vietnam, November 27–29, 2019, Proceedings 6, pp. 506-525. Springer International Publishing, 2019.

[14] Iqbal, Junaid, Ratinder Kaur, and Natalia Stakhanova. "PoliDOM: Mitigation of DOM-XSS by detection and prevention of unauthorized DOM tampering." In Proceedings of the 14th International Conference on Availability, Reliability and Security, pp. 1-10. 2019.

[15] Shetty, Roshan Ramprasad, Akalanka Mailewa Dissanayaka, Susan Mengel, Lisa Gittner, Ravi Vadapalli, and Hafiz Khan. "Secure NoSQL based medical data processing and retrieval: the exposome project." In Companion Proceedings of the10th International Conference on Utility and Cloud Computing, pp. 99-105. 2017.

[16] Xie, Mengfei, Jianming Fu, Jia He, Chenke Luo, and Guojun Peng. "JTaint: finding privacy-leakage in chrome extensions." In Information Security and Privacy: 25th Australasian Conference, ACISP 2020, Perth, WA, Australia, November 30–December 2, 2020, Proceedings 25, pp. 563-583. Springer International Publishing, 2020.

[17] Solomos, Konstantinos, Panagiotis Ilia, Nick Nikiforakis, and Jason Polakis. "Escaping the Confines of Time: Continuous Browser Extension Fingerprinting Through Ephemeral Modifications." In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pp. 2675-2688. 2022.

[18] Bian, Yan, Dechao Ma, Qing Zou, and Weirui Yue. "A Multi-way Access Portal Website Construction Scheme." In 2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 589-592. IEEE, 2022.

[19] Jairu, Pankaj, and Akalanka B. Mailewa. "Network Anomaly Uncovering on CICIDS-2017 Dataset: A Supervised Artificial Intelligence Approach." In 2022 IEEE International Conference on Electro Information Technology (eIT), pp. 606-615. IEEE, May 2022. (DOI:10.1109/eIT53891.2022.9814045)

[20] Fass, Aurore, Dolière Francis Somé, Michael Backes, and Ben Stock. "Doublex: Statically detecting vulnerable data flows in browser extensions at scale." In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, pp. 1789-1804. 2021.

[21] Wang, Xinyu, Yuefeng Du, Cong Wang, Qian Wang, and Liming Fang. "Webenclave: protect web secrets from browser extensions with software enclave." IEEE Transactions on Dependable and Secure Computing 19, no. 5 (2021): 3055-3070.

[22] Kaja, Durga Venkata Sowmya, Yasmin Fatima, and Akalanka B. Mailewa. "Data integrity attacks in cloud computing: A review of identifying and protecting techniques." Journal homepage: www. ijrpr. com ISSN 2582 (2022): 7421. (DOI:10.55248/gengpi.2022.3.2.8)

[23] Mailewa, Akalanka, and Kyle Rozendaal. "A Novel Method for Moving Laterally and Discovering Malicious Lateral Movements in Windows Operating Systems: A Case Study." Advances in Technology (2022): 291-321, ISSN 2773-7098. (DOI:10.31357/ait.v2i3.5584)

[24] Diamantaris, Michalis, Elias P. Papadopoulos, Evangelos P. Markatos, Sotiris Ioannidis, and Jason Polakis. "Reaper: real-time app analysis for augmenting the android permission system." In Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy, pp. 37-48. 2019.

[25] Razali, Muhammad Amirrudin, and Shafiza Mohd Shariff. "Cmblock: In-browser detection and prevention cryptojacking tool using blacklist and behavior-based detection method." In Advances in Visual Informatics: 6th International Visual Informatics Conference, IVIC 2019, Bangi, Malaysia, November 19–21, 2019, Proceedings 6, pp. 404-414. Springer International Publishing, 2019.

[26] Hajli, Nick, Farid Shirazi, Mina Tajvidi, and Nurul Huda. "Towards an understanding of privacy management architecture in big data: an experimental research." British Journal of Management 32, no. 2 (2021): 548-565.

[27] Mailewa Dissanayaka, Akalanka, Roshan Ramprasad Shetty, Samip Kothari, Susan Mengel, Lisa Gittner, and Ravi Vadapalli. "A review of MongoDB and singularity container security in regards to hipaa regulations." In Companion Proceedings of the10th International Conference on Utility and Cloud Computing, pp. 91-97. 2017.

[28] Sjösten, Alexander, Steven Van Acker, Pablo Picazo-Sanchez, and Andrei Sabelfeld. "Latex Gloves: Protecting Browser Extensions from Probing and Revelation Attacks." In NDSS. 2019.

[29] Khan, Muhammad Maaz Ali, Enow Nkongho Ehabe, and Akalanka B. Mailewa. "Discovering the Need for Information Assurance to Assure the End Users: Methodologies and Best Practices." In 2022 IEEE International Conference on Electro Information Technology (eIT), pp. 131-138. IEEE, May 2022. (DOI:10.1109/eIT53891.2022.9813791)

[30] Ghosal, Sandip, and R. K. Shyamasundar. "Preventing Privacy-Violating Information Flows in JavaScript Applications Using Dynamic Labelling." In Information Systems Security: 18th International Conference, ICISS 2022, Tirupati, India, December 16–20, 2022, Proceedings, pp. 202-219. Cham: Springer Nature Switzerland, 2022.

[31] Khan, Saad, and Akalanka B. Mailewa. "Discover Botnets in IoT Sensor Networks: A Lightweight Deep Learning Framework with Hybrid Self-Organizing Maps." Microprocessors and Microsystems (2023): 104753. (DOI: https://doi.org/10.1016/j.micpro.2022.104753)

[32] Gao, Yun, Kai Luo, Chongrong Fang, and Jianping He. "Fragility-Aware Stealthy Attack Strategy for Multi-Robot Systems against Multi-Hop Wireless Networks." In 2022 IEEE 61st Conference on Decision and Control (CDC), pp. 4827-4832. IEEE, 2022.

[33] Luo, Yukui, Cheng Gongye, Shaolei Ren, Yunsi Fei, and Xiaolin Xu. "Stealthy-shutdown: Practical remote power attacks in multi-tenant fpgas." In 2020 IEEE 38th International Conference on Computer Design (ICCD), pp. 545-552. IEEE, 2020.

[34] Mailewa, Akalanka, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Mechanisms and techniques to enhance the security of big data analytic framework with mongodb and Linux containers." Array 15 (2022): 100236. (DOI:10.1016/j.array.2022.100236)

[35] Chinprutthiwong, Phakpoom, Jianwei Huang, and Guofei Gu. "{SWAPP}: A New Programmable Playground for Web Application Security." In 31st USENIX Security Symposium (USENIX Security 22), pp. 2029-2046. 2022.

[36] Tramèr, Florian, Pascal Dupré, Gili Rusak, Giancarlo Pellegrino, and Dan Boneh. "Adversarial: Perceptual ad blocking meets adversarial machine learning." In Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, pp. 2005-2021. 2019.

[37] Gamnis, Steven, Matthew VanderLinden, and Akalanka Mailewa. "Analyzing Data Encryption Efficiencies for Secure Cloud Storages: A Case Study of Pcloud vs OneDrive vs Dropbox." Advances in Technology (2022): 79-98. (DOI:10.31357/ait.v2i1.5526)

[38] Lin, Xu, Panagiotis Ilia, and Jason Polakis. "Fill in the blanks: Empirical analysis of the privacy threats of browser form autofill." In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 507-519. 2020.

[39] Olaosebikan, Ayodeji, Thivanka PBM Dissanayaka, and Akalanka B. Mailewa. "Security & Privacy Comparison of NextCloud vs Dropbox: A Survey." In Midwest Instruction and Computing Symposium (MICS). 2022.

[40] Calzavara, Stefano, Alvise Rabitti, Alessio Ragazzo, and Michele Bugliesi. "Testing for integrity flaws in web sessions." In Computer Security–ESORICS 2019: 24th European Symposium on Research in Computer Security, Luxembourg, September 23–27, 2019, Proceedings, Part II 24, pp. 606-624. Springer International Publishing, 2019.

Survey on Security and Privacy of Cloud Computing Paradigm: Challenges and Mitigation Methods

Akhtar Hussain
akhtar.hussain@und.edu

Jun Liu
jun.liu@und.edu

Eunjin Kim
eunjin.kim@und.edu

School of Electrical Engineering and Computer Science
College of Engineering and Mines
University of North Dakota
Grand Forks, 58202

Abstract

The success of the Internet has significantly increased the volume of users and data. This has also raised the new requirements for accessing computational resources anywhere and at any time. Traditional computing infrastructure is difficult to meet the new requirements due to inflexible configurations and expensive maintenance and operations. Cloud computing provides a new paradigm of providing a large variety of computing services to large groups of users anywhere and at any time. While cloud computing emerged as a computing model that brought great deals of beneficial services, at the same time, it raised the possibility of risks. The most significant issues that this magnificent phenomenon faces are privacy and security that lead to illegal access of data, data leakage, the disclosure of confidential information, and privacy exposure. In this paper we will systematically assess and review the cloud security and privacy issues and their solution as well as present the systematic model of cloud computing and various types of security threats to this paradigm. Furthermore, this work will also discuss and analyze the data security and privacy protection for cloud storage. Moreover, our paper summarizes several new cryptographic technologies for security protection in cloud-computing paradigm, which include Attribute-Based Encryption (ABE), Homomorphic Encryption (HE), and Searchable Encryption (SE). Our paper also summarizes the open problems and future directions of security protection in cloud computing.

Keyword: Cloud Computing, Cloud actors, Security and Privacy, Encryption, Confidentiality, Security attacks and threats,

1. Overview of the system model of Cloud Computing

In the recent years, advancement in computing architecture and data processing mechanism has totally changed the computing paradigm. Due to this advancement, cloud computing system has turned out to be a necessity for external data storage and resource management. The success of global Internet has drastically increased the volume of users and new requirements for accessing computational resources anywhere and at any time. Traditional computing infrastructure is difficult to meet the new requirements due to inflexible configurations and expensive maintenance and operations. Cloud computing delivers data storage, processing power, databases, networking, and a large variety of software applications over the Internet with flexibility and reliability at much reduced costs. Cloud computing provides numerous advantages to both individuals and companies, particularly in terms of reducing capital expenses and cutting operational costs. By outsourcing on-premises computing resources to cloud service providers which provide quality guarantee on maintaining the operations of computing resources, users can be relieved from paying high procurement and maintenance costs and keeping a team of qualified IT professionals [1][2][26]. The procedure Pay-as-You-Go (PAYG) model gives the ability and flexibility to customize the computing resources, application, data storage, development platform as per the need of client [3]. The main aspects of cloud computing are manageability, scalability, and availability [1]. According to National Institute of Standards and Technology (NIST) [4], cloud computing is defined as a model for facilitating convenient, pervasive, on-demand network access to a shared pool of configurable computing resources (e.g., networks, services, applications, storage, and servers) that can be accessed and released efficiently with least managerial effort and with no interaction of service provider. This model consists of five characteristics, three service models, and four deployment models as summarized in Table 1.

Cloud computing model		
Service Models	Deployment Models	Essential characteristics
1. Software as a Service (SaaS) 2. Platform as a Service (PaaS). 3. Infrastructure as a Service (IaaS)	1. Private cloud 2. Community cloud 3. Public cloud 4. Hybrid cloud	1. On-demand self-service 2. Broad network access. 3. Resource pooling 4. Rapid elasticity 5. Measured service

Table 1 Cloud Computing Model

This paper is to discuss the cloud and security issues in cloud computing paradigm, for better understanding of security issues, it is necessary to understand the cloud computing service models first. Cloud service model is crucial because it lays down the foundation for comprehending the various responsibilities and risks connected to each service model. Different cloud service models give the cloud customer and the cloud provider distinct degrees of control and responsibility. Customers using the cloud can typically operate a variety of operating systems and applications in their virtual machines. Due to their possible size and complexity, the operating systems and applications used by cloud users may have security flaws[31]. Understanding these service models helps in identifying the different security concerns that arise and the corresponding measures that need to be taken. Identifying cloud security without explaining the cloud service model can therefore result in misunderstandings, poor communication, and insufficient security measures.

1.1 Cloud Computing Service Models

As per NIST [4] The basic service delivery models provided by cloud computing are software-as-service (SaaS), infrastructure-as-a-service (IaaS), and platform-as-a-service (PaaS). Depending on the IT requirements and budget of a company, each of the three models have a specific purpose. IaaS is most flexible of the three models. It gives complete control over a company's infrastructure. It is scalable and easily customizable. Computing capabilities, vital storage as standardized services is provided by IaaS. The main example of IaaS are Amazon Web services, Microsoft Azure, and Google Compute Engine (GCE). PaaS provides a layer of environment in which customers can develop their own application without installation underlaying development framework. PaaS also ensures the data protection using encrypting techniques while storing data on third party platform. AWS Elastic Beanstalk, Google App Engine, and Adobe Commerce, LAMP platform (Linux, Apache, MySQL, and PHP) are some examples of PaaS. Policy control management, access to application software and database are provided by SaaS. SaaS provides the ability for single service, that is available on the cloud, to be easily accessed by multiple users. SaaS services are provided by companies like Google, Microsoft Office 365, Dropbox etc.

1.2 Actors and their Roles in cloud computing

Five major actors with their functions and duties using the newly developed Cloud Computing Taxonomy are explained by NIST Cloud Computing Reference Architecture [5]. These contributing actors are explained below.

a) **Cloud Consumer** A principal stakeholder or group for cloud computing services that maintains to do business with cloud providers and makes use of their services. Activities and usage scenarios may differ from one another based on the services they require. A few examples of cloud services that a cloud user can choose from are shown in Figure1 [5].

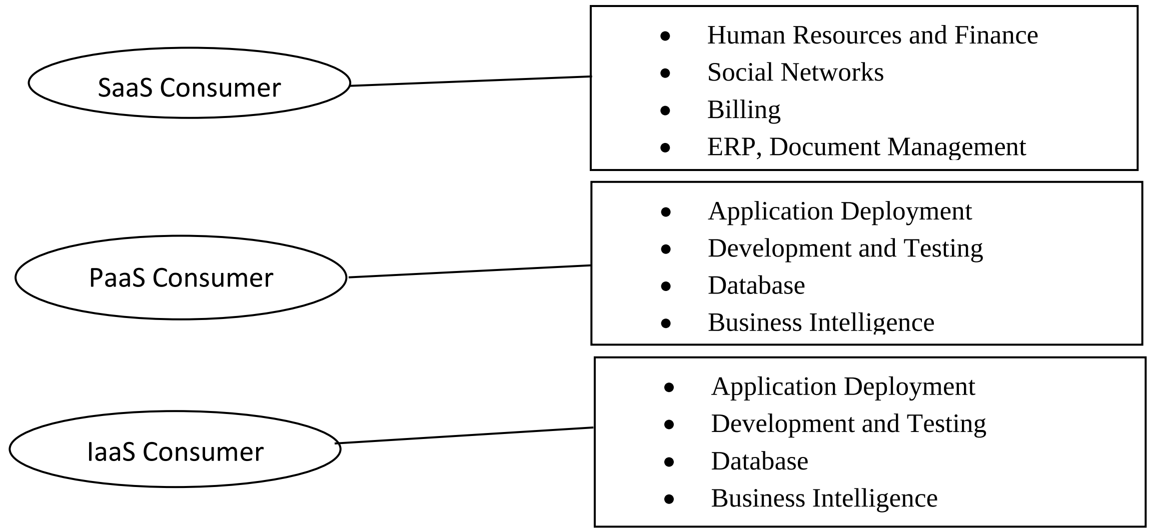


Figure1: Example Services Available to a Cloud Consumer

b) **Cloud Provider** A cloud provider could be an entity, person, or company that is liable for making services available to interested parties and provides the computing infrastructure necessary for offering the services that run the cloud software. The activities conducted by cloud provider are described in five major areas as shown in Figure 2 [5].

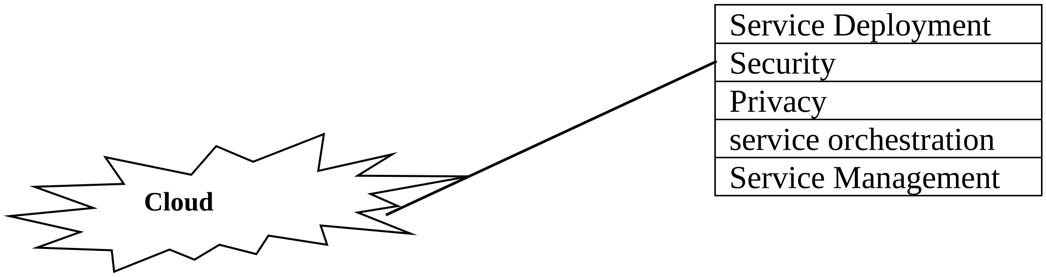


Figure 2: Major Activities of Cloud Provider

c) **Cloud Broker** A cloud broker is an individual that controls the use, execution, and distribution of cloud services and collaborates associations between cloud providers and cloud consumers. Instead of communicating directly to the cloud provider, a cloud consumer requests the services from a cloud broker. The services provided by a cloud broker are separated into three categories below.

- Service Intermediation: The cloud broker upgrades the service by improving some specific capabilities like performance reporting, enhanced security, identifying management, and offering value-added services to cloud consumers.
- Service Aggregation: Multiple services are combined and integrated to get new services. A cloud broker combines and integrates multiple services into one or more new services. Some other responsibilities of a cloud broker are to keep data movements secure and offering data integration.
- Service Arbitrage: Due to this, the cloud broker attains the flexibility to select services from multiple sources.

d) **Cloud Auditor** An independent entity which does evaluation of cloud services, its operation, and security of cloud execution. Interaction between a cloud provider and a cloud consumer may be involved by such assessment. Such audits assure the confidentiality, integrity, and availability of an individual's personal information at every step of creation and operation. This may assist Federal agencies complying with applicable privacy laws and regulations [6].

e) **Cloud Carrier** A cloud carrier behaves as an intermediary that offers connectivity and transfers cloud services between cloud consumers and cloud providers. Cloud providers participate in a way to have two service-level agreements: one with the cloud carrier and one with the cloud consumer. To ensure that the cloud services are used at a consistent level in accordance with the contractual responsibilities of the cloud consumers, a cloud provider negotiates service level agreements (SLAs) with a cloud carrier and may ask for resolved and encrypted connections.

2. Security Concerns in Cloud Computing

Cloud computing is an increasingly popular model for delivering and accessing IT resources over the internet, but at the same time, it also evolves security threats and attacks. Any possible risk to a computer system that could result in significant harm is referred to as a threat in the context of computer security. These threats lead to attacks. All the information and data must transfer through the network and be stored on the cloud, and malicious actors always try to manipulate different liabilities. Hardware and software components of cloud computing face serious threats like viruses, trojans, and inside and outside hackers that can lead to attacks on the whole cloud system [1]. Existing security solutions are not sufficient to secure the cloud infrastructure. In this section, major cloud computing attacks and threats will be explored. Some of the most common security attacks and threats in cloud computing are Spoofing, tampering, repudiation, information disclosure, denial of service, and elevation of privilege, or STRIDE for short, is a set of criteria developed in 1999 by Loren Kohnfelder and Praerit Garg to help companies spot potential weaknesses and threats to their products [9]. Each STRIDE class captures individual characteristics of attacks that represent a specific sort of threat [1]. Recent studies describe the current challenges and provide a useful roadmap for current cloud security. This research indicates that threats and related risks are becoming more prevalent. Below is a summary of the threats that this research highlights. While cloud computing emerged as a computing model that brought great deals of beneficial services, at the same time this valuable phenomenon suffers from both inside and outside attacks. Given that cloud computing is becoming more widespread, security attacks are apparent. Based on the Open Web Application Security Project (OWASP), attacks on the cloud [12] are classified in top-down mode.

2.1 Taxonomy of security threats targeting cloud computing

The use of cloud computing has become an imperative part of modern technology due to its advantages, such as its flexibility, scalability, and cost-effectiveness. Nevertheless, the transfer of sensitive data and applications to the cloud by numerous organizations also brings about security concerns. The potential sources of threats to cloud security are diverse and can originate from cybercriminals, insiders, or external causes. Threat taxonomy is essential to complete to deal with these security problems. This procedure involves locating, classifying, and ranking potential threats that might have an impact on the cloud environment. Threat classification allows organizations to prioritize their security efforts and distribute resources in accordance with the need to mitigate and prevent each type of threat. Common categories of cloud computing threats are related to network threats, user related security threats, software application threats and most importantly data related threats [32][1].

2.1.1 Network related Security Threats

Parallel to hardware and software applications, networks play a big role in cloud computing. Due to the difficulty of achieving end-to-end protection, cloud networks present a greater security challenge than conventional IT networks within organizational perimeter limits. The availability of the cloud may be impacted by an assault on the cloud network that reduces or intercepts network bandwidths. Customers who heavily rely on cloud services for their day-to-day company operations may experience widespread disadvantages as a result of the negative effect on cloud availability. For example, in the past GitHub was targeted by a DoS attack, causing widespread service outages. Cloud computing's network infrastructure is vulnerable to various attacks such as Distributed Denial of Service (DDoS), Man-in-the-Middle (MitM), and eavesdropping. DDoS attacks can cripple cloud services by overwhelming the network infrastructure with an enormous volume of traffic Significant threats to connection availability have been observed in network security, including denial of service (DoS), distributed denial of service (DDoS), flooding attacks, and Internet protocol weaknesses. The risk come from external users trying to launch a DoS assault on the network of the cloud service provider with the intention of blocking access to corporate and individual users' computer resources. Network administrators must therefore implement suitable security rules and make use of preventative tools and services to safeguard data and cloud infrastructure. Using firewalls is one of the most popular and efficient ways to stop these threats. There are external and internal network security threats, user related threats that can be performed on physical or virtual networks. Sensitive data is obtained from the businesses, processed by the SaaS application, and then stored at the SaaS vendor end in a SaaS deployment paradigm. To stop the leakage of sensitive data over the network, all data movement must be secured. To ensure security, this calls for the use of powerful network traffic encryption methods like Secure Socket Layer (SSL) and Transport Layer Security (TLS) [19][20][27][33][43].

- **Abuse of Functionality**
  To congest a network link or make cloud system to fail, attackers perform excessive malicious activities. For example, a denial-of-service attack that locks out genuine users by flooding a login system of web with legitimate usernames and random passwords [13]. The consequences of such attack are consuming resources, unauthorized access control and leakage of confidential information. Similarly, MitM attacks intercept data between the cloud provider and the user, making it possible for the attacker to access sensitive information. Eavesdropping, on the other hand, involves monitoring data traffic on the cloud network. Attackers can use this technique to steal sensitive information and exploit it for malicious purposes.
- **Spoofing attacks**
  A spoofing attack is a compilation of incidents in the field of cloud security, where a person or program efficiently mimics another to obtain an unethical advantage. The example of such attacks is DNS spoofing, IP spoofing, phishing, spoofing metadata by imitating a reliable email sender [18].

2.1.2 User related Security Threats

Security measures that cloud providers takes to safeguard the data of their consumers is stated as user centric security. That involves implementing security measures that are developed according to requirements and predilections of users, access control and encryption techniques. This includes implementing access controls, encryption, and other security measures that are tailored to the needs and preferences of individual users. Cloud computing providers can increase consumer confidence and guarantee the secure processing and storage of their clients' sensitive data by giving priority to user security. Malicious Insider, identity theft and unauthorized activities could be user related security threats and they accomplished by account hijacking and it is normally done by the stolen credentials by which the confidentiality, integrity, and availability of critical part of cloud services are compromised. Muti-factor authentication and data security platform such as end-to-end encryption can be used to avoid such hijacking threat. Authorization, authentication, and Identity and access management issues are main attributes of user-oriented security. These three attributes are explained below [34-38].

- **Authorization** The process by which a system decides what degree of access a specific authenticated user should have to secured resources under its control is known as authorization. To ensure that only authorized parties can interact with data in a cloud setting due to the increased number of entities and access points, authorization is essential to get access to databases, resources, and information systems and it is based on the responsibilities and permissions of the user. Various authorization control mechanisms are provided. These control models are DAC (Discretionary Access Control), MAC (Mandatory Access Control), RBAC (Role based access control), and ABAC (Attribute Based Access Control). These controls have advantages as well disadvantages.
- **Authentication** Determining a person's endorsement to conduct an action on data, such as reading or writing, is the act of doing so. Before engaging in the action, they are authorized to, users must authenticate. The cloud service provider presents and implements access control policies through the cloud environment, such as the services and resources should only be accessed by the authorized users . Authentication has two methods; they are physical security

mechanism and digital security mechanisms. Physical security mechanism includes retina recognition, face recognition and fingerprint recognition. Whereas digital security mechanism is simple credential like (username, passwords), Multifactor authentication and single sign-on (SSO).

- **Identification and Access Management** According to an IBM security framework for a typical company, identity and access management is one of the key security controls that should guide the organization's security policy. It should ensure that only legitimate users should be permitted entry to the corporate data that may be present across applications. Administrative, discovery, maintenance, policy enforcement, management, information sharing, and authentication tasks can all be handled by Identity and access Management. Identity and Access Management (IAM) validates the use of a single identity that is managed across all apps while also ensuring security. It is used to give or restrict access to data and other system resources as well as to authenticate users, devices, or services.

2.1.3 Software Application related Security

Software application is one the most vital component of cloud computing. Its significant importance makes application security as one of the most vulnerable areas of information security [1]. Software security is a major and crucial issue when creating a cloud system. It has numerous security flaws, such as implementation errors, buffer overflows, flaws in the way it was built, broken error handling promises, and more [7]. Many application developers today use programming languages with built-in classes and methods that have a variety of security flaws. like HTML/CSS/PHP/JS to mitigate injection masked code. Similarly, backend application's weaknesses are abused by SQL injection. To prevent such concerns and tackle application related security challenges there is need to train and make it to necessary for developers to concentrate on some areas like encryption identity management services, authentication services, and identity and access management services [21].

- **Hacked interface and application program interfaces** other threat which compromised security of cloud is called Hacked interface and application program interfaces. As the API is main entrance point for cloud customer and it help to hack the interface. Regular software patch update could only prevent from such attack [1].
- **Elevation of Privilege** In such threat a user exploits a buffer overflow to take control of the cloud system at the core level. An attacker who can breach all system defenses and enter the trusted system itself. The attackers get the elevated access privileges to secured resources. This is done by manipulating configuration fault in any application, design defect, bug, or system trickle [11][7]. With proper quality assurance check on implanted security techniques before deployment of cloud system could avoid such threat.
- **Buffer Overflow attack** Buffer overflows are a frequent occurrence in today's cloud systems, and vulnerability is created when memory close to a buffer is overwritten, which shouldn't be done either on purpose or carelessly in a program. Such attacks usually target to eliminate memory, which includes elements like the stack that hold local variables like those used as arguments and parameters inside of methods. Buffer overflow attacks should be avoided by risk managers by eliminating and distinguishing them before the software system is employed in cloud computing system [14].
- **Embed malicious code attack** One type of web-based assault is called a malware injection attack, in which attackers take lead of shortcomings in a cloud-based web application

by embedding malicious code that modifies the way the application typically runs. The malicious code may visibly undermine the application's security. It is always feasible for a developer to add malicious code with the goal of compromising an application's security, either now or in the future. Target cloud service models SaaS, PaaS, and IaaS are each infected with a malicious application, program, and virtual computer by hackers [15].

- **Malware injection attacks** SQL injection and cross-site scripting (XXS) attacks are the two main categories of malware injection attacks that are most repeatedly used to manipulate online application exposures in cloud computing. The most frequent assault for obtaining data from user cookies is cross-site scripting, which can result in a security issue. The primary target of SQL injection attacks is SQL servers hosting weak database apps. Generally, this attack launched with the help of multiple bots that are equipped with a SQL injection kit to fire a SQL injection attack and if the attacker launches it successfully then attacker can remotely retrieve sensitive data, manipulate the content of the database, and by executing the system commands take the control of the web server [16].

2.1.4 Data related Security threats

Cloud computing has transformed the way businesses store, process, and access data. However, it has also brought along new security challenges. One of the biggest concerns is the risk of data-related security threats. Malicious actors may try to exploit susceptibilities in the cloud infrastructure or steal login credentials to gain illegal access to data. In addition, cloud providers may face insider threats from employees who have access to confidential information. To mitigate these risks, businesses need to implement strong security measures such as encryption, access control, and regular monitoring and auditing of their cloud environments. It is also crucial to choose a trustworthy and dependable cloud service provider that follows best procedures for data security. Since cloud computing requires the storage and processing of sensitive information in remote servers, it can be susceptible to attacks such as data breaches, theft, loss, and reputational loss and disclosure of customer data.

- **Data Breaches** First important threat which is one of the critical for cloud customer is data breaches. In that threat personal critical information like credit card number, social security number etc. could be sneaked, viewed, or released to unauthorized users. A data spill or data leak is another name for a data breach. From a security perspective, data leakage has emerged as one of the biggest organizational dangers. Data breaches can be avoided by apply basic security measures like administering susceptibility, penetration testing and by applying strong protection against malwares in addition to strong password implementation. Sometime encryption technique could also prevent the data from thread actors [1][28].
- **Data Manipulation** Another issue that can occur when moving data to and from the cloud is data manipulation. This requires data insertion, alteration, and data removal. This compromises the availability, integrity, auditability of security. This type of threat can be managed by the simple methodology is to encrypt and/or sign all data that is being transferred backward and forward [7][10]. A provider may keep extra copies of the data fraudulently to sell them to interested third parties. The data leakage impacts the web application and attacker take off benefit of configured permission in cloud operations [7]. Man-In-The-Middle Cryptographic Attacks, Brute Force Attacks, Dictionary Attack imply to as probabilistic-based attacks which consider the possibility that attack would be successful. The attackers used different statistical and analytical tools to exploit the weak cryptographic cloud system [17].

- **Reputational loss and disclosure of customer data** Reputational loss and disclosure of customer data are Significant risks to customer and provider that are caused by a threat called Distributed denial-of-service attacks (DDoS). Source such threats are large number of internet bots which attach the cloud platform altogether and make the denial-of-service situation. Such threats could be avoided by deploying proper denial of-service response plan and proper management plan. The major cause of such threat happens when proper log mechanism for user's action on application or system is not implemented properly. So, it creates a chance that a person will commit a crime in a system that cannot track them. These threat compromises the Auditability, Trust Privacy, Cryptography [1][7]. Proper log tracking system can avoid this threat.

3. Overview of data security and privacy issues in cloud storage system

This study seeks to address the issues covering the security and privacy of cloud storage after reviewing the basic framework of cloud computing and the typical security risks associated with it in earlier sections. Cloud storage systems pose a number of data security and privacy concerns that must be resolved in order to maintain the protection of confidential data. Organization and users of cloud storage take the data security as important concern. Data access through a storage application for cloud computing must be highly available, while high speed and maximum scalability must also be maintained. Users moved to cloud data storage due to huge pool of shared resources provided by it. Due to the nature of cloud storage, problems with data security and privacy are unavoidably created during this process. Data storage provided by cloud storage providers offer this service with a guarantee of security Confidentiality, integrity, and Availability [8][25]. Cloud storage systems highlight a number of data security and privacy concerns that must be resolved in order to maintain the protection of private data. Data storage and its security becomes a most prominent issue after the active migration of government's departments, enterprises, and individual users. The protection of data from unauthorized modification, addition, or deletion in information system is critical [1]. The assurance integrity, confidentiality and availability of data can be achieved by ACID property. Atomicity, consistency, isolation, and durability are all abbreviated as ACID. The primarily challenges to maintaining data security and anonymity in cloud storage systems are as follows [22]:

- **Security Provider** Many customers are afraid about how effortlessly hackers and criminals can get into distant data. Cloud service providers pay special consideration to this challenge and devote a lot of resources to confront it.
- **Privacy Preserving** Virtual computing is utilized in cloud computing an data is spread over multiple virtual centers, due to that various legal systems will have differences over data privacy protection
- **Rights of Ownership** After data is transported to the cloud, some people are worried that they will lose their rights or won't be able to sustain the rights of their clients. Such issue could be resolved by well-experienced user-sided contracts.
- **Data Mobility** Data transportability is very high with cloud computing. Customers may not always be informed of where their data is situated.
- **Multiplatform Support** How the cloud-based service incorporates across numerous platforms and operating systems, such as Linux, Windows, OS X, and thin clients, is more of a problem for IT teams using managed services. The need for multiplatform assistance will decrease as more user interfaces move around to the web.

- **Recovery of Data** Cloud storage systems rely on complex infrastructure and may suffer from hardware or software failures, leading to data loss. Data should be backed up so it can be retrieved in the future to prevent this. Users of the cloud can maintain an offline backup of crucial data.
- **Data Portability and Transition** Some cloud users be concerned that if they shift service providers, their data may be difficult to transfer. Data conversion and porting rely heavily on the type of data retrieval format used by the cloud provider, especially when that format is incomprehensible.

3.1 Data encryption technologies and data protection method

Massive data generation and outsourcing of data to the cloud make the data insecure. An effective technique is used to protect the data is called encryption. That encode the data into other form by using some cipher algorithms. Three main aspects of cloud storage challenges are confidentiality, integrity, and availability. Data confidentiality refers to preventing active attacks on users' data by unauthorized parties. The data receiver complies exactly with the information transmitted by the sender. The reliability of the data is known as data integrity i.e., the data cannot be altered at choice. The term "data availability" highlights the ease with which users can access, download, or modify data at any time as soon as they require it, the cloud. Other than these other requirements of data security are Fine- Grained Access Control, Secure Data Sharing in Dynamic Group, Leaking Resistant, Completely Data Deletion.

At this point, encryption is still the primary remedy for cloud computing's problems with data security. The part that follows introduces some encryption technologies that are frequently used in cloud storage systems [23][29][30].

3.1.1 Identity-Based Encryption

Identity-Based Encryption (IBE) public-key cryptographic method that empowers users to encrypt and decrypt data using an identifier as the public key, such as an email address or a username. In traditional public-key cryptography, users must obtain a public key certificate from a trusted third party or certificate authority (CA) before they can use public-key encryption. However, with IBE, a user's identity can act as their public key, eradicating the need for a CA and making easier the key management process. In an IBE system, a trusted entity called the Private Key Generator (PKG) generates a master secret key and public parameters. The PKG uses the master secret key to generate private keys for each user in the system based on their identity. To encrypt a message for a user, the sender obtains the user's public parameters, which are typically available in a public directory, and uses them to encrypt the message. The recipient can then use their private key, which is derived from their identity and the public parameters, to decrypt the message. It also enables more flexible access control and allows for fine-grained encryption based on the identity of the user. However, IBE also has some security risks, such as the possibility of a PKG compromising the system by generating private keys for unauthorized users. Therefore, careful consideration of the security risks and appropriate safeguards should be taken when implementing an IBE system.

In conventional Public Key Infrastructure (PKI) process there is weakness that enhanced the workload of sender when it shares its data with multiple receivers. To resolve this shortcoming the concept of IBE was introduced. The concept is to link the user's identity. The basic idea of IBE is illustrated in scenario, When Alice sends an email to Bob at b@ho.com, she simply encrypts her communication using the public key string "b@ho.com". Alice does not need to obtain Bob's public key certificate. After receiving the encrypted message, Bob contacts a third-party Private Key Generator (PKG) and authenticates himself to it in the same way as he would to a Certificate

Authority (CA). The PKG generates Bob's private key, allowing him to decrypt and read the email. Notably, Alice can transmit encrypted email to Bob even if he has not yet configured his public key certificate, unlike the current secure email infrastructure. [24][2][39].

3.1.2 Attribute-Based Encryption

Attribute-Based Encryption (ABE) is enhanced version that replaces the identity of IBE with the set of attributes. ABE is based on user attributes; it is a form of public key encryption that enables users to encrypt and decrypt messages or data. ABE is a type of encryption method that grants access to encrypted data based on certain attributes or criteria rather than using traditional cryptographic keys. In ABE, data is encrypted by means of a set of attributes or policies that are defined by the owner of the data, rather than using a single key. ABE is a valuable tool for securing data in circumstances where traditional encryption methods may not be practical. For example, in a cloud computing environment, users may need access to data from multiple locations, and traditional encryption keys may not be sufficient for granting access to the data. With ABE, access to the data can be granted based on specific policies or attributes, making it easier to manage access control in complex environments. ABE works in four steps namely setup, key Generation, Encryption, Decryption phase. Firstly, relevant security parameters are entered, and the associated master key (MK) and public parameters (PK) are generated. The second step involves the data owner providing the system with their own attributes to acquire the private key related to those attributes. In the third stage, the data owner encrypts the data using his or her public key to produce the ciphertext (CT), which is then sent to the recipient or to a public cloud. Users of decryption finally receive ciphertext and can decrypt it using their own secret key. The data owner can designate who can access the encrypted data due to ABE's claim to offer fine-grained access control over encrypted files in data-sharing tools. Key-Policy Attribute-Based Encryption (KP-ABE) and Ciphertext-Policy Attribute-Based Encryption (CP-ABE) are the two major classes [2][38][40].

- **Key-Policy ABE (KP-ABE)**
  In KP-ABE, data is encrypted using a set of attributes or policies, and access to the data is conferred to a user who has a private key that matches the specified attributes or policies. This type of ABE is typically used for securing data in cloud environments or for data sharing applications.
- Ciphertext-Policy **Attribute-Based Encryption (CP-ABE)**
- In CP-ABE, data is encrypted using a set of attributes, and access to the data is granted to a user who has a set of attributes that match the attributes used to encrypt the data. This type of ABE is typically used for securing data in IoT applications or for securing communications between devices.

3.1.3 Homomorphic Encryption

Homomorphic Encryption was designed to overcome the concerns raised by IBE and ABE. Homomorphic encryption is a type of encryption that enables particular computations on ciphertexts to generate an encrypted result that, when decrypted, is indistinguishable to the outcome of operations carried out on the plaintexts. It effectively protects the security of data that is sent. The file is homomorphically encrypted by the data owner and sent to the cloud server. With the appropriate private keys, the authorized users can decrypt the ciphertext. Receiver simply needs to submit the functions that correspond to the operations to the cloud server if he wishes to perform certain operations on the ciphertext. This means that the data remains encrypted while computations are being performed, providing a high level of privacy and security. The security of data that is outsourced is adequately protected by homomorphic encryption. Homomorphic encryption could be enormously beneficial in circumstances where data confidentiality is crucial, for instance, in domains such as finance, healthcare, or government. On the other hand, the adoption of homomorphic encryption is confined by its high computational complexity and limited capabilities when compared to conventional encryption methods. However, ongoing research aims to enhance the efficiency and functionality of homomorphic encryption methods. There are three main types of homomorphic encryption [2][41].

- **Fully Homomorphic Encryption (FHE)**
  FHE is the most effective form of homomorphic encryption, which grants arbitrary calculations to be performed on ciphertext, including addition and multiplication.
- **Partially Homomorphic Encryption (PHE)**
  PHE grants only one type of computation to be performed on the ciphertext, either addition or multiplication.
- **Somewhat Homomorphic Encryption (SHE)**
  SHE is a settlement between FHE and PHE, allowing a limited number of calculations to be performed on the ciphertext.

3.1.4 Searchable Encryption

Normally data is uploaded on the cloud in encrypted form. To search the encrypted data over the cloud searchable encryption technique is used. SE (Searchable Encryption) is a good method for protecting users' confidential details while retaining server-side search functionality. The server can scan encrypted data using SE without disclosing any plaintext data. SE requires the encryption of the data and the formation of a searchable index over the encrypted data. There are several types of SE techniques, such as Symmetric Searchable Encryption (SSE), Public-key Searchable Encryption (PEKS), and Homomorphic Encryption (HE). SSE and PEKS agree to users to operate precise matching and range queries over encrypted data, whereas HE permits for more complex procedures, such as addition and multiplication over encrypted data. In contrast to PEKS, which allows multiple users who have access to the public key to produce ciphertexts, SSE only permits private key holders to produce ciphertexts and to construct trapdoors for search [2][42].

4. Direction of Future Research on Cloud Security

The popularity of cloud computing is anticipated to increase in the coming years as it has established itself as a crucial part of contemporary computing infrastructure. Research on cloud security is required as cloud computing develops further to guarantee the privacy, accuracy, and accessibility of data and services therein. A privacy protection system should be deployed to protect the private information that is embedded in shared data, particularly data containing highly private information like government and medical records. Recent years have also seen a significant increase in the popularity of modern machine learning and deep learning, particularly for image processing, DNA sequencing, and medical diagnosis. These algorithms require the creation and development of effective and secure outsourced data protection techniques. Future studies should focus on the problem of how to handle various data types using the concepts underlying various encryption techniques. How to find structured social network data that contains encrypted media data, such as an image or video data, for instance.

To protect against insider threats, cloud computing also requires a security solution. Numerous options still work with the cloud. But the insider danger cannot be resolved with the current solutions. Future studies could concentrate on creating mechanisms to guarantee the dependability and accountability of cloud-based services to boost trust in cloud computing. This might entail creating methods for auditing cloud-based services as well as building tools for service-level agreement enforcement. (SLAs). As cloud-based data analytics gain popularity, it is necessary to create methods for performing data analysis while protecting people's privacy. The development of data analytics algorithms that safeguard sensitive data while protecting privacy could be the main topic of future research.

There is a need for efficient intrusion detection methods in the cloud as cyber threats continue to change. Future studies might concentrate on creating real-time assault detection and response cloud-based intrusion detection systems. Deep learning methods gained popularity recently, and deep learning algorithms have made enormous advances in the fields of finance, defense, medical diagnosis, and academia. There is a need to develop techniques that can use these modern techniques while preserving the privacy of individuals, Future research could focus on developing privacy-persevering deep learning algorithms that protect sensitive data and help to identify threats and attacks in cloud computing. In addition, another important technology Blockchain could be ideal mitigation for many cloud security concerns due to its features such as immutability, accountability, efficiency, and privacy preservation. There are some research studies required that can address concerns related to malfunctioning machines, trust, accountability, compliance, integrity, and malicious insider behaviors [1][2][43].

5. Conclusion

Cloud computing provides a new paradigm of providing a large variety of computing services to large groups of users anywhere and at any time. While cloud computing emerged as computing model that brought great deals of beneficial services, however at the same time, this raises the possibility of risks. The most critical issues that this magnificent phenomenon faces are privacy and security that leads to illegal access of data, data leakage, the disclosure of confidential information, and privacy exposure. In this paper we reviewed the cloud security and privacy issues and their possible solution as well as present the systematic model of cloud computing and various types of security threats and attacks to this paradigm. Furthermore, we also discussed the data security and privacy protection for cloud storage. Moreover, and summarized several new cryptographic technologies for security protection in cloud-computing paradigm, which include Attribute-Based Encryption, Homomorphic Encryption, and Searchable Encryption. We also summarized open problems and future directions of security protection in cloud computing.

References

[1] Tabrizchi, H., & Kuchaki Rafsanjani, M. (2020). A survey on security challenges in cloud computing: issues, threats, and solutions. The journal of supercomputing, 76(12), 9493-9532.
[2] Yang, P., Xiong, N., & Ren, J. (2020). Data security and privacy protection for cloud storage: A survey. IEEE Access, 8, 131723-131740.
[3] Abdulsalam, Y. S., & Hedabou, M. (2022). Security and privacy in cloud computing: technical review. Future Internet, 14(1), 11.
[4] Mell, P., & Grance, T. (2011). The NIST definition of cloud computing
[5] Liu, F., Tong, J., Mao, J., Bohn, R., Messina, J., Badger, L., & Leaf, D. (2011). NIST cloud computing reference architecture. NIST special publication, 500(2011), 1-28.
[6] Chief Information Officers Council, "Privacy Recommendations for Cloud Computing", http://www.cio.gov/Documents/Privacy-Recommendations-Cloud-Computing-8-19-2010.docx

[7] Singh, S., Jeong, Y. S., & Park, J. H. (2016). A survey on cloud computing security: Issues, threats, and solutions. Journal of Network and Computer Applications, 75, 200-222.

[8] Prajapati, P., & Shah, P. (2022). A review on secure data deduplication: Cloud storage security issue. Journal of King Saud University-Computer and Information Sciences, 34(7), 3996-4007.

[9] Shevchenko, N. (2018, December 3). Threat Modeling: 12 Available Methods. Retrieved March 1, 2023, from https://doi.org/None.

[10] Lagesse, B. (2011, March). Challenges in securing the interface between the cloud and pervasive systems. In 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops) (pp. 106-110). IEEE.

[11] Yan, G., Wen, D., Olariu, S., & Weigle, M. C. (2012). Security challenges in vehicular cloud computing. IEEE Transactions on Intelligent Transportation Systems, 14(1), 284-294.

[12] Huang, Y. W., Huang, S. K., Lin, T. P., & Tsai, C. H. (2003, May). Web application security assessment by fault injection and behavior monitoring. In Proceedings of the 12th international conference on World Wide Web (pp. 148-159).

[13] Fernandes, G., Rodrigues, J. J., Carvalho, L. F., Al-Muhtadi, J. F., & Proença, M. L. (2019). A comprehensive survey on network anomaly detection. Telecommunication Systems,70,447-

[14] Alzahrani, S. M. (2021). Buffer Overflow Attack and Defense Techniques. Int. J. Comput. Sci. Netw. Secur, 21, 207-212.

[15] Chou, T. S. (2013). Security threats on cloud computing vulnerabilities. International Journal of Computer Science & Information Technology, 5(3), 79.

[16] Sharma, N., Alam, M., & Singh, M. (2015). Web based XSS and SQL attacks on cloud and mitigation. Journal of Computer Science Engineering and Software Testing, 1(2), 1-10.

[17] Murugan K, Suresh P (2018) Efcient anomaly intrusion detection using hybrid probabilistic techniques in wireless ad hoc network. Int J Netw Secur 20(4):730–737

[18] Gumaei A, Sammouda R, Al-Salman AMS, Alsanad A (2019) Anti-spoofng cloud-based multispectral biometric identifcation system for enterprise security and privacy-preservation. J Parallel Distrib Comput 124:27–40

[19] Somani G, Gaur MS, Sanghi D, Conti M, Buyya R (2017) DDoS attacks in cloud computing: issues, taxonomy, and future directions. Comput Commun 107:30–48

[20] Subashini, S., & Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. Journal of network and computer applications, 34(1), 1-11.

[21] Prokhorenko V, Choo K-KR, Ashman H (2016) Web application protection techniques: a taxonomy. J Netw Comput Appl 60:95–112

[22] Dinadayalan, P., Jegadeeswari, S., & Gnanambigai, D. (2014, February). Data security issues in cloud environment and solutions. In 2014 World Congress on Computing and Communication Technologies (pp. 88-91). IEEE.

[23] Yang, P., Xiong, N., & Ren, J. (2020). Data security and privacy protection for cloud storage: A survey. IEEE Access, 8, 131723-131740.

[24] Boneh, D., & Franklin, M. (2001, August). Identity-based encryption from the Weil pairing. In Advances in Cryptology—CRYPTO 2001: 21st Annual International Cryptology Conference, Santa Barbara, California, USA, August 19–23, 2001 Proceedings (pp. 213-229). Berlin, Heidelberg: Springer Berlin Heidelberg.

[25] Hashizume, K., Rosado, D. G., Fernández-Medina, E., & Fernandez, E. B. (2013). An analysis of security issues for cloud computing. Journal of internet services and applications, 4, 1-13.

[26] Subramanian, N., & Jeyaraj, A. (2018). Recent security challenges in cloud computing. Computers & Electrical Engineering, 71, 28-42.

[27] Agarwal, A., & Agarwal, A. (2011). The security risks associated with cloud computing. International Journal of Computer Applications in Engineering Sciences, 1(Special Issue on), 257-259.

[28] Barona, R., & Anita, E. M. (2017, April). A survey on data breach challenges in cloud computing security: Issues and threats. In 2017 International conference on circuit, power and computing technologies (ICCPCT) (pp. 1-8). IEEE.

[29] Liu, Y., Sun, Y. L., Ryoo, J., Rizvi, S., & Vasilakos, A. V. (2015). A survey of security and privacy challenges in cloud computing: solutions and future directions. Journal of Computing Science and Engineering, 9(3), 119-133.

[30] Seth, B., Dalal, S., Jaglan, V., Le, D. N., Mohan, S., & Srivastava, G. (2022). Integrating encryption techniques for secure data storage in the cloud. Transactions on Emerging Telecommunications Technologies, 33(4), e4108.

[31] Varadharajan, V., & Tupakula, U. (2014). Security as a service model for cloud environment. IEEE Transactions on network and Service management, 11(1), 60-75.

[32] Swathy Akshaya, M., & Padmavathi, G. (2019). Taxonomy of security attacks and risk assessment of cloud computing. In Advances in Big Data and Cloud Computing: Proceedings of ICBDCC18 (pp. 37-59). Springer Singapore.

[33] Xia, H., & Brustoloni, J. C. (2005, May). Hardening web browsers against man-in-the-middle and eavesdropping attacks. In Proceedings of the 14th international conference on World Wide Web (pp. 489-498).

[34] Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. Future Generation computer systems, 28(3), 583-592.

[35] Kumar, P. R., Raj, P. H., & Jelciana, P. (2018). Exploring data security issues and solutions in cloud computing. Procedia Computer Science, 125, 691-697.

[36] Indu, I., Anand, P. R., & Bhaskar, V. (2018). Identity and access management in cloud environment: Mechanisms and challenges. Engineering science and technology, an international journal, 21(4), 574-588.

[37]https://www.ibm.com/cloud/architecture/architectures/securityArchitecture/security-policygovernance-risk-compliance/

[38] Dubey, S., & Rai, P. K. (2021). A Review of Cloud Service Security with Various Access Control Methods.

[39] Anand, D., Khemchandani, V., & Sharma, R. K. (2013, September). Identity-based cryptography techniques and applications (a review). In 2013 5th international conference and computational intelligence and communication networks (pp. 343-348). IEEE.

[40] Goyal, V., Pandey, O., Sahai, A., & Waters, B. (2006, October). Attribute-based encryption for fine-grained access control of encrypted data. In Proceedings of the 13th ACM conference on Computer and communications security (pp. 89-98).

[41] Yi, X., Paulet, R., Bertino, E., Yi, X., Paulet, R., & Bertino, E. (2014). Homomorphic encryption (pp. 27-46). Springer International Publishing.

[42] Wang, Y., Wang, J., & Chen, X. (2016). Secure searchable encryption: a survey. Journal of communications and information networks, 1, 52-65.

[43] Akello, P., Beebe, N. L., & Choo, K. K. R. (2022). A literature survey of security issues in Cloud, Fog, and Edge IT infrastructure. Electronic Commerce Research, 1-35.

Survey on Security and Privacy Issues in Cloud-based Big Data Applications

Vedant Kharche
vedant.kharche@und.edu

Jun Liu
jun.liu@und.edu

School of Electrical Engineering and Computer Science
College of Engineering and Mines
University of North Dakota
Grand Forks, 58202

Abstract

Big data has emerged as a new paradigm for data applications and has attracted interest in many industry fields. The widespread usage of big data relies not only on the promising solutions and mechanisms of data analytics but also on security protection and privacy preservation. Big data are widely adopted in many security-sensitive sectors to support decision-making, which includes finance, marketing, politics, and healthcare systems. The five basic characteristics of Big Data (Volume, Variety, Velocity, Value, and Veracity) have been targeted raised security and privacy issues. It is critical to identify and analyze the various security and privacy issues targeting each of the five basic characteristics. Security protection for big data is even more important as distributed frameworks and cloud-based storage solutions become increasingly incorporated in big data applications. The integration of big data and cloud-based systems with smartphones has resulted in the development of numerous applications that generate real-time and sensitive data, raising concerns about data privacy. New security and privacy concerns arise with the adoption of Hadoop and MapReduce in storing and processing huge volumes of data. It is necessary to identify security and privacy issues associated with the essential components of the infrastructure used for hosting cloud-based Big Data applications, ranging from devices used for supporting data flows to access control policies. This paper aims to provide a survey on security and privacy issues associated with the five basic characteristics of Big Data applications, together with a detailed review of the new security and privacy changes for cloud-based Big Data applications.

1. Introduction

"Big data refers to large and complex datasets that typical software is inadequate for managing [1]. Big Data can be described based on five basic characteristics: Volume, Variety, Velocity, Value, and Veracity, which are commonly called 5V characteristics. 5Vs are typically used to characterize of Big Data as volume, velocity, variety, veracity, and value [2,3]. Volume is the size of data; velocity is the high speed of data; variety indicates heterogeneous data types and sources; veracity describes consistency and trustworthiness of data; and value provides outputs for gains from large data sets." Big Data can be generally described by 5 characteristics of volume, veracity, velocity, variety, and value. These characteristics significantly impact big data security. Before establishing security solutions, we must understand how these characteristics impact big data security by analyzing the challenges it poses:

1) Volume: The sheer volume of data in big data contexts makes implementing effective security controls difficult. Scalable security procedures that can be deployed consistently across the entire system are required to manage the security of such a massive amount of data.

2) Velocity: Because big data processing is sometimes required in real-time, it can be difficult to impose security measures without compromising performance. To keep up with the increasing data processing speeds, security measures must be quick and efficient.

3) Variety: Big data is derived from a wide range of sources, including organized and unstructured data. Protecting this various data involves different security methods based on the data type, which makes implementing a consistent security approach difficult.

4) Veracity: Big data is frequently plagued by data quality challenges such as mistakes, inconsistencies, and inaccuracies. This can make implementing security measures that rely on correct data, such as access control, difficult.

5) Value: The importance of big data makes it an appealing target for hackers. Protecting important data necessitates strong security procedures that guard against a wide range of security risks, including as cyberattacks, insider threats, and data breaches.

Due to the challenges posed by the characteristics of big data organizations need specific tools and technology such as Hadoop, Spark, NoSQL databases, and data warehouses that can efficiently store and process huge volumes of data to handle big data. These technologies offer distributed processing and storage capabilities that let businesses scale their data processing capacity as necessary. Big data is playing a vital role in business, healthcare, finance, and government, use big data in diverse ways. Big data is also driving innovation in industries, including artificial intelligence, machine learning, and data analytics. This innovation enables businesses to process, analyze, and process huge volumes of data in real time, allowing them to make data-driven decisions and helping businesses to better understand client behavior, run more efficiently, develop fresh goods and services, and make informed decisions. Ensuring appropriate security measures is crucial with the widespread use of big data technologies across various industries. These technologies allow organizations to collect, analyze and store vast amounts of data from sources such as mobile devices, social media apps, and IoT devices, including sensitive information like financial, health, personal identity, and company secrets. Breaches at different data lifecycle stages can lead to fraud, identity theft, and cyberattacks. Big data environments are complex and distributed, making data protection challenging. Enterprises must implement strong security measures, such as access controls, encryption, monitoring, and threat detection, to secure data confidentiality, integrity, and availability.

2. Security Challenges for Big data

Data, devices, networks, and cloud infrastructure all play a critical role in big data management and analysis. Data is at the heart of big data as insights are gained through collecting and analyzing large and complex datasets. This data is gathered in the form of structured or unstructured data from large number of devices such as mobile phones, computers, servers. The data collected from devices is transferred across networks, allowing organizations to share data between their teams. Big data technologies have also integrated cloud infrastructure as it provides scalable and flexible resources for storage, processing, and analysis, allowing organization to handle large volumes of data. Data, devices, networks, and cloud technologies all work in tandem to collect, store, process and analyze volumes of data, ultimately providing insights that can help organizations make informed decisions. It is, therefore, necessary to understand the various security challenges concerning these components.

2.1 Data Related Security Challenges

The primary goal of big data is to extracts insights from massive volumes of data. This leading to big data platforms holding vast volumes of sensitive data, which are appealing targets for hackers. Insiders with privileged access can either purposefully or inadvertently misuse or steal data. Big data systems must adhere to a variety of standards, including those governing data protection, privacy, and security. Many big data systems are cloud-based, which brings new security vulnerabilities such as data breaches, account hijacking, and denial of service assaults. Organizations must establish a complete security strategy that includes robust authentication and access control, encryption, monitoring and auditing, disaster recovery, and business continuity planning to handle these security problems.

2.2 Device Related Security Challenges

Massive volumes of data gathered from many sources, including sensors, IoT devices, and mobile devices, are processed and analyzed in big data settings. Device vulnerabilities, data privacy, authentication, access control, monitoring and auditing, and physical security are all security considerations in big data contexts. Cyberattacks, viruses, and other security dangers can befall devices utilized in big data contexts. It is vital to ensure the privacy and security of the data acquired by these devices to preserve user confidence and comply with data protection requirements. Unsecured devices can be used to gain unwanted access to the system, potentially leading to data breaches and other security concerns. It is critical to monitor and audit devices used in big data settings to detect and prevent security problems. Physical security of equipment used in big data contexts is crucial for preventing data theft, manipulation, or destruction. By addressing device security issues in big data settings, businesses may reduce the risk of security events while also protecting the confidentiality, integrity, and availability of their data.

2.3 Network Related Security Challenges

To move, store, and analyze the enormous volumes of data required for big data activities, a dependable and fast network is crucial. Finding insights and value in the data is impossible without a reliable and scalable network. As a result, network security considerations are important in large data contexts. Some network security considerations in big data contexts include network vulnerabilities, data privacy, authentication, and access control, monitoring and auditing, and cloud security. Because networks can be exposed to assaults, viruses, and other security concerns, network vulnerabilities are a major problem in big data contexts. In large data contexts, data transported over networks can be intercepted, hacked, or modified, making data privacy a key problem. Unsecured networks can be used to gain unwanted access to the system, potentially resulting in data breaches and other security concerns. In large data settings, monitoring and auditing network activity is critical for detecting and preventing security problems.

2.4 Cloud Integration Related Security Challenges

Cloud computing is becoming more popular for storing, processing, and analyzing large amounts of data. Its design, however, has specific security concerns that must be addressed to assure data privacy, confidentiality, and availability. One of the primary security issues in cloud computing is the possibility of data breaches. Because data is hosted on remote servers held by third-party cloud providers, attackers may be able to obtain unauthorized access to sensitive data via weaknesses in the cloud architecture or apps. Cloud companies also have access to consumer data, prompting issues about data ownership and management. Another major security risk is data privacy. Data sent to the cloud can be intercepted, hacked, or modified, possibly revealing critical information. Cloud computing also raises questions regarding access control. It is critical to ensure that access to data and cloud resources is correctly authorized and verified, with appropriate rights provided to individuals based on their roles and responsibilities. Maintaining the security of cloud infrastructure and services is crucial to ensuring data confidentiality, integrity, and availability in big data settings.

2.5 User Related Security Challenges

Big data security presents considerable difficulties for enterprises, especially when it comes to security threats associated with users. Insider attacks, which can happen when employees or contractors with access to sensitive data inadvertently or intentionally compromise security, are one of the most important user-related security challenges. This may occur because of data theft or the unintentional release of private information. Weak passwords pose a serious concern as well because they can be quickly cracked or deduced, providing illegal access to massive data platforms. Another popular strategy used by hackers to access large data systems is phishing assaults, which frequently fool users into disclosing important information or downloading malware. Those that purposefully or inadvertently divulge private information outside of the company may be causing data leakage.

3. Data Related Security Solutions

From a data perspective, securing big data entails safeguarding the data as well as the systems and procedures that handle and process the data. With the increasing amount of data being generated and processed, it is critical to protect sensitive data from unauthorized access or exposure. Organizations need to ensure that data is protected at all stages of the data lifecycle, from storage and transmission to processing and analysis. By using techniques such as data masking, data anonymization, data encryption, and access controls, organizations can reduce the risk of data breaches, protect sensitive information from cyber threats, and comply with regulatory requirements. Therefore, it is important to understand these techniques and how they can be implemented to ensure effective data-related security.

3.1 Access Control

Access control security solutions in big data environments are focused on ensuring that only authorized users have access to sensitive data. There are several mechanisms organizations can employ to ensure security. Big data systems must have reliable access control measures to guarantee their confidentiality, integrity, and availability. To make sure that only authorized individuals may access sensitive data and that unauthorized access attempts are immediately discovered and dealt with, access restrictions should be routinely evaluated and updated.

- **Role-Based Access Control (RBAC)**
  RBAC is a popular access control technique in which individuals are assigned roles based on their job duties and responsibilities. Each role relates to a set of permissions that govern the actions that a user may do within the system. RBAC is a versatile mechanism that may be controlled and scaled as the company expands.

- **Attribute-Based Access Control (ABAC)**
  ABAC is an access control method that leverages characteristics such as job title, department, or location to determine access capabilities. This technique enables businesses to set more detailed access controls based on specific qualities rather than just roles.

- **Discretionary Access Control (DAC)**
  A sort of access control method in which data owners or administrators have the power to give or prohibit access to their data. DAC is often utilized in smaller situations where data owners can successfully regulate access to their data.

- **Mandatory Access Control (MAC)**
  MAC is a type of access control technique in which access to data is regulated by security labels applied to the data and users. MAC is often utilized in government and military situations where tight security requirements are required.

In large data settings, a variety of access control measures should be utilized to guarantee that only authorized individuals have access to sensitive data. The technique utilized will be determined by the organization's security needs, the scale of the environment, and the complexity of the access control regulations.

3.2 Data Masking

Data masking and anonymization are two related approaches used in big data security to safeguard sensitive information while keeping its value for legitimate reasons. Both strategies are intended to safeguard the confidentiality and privacy of sensitive data by blurring or masking the actual data. Data masking is a technique for disguising or hiding sensitive data in a dataset. The purpose of data masking is to restrict unwanted access to sensitive information while retaining the data's overall usefulness. Data masking techniques include:

- **Substitution**
  With this strategy, sensitive data is replaced with non-sensitive data. For instance, substituting a social security number with a number produced at random. When the data is not required for analysis or processing, this strategy is beneficial.

- **Redaction**
  This approach includes limiting sensitive data from a dataset. For instance, eliminating all personally identifying information from an email message. When the data is not required for analysis or processing, this strategy is beneficial.

- **Format-Preserving Encryption**
  This technology encrypts sensitive data while retaining its original format. Encrypting a social security number, for example, while keeping its length and structure. This approach is useful when data must be handled or evaluated while remaining secure.

3.3 Data Anonymization

Anonymization is a process that includes eliminating or changing identifying information from a dataset. The purpose of anonymization is to make it difficult or impossible to identify individual users or sensitive information. Anonymization techniques include:

- **Generalization and Suppression**
  This mechanism reduces the amount of information that is shared or released and thus reduces the risk of breach of privacy. Generalization makes changes in the sensitive data by replacing it with less precise data while still maintaining the original meaning of the dataset. This reduces the information that is shared while still being useful to the application. In case of suppression the data is removed entirely when the application does not require extremely sensitive data for its analysis. Both generalization and suppression are used in various privacy-preserving techniques such as anonymization, differential privacy, and k-anonymity.

- **Perturbation**
  Perturbation refers to adding of noise or random data in a dataset. Perturbation allows the privacy to be preserved by adding noise or distortions while still being useful for data analysis. Some common methods of perturbation include rounding values, adding noise to data or aggregating data to higher level of abstraction.

- **Differential privacy**
  In differential privacy mechanisms Individual privacy is protected by adding random noise to the data in a controlled manner while maintaining the probability distribution of the data. This random noise makes determining the values of individual data items difficult for an attacker while still allowing meaningful analysis to be performed on the data.

3.4 Data Encryption

Data encryption is critical in big data security because it protects sensitive data from illegal access, theft, or change. Given the heightened danger of data breaches and cyber-attacks in today's data-driven society, encryption is critical for protecting sensitive data. Encryption adds an extra layer of protection to help prevent data breaches and protect organizations from reputational and financial harm.

- **Transparent Data Encryption**
  Some relational database management systems (RDBMS) employ transparent data encryption to encrypt data at rest. TDE encrypts data in database data files and safeguards it when the files are backed up. The database engine handles the encryption and decryption processes, so no modifications to the application or user interface are required. This implies that once TDE is configured, the application and user interface do not require any changes.

- **File-Level Encryption**
  Encrypting individual files or directories on a file system is referred to as file-level encryption. This can be accomplished using encryption software or tools, such as PGP (Pretty Good Privacy), which can encrypt and decode data using a public-private key pair. File-level encryption is an excellent choice for large data situations where data is stored in files, such as log files or CSV files. It allows you

to encrypt data at rest and secure specific files or directories with various encryption keys.

- **Application-level Encryption**
  Encrypting data within an application is what application-level encryption entails. This can be accomplished using encryption libraries or APIs (Application Programming Interfaces) like OpenSSL or Bouncy Castle, which give developers with encryption and decryption methods that can be implemented into their programs. Application-level encryption is a viable choice for large data settings that use bespoke programs to handle data. It enables developers to include encryption and decryption functions directly into their apps, adding another degree of protection.

- **Database-level Encryption**
  Encrypting certain columns or tables within a database is known as database-level encryption. This can be accomplished using database-specific encryption capabilities or third-party encryption solutions that connect with the database. Database-level encryption is an excellent choice for large data settings that contain sensitive data in databases. It enables the encryption of columns or tables, preserving important data even if the database is compromised.

4. Network Security Solutions

Network security solutions are critical in protecting big data systems. These solutions are intended to safeguard the network infrastructure, which consists of the hardware, software, and protocols that allow data to be delivered and received across the network. With the increasing reliance on computer networks and the internet for daily operations, the risk of cyber-attacks and data breaches has increased significantly. Organizations need to protect their networks from external and internal threats to ensure the confidentiality, integrity, and availability of their data. By implementing network security solutions such as firewall, network segmentation, IDS/IPS, and network access control, organizations can reduce the risk of cyber threats, prevent unauthorized access to their networks. To ensure successful network security solutions, it is important to understand these strategies and how to use them.

4.1 Firewall

A firewall is a network security device that is used to block unwanted network access. It serves as a firewall between the trusted internal network and untrustworthy external networks such as the Internet. Firewalls can be hardware- or software-based, and they operate by evaluating incoming and outgoing network traffic in accordance with pre-defined security standards. If the traffic does not comply with the security policy, it is denied or stopped. Several levels of security can be provided by firewalls. Simple firewalls merely allow or deny traffic based on IP addresses, but more sophisticated firewalls may scan packet content and do deep packet inspection. In addition to intrusion detection and prevention, content filtering, and antivirus protection, firewalls can provide other security functions.

4.2 Intrusion Detection and Prevention Systems (IDS/IPS)

IDS/IPS systems are intended to detect and block illegal network access. IDS systems monitor and analyze network traffic for signals of malicious activity, such as network scans, denial-of-service attacks, or efforts to exploit vulnerabilities. When an intrusion detection system detects an attack, it sends an alert to the network administrator. IPS systems go a step further by automatically identifying and blocking harmful traffic. IPS systems are more proactive than IDS systems because they may act automatically to prevent an attack from succeeding.

4.3 Network Segmentation

In big data network security, network segmentation is a crucial technique that involves dividing a network into smaller sub-networks or segments, each with its own set of security measures. By doing so, organizations can limit the attack surface and the impact of a security breach. One of the security technologies that can be used for network segmentation is a Virtual Private Network (VPN), which establishes a secure and encrypted tunnel between two endpoints, allowing two networks to communicate securely across the internet. VPNs can also be used to connect distant users to a business network or to link two geographically isolated networks. VPNs provide security by encrypting all network traffic, making it impossible for attackers to intercept and read the data. Other technologies such as VLANs, firewalls, and routers can also be used for network segmentation, ensuring that each network segment is separated from the others, and access between segments is governed by strict security standards.

4.4 Network Access Control

Network-related access control is an important aspect of big data security solutions, as it allows organizations to control access to their network and prevent unauthorized access, modification, or theft of data. Network access control (NAC) is a security tool which restricts access to a network depending on the user's identification, the device they are using, and whether they are following security guidelines. By limiting access to just conforming users and devices, NAC systems can be used to prevent illegal access to a network. NAC operates by defining rules that control network access. In accordance with these regulations, network permissions, device security settings, and user authentication requirements may all be necessary. The NAC solution will check a user's identification and device security status when they try to connect to the network to make sure they comply with the set policies. NAC solution can be implemented in different ways.

- **Endpoint NAC**
  Endpoint NAC solutions check the security status of devices attempting to connect to the network. These solutions can ensure that devices meet specific security requirements, such as having updated antivirus software, firewalls, or the latest security patches.

- **Network NAC**
  Network NAC solutions check the identity and security status of users attempting to connect to the network. These solutions can ensure that users have appropriate permissions to access specific network resources, and that they meet specific security requirements.

- **Hybrid NAC**
  Hybrid NAC solutions combine endpoint and network NAC capabilities to provide a comprehensive security solution. These solutions can ensure that both users and devices meet specific security requirements before being allowed to access the network.

5. Device Security Solutions

Solutions for device security are crucial for safeguarding the availability, confidentiality, and integrity of data held on devices. Implementing efficient device security solutions is essential given the rise in the number of devices used to store and process sensitive data. Device security can be ensured via a variety of methods, such as the usage of trust certificates, device management, and corruption-free software. Only trusted software is loaded on devices thanks to the usage of trust certificates, which are used to confirm the reliability and authenticity of software. Device management entails keeping track of devices to make sure they are current, correctly set up, and in line with security regulations. In order to guarantee that the software operating on devices is devoid of bugs, malware, or other vulnerabilities that could be exploited by attackers, it's crucial to use corruption-free software solutions. To effectively defend against cyber-attacks, firms must understand these tactics and how to put them into practice.

5.1 Trust Certificates

In a big data system consisting of many connected devices, it's crucial to make sure that each one is permitted and authenticated to access the data on the network. The identification of every device on the network can be confirmed using trust certificates. By authenticating and confirming the identification of devices on a network, these certificates guard against data breaches and unwanted access. When a device attempts to access data on the network, the device's trust certificate, which is specific to that device and provides information about the device's identification, is verified. Trust certificates can also be used to monitor network activities, regulate access to data on the network, and encrypt data sent between devices. When it comes to protecting devices in large data systems, trust certificates are a potent option. Trust certificates can aid in preventing unauthorized access, safeguarding against data breaches, and ensuring the integrity of the data on the network by offering authentication, encryption, access control, and monitoring capabilities.

5.2 Utilization of Corruption-Free Software

Maintaining security in a big data system depends on the devices' software not being corrupted. System outages, security breaches, and other security risks can be caused by corrupted software. Use software solutions that are created to be secure and free from corruption to prevent these issues. There are numerous ways to accomplish this. First off, selecting reputable software vendors can assist guarantee that the offered software is dependable and safe. Second, keeping software up to date can shield it against flaws and guarantee that the most recent security fixes are used. Thirdly, routine software testing can spot potential flaws or places where corruption might happen. Fourthly, putting security measures in place helps stop illegal access and safeguard sensitive data. These protocols include access control, encryption, and multi-factor authentication. By taking these measures, organizations can ensure the integrity of their big data system, prevent security breaches, and protect sensitive data.

5.3 Device Management Policies

In big data systems, device management policies ensure devices are managed, updated, and secured properly through centralized management, access control, monitoring, patch management, and encryption policies. These policies prevent security breaches, protect sensitive data, and ensure system integrity by controlling devices, tracking device behavior, and constantly updating security patches and software. Organizations can enforce these policies to stop unauthorized users from accessing critical data and prevent potential security holes and data loss. By doing so, they can guarantee uniform application of security requirements and comply with legal requirements.

6. Distributed Integration Oriented Security Solutions

In order to guarantee the security and protection of data that is shared across numerous systems and networks, distributed integration-oriented security solutions are crucial. The demand for efficient security solutions has grown more critical than ever with the rise of distributed systems and cloud computing. Data governance regulations, monitoring, and logging are a few of the strategies that can be used to guarantee the security of distributed systems. Organizations may manage their data efficiently and identify and address possible security issues in real-time by putting these approaches into practice. By implementing these techniques, organizations can establish clear policies for managing data, monitor and analyze data activity to identify potential security risks, and respond promptly to any security incidents.

6.1 Data Governance Policies

Effective management of data assets in a distributed environment requires a strong data governance policy. This policy should cover the management of data assets, including policies and processes, to ensure responsible data maintenance and access. To maintain data accuracy, completeness, and consistency, compliance rules must be created and adhered to in accordance with relevant laws, regulations, and industry standards. To achieve this, collaboration across multiple teams and departments and the use of technology tools to monitor and enforce policies is essential. A comprehensive data governance strategy can help ensure the successful use of data assets in a distributed environment and mitigate the risks of data breaches.

6.2 Monitoring and Logging

Monitoring and auditing systems are crucial for maintaining the security of a distributed environment. These systems help to detect and respond to security breaches, track network activity, and ensure compliance with regulatory requirements. There are several components to a robust monitoring and auditing system in a distributed environment:

- **Security Information and Event Management (SIEM) Systems**
  SIEM systems gather, analyze, and report on security data from diverse sources in real time. These sources might include network devices, servers, apps, and security devices. SIEM systems correlate data from these sources to identify security risks and provide alerts to warn security professionals of any questionable behavior. SIEM systems analyze data using a variety of methodologies such as signature-based detection, behavioral analysis, and machine learning. Signature-based detection includes comparing data to established patterns of malicious activity, whereas behavioral analysis searches for abnormalities in user behavior. Machine learning algorithms may be used to detect patterns of behavior that signal a security danger, even if the threat has not previously been discovered.

- **Data Loss Prevention (DLP) Systems**
  DLP systems are used to monitor data while it is accessed and to prevent unwanted access or data exfiltration. DLP systems may monitor data while it is at rest, in use, or in transited systems monitor data using a variety of methodologies, including content-based inspection, context-based inspection, and behavior-based inspection. The process of evaluating the content of data to identify sensitive information, such as credit card numbers or social security numbers, is known as content-based inspection. Context-based inspection is examining the environment in which data is utilized, such as the location or device used to access the data. The process of studying user behavior to identify security concerns is known as behavior-based inspection.

7. User Related Security Solutions

User-related security solutions are critical to mitigating user-related security risks in big data. User Behavior Analytics (UBA), authentication, employee training, and audit trails can be implemented. These solutions are designed to ensure the privacy and security of users' personal information, as well as to prevent unauthorized access to their accounts. By utilizing UBA, organizations can monitor user behavior and detect potential threats before they cause damage. Strong authentication mechanisms are essential for preventing unauthorized access to user accounts. Providing regular employee training can help ensure that employees understand how to identify and prevent security threats. Audit trails are important for tracking user activity and identifying potential security vulnerabilities. By implementing above solutions organizations can protect user data and prevent security breaches.

7.1 Authentication

Authentication is a critical aspect of security in big data environments. Authentication procedures in big data security relate to the processes, protocols, and technologies used to validate the identity of persons or systems accessing the data. These procedures are critical for maintaining data confidentiality, integrity, and availability in large data situations. Some of the authentication mechanisms are described as follows.

- **Multifactor authentication (MFA)**
  MFA is an authentication technique that requires users to give multiple authentication factors to access the system. These elements can include a password, a security token, or biometric data such as a fingerprint or face recognition. This approach provides an extra layer of protection to the authentication process, making it more difficult for unauthorized users to obtain access to the system.

- **Single Sign On (SSO)**
  SSO is an authentication technique that allows users to utilize a single set of credentials to access different systems. By minimizing the number of passwords that users must remember, this approach streamlines the authentication process and enhances security. It also minimizes the chance of password reuse and password-related security problems.

- **Public key infrastructure (PKI)**
  PKI is a system that employs public key cryptography to deliver authentication and encryption services. It employs digital certificates to authenticate users' identities and protect communications between them. PKI can be used in a big data environment to authenticate users accessing Hadoop clusters or to secure communication between nodes in a Hadoop cluster.

7.2 User Behavior Analytics (UBA)

UBA technologies monitor user activity to discover abnormalities and possible security issues, such as illegal access attempts or odd data access patterns. UBA tools may be used to monitor many sorts of data, such as network traffic, system logs, and application logs. UBA tools evaluate user behavior using a variety of methodologies such as machine learning, statistical analysis, and rule-based analysis. Machine learning algorithms may be used to find patterns of activity that suggest a security problem, whilst statistical analysis can be used to identify abnormalities in user behavior. Rule-based analysis entails defining rules or thresholds that signal a security issue. UBA is a valuable tool for organizations looking to improve their big data security. By analyzing user behavior in real-time, UBA can provide insights that help organizations detect and respond to potential threats before they become major security incidents.

7.3 Employee Training

A key security measure for large data protection is employee training. Data breaches and cyberattacks are on the rise in today's increasingly digital world, and enterprises must take preventative measures to safeguard their data. Organizations may lower the risk of security incidents and safeguard sensitive data from unauthorized access by training employees on security best practices and how to handle data securely. A wide range of subjects can be covered in employee training, including as password policies, phishing awareness, data handling protocols, compliance rules, and incident response. There are numerous ways to give this training, including in-person classes, online learning modules, and role-playing security situations. Maintaining employee knowledge of the most recent security dangers and best practices requires ongoing training. By investing in employee training, organizations can improve their overall security posture and create a culture of security awareness and accountability among employees.

7.4 Audit Trails

Audit trails are crucial in big data for monitoring user activity and detecting potential security incidents. They record system events and activities, including data access and user behavior, allowing organizations to quickly identify anomalies and investigate potential security problems. Audit trails also help organizations comply with legal requirements and prove they are taking the necessary precautions to secure sensitive data. To develop an effective audit trail, organizations should ensure proper logging methods and retention periods, establish policies and procedures for reviewing audit logs, and secure audit trails against illegal access or manipulation.

8. Conclusion

The emergence of big data has brought about unprecedented benefits to organizations, but it has also given rise to new security challenges. This paper has explored various security challenges related to data, devices, networks, cloud integration, and user-related issues. Moreover, we have reviewed several solutions for these challenges that are aimed at protecting the confidentiality, integrity, and availability of big data. It is important for organizations to take a holistic approach to address big data security challenges by implementing appropriate technical and organizational solutions, and to ensure that security measures are regularly reviewed and updated to mitigate emerging threats. Ultimately, the successful implementation of security solutions will enable organizations to leverage the full potential of big data while maintaining the confidentiality, integrity, and availability of sensitive information.

References

[1] D. S. Terzi, R. Terzi, et al., *A Survey on Security and Privacy Issues in Big Data*[C], in Proc. 2015 IEEE International Conference on Internet

[2] ISO/IEC JTC 1, *Big Data* [R], Preliminary Report, 2014.

[3] Technology and Secured Transactions(ICITST' 2015), 2015.

[4] Kaustav Ghosh, and Asoke Nath. "Big Data: Security Issues, Challenges and Future Scope." International Journal of Research Studies in Computer Science and Engineering 3.3 (2016): 1-9.

[5] A.A. Hussein (2020) Fifty-Six Big Data V’s Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). Journal of Information Security, 11, 304-328.

[6] Dongpo Zhang. "Big data security and privacy protection." 8th international conference on management and computer science (ICMCS 2018). Atlantis Press, 2018.

[7] Sitalakshmi Venkatraman, and Ramanathan Venkatraman. "Big data security challenges and strategies." AIMS Math 4.3 (2019): 860-879.

[8] Taran Singh Bharati. “Challenges, Issues, Security And Privacy Of Big Data.” *International Journal of Scientific & Technology Research* 9 (2020): 1482-1486.

[9] M. Binjubeir, A. A. Ahmed, M. A. B. Ismail, A. S. Sadiq and M. Khurram Khan, "Comprehensive Survey on Big Data Privacy Protection," in *IEEE Access*, vol. 8, pp. 20067-20079, 2020, doi: 10.1109/ACCESS.2019.2962368.

[10] M Kantarcioglu and E Ferrari (2019) Research Challenges at the Intersection of Big Data, Security and Privacy. Front. Big Data 2: 1.doi: 10.3389/fdata.2019.00001

[11] S. Subbalakshmi, and K. Madhavi. "Security challenges of Big Data storage in Cloud environment: A Survey." *International Journal of Applied Engineering Research* 13.17 (2018): 13237-13244.

[12] J. Koo, G. Kang, Y.-G. Kim, Security and Privacy in Big Data Life Cycle: A SurveyandOpenChallenges. *Sustainability* 2020, *12*,10571.https://doi.org/10.3390 /su122410571

[13] A Comprehensive Survey on Security and Privacy for Electronic Health Data" (2021) by Se-Ra Oh, et. al.

[14] Big Data Security Issues with Challenges and Solutions (2023) by Santanu Koley

[15] A Survey of Security and Privacy in Big Data (2016) by Haina Ye, et. Al

[16] Danda B. Rawat, Ronald Doku, and Moses Garuba. “Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security.” IEEE transactions on services computing 14.6 (2021): 2055–2072. Web.

[17] Bardi Matturdi et al. “Big Data Security and Privacy： A Review.” China communications 11.2 (2014): 135–145.

[18] KamtaNath Mishra et al. “Cloud and Big Data Security System’s Review Principles: A Decisive Investigation.” Wireless personal communications 126.2 (2022): 1013–1050.

[19] Adel Al-Zahrani, and Mohammed Al-Hebbi. “Big Data Major Security Issues: Challenges and Defense Strategies.” Tehnički Glasnik 16.2 (2022): 197–204.

[20] Shahab KAREEM et al. “Big Data Security Issues and Challenges.” Journal of applied computer science & mathematics 15.2 (2021): 28–36. Web.

[21] Julio Moreno, Manuel Serrano, and Eduardo Fernández-Medina. “Main Issues in Big Data Security.” Future internet 8.3 (2016): 44–. Web.

Interactive Mood Boards to Teach User Experience (UX) Principles as Part of an Agile Methodology

Tim Krause

Computing and New Media
Technologies

UW – Stevens Point

Stevens Point, WI 54481

tkrause@uwsp.edu

Abstract

A consideration of interactive mood boards using multimedia elements such as visual, video, and audio to help students understand core concepts of the user experience (UX) across a variety of web-based and mobile applications within the context of an Agile development process. Students consider the interplay of font, color, text, image, and other aspects of rich media experiences while designing for an 'accessibility first' mindset in consideration of issues like low-sight and color blindness. Students conclude with in-class presentations that simulate the integration of UX within an Agile development environment emphasizing 'just in time' iterative design to facilitate rapid deployment.

1 Overview of Interactive Mood Boards

The incorporation of User Experience (UX) methodologies into a lean or Agile approach to software development is not a new concept. Shamp Winstel [1] and others typically focus on personas, wireframes, and prototypes as central artifacts to those processes. While this paper does not argue against any of those uses, a 'just in time' approach to UX would also suggest that development teams need specific artifacts and design standards to begin their work. Those artifacts, which may include logos, images, and font treatments (to name a few), are best exemplified through the use of media rich mood boards.

This paper contends that the most useful mood boards not only include those elements, but also begin the work of developing robust information for the overall user experience—often employed through the separate development of personas. While Getto and Flanagan [2] have effectively argued for the amplification of user agency, advocacy, and accountability through the use of personas, this paper argues that in a lean UX environment, thoughtfully constructed mood boards even further extend the notion of user agency to accommodate for additional user needs centered around accessibility and control starting from the beginning of the design and development process.

The word *interactive*, framed through a lens of accessibility, therefore, foregrounds issues associated with limited motor skills, sight, and other issues in designing accessible user experiences in a lean UX environment. More as a matter of practicality, this paper, and assignment, focused on low and varied experiences with sight to illustrate how accessibility issues may be incorporated into mood board design.

1.1 Learning Outcomes

Interactive mood boards that accommodate for accessibility issues are a multilayered, complex assignment for students because the assignment requires that students understand and can demonstrate:

- basic elements of a traditional mood board including fonts, images, and color.
- relationships between those elements in bringing together a cohesive, consistent user experience, rather than merely a collection of design artifacts and concepts.
- development of a mood or theme, consistent with a brand and user experience.
- appropriate and consistent development of that experience for individuals with as varied experiences and abilities (as evidenced through attention to accessibility) as reasonably practical.
- negotiation skills in providing, and accepting, feedback to other members of their design teams.
- communication skills necessary to convey those design decisions and recommendations to the development team and other key stakeholders.

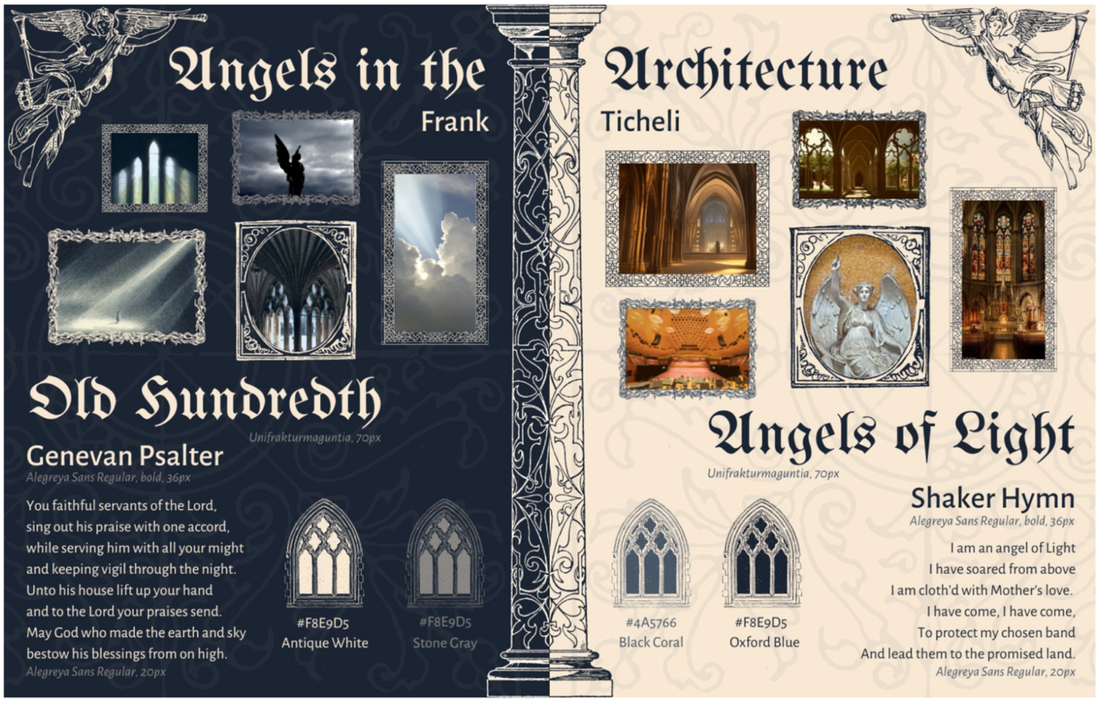
1.2 The Assignment

The core assignment for interactive mood boards starts deceptively simple in the requirement to:

1) Create a mood board based on a song lyric, poem, or selection from another creative piece of your choosing.

This assignment is designed intentionally to ask students do something different from what they are accustomed to in most software analysis and design classes: there is no client or simulated business case in the prompt. The goal is to also simplify at least some aspects of this assignment by allowing students to start with something of interest to them, rather than a piece of software, a brand, or other business problem that may only add un-necessarily to the complexity of the task at hand.

Figure 1 provides an example of a solution that meets the first requirement of the assignment and includes all of the elements one might expect of a typical mood board: font families, sizes, text treatment, imagery, and color palettes (for example).



You faithful servants of the Lord,
sing out his praise with one accord,
while serving him with all your might
and keeping vigil through the night.
Unto his house lift up your hand
and to the Lord your praises send.
May God who made the earth and sky
bestow his blessings from on high.

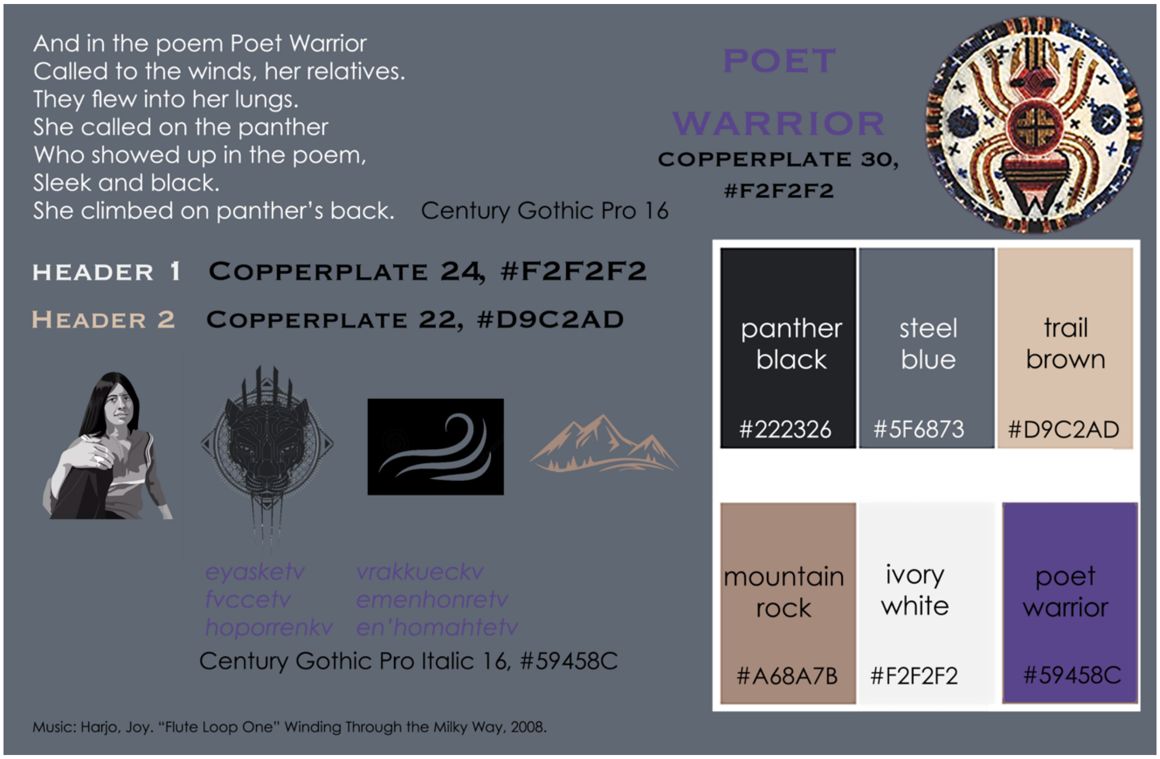
I am an angel of Light
I have soared from above
I am cloth'd with Mother's love.
I have come, I have come,
To protect my chosen band
And lead them to the promised land.

Figure 1: Sample Mood Board

The second part of the assignment:

2) Your mood board should consider accessibility best practices and also include at least one aspect or approach you believe is unique, but useful, to your mood board.

In past teaching experience, a combination of Sharp, Preece, and Rogers [3], and Rosenfeld, Morville, and Arango [4] provide more than sufficient background in helping students learn the basics of mood board creation. However, as is often the case, while textbook examples are useful in teaching basic principles, they are by nature poorly suited to specific application in the real-world. The two-fold design of the assignment is intentional in ensuring that students first understand the basic structure and use for a mood board in setting a design direction for the development team before secondarily understanding how to adapt it situationally to specific user needs—whether through general accessibility principles, incorporation of multimedia elements or other atypical requirements.



And in the poem Poet Warrior
Called to the winds, her relatives.
They flew into her lungs.
She called on the panther
Who showed up in the poem,
Sleek and black.
She climbed on panther's back.

Music: Harjo, Joy. "Flute Loop One" Winding Through the Milky Way, 2008.

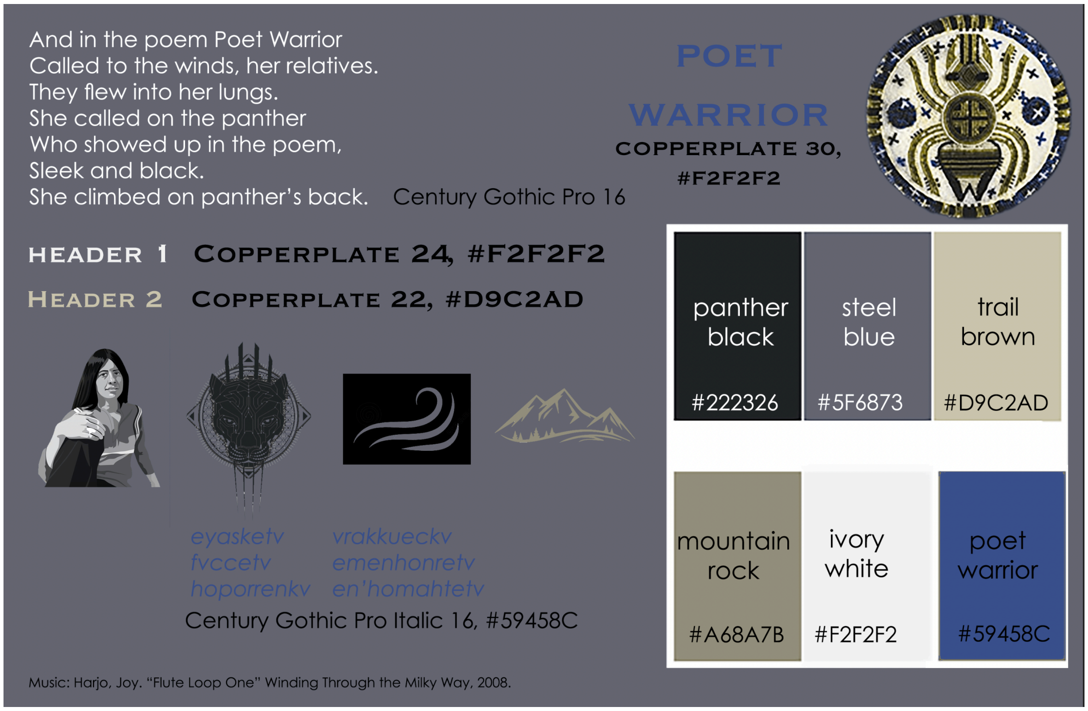
Figure 2: Interactive Mood Board – Base Design

It is difficult to represent some of the multimedia and interactive elements of a mood board in a print paper, however, Figure 2 includes a foot note acknowledging that it includes a song from Joyce Harjo's "Flute Loop One" that auto-plays when the mood board is opened. The piece is essential to the essence of the mood board because, in Harjo's own autobiographical words, "musicians here what can't be heard" (p. 145) [5] and is cornerstone to understanding her self-described work as poet, storyteller, and musician.

While students found ways to stretch the boundaries of the artifacts one might include in mood boards beyond iconography, images, color palettes, and logos to music and video, the project also encouraged them to use the mood boards as an early space to plan for accessibility. Grayson [6] and others have advocated for an "Accessibility-First" approach; however, those recommendations often suggest early usability testing as the solution.

While usability should not be overlooked, it is also a time-consuming and expensive process that may also not be appropriate in the early stages of a project.

By contrast to usability testing, the mood board offers an even early opportunity for students and designers to begin thinking about how their design choices impact accessibility in advance of more formalized testing methodologies. Figure 3 offers just one example of testing for deuteranopic colorblindness on the elements from a mood board in advance of more formalized user testing. It is not intended to serve as the only possibility for early design work in accessible design and testing, but to illustrate the potential. It was also selected because of the relative ease of testing: many common design software applications and third-party web sites facilitate testing for a variety of ways in which color blindness might manifest itself. In considering Figure 3, early testing points to potentil issues with muted colors and lower contrast with some of the color choices in the mood board. The test and the mood board don't themselves, evaluate or solve those issues but offer the design and development teams early insight into design challenges they will need to consider as they move their work forward.



And in the poem Poet Warrior
Called to the winds, her relatives.
They flew into her lungs.
She called on the panther
Who showed up in the poem,
Sleek and black.
She climbed on panther's back.

Century Gothic Pro 16

POET WARRIOR

COPPERPLATE 30, #F2F2F2

HEADER 1 COPPERPLATE 24, #F2F2F2

HEADER 2 COPPERPLATE 22, #D9C2AD

eyasketv *vrakkueckv*
fvccetv *emenhonretv*
hoporrenkv *en'homahtetv*

Century Gothic Pro Italic 16, #59458C

panther black #222326	steel blue #5F6873	trail brown #D9C2AD
mountain rock #A68A7B	ivory white #F2F2F2	poet warrior #59458C

Music: Harjo, Joy. "Flute Loop One" Winding Through the Milky Way, 2008.

Figure 3. Interactive Mood Board – Deuteranopia Perspective

In this assignment, students additionally focused their mood boards on other aspects of design not typically considered in mood boards that included:

1) representation of non-Western language alongside English language.
2) video.
3) virtual reality interfaces.
4) advanced media players; and
5) atypical design through experimentation with color and negative space.

Figure 4 illustrates one mood board that employed a variety of experiments with color, typography and use of white space that violated many of the established norms for UX design. This is not an insignificant point: asking students to be creative with their mood boards in this manner requires an additional level of explanation, through an oral presentation and written retrospective. Both are opportunities for the student to express intentionality in their work that extends beyond a mere desire to "break the rules". Both are also an essential reason for including oral presentations, one-on-one meetings with the instructor and a final retrospective as part of the design process. Additionally, the grading heuristic also balances the requirements for demonstrating a base level of competency with an understanding of the broader contexts in which UX functions in the development process.

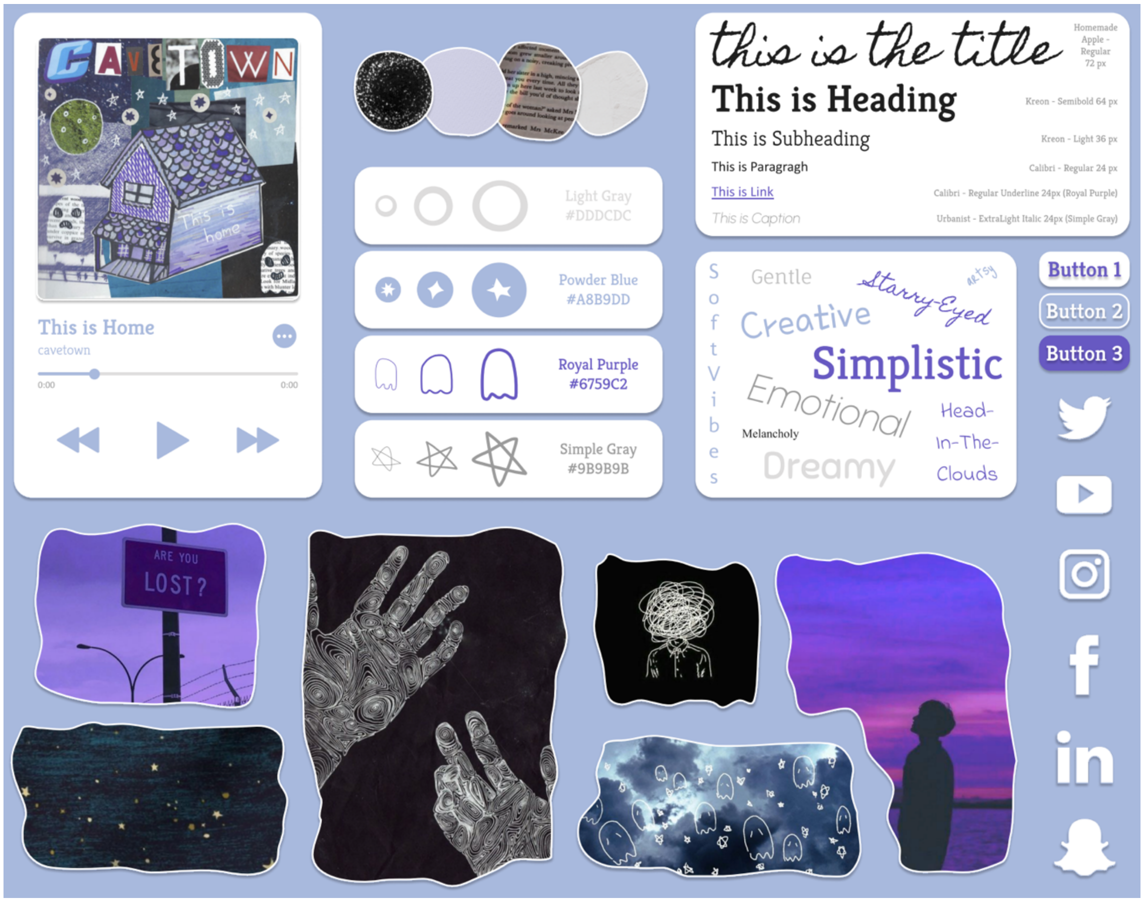


Figure 4. Alternative Control Structures, Design Spaces, and Use of Design

5

1.3 Oral Presentations

A vital aspect of UX design is the presentation of one's work in a variety of institutional settings, and that often starts with one's peers and immediate supervisor. Prior to the oral presentations, students were given the opportunity to practice providing each other with feedback using an online discussion forum in Canvas, our online learning management system (LMS). The discussion was less formal, and lower stakes (from a grading perspective) than the final oral presentation.

Oral presentations of the mood boards for this assignment were short, five-to-seven-minute opportunities for students to present their work in class and construct an oral argument in support of their choices. The presentations were required to include multiple sources, justifying their design decisions, and involved a brief period for questions and feedback from their peers and the instructor.

It is worth noting that students were given the opportunity to then revise their mood boards, based on this presentation and peer review, before turning them in for final grades. The last time this assignment was used in a UX course, roughly half of the students revised their work before final submission.

1.4 Grading Heuristic and Design Retrospectives

The final grading heuristic for the mood boards (not the final presentation) is offered in the Appendix. It includes grading standards for a variety of aspects of this assignment including:

1) Demonstrating an understanding and corresponding implementation of basic mood board components (Fonts, Images and Other Supporting Elements). 60% of grade
2) Providing active and substantive peer review on each other's work (Discussion Posting and Discussion Response), and in an informal one-on-one feedback session with the instructor. 20% of grade
3) Explaining, in writing, the research and rationale they employed across the entire design process employed in their mood boards (Overall Synthesis and Design Rationale). 16% of grade
4) Innovating in at least one aspect of their mood board design (Unique Approach). 4% of grade

It is worth noting that the student's explanation of their research and rationale (item 3, above) accounted for 16% of their final grade, emphasizing the importance of the student-designer's ability to explain the reasoning behind their work. Every student completed this work to varying degrees of success.

Additionally, 4% of the final grade was available based on the student's attempt (item 4, above) at innovating, with 2% awarded for the attempt, and the remaining 2% for relative success. All but one student at least attempted something unique and innovative with their mood board.

1.4 Broader Considerations and Next Steps

The combination of structured feedback from peers and the instructor helped to establish a level of trust that facilitated a space for students to simultaneously demonstrate the basic competencies required in creating a mood board as part of a broader UX design process, while also experimenting with more interactive and multi-media-oriented aspects of those mood boards. It also introduced students to the notion that it is not only possible, but desirable, to begin the design process with a consideration of accessibility issues—as opposed to leaving them to the end of the process.

Future assignments might consider:

1) Additional accessibility topics beyond researching aspects of color blindness, and low vision in UX design; and
2) The application of principles in a business, rather than a more purely creative, setting as they might apply to issues of brand and image.
3) Integration of mood boards into a more complete, Agile UX design process.

Nevertheless, students demonstrated competency in using a mood board to simultaneously deliver key artifacts (logos, fonts, color palettes) in an efficient manner that would allow Agile development teams to begin early development efforts while also using those same mood boards as an opportunity to creatively explore those same design elements from a variety of other perspectives and technologies (multiple languages, multimedia, and through the lens of accessibility).

References

[1] Bonnie J. Shamp Winstel. 2014. "UX for Lean Startups: Faster, Smarter, User Experience Research and Design." *Technical Communication*, *61(2)*, page 134.

[2] Guiseppe Getto and Suzan Flanagan. Nov. 2022. "Localizing UX Advocacy and Accountability: Using Personas to Amplify User Agency." *Technical Communication, 69 (4)*, pages 97-113.

[3] Helen Sharp, Jennifer Preece, and Yvonne Rogers. 2019. *Interaction Design: Beyond Human Computer Interaction*. (5th Ed). New York: Wiley.

[4] Louis Rosenfeld, Peter Morville, and Jorge Arango. 2015. *Information Architecture for the World Wide Web*. (4th Ed.). Sebastopol: O'Reilly.

[5] Joy Harjo. 2021. *Poet Warrior: A Memoir*. New York: WW Norton.

[6] Kathryn Grayson Nan. Jan. 2022. "Rethinking Tech Design with an Accessibility-First Approach." *eWEEK.*

Appendix 1: Grading Heuristic

Criteria	Ratings		
Discussion Posting	5 pts – Full Marks Contributed discussion post that included both the item for your mood board, and reason for selecting it	3 pts – Meets Missing reason for selecting or not enough detail on the item selected	0 pts – No Marks Did not complete, or did not complete on time.
Discussion Response	5 pts – Full Marks Contributed a response that offered feedback or suggestions that were actionable and thoughtful.	3 pts – Meets Contributed a response, but with little meaningful or actionable contribution.	0 pts – No Marks Did not complete, or did not complete on time.
Fonts	10 pts – Full Marks Font treatment, as appropriate, provided examples for body, headings, links and other elements, including meta-data (font family, size, weight, color, etc.). If an element is not included, mention why in your design rationale.	5 pts – Meets Fonts may be missing meta-data font family, etc.) or may not support overall approach.	0 pts – No Marks Fonts missing with no rationale.
Images	10 pts – Full Marks Images, illustration, and textures, as appropriate, and consistent with the overall mood or theme of your board. If this	5 pts – Meets Visual elements may be missing or inconsistently support your mood board.	0 pts – No Marks Images missing with no rationale.

	element is not included mention why in your design rationale.		
Sample Text and Other Supporting Elements	10 pts – Full Marks Sample text should be supportive and evocative of your mood board. If an element is not included mention why in your design rationale.	5 pts – Meets Sample text and/or other supporting elements may be missing or inconsistently support your mood board.	0 pts – No Marks Sample text and support elements are missing.
Overall Synthesis and Design Rationale	8 pts – Full Marks All elements of your mood board cohesively, and thoroughly support the piece of creative work that you selected.	4 pts – Meets Elements may work fine as stand-alone elements, but may not fully come together as a mood board.	0 pts – No Marks Either missing mood boards, or mood boards that appear to be collages rather than an effort at a cohesive piece.
Unique Approach	2 pts – Full Marks Your mood board contains at least one element not seen elsewhere in another mood board.	1 pts – Meets You found an element at least rarely employed in another mood board.	0 pts – No Marks Not attempted.

Tutorial on TensorFlow Spark
for BCI Augmented Robotics

Adriano Cavalcanti
Department of Computer Science and Information Technology
Interdisciplinary Computer Hub
Brain Computer Interface Laboratory
Saint Cloud State University
St. Cloud, Minnesota, 56301
adriano.cavalcanti@stcloudstate.edu

Abstract

This work presents the current approach at SCSU for creating and offering a framework to give students a hands-on experience with Brain Robotics interfaces. The proposed work describes the overall steps in hardware architecture and system platform, thus providing valuable information for implementing an enterprise solution or a practical procedure for higher education institutions interested in starting development in the new technology field known as Brain-Computer Interaction and Interface.

Keywords: Avatar, Brain-Computer Interface, Hardware Architecture, Spark, TensorFlow.

1 Introduction

The field of HPC High-Performance Computing has been using the current trends in Cloud Computing with easy scalability. The creation of cluster computing has successfully led the way for advanced distributed systems. Ongoing trends in Brain-Computer Interface offer a cutting-edge approach to Human-Computer Interaction, using pattern identification techniques with real-time sensors to harvest brain wave control to perform brain system connectivity. For steering and control of real-time agents, as a UAS Unmanned Aerial System, the challenge is to build a reliable architecture platform to address low latency at low cost with high stability and load balance.

2 Motivation

Sci-Fi movies and literature have featured the concept of controlling devices, systems, and machines through the brain's thoughts for many years. In the early '80s, the scientific community presented initial prototypes for the Brain-Computer Interface. Back then, the first BCI devices were invasive until the end 1960s, meaning that EEG sensors needed to be implanted in the person's head to allow precise reading of the neurons and brain signals. In the early 70s, through the BCI challenge initiative, the first version of EEG that was not invasive was presented. Despite that, initial EEG-BCI was still quite expensive, limiting the development of several systems based on this type of technology. However, in recent years the cost involved to such devices decreased significantly. Thus, we can observe many projects to integrate BCI more commercially towards developing market-oriented solutions that rely on BCI as the next generation of the Human-Computer Interaction and Interface paradigm. Several startup companies are currently developing many applications intended to work solely using BCI. In this race to commercialize the first wave of BCI applications, giant high-tech companies invest their resources and development teams towards this objective. Examples are NeuraLink from Tesla and Facebook Meta.

Micro and nanoelectronics have continuously paved the way for a growing number of advanced devices at low cost with large-scale manufacturing and distribution [1]. That resulted in our society's ability to access complex sensors at a low price, with high precision and lower costs.

3 Methodology

The support received by the SCSU and the teamwork started with a faculty board discussing establishing a workspace for research and development in the Engineering and Computing Center building; after several meetings, the R&D committee member came out with naming the workspace Interdisciplinary Computer Hub. A few months later, the Dean’s office administration approved a faculty start-up package to fund the initiative that we used to have a thing to start moving. The first most noticeable acquisition of 4 sets from the: OpenBCI All-in-one Biosensing R&D Bundle, followed by additional gears and gadgets. The next step was to reorganize the layout of the room used for project: ECC102E. Besides the infrastructure, the other important aspect of the approach is using open-source technologies, which comprise 100% of the project system implementation. The ability to use readily available libraries for cloud computing performance, automation, and machine learning algorithms with a framework for real-time data streaming is paramount for fast prototyping and implementation.

Last but not least, and most importantly, the teamwork and promotion: the Cloud Computing Club develops most activities remotely using a high-performance distributed cluster approach. For the development team coordination, we use Github and Atlassian Jira. While all contributors use Github, once a member becomes a more frequent contributor, if that member becomes a committer or one of the project administrators, then that implies also having access to the Atlassian Jira repository used for starting a new patch in tandem with Github, for such as fixing a bug or adding a new feature to the code base.

4 Hardware Architecture

The current development of low-cost mass production of electronics and components for integrating advanced architecture with hardware and software solutions is making possible the creation of immersive reality and telemetric control of avatars [2]. In our approach, we establish a suitable platform that can effectively provide the application to various control and haptics interactions [3]. The user can apply his thoughts and activate the system (Figure 1). In the sequence, we describe each component's technical specification and how it integrates with our implemented solution.

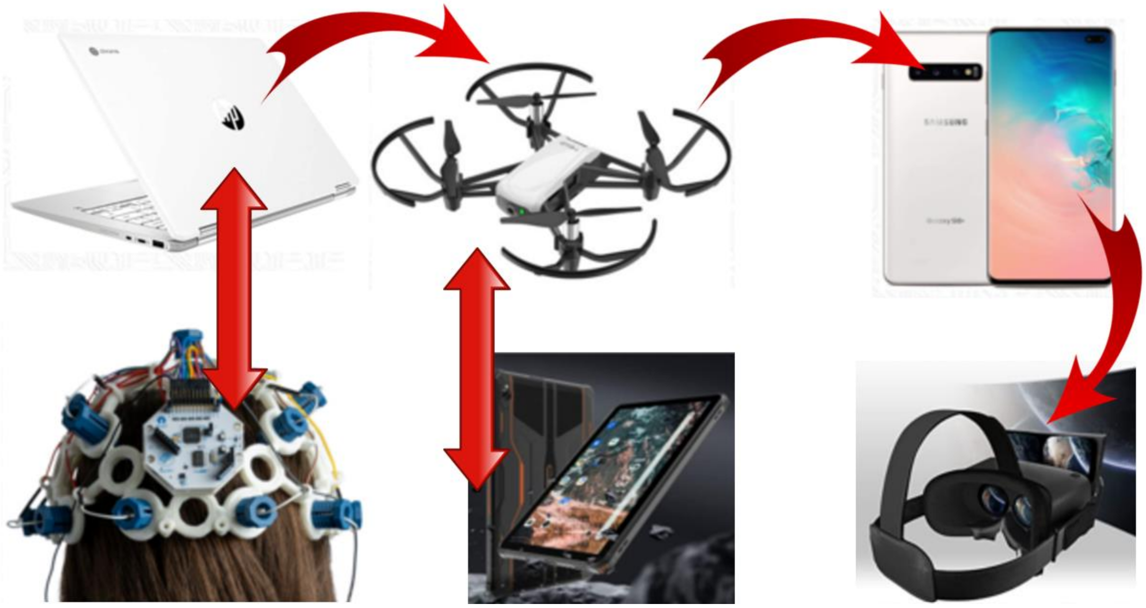


Figure 1: Dataflow - connectivity and data streaming in real time for BCI drone augmented robotics.

4.1 EEG-BCI

The use of advanced electroencephalography (EEG) for Brain-Computer Interface (BCI) makes possible the implementation of a new framework to control systems and devices using machine learning to automate teleoperated robots with your mind (Figure 2). Initially, the development of EEG-BCI centered around partially impaired patients [4].

Figure 2: Calibrating to run a data collection with the BCI takes about a minute or two.

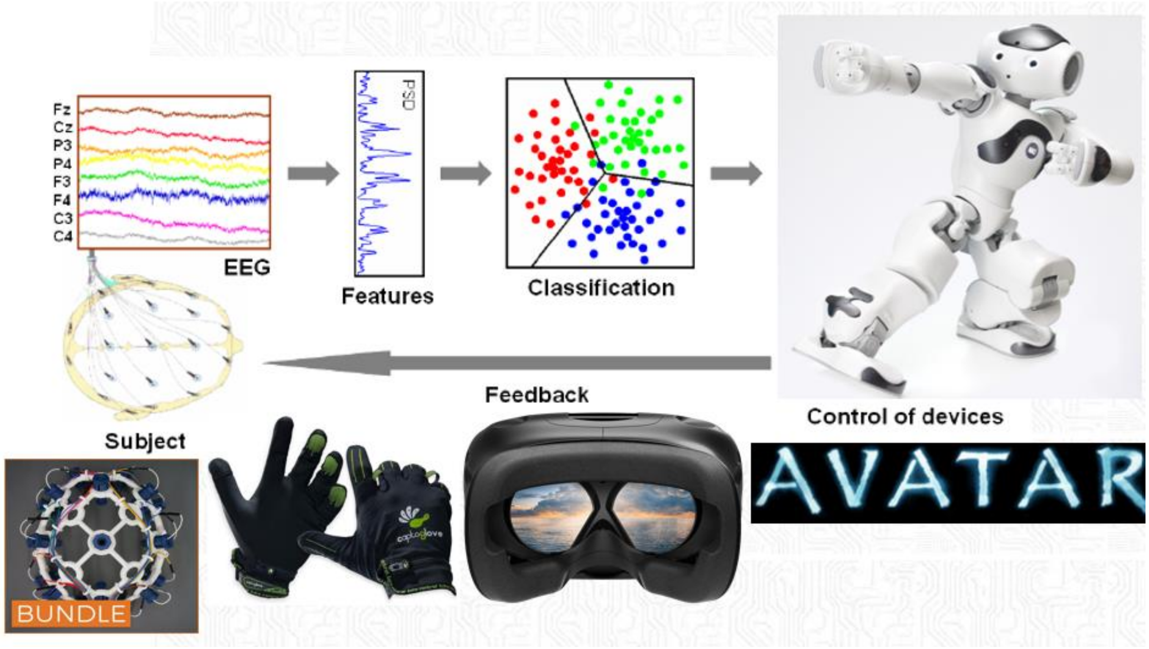


Figure 3: Dataflow with real-time connectivity and data streaming for BCI NAO6 augmented robotics.

With the fast-growing development of EEG-BCI, commercial applications for such technologies are expected to grow in the coming years (Figure 3). Several big tech companies are betting on this new technological frontier as the next step in Human-Computer Interaction. Some most known companies doing this kind of development include Meta Facebook, Tesla NeuraLink, and some other 110 start-ups around the globe. The fact that the currently in development technologies allows the EEG-BCI to be non-invasive plays an essential role in making this technology for UI and UX more commonly accepted and desirable.

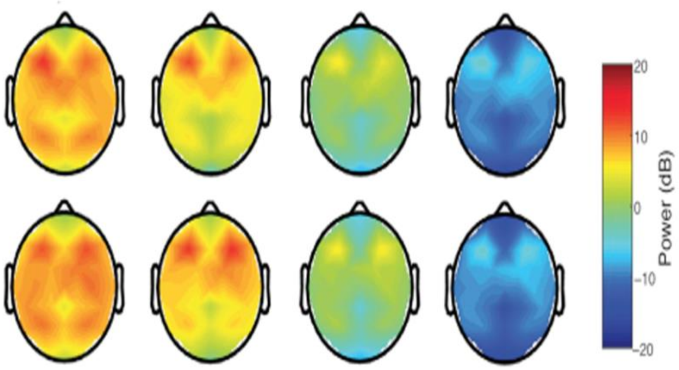


Figure 4: Brain activity - the EEG readings show the difference when certain parts of the brain might be more active than others.

Topographic EEG maps can read the brain wave activities from lower to highest frequencies (Figure 4). Typical brain wave frequencies are associated with specific types of activities (Figure 5). As our framework and activities in the BCI Lab, we concentrate on brain reading frequencies, so we want your brain waves in our cloud database [5].

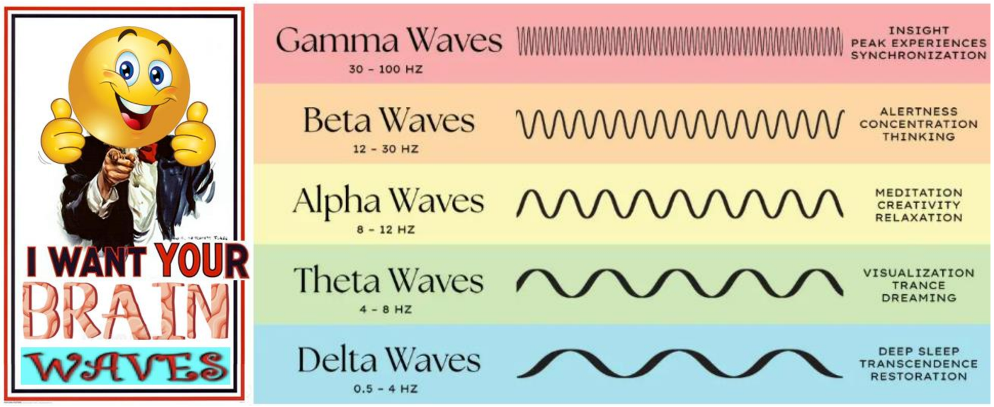


Figure 5: Brain waves - all the Cloud Computing and BCI Lab want.

Source code implemented in MATLAB is available as open source for computing BCI applications [6].

4.2 Drone

Ryze Tello Edu provides a programmable drone at a very accessible price (Figure 6). It has a broad connectivity capability and offers the option to program the device in Python [7]. In our case, we considered a compelling approach to have the bird's eye view of the footage captured by the drone and projected onto the smartphone inside the VR headset and simultaneously transmit the duplicate footage in the large screen ROKU TV. The Smart-TV just arrived this week in our lab.

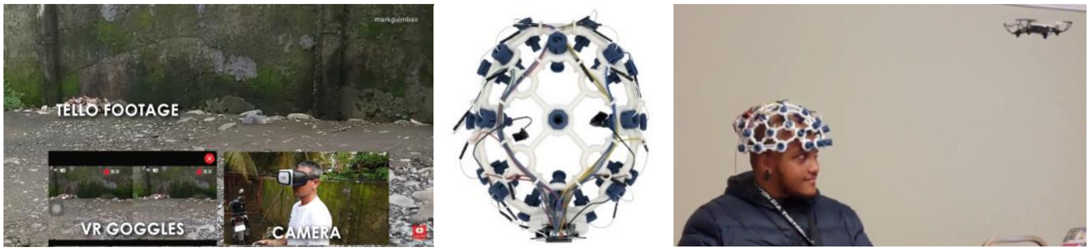


Figure 6: DJI Tello Edu Programmable Drone provides BCI and VR headsets with a mind-teleoperated flying avatar.

The current Tello EDU runs on version SDK 2.0 and is also compatible with Scratch or Swift programming. Brain-Computer Interface from OpenBCI Headset reads brain waves generated by thoughts that create related signal patterns for our machine learning platform [8].

4.3 Robotics

Currently, many robot models are available for implementing augmented platforms to facilitate avatar instrumentation. One interesting approach that our development considers besides using drones for flying avatars is to use a humanoid robot for evident reasons.

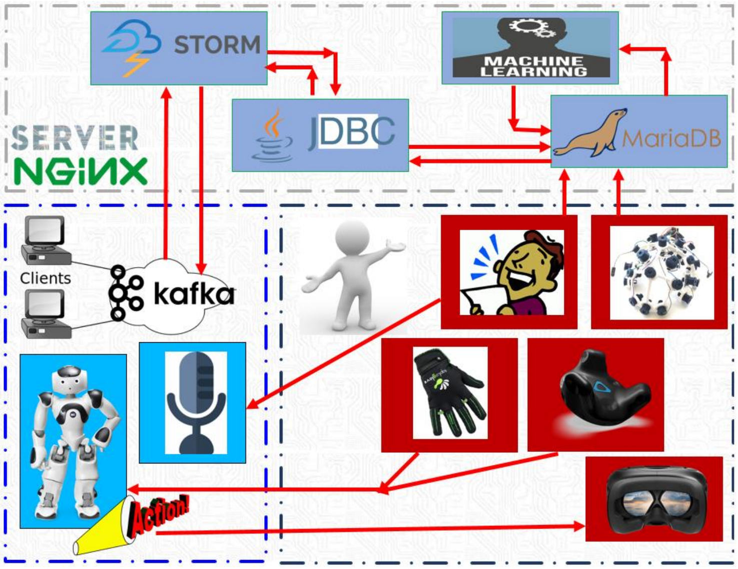


Figure 7: BCI-NAO6 Implementation done for a humanoid robot with alternative pathways for control.

It has many haptic sensors and a very similar human-like design. NAO6 is an option that we are considering acquiring soon to integrate into our current framework (Figure 7). The implementation types using a similar architecture can apply to connecting your mind to perform big data analytic queries to social media searches and much more.

4.4 Tablet

We got the Hotwav R5 Rugged Tablet 10-inch Tablet Android 12 2023, which is used for steering control of the drone (Figure 8). For the testing and programming of the drone, first, we define the type of steering actions using the Python SDK for the Ryze Tello Edu Quadricopter. Although the final goal is to control the drones solely with our minds, for the prototyping and testing phases, using the drone app helps define control strategies that best fit the BCI low latency and cloud-based machine learning processing steps.

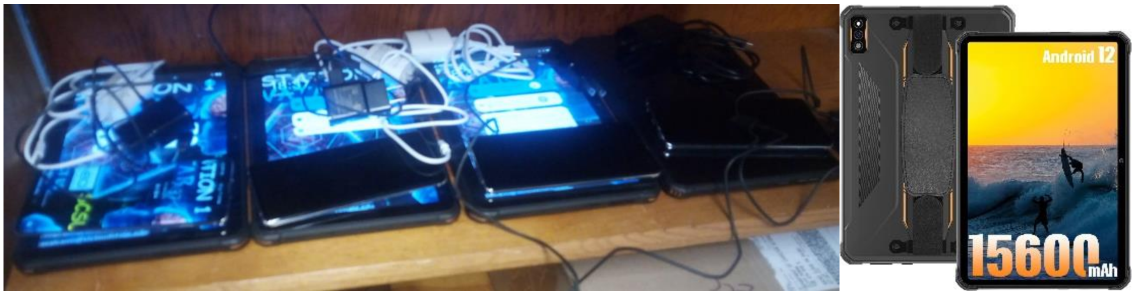


Figure 8: Tablet with MediaTek MT6762 processor, 6GB RAM, 4GB+64GB (SD up to 1TB), Cortex A73 2.0GHZ×4+A53 2.0GHZ×4,12nm process technology.

4.5 Smartphone

The smartphone provides the Bird's Eye view from the Tello Edu drone (Figure 9). Although we were initially planning to acquire rugged phones, we had to adjust our initial expectations and pick another phone model. The reason for the limitation is those rugged ones are more durable on the one side, and with long-lasting batteries, they are bulky and, therefore, cannot fit inside the FEEBZ headsets. Therefore, we switch to a smartphone that effectively fits within the specifications that work with the VR headset. The phone we purchased for our workstations was the Samsung Galaxy S10+ Plus (Dimensions 6.2 x 0.3 x 2.9 inches).

Figure 9: Smartphone with 128GB Storage, 8GB RAM, Up to 1TB microSD Card slot, Qualcomm SDM855 Snapdragon 855 (7 nm), CPU: Octa-core.

4.6 Augmented Reality Goggles

We have compared several VR headsets that would effectively fit the context of the type of application we are developing. After considering several factors, including cost-benefit, connectivity, and reviews, we decided on the FEEBZ VR headset that gives the augmented reality feeling when someone is flying the drone controlled by the pilot's mind (Figure 10).

Figure 10: Augmented Reality - the Feebz VR headsets design fits almost any iPhone or Android phone with a screen size of up to 6.5".

4.7 Workstation

For development, we chose Chromebook compared to other options. Some objective motivations for this choice: ChromeOS is an open-source operating system based on Gentoo Linux (Figure 11). Gentoo Linux offers the flexibility of customizing the system source code, thus allowing optimized configuration for targeted computer hardware components and architecture. The model we got was the HP Chromebook X360 2-in-1 14.0” HD (1366 x 768).

Figure 11: ChromeOS - it uses Gentoo Linux as it offers a flexible range of customization, allowing optimized hardware adaptability and configuration.

There are several advantages to using a Chromebook (Figure 12). One of them, of course, is the price tag. Because the operating system is free and open source, you can get a better machine configuration for half price than if you get the same computer running on a proprietary closed-source operating system with the same hardware configuration. For our initial setup, instead of two workstations as laptop computers, we got four workstations for the same amount of available budget.

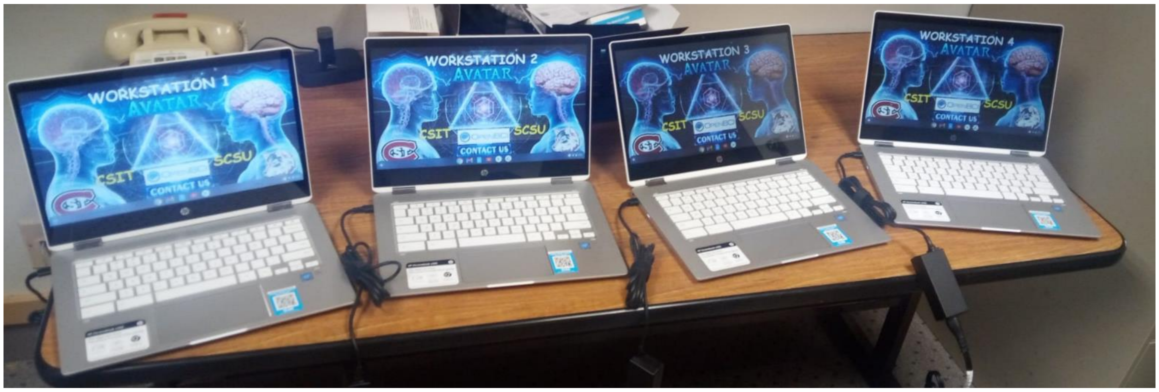


Figure 12: Chromebook with Intel Celeron N4000 (1.1 GHz base frequency, up to 2.6 GHz burst frequency, 4 MB L2 cache, 2 cores) & Intel UHD Graphics 600, 4GB LPDDR4-2400 MHz SDRAM, 32GB eMMC+32GB SD card.

4.8 Cloud Computing VPS

We have compared many options: Google Cloud, AWS, Azure, CloudSurph, and IONOS. After reaching and trying some of them, we ultimately used the IONOS. Our project currently has two VPS servers running on the cloud, plus a data repository integrated with Delta Lake. One server is located geographically in US CT; meanwhile, the second mirroring server is in Germany.

4.9 Budget and Funding

Our project received a faculty grant from the university as part of the called start-up fund. The department also provided additional support to help to complete some of the components and devices needed to integrate a set of 4 complete permanent workstations for Brain-Computer Interface. The Cloud Computing Club directly benefits from the opportunity to access a development framework environment on the edge of advanced technologies. The same infrastructure for development and programming is also helping the senior student and master students taking the course SE475/575 Brain Robotics Interface, which is a course entirely dedicated to teaching and offering the students the valuable skillset to aid their professional development with know-how currently considered the next step on Human-Computer Interaction interfaces. The students taking the SE413/513 Big Data Organization and Management Systems will also have Brain-Computer Interface technologies aided to their Big Data Analytics activities as a new way to query big data with your mind. And the student taking the class SE350 Software Engineering and Human-Computer Interaction are also equally offered the opportunity to integrate BCI into their Human Computer Interaction control applications.

Additionally, passed three months have since the equipment arrived, and we already have it set up and ready for an Open Lab event that is part of the Math Contest at the end of March 2023. We had the opportunity to see all students from High School exceedingly motivated after watching live the system running and the possibilities of control of systems and devices with the power of their minds. For the Math Contest event, we expect to have at least 140 students visiting the Lab from 11 AM to Noon at the of the month. Undoubtedly, the most expensive part of the workstations is the OpenBCI combo, which costs around 3K each. But the investment is worth it. So far, we have been delighted with the support from OpenBCI and the quick post-sale response on any eventual needs in keeping our Lab running smoothly and homogeneously.

5 Learning and Development Workspace

Working as a team towards transforming the ECC102E into a development hub to strengthen the CSIT department has been excellent. All faculty and students can benefit from an R&D space for collaboration so that ideas and development can flourish. We use in-person, distributed, and virtual tools that help boost cooperation and speed development and integration. Among other technologies for collaboration, it is worth mentioning: Zoom, Atlassian Jira, Github, VPS, Spark, Dockers, and Kubernetes. These tools provide the ideal scalability, automation, and remote collaboration framework.

5.1 Interdisciplinary Computer Hub

Creating a physical space for uplifting creativity and collaboration is essential (Figure 13). Therefore, establishing an environment that supports students in a continued implementation practice significantly impacts students' engagement in teamwork development.

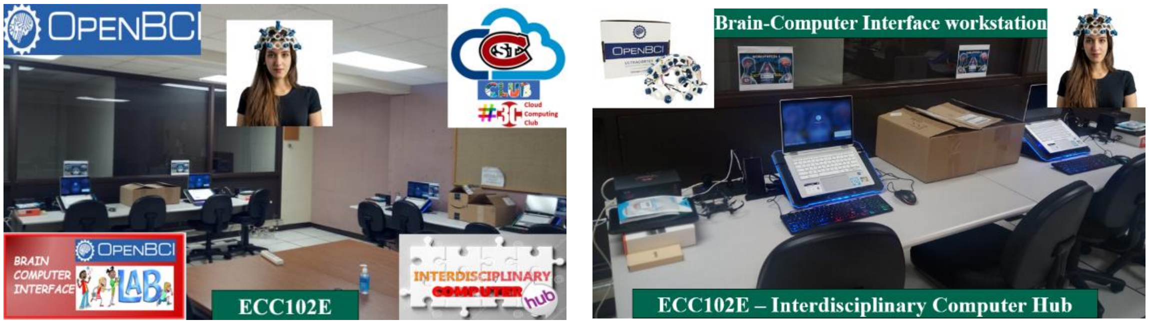

OPENBCI
CLUB
Cloud Computing Club
BRAIN COMPUTER INTERFACE
OPENBCI
LAB
ECC102E
INTERDISCIPLINARY COMPUTER hub
Brain-Computer Interface workstation
OPENBCI
ECC102E – Interdisciplinary Computer Hub

Figure 13: ECC102E - Interdisciplinary Computer Hub comprising the Brain Computer Interface Lab, and it also serves as the physical headquarters for the Cloud Computing Club.

For that, a committee comprising three faculties had several meetings last year to establish a clear and transparent policy about how a new workspace for research and development could follow guidelines that will bring together collaboration towards supporting our students' success. Thus, some guidelines were drafted and voted. It was afterward presented and approved by all faculty during our department meetings.

The term Interdisciplinary Computer Hub (ICH) is a name that encompasses all activities of all faculty that belong to the CSIT. Additionally, the reason to call it a Hub instead of a

Center is that the ECC102E locates in the Engineering and Computing Center building. This new workspace for development and research supports undergraduate and postgrad students in all areas of their computing-related creativity and imagination. The students have the opportunity to translate their ideas into applied systems.

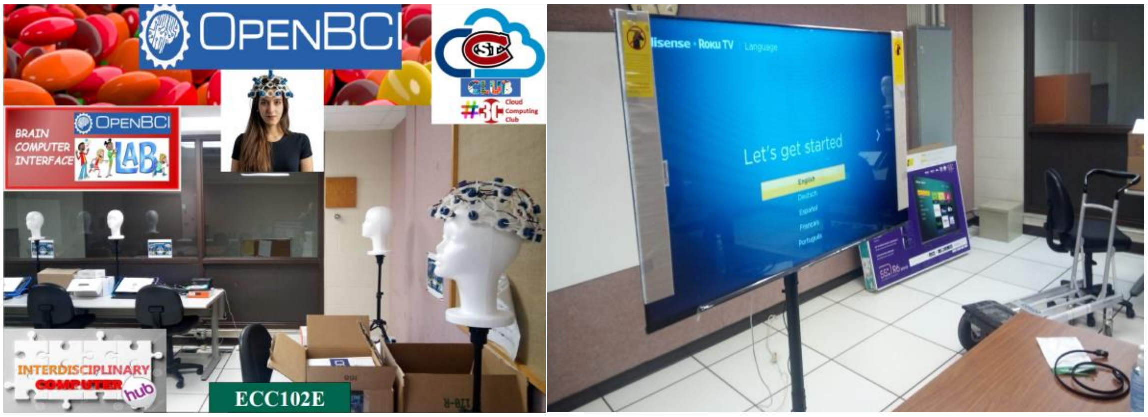


Figure 14: Multimedia - the stand with a wheel holds the TV, making it easy to move around for open lab demos and events.

The Brain-Computer Interface Lab (Figure 14) received most of the equipment in January and early February this year, yet some other components arrived just recently. As one can see, we just assembled and mounted the smart TV to allow students to visualize the drone bird's eye view on the big screen [9].

5.2 Cloud Computing Club

The club brings together all students interested in implementing distributed high-performance cloud computing. The club started on March 2022, initially with about eight active members. Discussions about it started in November 2021, and it became an official student's club on 2/23/2022. The Cloud Computing Club (#3C) provides the following:

- A forum for discussion on the newest technologies and trends.
- Providing a bright space for joint development.
- Discussion on professional development strategies and mentoring.

For team communication and new members, we have the Cloud Computing Club online discord permanent link that allows anyone interested in the ongoing development and news to join in. Currently, the #3C Discord community counts 35 subscribed members, and the official #3C university member's registration log counts presently as having 18 students. Thus, #3C has 53 students participating in the club activities. Those activities include professional mentoring, participation in events like workshops and conferences, and hands-on development of cluster cloud computing technologies applied to high-performance distributed computing for brain-computer interface integration. One logical explanation for that remarkable growth is the fact that the students are interested and motivated in having the opportunity to have their technical skills and abilities tested in a practical approach. We have patches, and only the best implementation is selected, tested, and approved to enter our source code base. Having a code peer-reviewed by a team and chosen as the best solution among several code version submissions is an astonishing thing to add to the student's Portfolio when going for a job interview.

5.3 Course Curriculum

The program proposed and approved the SE475 Brain Robotics Interface course. It provides a feasible approach to allow the students a comprehensive pathway into the Brain Computer Interface realm and technologies. The fact that we have been able to organize infrastructure in a relatively fast way, the created workspace was also of remarkable value, which indeed helped the students to start learning how the different brain wave bands work using machine learning to put into practice ways to implement direct control in real time of systems and devices through the power of your mind. After the concept and different hardware components and parts came together, we concluded that the created infrastructure could integrate as part of other course curricular activities, which we did as detailed further down in this article.

5.4 Security

The students learn the importance of taking the necessary steps to avoid hacking. They use the Discord Forum to collaborate with real-time remote development discussions. Usually, any broadly popular forum or app like Discord would be safe. However, Discord is not peer-to-peer encrypted. The students didn't check on that, so while discussing, one of them inadvertently texted the root password in the forum and deleted it a few minutes later. A week later, the VPS host informed us that our suspended server became a tool for DNS attacks as a zombie server. So, nobody got any losses or liability, as we quickly took the necessary actions to solve the problem. It was a real-life good lesson for the students about being extra careful. Since then, we have adopted many steps to ensure that our server stays safe, avoiding becoming targeted again. Actions taken, among other guidelines, include that data transmission from local BCI workstations to the VPS servers must be encrypted. The encryption methods adopted and recommended by #3C include Rsync, Tokio, SSH2, and Blockchain.

We were additionally setting up the server to be accessed only using SRA private and public keys instead of a simple remote connection with a password and user name. Noteworthy to mention that last year, we had just started creating our LAMP (Alma Linux, NGNIX, Apache SQL, Rust). So, there were no data britches. Besides the data security aspects of the distributed system implementation and data transmission, the other factor that we also aided in the security of the infrastructure had the place to be under 24-hour surveillance. We tried to find a reliable yet secure and easy-to-use camera recording monitoring system. After doing some search and several comparisons, we purchased the system: LaView 4MP 2K 4 Security Cameras Outdoor Indoor Wired, IP65, Starlight Sensor. This system addressed our needs for our ICH/BCI Lab workspace. For video recording storage, we have a remote Chromebook used as a machine dedicated to monitoring and visualizing ongoing activities inside the laboratory. We also placed a sign at the entrance door advising that 24-hour surveillance cameras monitor the place.

5.5 Human Subject IRB

To collect and save brain waves as part of our dataset, we must abide by the Human Subject Regulations. Thus, we submitted a proposal to the Institutional Research Board with a detailed description of the challenges involved with current activities we are developing with the participation of the Cloud Computing Club and other students that use the ECC102E Interdisciplinary Computer Hub as part of the BCI Lab. The IRB asked us to have the Principal Investigator take a course on Human Subjects and take the test to demonstrate an understanding of the IRB Human Subject regulation. They also asked us to draft and submit the form: Informed Consent for approval. All volunteers wishing to participate and contribute to the project should sign this form, especially in donating their brain waves to our development framework database. All data collected remain unidentifiable and securely stored in the cloud.

6 System Platform Development

Many open-source packages and libraries are available for implementing distributed high-performance computing. The software industry has evolved into the current era, where companies want to optimize the development life cycle of any system. Thus, working towards open source and collaborative implementation is of paramount importance.

Spring 2023		
Category	**Activity or Course name**	**# of students - capacity**
#3C	Cloud Computing Club	50
SE475/575	Brain Robotics Interface	20
SE413/513	Big Data Management Systems	40
SE350	Human Computer Interaction	20
SE610	Operating Systems	20
	TOTAL	**150**

Table 1: Current student with access upon request to the Brain Computer Interface Lab activities and respective four workstations.

6.1 Open Source

The students can experience the development and the type of technologies currently being put together through various approaches. For example, as optional extra credit activities, knowingly that their contributions will also have to be reviewed and approved by the Cloud Computing Club. Including all students from several courses that now have access to the activities related to the project that works to implement Flying Avatars with Brain Computer Interface, we can currently count on a broad engagement (Table 1).

The #3C does the code review and testing, then once the best patch is selected, that program or feature contribution done by the student interested in participating will enter the project code base. For the student, it also shows that his coding skills are peer-reviewed and that the student knows how to implement a high-quality code awarded as the best contribution for the specific feature implementation or bug that was published and up for fixing.

6.4 System Integration

The entire implementation uses open-source packages. The exciting aspect of our development framework and architecture is that the same architecture can be effortlessly reusable for another type of BCI application. From controlling drones to humanoid robots, the overall building blocks are the same (Figure 15). The critical thing that will only change is the dataset used to identify brain wave patterns associated with specific thought commands read by sensors for the avatar's control.

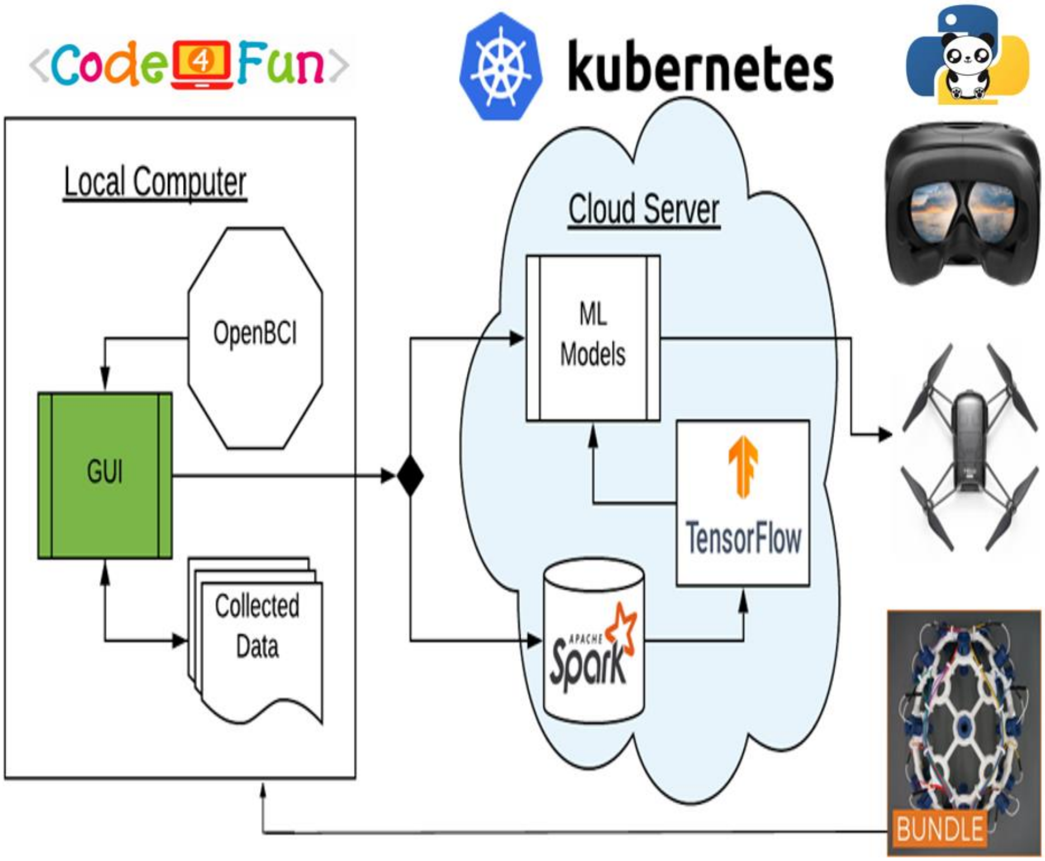


Figure 15: System architecture.

6.5 Spark

We define all necessary configurations to Spark for the 16 sensors reading from the OpenBCI headset. For such, we define Dataframe and RDD, set by default as part of Spark 2 once you create a SparkSession. The SparkSession import with SparkContext has most of the needed capabilities, including SparkSQL [10]. Pandas Dataframe allows easy data format compatibility for importing and exporting the data we connect from the cloud.

```
DF1 = InputFrame.input("TakeOFF")
DF2 = InputFrame.input("Higher")
DF3 = InputFrame.input("Landing")

set_df = df.concat([DF1, DF2, DF3], ignore_index=True)
```

Then, input all sensors data to the DenseVector:

```
dataset = []
for x in range(0, len(set_df)):
        vec = []
        for y in range(1, 10):
                if (y != 1):
                        vec.append(Vectors.dense(set_df[y][x]))
                else:
                        vec.append(set_df[y][x])
        dataset.append(vec)

        df = spark.createDataFrame(dataset,
                ["label","ch1","ch2","ch3","ch4","ch5","ch6","ch7","ch8",
                        "ch9","ch10","ch11","ch12","ch13","ch14","ch14","
                ch16"])
```

6.6 TensorFlow

For machine learning, clustering, and classification, some of the most popular open-source libraries are respectively: TensorFlow, ScikiLearn, PyTorch, and Keras. Among those alternatives, we choose TensorFlow, which integrates with Keras. Additionally, TensorFlow is considered a robust framework for Big Data processing. Specifically for Brain Computer Interface, using the algorithm derived from the Neural Networks is the most known effective way of supervised learning training your model applicable to EEG brain signal reading and pattern identification processing.

Organize the data for classification and analysis:

```
supervised_learn = []
for action in dataset:
        for data in all_data[action]:
                if action == ‘TakeOFF’:
                        supervised_learn.append([‘TakeOFF’, data])
                elif action == ‘Higher’:
                        supervised_learn.append([‘Higher’, data])
                elif action == ‘Landing’:
                        supervised_learn.append([‘Landing’’, data])
```

Apply the TensorFlow learning process to clustering classification to a dense vector:

```
def data_normalization(data):
        set = []
        set.append(dataset[0])
        for sensorBCI in dataset[1]:
                set.append(Vectors.dense(sensorBCI))
        return set
```

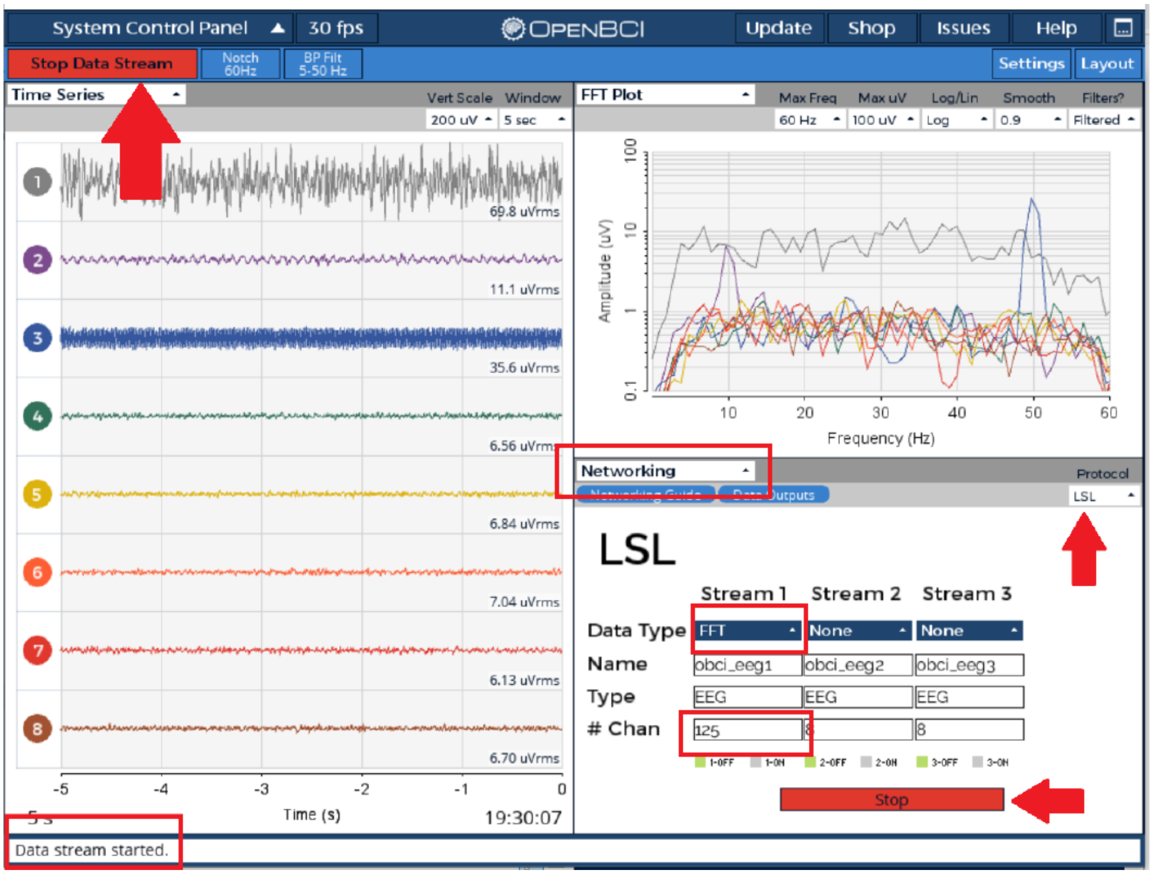
Define and import sensor data into Dataframe:

```
Rdd = sc.parallelize(supervised_learn).map(dataset)

sc.parallelize(supervised_learn.collect())
df = rdd.toDF(["label","ch1","ch2","ch3","ch4","ch5","ch6","ch7","ch8",
                "ch9","ch10","ch11","ch12","ch13","ch14","ch15",
                "ch16"])
```

blockSize – the larger the size, the faster will be the training times. One must ensure to set a value between 100-1000, and the default is 128. For better performance, we put it to 256.

_ layers – define the network configuration. For example, our BCI headset with sixteen sensors can look like this [960, 480, 3].

```
NNM = MultilayerPerceptronClassifier(labelCol="labelIndex",
          featuresCol="features",
          layers=[960,480,3],
          upper_interaction=200,
          blockSize=256)
```



LSL

	Stream 1	Stream 2	Stream 3
Data Type	FFT	None	None
Name	obci_eeg1	obci_eeg2	obci_eeg3
Type	EEG	EEG	EEG
# Chan	125	8	8

Data stream started.

Figure 16: Data transmission.

After you have defined your machine learning and the spark session, you can then transmit the data using the open "Networking" window and set the networking fields as shown to allow data transmission (Figure 16). The information "Data stream started" will appear in the bottom left-hand corner if the transmission is processing without problems. If you eventually see no message indicating that the stream started, press the stop transmission button.

7 Conclusion

This paper described the key components and steps to assemble the necessary supporting environment to allow the effective learning and development of the Brain-Computer Interface for undergrad and post-grad students. At St. Cloud State University, this aim is effectively growing stronger once the Interdisciplinary Computer Hub was established, with the subsequent arrival of the equipment put in place. Additional activities helped to instigate interest and broader engagement toward contributing to the project code that is part of the Cloud Computing Club activities. The primary motivation for students getting involved is the prospect of aiding several skill sets to help them differentiate themselves in the job market. All technologies used in the current implementation framework also hint at why so many students have decided to join the Cloud Computing Club and actively participate in the activities in the Brain Computer Interface Lab at Interdisciplinary Computer Hub. The possibilities for machine learning, distributed scalable cluster computing, and using BCI for human-computer interaction have become a reality moving quickly toward system applications where the student's brain has the power.

References

[1] Cavalcanti, A., Shirinzadeh, B., Zhang, M., and Kretly, L. C. *Nanorobot Hardware Architecture for Medical Defense*, Sensors, 8(5), 2932-2958, 2008 DOI:10.3390/s8052932.

[2] Knudson, J., Doyle, W., Hayen, M., and Cavalcanti, A. *Kubernetes for High Performance BCI Flying Avatars*, Minnesota State Symposium, St. Paul MN, USA, March 2023; Available at: https://symposium.foragerone.com/2023-posters-at-st-paul/presentations/50654 (Accessed: March 10, 2023).

[3] Cavalcanti, A., Newhard, N., Harmsen, B., and Ruymen, I. *Cloud Computing for Humanoid Robotics Tele-immersion*, Sixteenth International Conference on Technology, Knowledge, and Society, Champaign IL, USA, October 2020.

[4] Värbu, K., Muhammad 1, N., and Muhammad, Y. *Past, Present, and Future of EEG-Based BCI Applications*, MDPI, 2022 DOI:10.3390/s22093331.

[5] Pearce, K. *Understanding Brain Waves: Beta, Alpha, Theta, Delta + Gamma*, DIY Genius, 2022; Available at https://www.diygenius.com/the-5-types-of-brain-waves (Accessed: March 10, 2023).

[6] Cornelissenm, L. et al. () *Electroencephalographic markers of brain development during sevoflurane anaesthesia in children up to 3 years old*, Br J Anaesth. Jun; 120(6):1274-1286; 2018 DOI:10.1016/j.bja.2018.01.037.

[7] Guimbao, M. *Flying FPV with DJI Ryze TELLO and VR Goggles in 2023*, YouTube. 2023; Available at: https://www.youtube.com/watch?v=43tC2WTOolg (Accessed: March 10, 2023).

[8] Zhang, S. *GearedUpTech's Guide to a Mind-Controlled Drone, community open BCI*. 2019; Available at: https://openbci.com/community/geareduptechs-guide-to-a-mind-controlled-drone (Accessed: March 10, 2023).

[9] Baldwin, D. *Streaming Video from Tello and Tello EDU Drones with Python*, Jan 4, 2019; Available at: https://www.youtube.com/watch?v=kcXN7CYgQ0g (Accessed: March 10, 2023).

[10] Cavalcanti, A., Huang, C., Nguyen, T., and Qin, S. *Spark Facebook for Cloud Brain Computer Interface*, International Conference of Advanced Research in Applied Science, Engineering and Technology, Houston TX, USA, March 2020.

Transforming MoonBoard Climbing Route Classification

Joshua Grant, Michael Kirkton, Aiden Miller, Aydin Ruppe, Benjamin Weber, and Ryan Kruk

Department of Electrical Engineering and Computer Science

Milwaukee School of Engineering

{grantj, kirktonm, milleraa, ruppea, weberbw, krukr}@msoe.edu

March 18, 2023

Abstract

This study endeavors to showcase the viability of leveraging machine learning to automate the procedure of rating rock climbing routes. The focus of this investigation lies in the MoonBoard, a globally standardized rock climbing wall, which simplifies the challenge of locating relevant data and reduces the inherent complexity involved in climbing. Our proposed methodology involves utilizing attention-based models to classify routes. In our experiments, we employed an encoder predicated on the transformer architecture, and have thus far attained an accuracy of 48.8%, alongside a ±1 accuracy of 85.3%. Our empirical findings propose that deep learning models exhibit the potential to predict difficulty ratings more effectively than humans, thereby opening avenues for future work in route generation and bouldering training via data-driven approaches.

1 Introduction

The emergence of machine learning in sports has witnessed a tremendous surge in its application, ranging from forecasting game outcomes to analyzing player performance. However, the utilization of machine learning in climbing remains largely unexplored. In the realm of rock climbing, route setting represents a daunting and highly subjective task that necessitates expertise and proficiency. Currently, experienced climbers manually set the routes, while the difficulty rating of the routes remains highly subjective, as it is contingent upon individual factors such as strength, height, and experience. This subjectivity poses a significant challenge in establishing an objective rating standard for climbing problems and accurately predicting ratings based on data. Furthermore, the complexity of climbing problems and the limited availability of data compounds the difficulty of automating route setting and difficulty rating. By leveraging the MoonBoard, a standardized rock-climbing wall utilized globally in gyms, we can make several of the issues above obsolete.

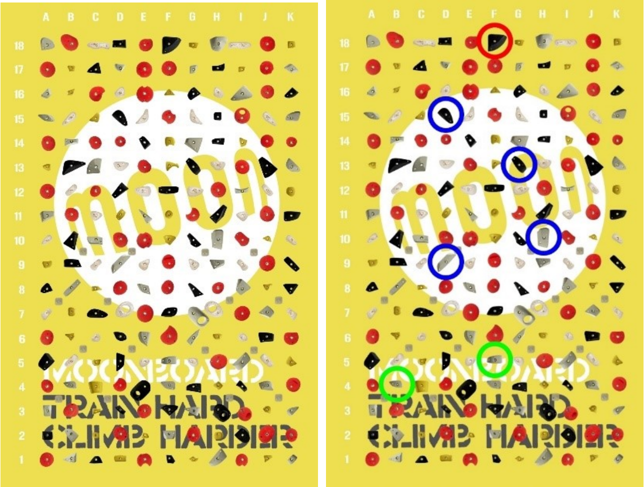


Figure 1: The 2017 MoonBoard's base configuration is on the left, with a route displayed on the right. The circles in green are the start holds, blue are the intermediate holds, and red is the end hold

The MoonBoard, a standardized training board with a fixed set of holds, has gained widespread popularity among climbers for enhancing their skills and strength. It offers three distinct variations, namely the 2016 MoonBoard, the 2017 MoonBoard, and the 2019 MoonBoard. The MoonBoard app has garnered a substantial community of climbers who have developed routes and labeled their difficulty levels, thus providing us with an abundant dataset to train a machine learning model. We utilized data from the 2017 MoonBoard as it presented the largest collection of routes at our disposal.

Several past models have represented the MoonBoard as an 11x18 matrix, with each cell indicating a hold's position. Initially, we adopted this approach; however, we observed that the information available to the learning model was insufficient, and models encountered challenges in establishing connections between distinct holds. The sparse nature of features in the matrix meant that CNN models struggled to extract features, while classical models found it challenging to establish associations between the features. Drawing upon previous model types, data representation methods, and their success (detailed further in section II), we propose leveraging a transformer-based model, entirely composed of attention mechanisms, to classify routes represented as a sequence, rather than a matrix, on the MoonBoard. The ability to classify routes presents practical applications for route setters and climbers, assisting with route setting and enabling climbers to choose routes that align with their skill levels.

2 Related Works

In recent years, there has been a growing interest in using machine learning techniques for climbing route classification on the MoonBoard. Dobles et al. [2] evaluated three different types of models, including Naive Bayes, Softmax Regression Classifiers, and a CNN to classify MoonBoard routes. They represented each route as an 11x18 matrix, and as previously mentioned, they struggled to produce satisfactory results. The use of stratified data and weighted training to combat class imbalance failed to improve their results.

In contrast, Duh and Chang [3] proposed that a sequence model is a more natural representation of a MoonBoard problem than the 11x18 matrix. They preprocessed their data into a list of vectors that represented a sequence of left and right-hand moves, combined with various other details about each hold. Their main model, GradeNet, is an RNN, specifically an LSTM, with an accuracy of 47.5% and a ±1 accuracy of 84.8%. In a similar vein, Tai et al. [6] used a GCN with attention mechanisms to classify rock climbing difficulties. They used oversampling and undersampling to limit class imbalance and achieved an accuracy of 21.9%, with a ±1 accuracy of 56.3%, which is on par with other models that used the matrix as a form of data representation.

3 Methodology

Our study extends prior research by utilizing a transformer-based encoder model, in contrast to LSTM models. The emergence of transformers in 2017 through the publication of "Attention is All You Need" [7] has led to their widespread use in deep learning. This model type has proven to be highly effective in addressing sequence-based problems, due to its self and global attention mechanisms, efficient parallelization, and rapid training. Given its success in natural language processing tasks, we opted for this model type to handle the processing of our route sequence data.

num_layers	3
d_model	128
dff	736
num_heads	8
dropout_rate	0.2
warmup_steps	3250
beta_1	0.79
beta_2	0.98
epsilon	6.3581e-08

Table 1: Hyperparamters

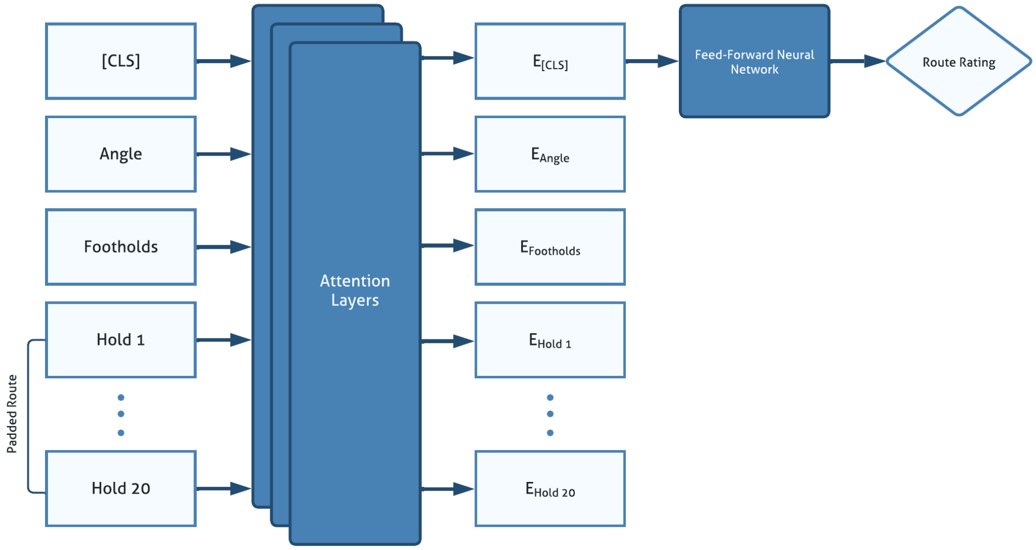


Figure 2: Model Architecture

3.1 Sequence

The sequence used in our model consisted of three distinct components. The first component was a singular token that denoted the angle of the MoonBoard. The second component was a token that captured the footholds that were available for use during the climbing route. Finally, the sequence was composed of a series of tokens that captured the specific holds present along the route. To ensure that the sequence was presented in a logical and comprehensible order, we applied a heuristic approach that approximated the order in which a climber would tackle the route. This approach was implemented with the aim of enhancing the overall effectiveness of the sequence by incorporating a degree of human insight into its design.

3.2 Class Token

As part of our model design, we implemented a class token inspired by BERT [1]. The class token is a special token that is prepended to the input sequence and represents the classification label of the input. The class token allows the model to incorporate the classification label into its training and prediction processes. This allows our model to use an attention-based architecture for the use of classifying the routes rating by running inference on the output of the attention encoding on the class token only.

4 Dataset

The class imbalance problem inherent in the dataset was addressed through a combination of weighted training as well as oversampling and undersampling techniques during the training process. Specifically, the benchmark routes, which were considered to be high-quality problems defined by expert climbers at MoonBoard, were oversampled at a higher rate than the other samples, and the lower-rated routes were undersampled to balance the distribution of ratings in the dataset. In addition, routes that had no repeated ascents were removed, as they were considered to have low quality, and routes with ratings ranging from V11 to V14 were also excluded due to the lack of accurately rated data points in those classes. The final grade distribution is shown in Appendix A2. The remaining 21695 problems were divided into training, validation, and test splits, with 17356, 2169, and 2170 problems, respectively before sampling.

4.1 Default Rating

A significant finding was that a subset of the MoonBoard database contained a high degree of error. Notably, the default rating used when a user creates a route is 6B+ on the font scale [5]. However, any user who is just playing around with the route creator or does not end up changing the rating creates a route with an improper classification [4]. Since both 6B and 6B+ from the font scale correspond to V4 on the V-scale, all 6B+ routes were removed from our dataset due to the inability to verify their rating. We believe that this issue has not been addressed in any previous studies.

5 Results

We evaluated our model using Accuracy (the percentage of true predictions out of the dataset), ±1 accuracy (the percentage of predictions within one grade of the true rating), and ±2 Accuracy (predictions within two grades). Other models have used the F1 score (a measure of accuracy and precision from 0-1.0, higher is better), and the AUC (Area under the curve from 0-1.0, with values over 0.8 considered good), which were not used in training, but have been calculated (found in Table 1), both exceeding any previous results. We found that our encoder has so far achieved an accuracy of 48.8%, and a ±1 accuracy of 85.3%, which is a little better than GradeNet, the most accurate classifier among previous studies for the MoonBoard.

	HLP	Attention (Ours)	GradeNet (LSTM)	GCN	Naive RNN	MLP	CNN
Accuracy	45%	48.8%	47.5%	21.9%	34.7%	35.6%	34%
±1 Accuracy	87.5%	85.3%	84.8%	56.3%	-	74.5%	-
±2 Accuracy	-	97.2%	-	75.6%	-	88%	-
F1 Score	-	0.362	0.242	0.310	0.165	-	-
AUC	-	0.875	0.764	0.73	-	-	-

Table 2: Our models vs. previous models

Our model's performance is on par with the best-performing models created thus far. Our model clearly outperforms the Naive Bayes, Softmax Regression, and CNN models, as well the GCN model. While our model's performance only slightly exceeds the LSTM GradeNet model, it is important to note that these models were evaluated using the 2016 MoonBoard, not the 2017 MoonBoard that we used. To compare our models completely accurately with the others, we would need to train a copy of our model on the 2016 dataset as well as tune all the hyperparameters. Despite this, we believe the accuracy demonstrated by our model is enough to prove that transformers are competitive with the best-performing models published thus far and are a viable option for MoonBoard route classification.

6 Conclusion

In conclusion, we have shown that an attention-based model applied to the MoonBoard 2017 dataset, with a combination of weighted training, oversampling and undersampling techniques and removal of unverified and high difficulty routes, achieved an accuracy of 48.8% and a ±1 accuracy of 85.3%. Our model outperformed previously explored models such as the Naive Bayes, CNN, and GCN models, and was slightly better than the GradeNet LSTM model.

Based on our findings, we suggest the adoption of attention-based models for MoonBoard climbing route classification, as they have the potential to offer improved accuracy and are more suitable for sequence modeling. Additionally, we recommend the continued implementation of sample weighting techniques, with a particular emphasis on benchmark routes. As a prospect for future research, we propose the investigation of route generation methods. Our model may be utilized to generate new climbing routes for the MoonBoard, an accomplishment that would be a valuable contribution to the climbing community. While we have commenced exploring the usage of attention mechanisms for route generation, the exploration of this calls for further research.

References

[1] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding — arxiv.org*. https://arxiv.org/abs/1810.04805. 2019.

[2] A. Dobles, J. C. Sarmiento, and P. Satterthwaite. *[PDF] Machine Learning Methods for Climbing Route Classification — Semantic Scholar — semanticscholar.org*. https://www.semanticscholar.org/paper/Machine-Learning-Methods-for-Climbing-Route-Dobles/dd282a66afd38b92698dd56ac812122403b1629a.

[3] Y.-S. Duh and J.-R. Chang. *Recurrent Neural Network for MoonBoard Climbing Route Classification and Generation — arxiv.org*. https://arxiv.org/abs/2102.01788.

[4] Remus Knowles. *Every Moonboard Problem Ever, Analysed — latticetraining.com*. https://latticetraining.com/2018/01/26/every-moonboard-problem-ever-analysed/.

[5] M. *Free Climbing Grades Conversion Chart + Complete Rating Guide — cruxrange.com*. https://www.cruxrange.com/blog/climbing-ratings-explained/.

[6] C.-H. Tai, A. Wu, and R. Hinojosa. *[PDF] Graph Neural Networks in Classifying Rock Climbing Difficulties — Semantic Scholar — semanticscholar.org*. https://www.semanticscholar.org/paper/Graph-Neural-Networks-in-Classifying-Rock-Climbing-Tai-Wu/77e2b78ed51b59558f9d48187a36bd56571553bd. 2020.

[7] A. Vaswani et al. *Attention Is All You Need — arxiv.org*. https://arxiv.org/abs/1706.03762. 2017.

A Appendix

A.1 Appendix A1: Human-level Performance

In a previous study conducted by Duh and Chang, the use of user-rated grades was found to be unfair as climbers generally adhere to the original grade, unless there is a significant discrepancy. As an alternative, the study sought to determine the accuracy of grades estimated by climbing experts who had not climbed the problems themselves. Specifically, three climbing experts were asked to estimate the grades of 40 climbing problems without actual climbing the route itself. The results demonstrated that even for experts, it is challenging to determine the grade without firsthand experience, with an accuracy of up to 87.5% within a tolerance of one grade off. The experts cited several reasons for the difficulty in accurate grading, including the challenge of grading without personal climbing experience, the suitability of problems for different body types, and the uncertainty when a problem falls between two grades.

	Accuracy	±1 Accuracy
Climbing Expert 1	47.6%	82.5%
Climbing Expert 1 (second try)	30%	77.5%
Climbing Expert 2	42.5%	87.5%
Climbing Expert 3	45%	87.5%
Estimated HLP	45.0%	87.5%

Table 3: Human level performance of estimating the grade of a MoonBoard problem without actually climbing it.

A.2 Appendix A2: Grade Distribution

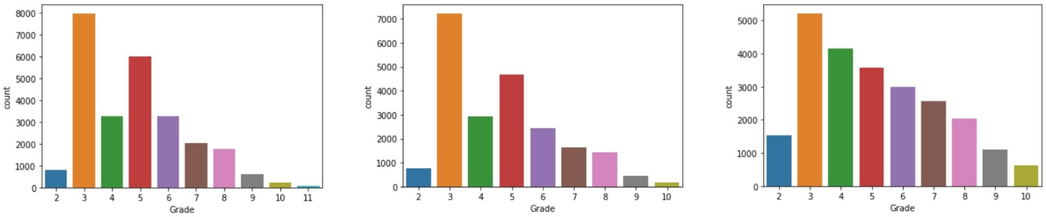


Figure 3: The distribution of our data. On the left is our initial dataset, in the center is after repeats have been removed, and on the right is after sampling has been performed.

A.3 Appendix A3: Training Plots

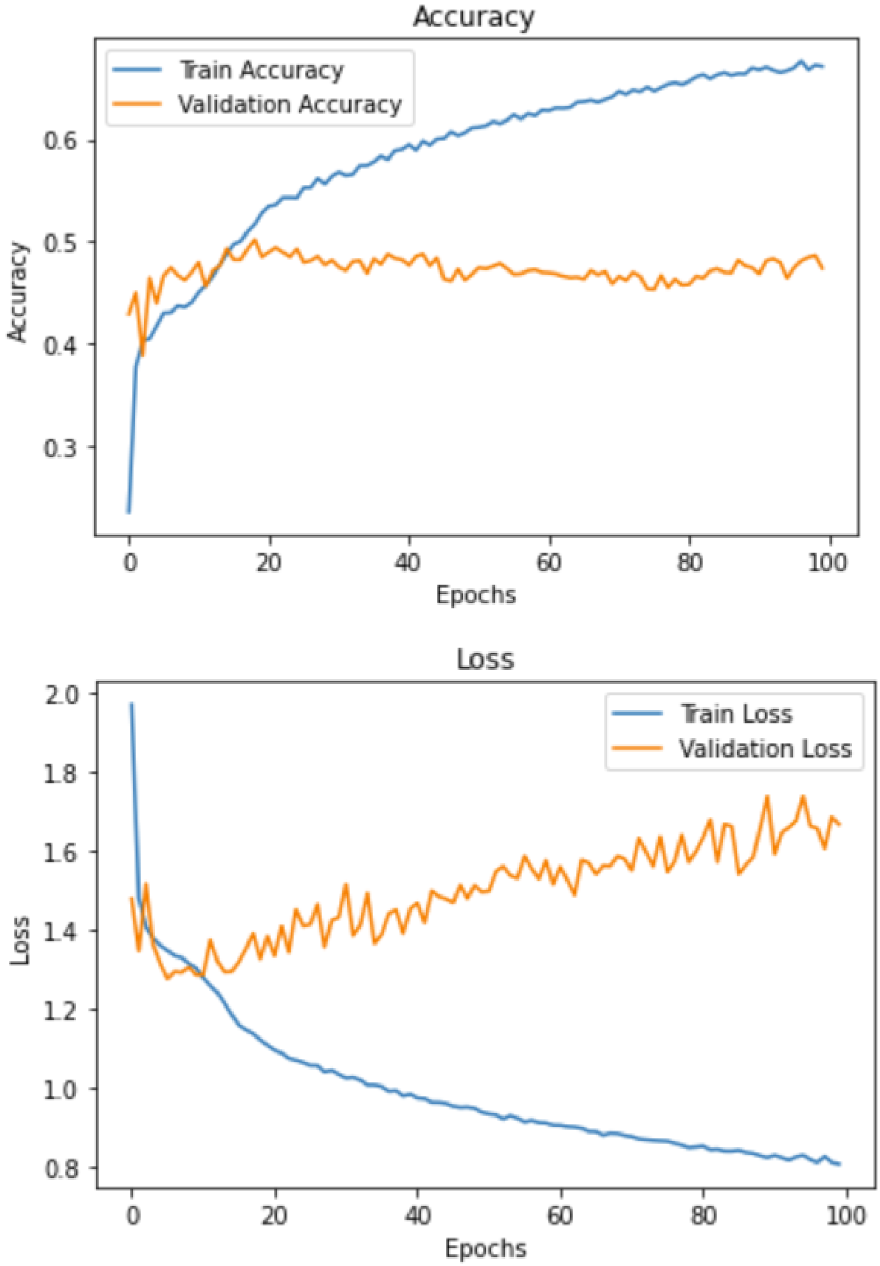


Figure 4: Training curves of the Attention model.

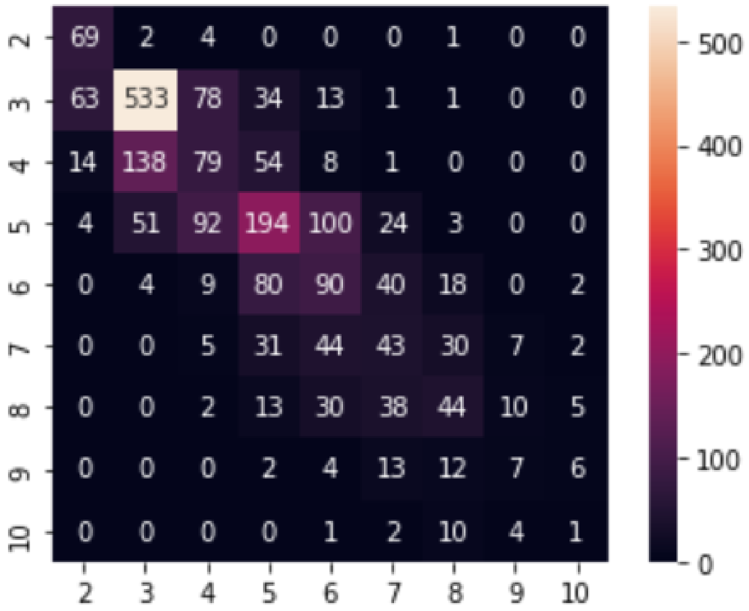
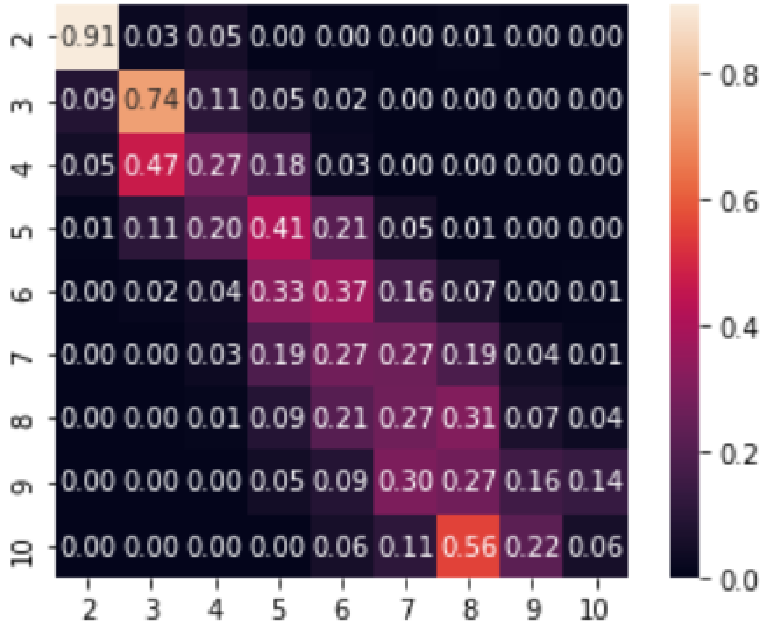


Figure 5: The confusion matrix of the Attention model. The predicted grade (x) and actual grade (y) are distributed along the main diagonal.



w

Figure 6: The normalized confusion matrix of Attention model. The predicted grade (x) and actual grade (y) are distributed along the main diagonal. Shows the percentage of each the predicted routes compared to the total number of routes belonging to that class.

Separating Spaces in Relative Attention for Music Generation

Michael Conner, Jonathan Keane, Josiah Yoder
EECS
Milwaukee School of Engineering
Milwaukee, WI 53203
{connerm, keanej, yoder}@msoe.edu

March 19, 2023

Abstract

Songwriters rely heavily on timing when creating structure to their music. Not just absolute timing but the relative distances between notes, motifs, phrases, and more are all important when creating a song. In current state-of-the-art transformers for music generation, a variant of self attention which efficiently captures relationships between tokens based on relative distance is used. The efficient method helps overcome challenges to memory requirements which restricted possible sequence lengths when relative relationships were first proposed. However, this method leaves too much room to get lost in the latent space of the attention matrix as the relative positional information is added straight into the attention matrix. We propose alternative methods for integrating the relative positional information instead to help keep a clear separation between semantic and positional information. While the complexity does increase with the proposed solution, it is still requires much less computation than the quadratic requirements of the original relative self attention system. We trained and quantitatively evaluated our transformer on the Piano-e-Competition dataset as well as qualitatively with human evaluations.

1 Introduction

When writing music, songwriters rely heavily on the timing between different musical elements, and often need to look far back at previous sections of a song in order to iterate and alter those past sections. In the past, recurrent neural networks have been used in artificial music generation, however they are restricted to storing all past elements in a fixed size state which does not lend itself well to long sequences like music or generalizing to sequences of any length. More recently, transformers have been used for this task and have show impressive performance compared to their RNN predecessors. Huang et al. [2] for instance, have achieved state-of-the-art results using a transformer with relative positional information added to the attention matrix in a similar fashion to [5].

While there are great results achieved by [2], they rely on the same latent space to capture both the positional and semantic information for the pairwise relationships between elements. We believe that this space will get too complex and confusing with the two different types of information pushed into that space. Therefore, we propose an alternative implementation for the relative information in order to keep the different types of information in completely different spaces to be mixed later in more complex ways than addition.

When the original transformers were introduced with global positional information added to the input embeddings, many attempted, and acheived better results, by concatenating the positional information to the inputs instead. This was done for the same reasoning as we stated above and serves as the inspiration for this work.

For data, the Piano-e-Competition dataset consists of human played piano performances represented as MIDI. As the performances are human, there are expressive dynamics which have very fine timings. Therefore, following [2], we use a similar representation to the one proposed by [4] where MIDI events are expressed as note-on, note-off, velocity change, and time change events and are then put into one-hot vectors.

2 Previous Work

2.1 Multi-Head Attention

An attention head takes in an input sequence $X = (\mathbf{x}_1, ..., \mathbf{x}_n)$, $X \in \mathbb{R}^{n \times d_x}$ of n elements where $\mathbf{x}_i \in \mathbb{R}^{d_x}$ is a row vector, and creates a new sequence $Z = (\mathbf{z}_1, ..., \mathbf{z}_n)$, $Z \in \mathbb{R}^{n \times d_z}$ where $\mathbf{z}_i \in \mathbb{R}^{d_z}$ and each row$\mathbf{z}_i$ is a weighted sum of a linear transformation of an input row:

$$\mathbf{z}_i = \sum_{j=1}^{n} \alpha_{ij}(\mathbf{x}_j W^V) \tag{1}$$

The weight coefficients α_{ij}, are computed via a softmax:

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \tag{2}$$

The element $e_{ij}$is computed using a matrix multiplication of two linear transformations of input elements and normalized with $\sqrt{d_z}$, also known as a scaled dot product:

$$e_{ij} = \frac{(\mathbf{x}_i W^Q)(\mathbf{x}_j W^K)^T}{\sqrt{d_z}} \tag{3}$$

Because $\mathbf{x}_i$ and $\mathbf{x}_j$ are row vectors, this is an inner product.

The parameter matrices $W^Q, W^K, W^V \in \mathbb{R}^{d_x \times d_z}$ are learned and unique for every attention head. Each head can be seen as getting a section of the embeddings, and heads learn to attend tokens to each other differently based on the section of data that they receive. Later, the data is aggregated back together to get single positions to attend.

2.2 Relative Self-Attention

The self-attention mechanism proposed by [5] alters the preexisting attention mechanisms in order to allow information on relational distances between items to be encoded directly into the attention matrix. They model the input as a labeled, directed, fully-connected graph where the edge between two input elements x_i and x_j is represented by the vectors $\mathbf{a}_{ij}^V, \mathbf{a}_{ij}^K \in \mathbb{R}^{d_a}$ where $d_a = d_z$. They use two different edge representations so that one can be used to modify eq. (1) and the other to modify eq. (3). Unlike the parameter matrices, the edge representations are shared across all attention heads. As they are meant to capture information about the relative distances between elements, it makes sense that they should be consistent in all scenarios for any context. In order to inject these representations into the attention matrices, they propose adding them to the results of their corresponding linear transformations $\mathbf{x}_j W^V$ and $\mathbf{x}_j W^K$ in eq. (1) and eq. (3) respectively. This results in the first new formula:

$$\mathbf{z}_i = \sum_{j=1}^{n} \alpha_{ij}(\mathbf{x}_j W^V + \mathbf{a}_{ij}^V) \tag{4}$$

The authors hypothesize that the relative information is important in the above equation for tasks where edge information is useful to future layers for information already being attended to by a given head, however it is the less important of the two places where edge information is injected.

This also results in the second, more important, new formula:

$$e_{ij} = \frac{(\mathbf{x}_i W^Q)(\mathbf{x}_j W^K)^T + (\mathbf{x}_i W^Q)(\mathbf{a}_{ij}^K)^T}{\sqrt{d_z}} \tag{5}$$

The idea behind adding the relative information here is to use that information when determining compatibility of queries and keys. They state the motivation for using simple addition in these cases is for an efficient implementation. Note as well that while this equation could factor out $(\mathbf{x}_i W^Q)$, multiplying before the addition eliminates the need to broadcast the relative position information. While this is more efficient, it is still quite costly with a space complexity of $O(n^2 d_x)$ as pointed out and improved by [2].

2.3 Memory Efficient Relative Self-Attention

To improve upon this, [2] first dispose of the relative position representation used for values (Eq. 4) entirely as [5] was unable to prove a significant improvement while including it with the query relative representations. Huang et al. [2] thus propose the new relative attention function:

$$RelativeAttention = Softmax(\frac{QK^T + S_{rel}}{\sqrt{d_h}})V \tag{6}$$

where $S_{rel} = QR^T$, $R_{ij} = \mathbf{a}_{ij}^K$, $Q = XW^Q$, $d_h = \frac{d_x}{h}$, and $K = XW^K, V = XW^V$ are their respective linear transformations. To reiterate, $Q \in \mathbb{R}^{h \times n \times d_h}$, $R \in \mathbb{R}^{h \times n \times d_h}$, and $S_{rel} \in \mathbb{R}^{h \times n \times n}$, where h is the number of heads and R is the matrix of relative positional representations. The benefit of using S_{rel} is that the memory requirements go from $O(n^2 d_x)$ to $O(nd_x)$, additionally, while the time requirements remain the same on paper, the authors found that there was about a 6x speedup as well. Huang et al. [2] achieve this by creating what they call the 'skewing' procedure.

The skewing procedure performs the operation $QR^T$where R is a matrix representing the embeddings of the relative positional information. While on paper the details of this and the original relative self attention look the same, [5] uses an implementation which adds an extra dimension to the data as an intermediate step in the computation of S_{rel}. As shown below, Q and R are matrices where each row is representing a query in the case of q, and a relative positional embedding in the case of r, r_{-2} denotes an embedding for the relationship between a query and the value that is two behind it relatively. After computing QR^T, the values that relate queries to non-existent values due to looking too far behind are removed with an upper left mask. These positions will eventually end up in the location where the look ahead mask in regular attention is. Now the task is to efficiently shift each row so that it's masked values are on the right instead of the left. In order to do this, one can pad a column onto the left of the matrix, reshape the matrix from $(n+1) \times n$ to $n \times (n+1)$, and slice out the desired lower portion of the result in order to get the final value of S_{rel}.

$$Q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix}$$

$$R = \begin{bmatrix} r_{-2} \\ r_{-1} \\ r_0 \end{bmatrix}$$

$$QR^T = \begin{bmatrix} q_1r_{-2} & q_1r_{-1} & q_1r_0 \\ q_2r_{-2} & q_2r_{-1} & q_2r_0 \\ q_3r_{-2} & q_3r_{-1} & q_3r_0 \end{bmatrix}$$

$$QR^T = \begin{bmatrix} 0 & 0 & q_1r_0 \\ 0 & q_2r_{-1} & q_2r_0 \\ q_3r_{-2} & q_3r_{-1} & q_3r_0 \end{bmatrix}$$

$$S_{rel} = \begin{bmatrix} q_1r_0 & 0 & 0 \\ q_2r_{-1} & q_2r_0 & 0 \\ q_3r_{-2} & q_3r_{-1} & q_3r_0 \end{bmatrix}$$

3 Experiments

In the relative positional embeddings used by [2] the embedding matrix is incorporated into the model by multiplying it by the query matrix, and then adding the result to the attention matrix. However, this leaves room for the model to intermingle and confuse semantic and positional information which was shown in a different context to hurt the perormance of a Transformer model [3]. In order to help alleviate this we propose different integrations of the relative positional information that aim to better fuse the information into the model instead of simple addition.

The hyperparameters were kept consistent between our experiments and the baseline model. The optimizer, and learning rate warmup and decay were the same as described in [6], however, we reset the learning rate every 400,000 steps in order to speed up training. The inputs were learned embeddings of size 512 and were then concatenated with sinusoidal position embeddings again as described in [6] resulting in inputs of size 1024.

Table 1: Hyperparameter configuration.

Hyperparameter	Value
Sequence Length	2048
Batch Size	8
Transformer Layers	6
Model Dimension	1024
Feed Forward Dimension	2048
Heads	8
Dropout	0.1
Learning Rate	0.1

3.1 Additional Attention

In order to better integrate the relative positional information into the attention mechanism, we propose adding an additional query and value transform of the input which require their own parameters $W_2^Q, W_2^V \in \mathbb{R}^{d_x \times d_z}$ and results in $Q_2, V_2 \in \mathbb{R}^{h \times n \times d_h}$. The relative self attention mechanism originally used the same values to query into the keys given by inputs and the keys that translate to relative positions. We plan to learn those queries and corresponding values again as functions of the input but separately:

$$RelativeAttention = Softmax(\frac{QK^T}{\sqrt{d_h}})V + Softmax(\frac{Q_2R^T}{\sqrt{d_h}})V_2 \tag{7}$$

As in previous works $R \in \mathbb{R}^{n \times d_h}$. R, unlike the other parameters in the attention mechanism, is consistent across attention heads, although still unique per transformer layer as this is at the level of the attention mechanism. In order to efficiently compute this across heads, an einsum is used. This increased our parameter count by 1.3x, and there was no noticeable difference in run time.

3.2 Data Augmentation

As previously stated, we used the same data augmentation methods as proposed by [4] except for cases where the vocabulary of notes would be extended. We also dropped time shift events. For implementation, we generated all of the avaliable pitch shifts in the range of [-3,3] ahead of time, and each epoch iterated over all of the new data.

4 Results

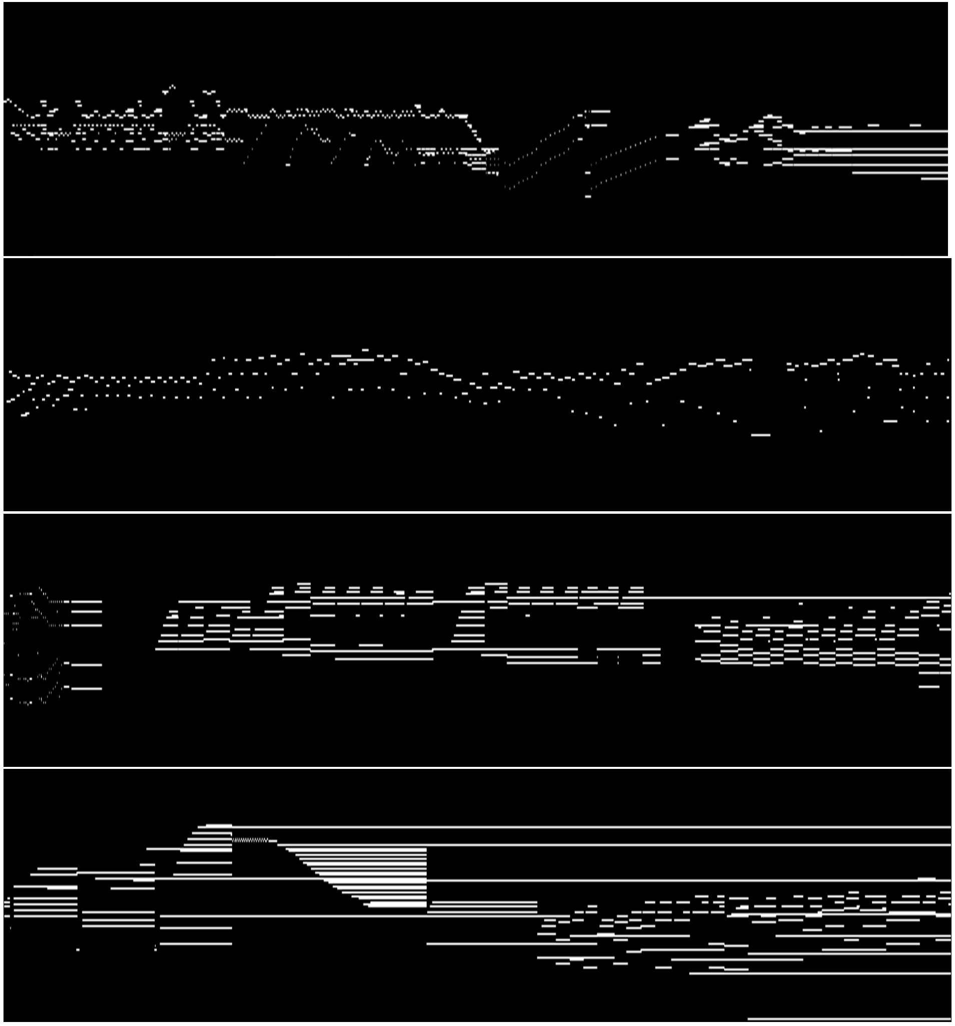
For training, 4 Tesla V100 GPU's were used. The compute was supplied by the the Milwaukee School of Engineering's super computer ROSIE. After hyperparameter tuning on each different experiment, two types of evaluation were used. First is the quantitative evaluation which was used for hyperparameter selection. The quantitative evaluation was the categorical cross entropy of a portion of the training data reserved for testing. The train/test split was 90/10 after shuffling the training data. After model selection for the individual experiments based on quantitative evaluation we will move on to qualitative evaluation. The qualitative evaluation will consist of blind participants listening to samples generated by a pair of models off of the same primer sequence or off of no primer and choosing which output they favored more.

For generating figures that were not given a primer, simply choosing the most likely event would result in the same generated sequence every time. In order to remedy this, we used an epsilon of $e = 0.6$ as a likelihood to not always pick the most likely event, and instead sample from the top-k likelihoods with probabilities equal to their model outputs with $k = 32$.

4.1 Data

The dataset used was the piano-e-competition. All of the years prior to 2018 were used as training and testing data, while the 2018 competition was withheld throughout the entirety of training as a validation set. That set was used for the

Figure 1: An example visualization of a sample generated from a primer in our testing dataset, along with three samples generated from no primer.



final quantitative evaluation below, as well as for generating primers to generate off of for the qualitative evaluation.

4.2 Quantitative Evaluation

Table 2: Qualitative evaluation of each model based on it's cross entropy of the validation dataset. The validation dataset was completely withheld during training. We also found that our model converged much faster than the baseline.

Model Variation	Validation Cross Entropy
[2] Music Transformer (Baseline)	2.0608
Our Additional Attention Model	1.9239

4.3 Qualitative Evaluation

For a qualitative evaluation we had participants listen to pairs of samples that were generated in batches and had them pick which of the samples they preferred per pair. Pairs were either generated off of the same primer or were generated with no primer as described in section 4.

Table 3: Quantitative evaluation of models compared pairwise in human listening tests. Wins and losses are from the perspective of Model A. The p-value was calculated using a binomial test. A statistically significant difference is denoted by an asterisk.

Model A	Model B	Win	Loss	p-value
Our Model	[2] (Baseline)	43	19	3.2e-3*

5 Conclusion

We have seen both a quantitative and a qualitative improvement over the Music Transformer [2] with the trade off of increasing memory requirements due to the extra parameters. While our quantitative improvement is small the results of our qualitative evaluation shows a significant improvement.

Bibliography

References

[1] Michael Conner, Lucas Gral, Kevin Adams, David Hunger, Reagan Strelow, and Alexander Neuwirth. Music generation using an lstm, 2022.

[2] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.

[3] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. 2020.

[4] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: Learning expressive musical performance, 2018.

[5] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations, 2018.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Appendix A: Related Work

For a long time, sequence models have been the tool of choice for single track music generation, from Recurrent Neural Networks, to Gated Recurrent Units, and Long Short Term Memory [4, 1]. The representation of music has taken many forms, such as a discretized multi-hot grid based on time [4] [2], or using one-hot representations with specific events for changing time [4] [2], or using multi-hot representations to capture the note turned on, velocity, duration, and time-shift as a single event [1].

In recent years the transformer has been used in many applications ranging from image generation, speech summation, and chat bots. Many of these applications all share the improvements from their predecessors by increasing the length of sequences that are able to be input or generated. While the improvements are substantial there is an issue, as memory requirements grow quickly with longer sequences. This is especially the case for some implementations which increase memory requirements already in order to capture more information such as the relative self attention mechanism proposed by [5].

Appendix B: Annotated Bibliography

Attention is All You Need

[6] describe a new alternative to RNNs, LSTMs, GRUs and other models that handle sequential information, which they named the Transformer. The key element to their work is the lack of hidden states encoded into other hidden states. This caused information to be lost as time went on making large sequence lengths impossible. As the name suggests, the authors suggest a model which relies almost exclusively on attention mechanisms and remove the use of hidden states as RNNs use them. As they attend to entire sequences and no longer have the issue of hidden states being drowned out as sequences increase in length, they are able to handle tasks with much larger sequences. Like many other models made for sequence to sequence tasks, they use an encoder and a decoder. Each

uses attention in a new mechanism they propose called multi-head attention. Then using regular dense layers and normalization, they are able to achieve state of the art results in translation tasks. Since then, the Transformer has been iterated upon and used in many other tasks.

Self-Attention with Relative Position Representations

[5] builds upon the attention mechanism by creating a new positional representation. Unlike the sinusoidal positional embeddings used by [6], the new positional representation is first learned, and also captures positional information about the relative distances between tokens rather than absolute. This is stored in a matrix of the same size of the attention matrix, and is then added pairwise to get the new attention matrix. Also unlike the original positional embeddings, the relative information will be the same between the first and third elements, as it is for the second and fourth elements.

This Time with Feeling: Learning Expressive Musical Performance

[4] introduces many new ways to capturing many different types of representations for music. The relevant sections of the paper are those which cover expressive performances of polyphonic piano captured as midi data, as well as their methods of data augmentation. They introduce a way to use one-hot encoding to capture midi events and their expressive timings for live performances. In addition, this proposal also captures more events per time period when music is more dense in content, this gives a better representation that past representations which use sequences that correspond to fixed time periods.

The data starts as MIDI files which are preprocessed to remove all events that are irrelevant to capturing the needed data. Then all of the relevant events are condensed to account for events like a change in sustain pedal usage. The sustain pedal events give a value in [0-127], the pedal is regarded as in the on state when the event's value is $>=$ 64 and off when the value is $<$ 64. The note off events during time frames where the sustain pedal is in the on state are shifted to be at the same time the sustain pedal transitions to it's off state, or when the same note is repeated again, whichever happens first.

The events are then one-hot encoded where [0-127] represents note on events, [128-255] represents note off events, [256-355] represents time shifts, and [356-387] represents velocities. Unlike [4] which proposes 125 time shifts of 8ms increments, we chose 100 time shifts of 10ms time shifts. MIDI velocity events are also represented as a value between [0-127], however binning these into 32 bins greatly reduces complexity and still captures the dynamics needed to represent the performances.

They propose two types of data augmentation for less and more augmenting. The less augmentation consists of transposing all examples up or down all intervals up to a major third which creates 8 new examples. In addition, all examples can be stretched by +-2.5% or +- 5.0% in time which creates 4 new examples.

Music Transformer

[2] makes the first introduction of using transformers for music generation. They show that transformers provided better results than their LSTM predecessors in both quality and output sequence length. In addition to their contributions to the field of music generation, they also provide algorithmic contributions to the relative self attention mechanism proposed in [5]. Their algorithmic contributions take the previous relative self attention mechanism which has memory requirements of $O(n^2 d_x)$ and reduces them to $O(nd_x)$. While the time complexity still remains at $O(n^2 d_x)$ they show improvements in model performance as well. This reduction is what allows them to handle sequences of higher lengths than previous works.

Additionally, the authors propose two different model architectures based on the type of music to be generated, continuations or accompaniments. In the case of accompaniments, which the authors generate for the JS-Bach dataset and introduce first, they use the original encoder-decoder type architecture as proposed in [6]. However, when the authors shift to generating continuations of given music pieces they instead opt to completely remove the encoder section of the network and put everything into a decoder. The decoder, and thus the core of the network, is then just an encoder with masking and a different attention mechanism followed by the final linear and softmax layers.

Encryption Methods and Key Management Services for Secure Cloud Computing: A Review

Tristan L. Moore
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
tlmoore@go.stcloudstate.edu

Samuel S. Conlon
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
ssconlon@stcloudstate.edu

Anushka U. Hewarathna
Department of MSIA
Saint Cloud State University
St. Cloud, Minnesota, 56301
auhewarathna@go.stcloudstate.edu

Thivanka B. M. Dissanayaka M.
Department of MSIA
Saint Cloud State University
St. Cloud, Minnesota, 56301
thivankabm@gmail.com

Akalanka B. Mailewa
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
amailewa@stcloudstate.edu

Abstract

Utilization of public Cloud Service Providers (CSPs) has increased drastically since its inception, with many businesses using a Software-as-a-Service (SaaS) business model, meaning their entire business is run on the cloud. Due to this business model's increase in popularity in recent years, CSPs need to make security of data in the cloud their top priority and give their customers the proper tools they need to protect their data while using their services. This paper intends to give a comprehensive overview of the Encryption Key Management Services offered by two of the most frequently use CSPs: Amazon Web Services (AWS) and Google Cloud Platform (GCP). In this research, AWS Key Management Service (KMS) and Google Cloud Key Management Service offerings were tested hands-on within each respective CSPs cloud console. Each Key Management Service was thoroughly tested to fully understand their capabilities, use cases, and faults. Based on the findings it can be observed that AES-256 was the clear winner for symmetric encryption key use and RSA for asymmetric encryption keys. In addition, this paper identifies and presents several open research problems in the field of cloud-based encryption as well.

Keywords: Risks; Privacy; Performance; Amazon Web Services (AWS); Encryption; Google Cloud Platform (GCP); Key Management Service; Could Security

1 INTRODUCTION

This paper reviews the encryption methods and key management services for secure cloud computing by using AWS and GCP. This section briefly overview the context and definitions that will give guidance to the overall subject of this discussion.

1.1 Cloud Service Provider (CSP)

A cloud service provider (CSP), is a company that offers some components of cloud computing [1] [2]. The most common methods to leverage CSP services are Infrastructure as a Service (IaaS), where a Company maintains on premises servers and datacenters, but pushes some of the workload and storage to the cloud. Platform as a Service (PaaS), where a Company maintains their own application engine, but leverages the cloud to deploy their application, and Software as a Service (SaaS) where a Company completely utilizes the cloud for hosting their application and storage. The most widely used public cloud service providers include Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure [3][4].

1.2 Symmetric Encryption

In symmetric encryption, only one key is used to encrypt and decrypt data [5] this is typically used for encrypting storage volumes, databases, and storage buckets as the computing overhead is generally much lower in symmetric encryption than asymmetric. The industry standard symmetric encryption algorithm utilized by all public CSPs is Advanced Encryption Standard with 256-bit length key (AES-256) [6].

1.3 Asymmetric Encryption

Asymmetric encryption, also known as public key cryptography, encrypts and decrypts data using two separate yet mathematically connected cryptographic keys. These keys are known as a “public key” and “private key.” Together, they’re called a “public and private key pair.” [5][7] Asymmetric encryption is typically used in end-to-end encryption between a client and server. On a website for example, the owner of the website maintains the private key with a certificate signed by a certified authority (CA) and distributes a public key to each client whenever they visit the website [7][8].

1.4 Key Management Service

The three most widely used CSPs (AWS, GCP, and Azure) all offer key management services [9]. A key management service allows customers to centrally manage all of their encryption keys within the cloud. Customers can create, manage access, delete, and rotate their encryption keys within their key management service dashboard. This is particularly useful as the service is integrated with Identity and Access Management (IAM) services, allowing for granular distribution and restriction of access to these encryption keys. These encryption keys can encrypt access to storage volumes, databases, storage buckets, and even different cloud-based services such as infrastructure monitoring logs [10].

2 BACKGROUND

Security is generally a large concern among previous works regarding the topic of key management services in the cloud, this is due to the fact that a single point of failure is created when customers become over-reliant on cloud services to manage resources and sensitive customer data [11]. Depending on the type of data stored within databases, storage volumes, and storage buckets in the cloud, customers want a full guarantee that their data will be safe utilizing the CSPs methods of encryption key management [12]. To assure the security in cloud based environments, this paper presents two frameworks as follows as prior work.

Ahmad, Shahnawaz, et al. [13], the authors go in depth about secure encryption key creation and processes that must be in place to ensure confidentiality, integrity, and availability are not compromised for these keys which are managed through a CSP. Random bit generation for keys must be truly random enough to avoid the key encryption algorithm from being compromised and keys being distributed must be protected during transmission. The authors then proposes a more generic solution to stop data loss by implementing a Cloud-Based Data Loss Prevention (DLP) framework. The proposed framework consists of six major categories:

1. Data Discovery and Classification – This category focuses on determining what data is worth protecting, as well as associating a level of protection for each document. For example, levels of sensitivity could vary from public information which requires no protection, to confidential information which must be stored encrypted at rest [13][14].
2. Advanced OCR & NLP Capability – This is the process of utilizing machine learning and artificial intelligence software to determine what data is stored in documents and how sensitive the information is. This is important as an IT security department doesn’t necessarily have the time to sift through all stored documents and classify them based on sensitivity of the data [13][15].
3. Fingerprinting & Tagging – This is the process of identifying IT assets, this can be managed using an asset management platform or mobile device management (MDM) tool. This process is important to identify mission-critical information assets and what levels of protection each asset may require to keep the business operational [13][16].
4. Clipboard Monitoring – This process involves logging all user interactions within the cloud-based environment. This is important for identifying potential disgruntled employees and insider threats, as well as unusual/malicious activity occurring in the environment [13][17].
5. Content-Based Policy & Rules – This process involves establishing baselines that all employees must adhere to, as well as give direction to restoring the environment in the event of a disaster or disruption. Other policies could include ethical behavior requirements for employees and an incident response plan in the event of a potential data breach [13][18].

6. Compliance Management – This process involves adhering to various compliance frameworks and requirements to assure user entities of the business that their data is being properly protected [13][19].

This framework establishes defense-in-depth by hardening security at all layers and facets a business may operate on.

Noor, Talal, et al. [20] the authors propose a "Trust Management Framework" for public cloud service providers (e.g., AWS, GCP, and Azure) to maintain a healthy trust relationship between the customer and CSP. This framework proposes three layers:

1. Trust Feedback Sharing Layer – This layer consists of different parties including cloud service consumers and providers, which give trust feedback to each other. The feedback is maintained via a module called the Trust Feedback Collector. The feedback storage relies on the trust management systems, in the form of centralized, decentralized, or within the cloud environment through a trusted cloud service provider. Within the Trust Feedback Sharing Layer lies four dimensions [20]:

1.1. Credibility – This dimension refers to the quality of the information or service that makes cloud service consumers or provides trust the information or the service [20][21].

1.2. Privacy – This dimension refers to the degree of sensitive information disclosure cloud service consumers may fall victim to during interactions with the trust management system [20][22].

1.3. Personalization – This dimension refers to the degree of autonomy cloud service consumers and providers adhere to within the rules of the trust management system [20][23].

1.4. Integration – This dimension refers to the ability to integrate different trust management perspectives and techniques [20][24].

2. Trust Assessment Layer – This layer represents the core of any trust management system: trust assessment. The assessment might contain more than one metric. TAL handles a huge amount of trust assessment queries from several parties through a module called the Trust Result Distributor. This typically involves checking the trust results database and performing the assessment based on different trust management techniques. TAL delivers the trust results to a database in the trust results distribution layer through the module of the trust result distributor. This procedure is taken to avoid redundancy issues in trust assessment. Within the Trust Assessment Layer lies six dimensions [20]:

2.1. Perspective – This dimension depends on the focus of the trust management approach of the framework, some may focus on the consumer's perspective, while others may focus on the CSPs perspective [20][25].

2.2. Technique – This dimension refers to the degree to which a technique can be adopted by the trust management system to manage and assess trust feedback [20][26].

2.3. Adaptability – This dimension refers to how quickly the trust assessment function can adapt to the changes of involved parties such as the cloud service consumer and CSP [20][27].

2.4. Security – This dimension refers to the robustness of the trust assessment function against malicious behaviors and attacks. This dimension is very important as a cloud service consumer is trusting the CSP to offer services to allow for the cloud service consumer to protect their data, as well as deferring other aspects of security such as physical data center access to the CSP [20][28].

2.5. Scalability – This dimension refers to one of the most fundamental aspects of the cloud. One of the most appealing aspects of the cloud is that it operates on a pay-as-you-go model, and only pay for the compute resources used. The cloud is heavily relied upon to be able to scale for any amount of network demand and to continuously grow its capabilities as time goes on [20][29].

2.6. Applicability – This dimension refers to the degree that the trust assessment function can be adopted to support trust management systems deployed for cloud services [20][30].

3. Trust Results Distribution Layer – Similar to TFSL, this layer consists of different parties including cloud service consumers and providers, which issue trust assessment inquiries about other parties (e.g., a cloud service consumer inquiry about a specific cloud service). All trust assessment inquiries are transmitted to the trust assessment function through the module of trust assessment and results distributor. The final results are maintained in a database where cloud service consumers and providers can retrieve. Within the Trust Results Distribution Layer lies four dimensions [20]:

3.1. Response Time – This is the time that the trust management system requires to handle trust assessment inquiries, access feedback, and distribute trust results [20][31].

3.2. Redundancy – This dimension refers to the degree of redundancy support that the trust management system maintains in order to manage and assess the trust feedback [20][32].

3.3. Accuracy – This dimension refers to the degree of correctness of the distributed trust results that can be determined through one or more accuracy characteristics such as the unique identification of feedback and using the proper assessment function security level [20][33].

3.4. Security – This dimension refers to the degree of protection that the trust assessments and results distributor has against malicious behaviors and attacks [20][34].

3 METHODOLOGY

This research uses a hands-on approach for the analysis of each key management service available from each of the two largest cloud service providers (AWS and GCP). The authors have utilized demo accounts within each cloud service and tested the capabilities of each service. The following investigation will provide figures to guide each of the testing steps in this analysis.

3.1 AWS KEY MANAGEMENT SERVICE (KMS)

AWS KMS allows customer to create, manage, rotate, and delete their customer-owned encryption keys. Each encryption key managed within KMS includes attached metadata that customers can view, such as the key ID, key spec, key usage, creation date, description (optional), key state, and key material. Key spec refers to the type of encryption the key utilizes, this could be either symmetric or asymmetric, as well as the type of algorithm the key supports. By default, AWS KMS keys use AES-256 encryption for symmetric keys. Asymmetric key algorithms are typically customer defined, as asymmetric encryption is rarely utilized in AWS KMS [35]. AWS supports RSA and Elliptic Curve (ECC) asymmetric key pairs [36]. Key usage determines how the encryption key is used, a key can either be used to encrypt and decrypt in most cases. However, in the case of asymmetric encryption, a key can also be used for signing and verifying signatures. Each KMS key can only have one type of usage associated with them. Key state determines the current status of the key, this could be enabled, disabled, or pending deletion. Key material refers to the string of bits that make up the encryption algorithm of the key, this must be kept secret to protect the cryptographic operations that use it. However, public key material is designed to be shared [37].

3.1.1 AWS KMS – Envelope Encryption

When a customer encrypts their data with an encryption key, the data is protected. However, the key remains exposed. To solve this issue, envelope encryption can be used within KMS. This concept involves encrypting the encryption key that encrypts the data (referred to as the 'data key') with another encryption key (referred to as the 'root key') [38]. An AWS customer can import their plaintext data key into KMS, which will encrypt it with a root key, the root key can never leave the KMS module unencrypted. To use the key, it must be called within KMS. Envelope encryption also allows for the combination of multiple algorithms, a symmetric root key can be used to encrypt an asymmetric data key. Reference the figure below which further describes the concept of envelope encryption [39].

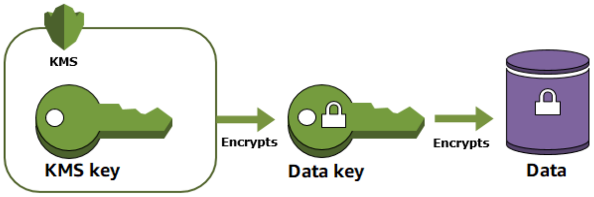


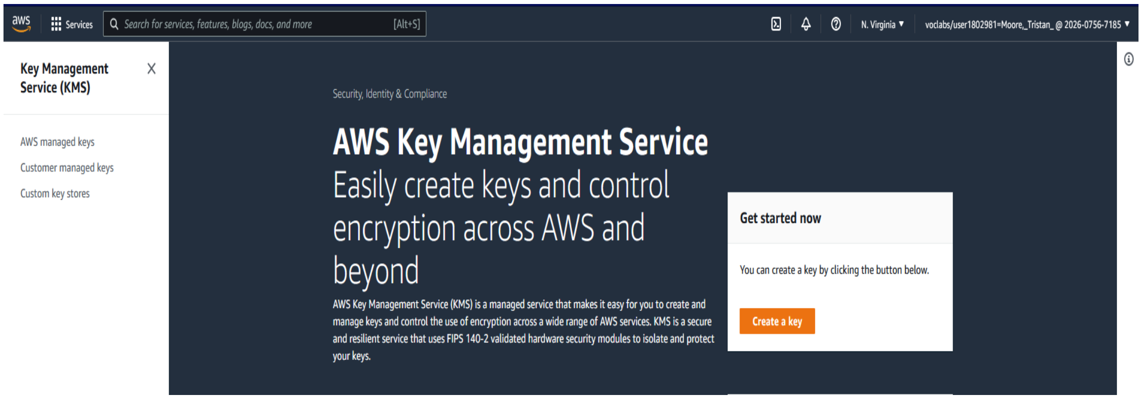
Figure 1: AWS KMS Envelope Encryption [39]

3.1.2 AWS vs Customer Managed Keys

Within KMS, there are two types of encryption key ownership, AWS managed, and customer managed. For customer managed keys, the customer retains full ownership of the key. The customer can enable, disable, rotate, and delete the key. AWS managed keys are KMS keys in a customer's account that are created, managed, and used on their behalf by an AWS service integrated with KMS top protect the customer's resources in the service. Some examples of services include AWS S3 (global storage buckets), AWS CloudTrail (infrastructure logging service), and AWS Inspector (vulnerability scanning service for containers and compute instances). Customers can still view their AWS managed key's policies and audit their use [39][40]. However, they cannot change the policies, rotate, or delete the keys. AWS managed keys are rotated on an annual basis.

3.1.3 AWS KMS Walk-Through Implementation

To further test the capabilities of AWS KMS, a demo account was created to test the service. AWS KMS was used to create a KMS key the figure below shows the AWS KMS dashboard.


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia ▼
voclabs/user1802981=Moore,_Tristan_@ 2026-0756-7185 ▼
Key Management Service (KMS)
X
AWS managed keys
Customer managed keys
Custom key stores
Security, Identity & Compliance
Get started now
Create a key

AWS Key Management Service

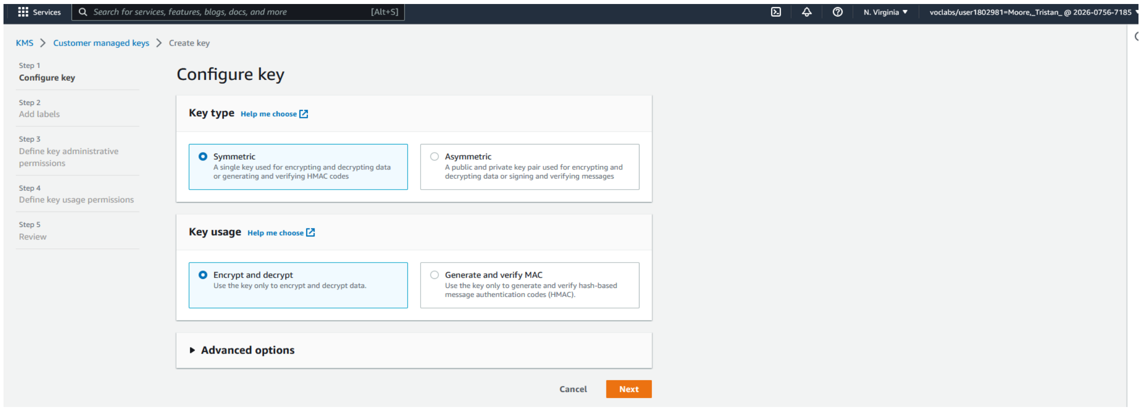
Easily create keys and control encryption across AWS and beyond

AWS Key Management Service (KMS) is a managed service that makes it easy for you to create and manage keys and control the use of encryption across a wide range of AWS services. KMS is a secure and resilient service that uses FIPS 140-2 validated hardware security modules to isolate and protect your keys.

You can create a key by clicking the button below.

Figure 2: AWS KMS Dashboard

We were then introduced to a wide variety of options such as the key type (symmetric or asymmetric), key usage, and other advanced options.


Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia ▼
voclabs/user1802981=Moore,_Tristan_@ 2026-0756-7185
KMS > Customer managed keys > Create key
Step 1
Configure key
Step 2
Add labels
Step 3
Define key administrative permissions
Step 4
Define key usage permissions
Step 5
Review
Help me choose
Help me choose
▸ Advanced options
Cancel
Next

Configure key

Key type

● Symmetric
A single key used for encrypting and decrypting data or generating and verifying HMAC codes

○ Asymmetric
A public and private key pair used for encrypting and decrypting data or signing and verifying messages

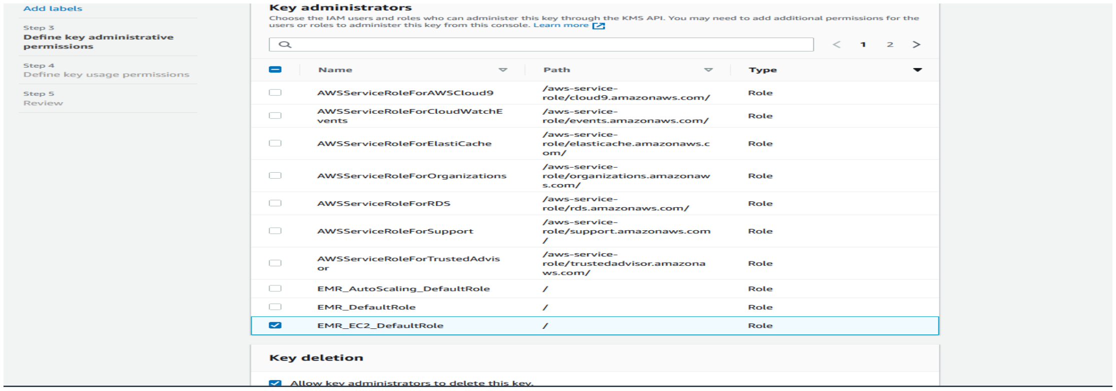
Key usage

● Encrypt and decrypt
Use the key only to encrypt and decrypt data.

○ Generate and verify MAC
Use the key only to generate and verify hash-based message authentication codes (HMAC).

Figure 3: AWS KMS Key Type and Key Usage

We were then able to allow specific AWS Identity and Access Management (IAM) groups and roles for users who were allowed have administrative permissions over this encryption key. Reference the figure below.



Add labels

Step 3
Define key administrative permissions

Step 4
Define key usage permissions

Step 5
Review

Key administrators

Choose the IAM users and roles who can administer this key through the KMS API. You may need to add additional permissions for the users or roles to administer this key from this console. Learn more

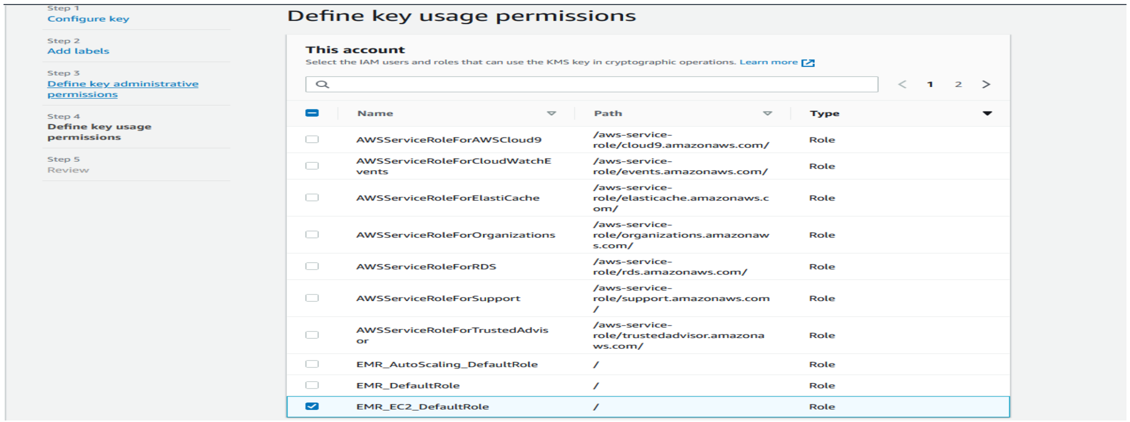
☐	Name	Path	Type
☐	AWSServiceRoleForAWSCloud9	/aws-service-role/cloud9.amazonaws.com/	Role
☐	AWSServiceRoleForCloudWatchEvents	/aws-service-role/events.amazonaws.com/	Role
☐	AWSServiceRoleForElastiCache	/aws-service-role/elasticache.amazonaws.com/	Role
☐	AWSServiceRoleForOrganizations	/aws-service-role/organizations.amazonaws.com/	Role
☐	AWSServiceRoleForRDS	/aws-service-role/rds.amazonaws.com/	Role
☐	AWSServiceRoleForSupport	/aws-service-role/support.amazonaws.com/	Role
☐	AWSServiceRoleForTrustedAdvisor	/aws-service-role/trustedadvisor.amazonaws.com/	Role
☐	EMR_AutoScaling_DefaultRole	/	Role
☐	EMR_DefaultRole	/	Role
☑	EMR_EC2_DefaultRole	/	Role

Key deletion

☑ Allow key administrators to delete this key.

Figure 4: KMS Key Admin Permissions

The next step in our configuration allowed us to assign AWS IAM groups and roles who were allowed to use the encryption key, but not have administrative privileges of the key.



Step 1
Configure key

Step 2
Add labels

Step 3
Define key administrative permissions

Step 4
Define key usage permissions

Step 5
Review

Define key usage permissions

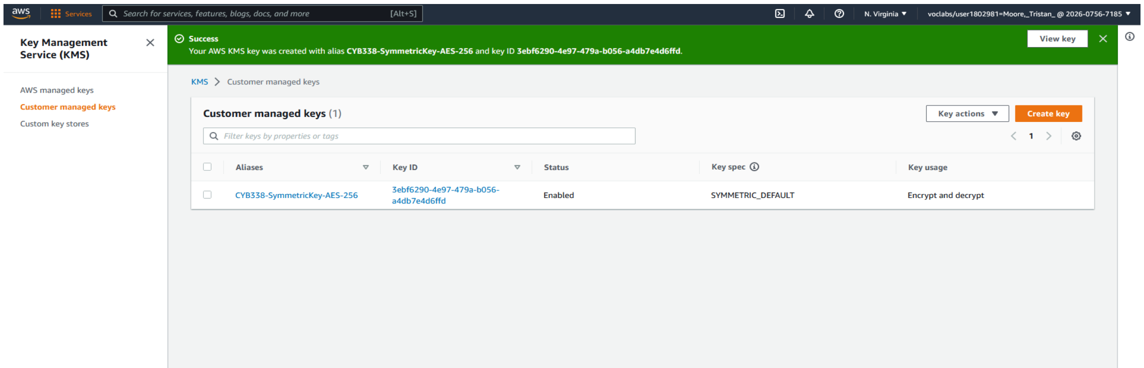
This account

Select the IAM users and roles that can use the KMS key in cryptographic operations. Learn more

☐	Name	Path	Type
☐	AWSServiceRoleForAWSCloud9	/aws-service-role/cloud9.amazonaws.com/	Role
☐	AWSServiceRoleForCloudWatchEvents	/aws-service-role/events.amazonaws.com/	Role
☐	AWSServiceRoleForElastiCache	/aws-service-role/elasticache.amazonaws.com/	Role
☐	AWSServiceRoleForOrganizations	/aws-service-role/organizations.amazonaws.com/	Role
☐	AWSServiceRoleForRDS	/aws-service-role/rds.amazonaws.com/	Role
☐	AWSServiceRoleForSupport	/aws-service-role/support.amazonaws.com/	Role
☐	AWSServiceRoleForTrustedAdvisor	/aws-service-role/trustedadvisor.amazonaws.com/	Role
☐	EMR_AutoScaling_DefaultRole	/	Role
☐	EMR_DefaultRole	/	Role
☑	EMR_EC2_DefaultRole	/	Role

Figure 5: KMS Key Usage Permissions

After this step, our key had been generated and is now visible within the AWS KMS dashboard.



aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
N. Virginia
voclabs/user1802981=Moore,_Tristan_ @ 2026-0756-7185
View key

Key Management Service (KMS)

AWS managed keys
Customer managed keys
Custom key stores

Success
Your AWS KMS key was created with alias **CYB338-SymmetricKey-AES-256** and key ID **3ebf6290-4e97-479a-b056-a4db7e4d6ffd**.

KMS > Customer managed keys

Customer managed keys (1)

Key actions
Create key
Filter keys by properties or tags

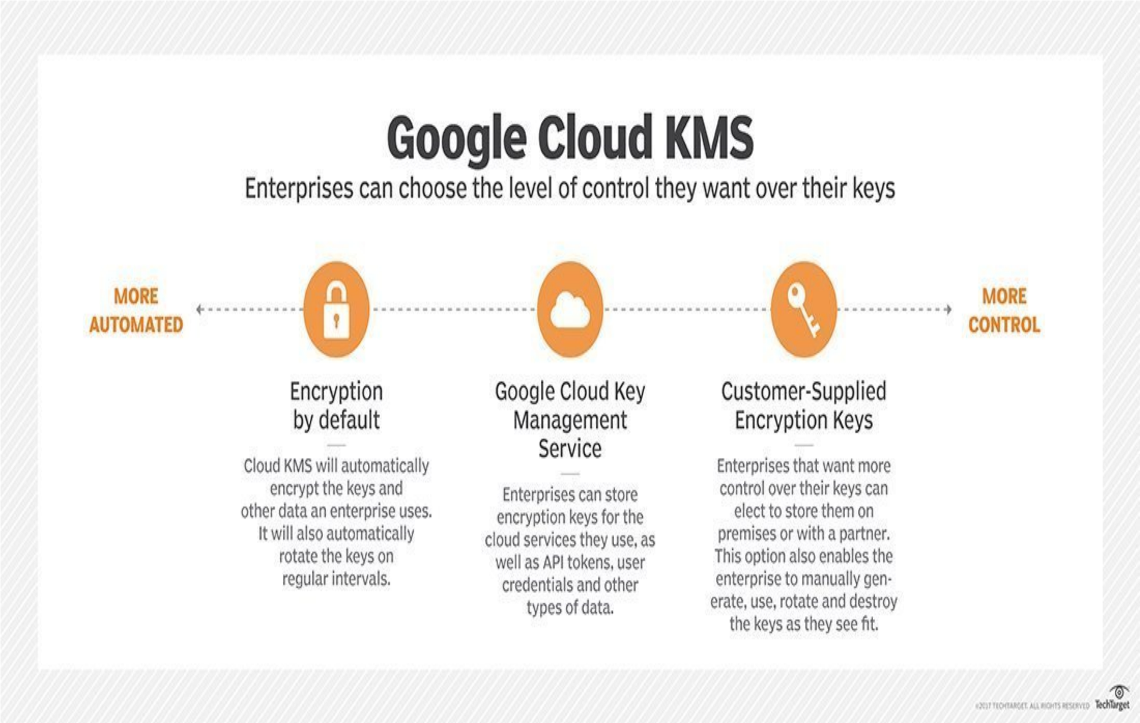
☐	Aliases	Key ID	Status	Key spec	Key usage
☐	CYB338-SymmetricKey-AES-256	3ebf6290-4e97-479a-b056-a4db7e4d6ffd	Enabled	SYMMETRIC_DEFAULT	Encrypt and decrypt

Figure 6: AWS KMS Dashboard after the KMS Key had been created

3.2 GOOGLE CLOUD PLATFORM (KMS)

Google Cloud Platform is a cloud hosted key management service that lets you manage symmetric and asymmetric cryptographic keys for cloud services. GCP can handle a variety of keys, these include AES 256, RSA 2048, RSA 3072, RSA 4096, EC P256, and EC P384 [41][42]. These can be protected via software or hardware based on user preference, and users can switch back and forth with a simple button click. Encryption keys can be managed by a third party as well using EKM (external key manager). These are deployed outside of Googles infrastructure and allows separation of data at rest and encryption keys. Using this EKM, users can request an encryption key, provide justification for said key, and a mechanism to either deny or approve the request [43].

Google KMS uses a five-level hierarchy. The top level is called GCP project which can be linked to an organization or company. After this comes keyrings, which hosts separate crypto keys. A key ring belongs to a certain project and therefore resides in a certain location. They also set permissions for the various keys they hold, so the keys within each key ring has the same permissions. These keys are subject to changes as the encryption changes. This is where the final tier comes in, CryptoKeyVersion. Google also offers a REST API as part of the KMS. This allows developers to access KMS functions to list, create, destroy, and update various encryption keys [44][45].



Google Cloud KMS

Enterprises can choose the level of control they want over their keys

MORE AUTOMATED
MORE CONTROL
Encryption by default
Google Cloud Key Management Service
Customer-Supplied Encryption Keys
©2017 TECHTARGET. ALL RIGHTS RESERVED TechTarget

Cloud KMS will automatically encrypt the keys and other data an enterprise uses. It will also automatically rotate the keys on regular intervals.

Enterprises can store encryption keys for the cloud services they use, as well as API tokens, user credentials and other types of data.

Enterprises that want more control over their keys can elect to store them on premises or with a partner. This option also enables the enterprise to manually generate, use, rotate and destroy the keys as they see fit.

Figure 7: GCP KMS summed up [44]

3.2.1 Google Cloud Platform Encryption

Google uses a very similar style of encryption for all their storage systems. While it follows the same style, the way it is implemented and rolled out varies with each system [46]. By default, GCP uses AES-256 encryption when data is at rest in storage. Data in transit is encrypted using TLS. By default, GCP uses Data Encryption Key (DEK) and Key-Encryption-Key (KEK). These two are paired and stored using Googles own Key Management Service (KMS). The KEK is used to encrypt the DEK, which was used to encrypt the actual data, to help increase security [47][48][49][50]. From here, the Google KMS is goes to work, using other services provided by Google to store the keys that are going to be used for decryption and further encryption on the cloud. To retrieve these keys, users must submit credentials/permissions proving they have rightful access. This is done using Identity and Access Management (IAM) [51].

GCP also provides another method to help further encryption. This is called Data Loss Prevention (DLP). DLP helps users identify possibly sensitive data and mask that data. Such data could include Personally Identifiable Information. By using the DLP method and combining it with the KMS, it is possible to use various encryption methods such as Format Preserving Encryption [52]. This means that the data is encrypted into an impossible to understand mess while the format is kept to the original plaintext. This method and many more can be accomplished using KMS, IAM, and DLP to help users further their encryption capabilities with their data. These services can also be setup to encrypt data automatically when uploaded to the Google cloud storage. Figure 8 shows how data is encrypted at Google. It begins by uploading the data, from there the data is chunked and encrypted separately. These encrypted chunks are then spread across Googles storage infrastructure.

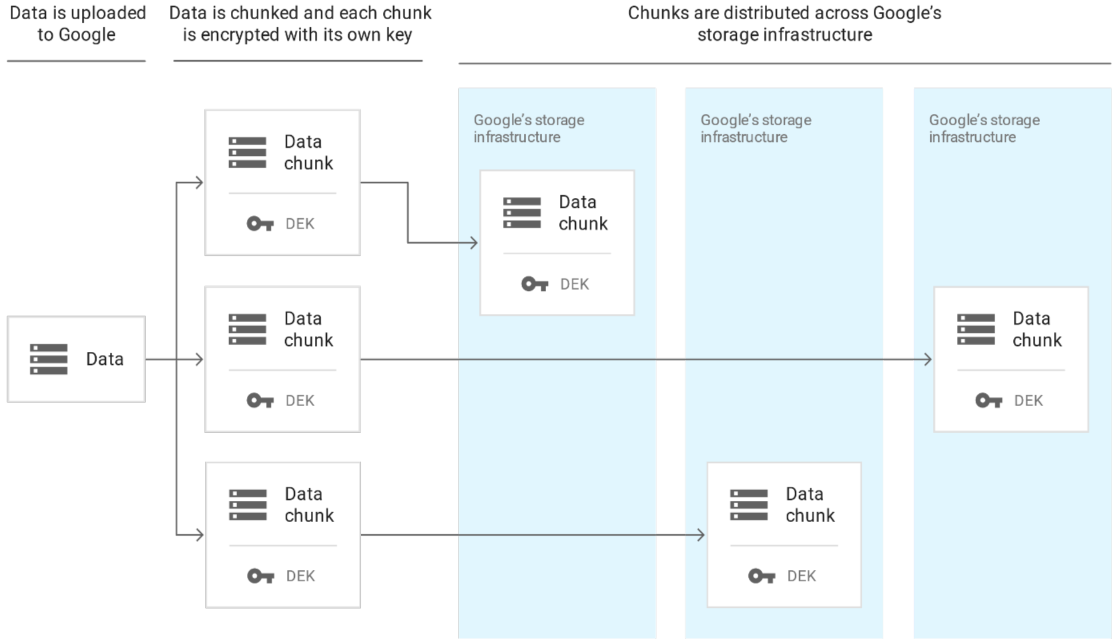
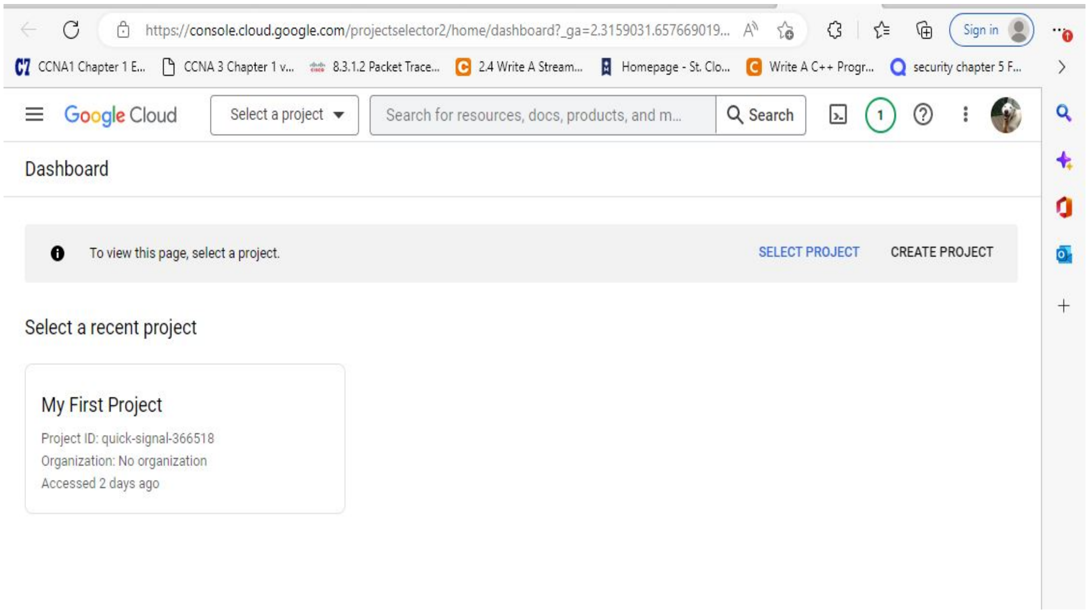

Data is uploaded to Google
Data is chunked and each chunk is encrypted with its own key
Chunks are distributed across Google's storage infrastructure
Google's storage infrastructure
Google's storage infrastructure
Google's storage infrastructure
Data
Data chunk
DEK
Data chunk
DEK
Data chunk
DEK
Data chunk
DEK
Data chunk
DEK
Data chunk
DEK

Figure 8: How data is encrypted on GCP [52]

3.2.2 Google Cloud Platform KMS Walk-Through

We are now going to show a walk-through of how to create a cryptographic key using GCP. The first step is to select the project from which the keyring will be held. We will be following the instructions found in [53].

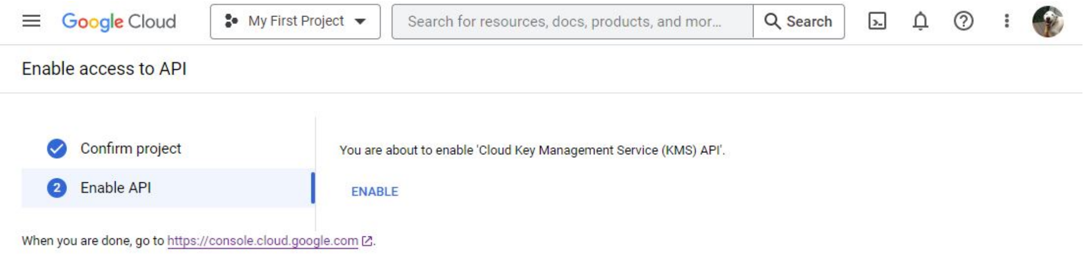


To view this page, select a project.

My First Project
Project ID: quick-signal-366518
Organization: No organization
Accessed 2 days ago

Figure 9: Selecting project to hold keys/keyring [53]

After the project has been selected, we need to make sure that some form of billing is enabled on the account to store the keys. We were using a 3-month trial so we were able to skip this step. After this we need to make sure the GCP KMS API is enabled.

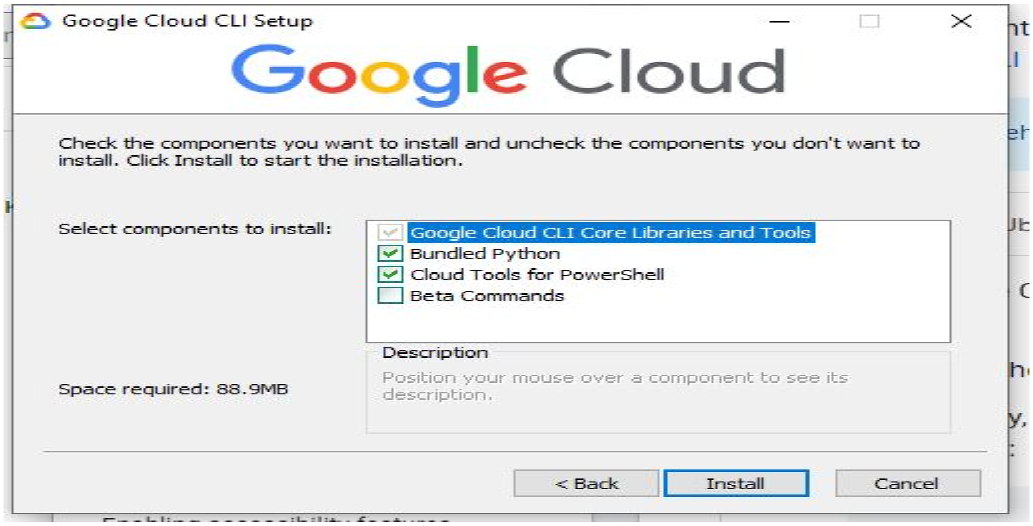


You are about to enable 'Cloud Key Management Service (KMS) API'.

When you are done, go to https://console.cloud.google.com.

Figure 10: Enabling the KMS API.

After enabling the API, we need to install and initialize GCP command line interface. This can be done through command line or a setup wizard. We used the setup wizard as it initializes the command line with the base configuration. Another reason why we used the setup wizard is because it also installs other dependencies that Google command line requires [54].



Google Cloud CLI Setup

Check the components you want to install and uncheck the components you don't want to install. Click Install to start the installation.

Select components to install:

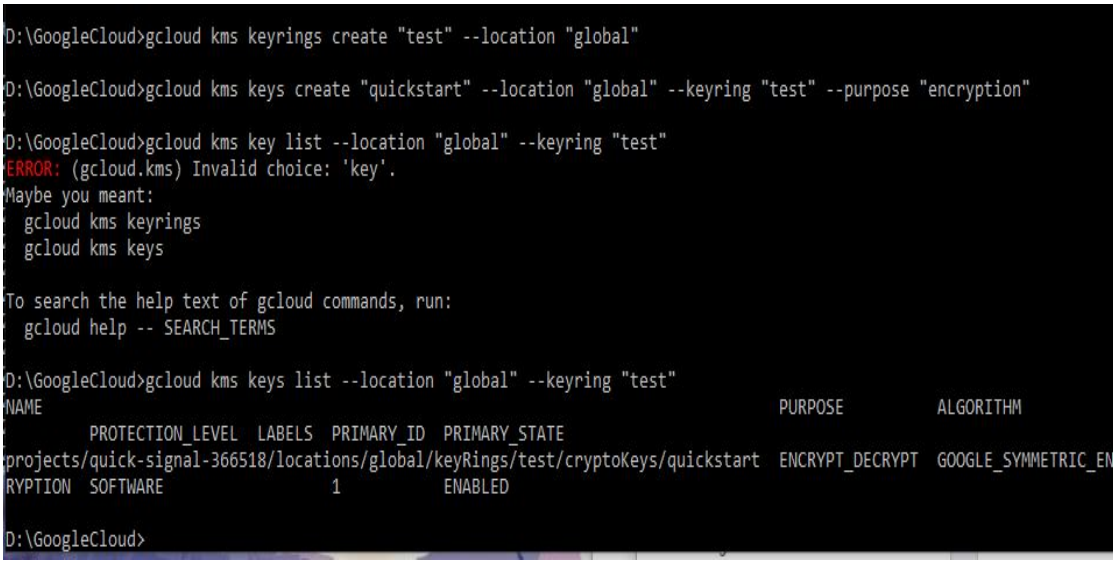
☑ Google Cloud CLI Core Libraries and Tools
☑ Bundled Python
☑ Cloud Tools for PowerShell
□ Beta Commands

Description
Position your mouse over a component to see its description.

Space required: 88.9MB

Figure 11: Using setup wizard to install GCP CLI.

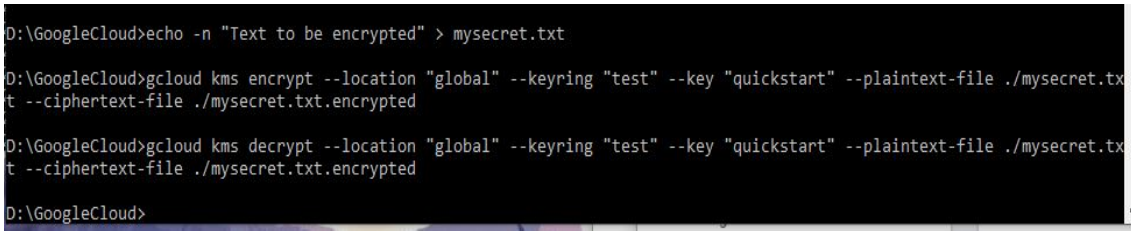
The next step is to create a keyring and key to encrypt data. We will make a keyring named test and a key named QuickStart. These are done using the following commands. The last gcloud command shows the metadata for the key just created. As shown in the path name, a keyring named test was created with a crypto key named QuickStart.



```
D:\GoogleCloud>gcloud kms keyrings create "test" --location "global"

D:\GoogleCloud>gcloud kms keys create "quickstart" --location "global" --keyring "test" --purpose "encryption"

D:\GoogleCloud>gcloud kms key list --location "global" --keyring "test"
ERROR: (gcloud.kms) Invalid choice: 'key'.
Maybe you meant:
  gcloud kms keyrings
  gcloud kms keys

To search the help text of gcloud commands, run:
  gcloud help -- SEARCH_TERMS

D:\GoogleCloud>gcloud kms keys list --location "global" --keyring "test"
NAME                                                                               PURPOSE          ALGORITHM
         PROTECTION_LEVEL  LABELS  PRIMARY_ID  PRIMARY_STATE
projects/quick-signal-366518/locations/global/keyRings/test/cryptoKeys/quickstart  ENCRYPT_DECRYPT  GOOGLE_SYMMETRIC_EN
RYPTION  SOFTWARE                  1           ENABLED

D:\GoogleCloud>
```

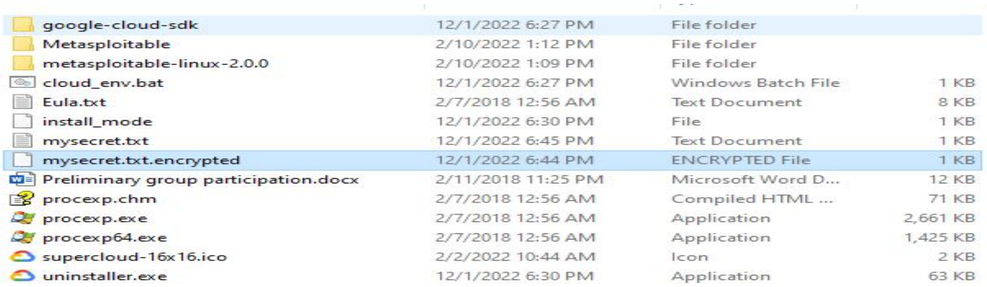
Figure 12: Creating a key ring and key named test and QuickStart respectively

Now we have created the keys, it is time to encrypt something. We used the echo command to send “Text to be encrypted” to a file named mysecret.txt. We then used the gcloud kms encrypt command, specifying the location, keyring, key, and plaintext/ciphertext file to input the unencrypted data from and output the encrypted data to. We finally encrypted something using the keys we created. Now it is time to decrypt the encrypted text. This is done through almost the same as the encrypt command, except we replace encrypt with decrypt. The rest of the parameters remain the same as the original encrypt command except for the cipher and plaintext locations are swapped.



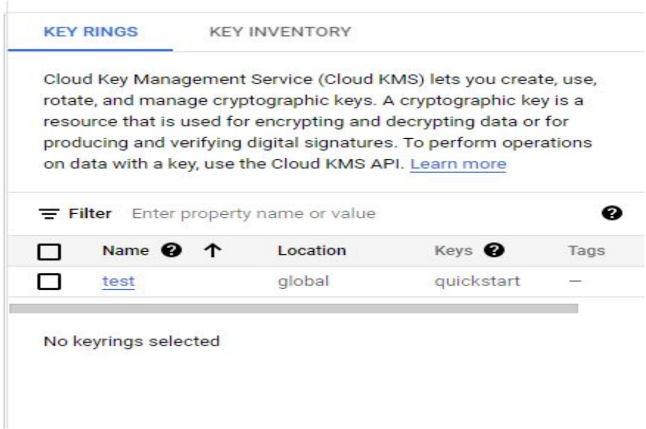
```
D:\GoogleCloud>echo -n "Text to be encrypted" > mysecret.txt

D:\GoogleCloud>gcloud kms encrypt --location "global" --keyring "test" --key "quickstart" --plaintext-file ./mysecret.tx
t --ciphertext-file ./mysecret.txt.encrypted

D:\GoogleCloud>gcloud kms decrypt --location "global" --keyring "test" --key "quickstart" --plaintext-file ./mysecret.tx
t --ciphertext-file ./mysecret.txt.encrypted

D:\GoogleCloud>
```

Figure 13: Creating text to be encrypted and encrypting it as well as decrypting



google-cloud-sdk	12/1/2022 6:27 PM	File folder	
Metasploitable	2/10/2022 1:12 PM	File folder	
metasploitable-linux-2.0.0	2/10/2022 1:09 PM	File folder	
cloud_env.bat	12/1/2022 6:27 PM	Windows Batch File	1 KB
Eula.txt	2/7/2018 12:56 AM	Text Document	8 KB
install_mode	12/1/2022 6:30 PM	File	1 KB
mysecret.txt	12/1/2022 6:45 PM	Text Document	1 KB
mysecret.txt.encrypted	12/1/2022 6:44 PM	ENCRYPTED File	1 KB
Preliminary group participation.docx	2/11/2018 11:25 PM	Microsoft Word D...	12 KB
procexp.chm	2/7/2018 12:56 AM	Compiled HTML ...	71 KB
procexp.exe	2/7/2018 12:56 AM	Application	2,661 KB
procexp64.exe	2/7/2018 12:56 AM	Application	1,425 KB
supercloud-16x16.ico	2/2/2022 10:44 AM	Icon	2 KB
uninstaller.exe	12/1/2022 6:30 PM	Application	63 KB

Figure 14: Encrypted file created



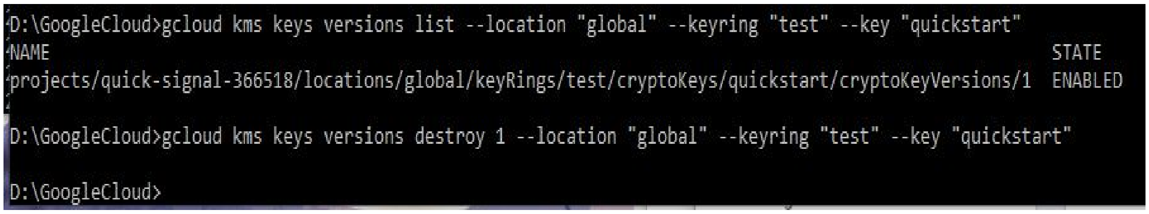
Cloud Key Management Service (Cloud KMS) lets you create, use, rotate, and manage cryptographic keys. A cryptographic key is a resource that is used for encrypting and decrypting data or for producing and verifying digital signatures. To perform operations on data with a key, use the Cloud KMS API. Learn more

□	Name ↑	Location	Keys	Tags
□	test	global	quickstart	—

No keyrings selected

Figure 15: Keyring and key shown on GCP dashboard

Now we have successfully decrypted the file, we need to think about our keyring and the matter of storing it. We don't have much use for these keys and more, so we don't want to get charged for storage fees. To remedy this, we need to delete or destroy the keys. To do that we need the key version. This can be found using the KMS keys versions command. Our version was 1 so we input that into the KMS keys versions destroy command and specify which key we want to target.



```
D:\GoogleCloud>gcloud kms keys versions list --location "global" --keyring "test" --key "quickstart"
NAME                                                                                      STATE
projects/quick-signal-366518/locations/global/keyRings/test/cryptoKeys/quickstart/cryptoKeyVersions/1  ENABLED

D:\GoogleCloud>gcloud kms keys versions destroy 1 --location "global" --keyring "test" --key "quickstart"

D:\GoogleCloud>
```

Figure 16: Destroying the key we previously made

4 RESULTS

We were able to successfully create encryption keys in both AWS KMS and GCP KMS. Amazon Web Services offers a simple, yet granular Key Management Service that allowed us to leverage and manage encryption keys easily within the cloud-based environment. The Google Cloud Platform was relatively easy to use to create keys, and the command line that it offers is used for all Google Cloud resources. So it is very versatile as well as straight forward. Everything provided on the Google Cloud Console is also able to be altered through this command line, though it may be easier in most instances to just use the console. Both services offer a way to create an asymmetric or symmetric key.

In addition, this research shows how similar both systems are in how they work. Both utilize different method to encrypt the data key and use some of the same services to accomplish a task, such as Identity and Access Management (IAM). On the other hand, one of the main differences noticed between the two is that AWS handles key creation entirely through a GUI on the cloud dashboard. Meanwhile, GCP uses a specific command line interface to create keys and encrypt data. These keys can also be created through the console, though it is not as straightforward as it was with AWS. Other than how things are implemented; the actual process is relatively the same except for the concept of keyrings, which is unique to GCP it seems. One reason why this could be is because AWS asks while a key is being created for its permissions and who has access to it. GCP handles this by creating different keyrings to hold keys having different permissions with different accesses. Overall AWS seems to handle key creation in an easier, more graceful, way. This is mostly due to the fact a GUI is used instead of having to enter commands as the fastest method.

When it comes down to volume sizes available, AWS offers 500GB to 16 TB while GCP offers 1 GB to 64 TB. AWS offers different types of keys, those being regular data keys and customer master keys (CMK's). These master keys can be used to encrypt and decrypt data and the data keys are generated, encrypted, and decrypted by the CMK's. These can be customer or AWS managed. They essentially serve the same purpose as GCP's Key-Encryption-Key. GCP and AWS offer different encryption types. AWS offers AES-GCM and RSA-OAEP; while GCP offers RSA PKCS#1v1.5 and RSA-OAEP. Both also offer the same asymmetric key lengths. These are 2048-bit, 3072-bit, and 4096-bit RSA. It is the same for symmetric key length, that being 256-bit AES.

Another noticeable difference involves the way each of the two CSPs operate. AWS KMS is used to encrypt storage, services, and other resources within the cloud, while GCP KMS is utilized to encrypt data elsewhere and manage the keys within the cloud. This is largely due to the fact that GCP by default encrypts all data at-rest for their databases, data warehouses, storage buckets, and other storage services to take the burden off the customer to do so themselves.

5 CONCLUSION

Throughout this paper we have compared two different yet uniquely similar KMS and encryption systems. Both these systems seem to follow a similar guideline or path for how they encrypt their data and how they manage their keys. Even though they are very similar they still have their own unique qualities. Both services do a phenomenal job in encrypting data and managing their keys. It really is a user's preference as to what service they choose, as both offer almost the same service with a few variations. While it is unfortunate that we were unable to test Microsoft Azure KMS, it is not farfetched to believe that it also operates along similar principals to AWS and GCP. Both of these services can confidently say that they protect both the integrity of that data users provide to it and the keys created within.

References

[1] Mukherjee, Subhodeep, Venkataiah Chittipaka, Manish Mohan Baral, and Sharad Chandra Srivastava. "Integrating the challenges of cloud computing in supply chain management." In Recent Advances in Industrial Production: Select Proceedings of ICEM 2020, pp. 355-363. Springer Singapore, 2022.

[2] Singh, Nicholas, Kevin Bui, and Akalanka Mailewa. "Robust Efficiency Evaluation of NextCloud and GoogleCloud." Advances in Technology (2021): 536-545. (DOI:10.31357/ait.v1i2.5392)

[3] Wulf, Frederik, Tobias Lindner, Susanne Strahringer, and Markus Westner. "IaaS, PaaS, or SaaS? The Why of Cloud Computing Delivery Model Selection: Vignettes on the Post-Adoption of Cloud Computing." In Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, pp. 6285-6294. 2021.

[4] Olaosebikan, Ayodeji, Thivanka PBM Dissanayaka, and Akalanka B. Mailewa. "Security & Privacy Comparison of NextCloud vs Dropbox: A Survey." In Midwest Instruction and Computing Symposium (MICS). 2022.

[5] Mailewa Dissanayaka, Akalanka, Roshan Ramprasad Shetty, Samip Kothari, Susan Mengel, Lisa Gittner, and Ravi Vadapalli. "A review of MongoDB and singularity container security in regards to hipaa regulations." In Companion Proceedings of the10th International Conference on Utility and Cloud Computing, pp. 91-97. 2017.

[6] Njuki, S., JIANBIAO ZHANG, EDNA TOO, and HAROLD BUKO DADYE. "Enhancing user data and VM security using the efficient hybrid of encrypting techniques." Journal of Theoretical and Applied Information Technology 97, no. 15 (2019).

[7] Ayuninggati, Tsara, Eka Purnama Harahap, and Raihan Junior. "Supply Chain Management, Certificate Management at the Transportation Layer Security in Charge of Security." Blockchain Frontier Technology 1, no. 01 (2021): 1-12.

[8] Ndri, Anna, Divya Bellamkonda, and Akalanka B. Mailewa. "Applications of Block-Chain Technologies to Enhance the Security of Intrusion Detection/Prevention Systems: A Review." In Midwest Instruction and Computing Symposium (MICS), vol. 2, p. 4. 2022.

[9] Kamal, Muhammad Ayoub, Hafiz Wahab Raza, Muhammad Mansoor Alam, and M. Mohd. "Highlight the features of AWS, GCP and Microsoft Azure that have an impact when choosing a cloud service provider." Int. J. Recent Technol. Eng 8, no. 5 (2020): 4124-4232.

[10] Deochake, Saurabh, and Vrushali Channapattan. "Identity and access management framework for multi-tenant resources in hybrid cloud computing." In Proceedings of the 17th International Conference on Availability, Reliability and Security, pp. 1-8. 2022.

[11] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Security assurance of MongoDB in singularity LXCs: an elastic and convenient testbed using Linux containers to explore vulnerabilities." Cluster Computing 23 (2020): 1955-1971.

[12] Shamseddine, Maha, Wassim Itani, Auday Al-Dulaimy, and Javid Taheri. "Mitigating rogue node attacks in edge computing." In 2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), pp. 1-6. IEEE, 2019.

[13] Ahmad, Shahnawaz, et al. “Cloud Security Framework and Key Management Services Collectively for Implementing DLP and IRM.” Materials Today: Proceedings, vol. 62, 2022, pp. 4828–36, https://doi.org/10.1016/j.matpr.2022.03.420.

[14] Khan, Saad, and Akalanka B. Mailewa. "Discover Botnets in IoT Sensor Networks: A Lightweight Deep Learning Framework with Hybrid Self-Organizing Maps." Microprocessors and Microsystems (2023): 104753. (DOI: https://doi.org/10.1016/j.micpro.2022.104753)

[15] Rozendaal, Kyle, and Akalanka Mailewa. "Neural Network Assisted IDS/IPS: An Overview of Implementations, Benefits, and Drawbacks." International Journal of Computer Applications 975: 8887. (DOI:10.5120/ijca2022922098)

[16] Muluve, Eva, Quentin Awori, Phanuel Owiti, Daniel Osuka, James Serembe, Paul Macharia, and Karen Austrian. "Using Mobile Biometrics and Management Information Systems to Enhance Quality and Accountability of Cash transfer in a Girls’ Empowerment Program in Rural and Urban Poor Settings." In 2020 IST-Africa Conference (IST-Africa), pp. 1-11. IEEE, 2020.

[17] Haghnegahdar, Lida, Sameehan S. Joshi, and Narendra B. Dahotre. "From IoT-based cloud manufacturing approach to intelligent additive manufacturing: Industrial Internet of Things—An overview." The International Journal of Advanced Manufacturing Technology (2022): 1-18.

[18] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Vulnerability prioritization, root cause analysis, and mitigation of secure data analytic framework implemented with mongodb on singularity linux containers." In Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis, pp. 58-66. 2020.

[19] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Dynamic & portable vulnerability assessment testbed with Linux containers to ensure the security of MongoDB in Singularity LXCs." In Companion Conference of the Supercomputing-2018 (SC18). 2018.

[20] Noor, Talal, et al. “Trust Management of Services in Cloud Environments: Obstacles and Solutions.” ACM Computing Surveys, vol. 46, no. 1, 2013, pp. 1–30, https://doi.org/10.1145/2522968.2522980.

[21] Alshammari, Salah T., and Khalid Alsubhi. "Building a reputation attack detector for effective trust evaluation in a cloud services environment." Applied Sciences 11, no. 18 (2021): 8496.

[22] Gheisari, Mehdi, Hamid Esmaeili Najafabadi, Jafar A. Alzubi, Jiechao Gao, Guojun Wang, Aaqif Afzaal Abbasi, and Aniello Castiglione. "OBPP: An ontology-based framework for privacy-preserving in IoT-based smart city." Future Generation Computer Systems 123 (2021): 1-13.

[23] Khan, Muhammad Maaz Ali, Enow Nkongho Ehabe, and Akalanka B. Mailewa. "Discovering the Need for Information Assurance to Assure the End Users: Methodologies and Best Practices." In 2022 IEEE International Conference on Electro Information Technology (eIT), pp. 131-138. IEEE, May 2022. (DOI:10.1109/eIT53891.2022.9813791)

[24] Kaja, Durga Venkata Sowmya, Yasmin Fatima, and Akalanka B. Mailewa. "Data integrity attacks in cloud computing: A review of identifying and protecting techniques." Journal homepage: www. ijrpr. com ISSN 2582 (2022): 7421. (DOI:10.55248/gengpi.2022.3.2.8)

[25] Crișan-Mitra, Cătălina Silvia, Liana Stanca, and Dan-Cristian Dabija. "Corporate social performance: An assessment model on an emerging market." Sustainability 12, no. 10 (2020): 4077.

[26] Landi, Giovanni Catello, Francesca Iandolo, Antonio Renzi, and Andrea Rey. "Embedding sustainability in risk management: The impact of environmental, social, and governance ratings on corporate financial risk." Corporate Social Responsibility and Environmental Management 29, no. 4 (2022): 1096-1107.

[27] Kumar, Rakesh, and Rinkaj Goyal. "Performance based Risk driven Trust (PRTrust): On modeling of secured service sharing in peer-to-peer federated cloud." Computer Communications 183 (2022): 136-160.

[28] Jairu, Pankaj, and Akalanka B. Mailewa. "Network Anomaly Uncovering on CICIDS-2017 Dataset: A Supervised Artificial Intelligence Approach." In 2022 IEEE International Conference on Electro Information Technology (eIT), pp. 606-615. IEEE, May 2022. (DOI:10.1109/eIT53891.2022.9814045)

[29] Sapkota, Bhumika, and Akalanka B. Mailewa. "A Scalable Framework to Detect, Analyze, and Prevent Security Vulnerabilities in Enterprise Software-Defined Networks." Journal homepage: www. ijrpr. com ISSN 2582: 7421. (DOI:10.55248/gengpi.2022.3.2.1)

[30] Khayer, Abul, Md Shamim Talukder, Yukun Bao, and Md Nahin Hossain. "Cloud computing adoption and its impact on SMEs' performance for cloud supported operations: A dual-stage analytical approach." Technology in Society 60 (2020): 101225.

[31] Chahal, Rajanpreet Kaur, Neeraj Kumar, and Shalini Batra. "Trust management in social Internet of Things: A taxonomy, open issues, and challenges." Computer Communications 150 (2020): 13-46.

[32] Alemneh, Esubalew, Sidi-Mohammed Senouci, Philippe Brunet, and Tesfa Tegegne. "A two-way trust management system for fog computing." Future Generation Computer Systems 106 (2020): 206-220.

[33] Jorquera Valero, José María, Pedro Miguel Sánchez Sánchez, Manuel Gil Pérez, Alberto Huertas Celdrán, and Gregorio Martinez Perez. "Cutting-Edge Assets for Trust in 5G and Beyond: Requirements, State of the Art, Trends, and Challenges." ACM Computing Surveys 55, no. 11 (2023): 1-36.

[34] Gamnis, Steven, Matthew VanderLinden, and Akalanka Mailewa. "Analyzing Data Encryption Efficiencies for Secure Cloud Storages: A Case Study of Pcloud vs OneDrive vs Dropbox." Advances in Technology (2022): 79-98. (DOI:10.31357/ait.v2i1.5526)

[35] Garfinkel, Simson L., and Philip Leclerc. "Randomness concerns when deploying differential privacy." In Proceedings of the 19th Workshop on Privacy in the Electronic Society, pp. 73-86. 2020.

[36] Saarinen, Markku-Juhani O. "Mobile energy requirements of the upcoming NIST post-quantum cryptography standards." In 2020 8th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), pp. 23-30. IEEE, 2020.

[37] Mailewa, Akalanka, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Mechanisms and techniques to enhance the security of big data analytic framework with mongodb and Linux containers." Array 15 (2022): 100236. (DOI:10.1016/j.array.2022.100236)

[38] Raso, Emanuele, Lorenzo Bracciale, Pierpaolo Loreti, and Giuseppe Bianchi. "ABEBox: A data driven access control for securing public cloud storage with efficient key revocation." In Proceedings of the 16th International Conference on Availability, Reliability and Security, pp. 1-7. 2021.

[39] Jarecki, Stanislaw, Hugo Krawczyk, and Jason Resch. "Updatable oblivious key management for storage systems." In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 379-393. 2019.

[40] Kamaraju, Ashvin, Asad Ali, and Rohini Deepak. "Best Practices for Cloud Data Protection and Key Management." In Proceedings of the Future Technologies Conference (FTC) 2021, Volume 3, pp. 117-131. Springer International Publishing, 2022.

[41] Guptha, Ashwin, Harshaan Murali, and T. Subbulakshmi. "A Comparative Analysis of Security Services in Major Cloud Service Providers." In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 129-136. IEEE, 2021.

[42] Mailewa, Akalanka, and Kyle Rozendaal. "A Novel Method for Moving Laterally and Discovering Malicious Lateral Movements in Windows Operating Systems: A Case Study." Advances in Technology (2022): 291-321, ISSN 2773-7098. (DOI:10.31357/ait.v2i3.5584)

[43] ACHAR, SANDESH, HRISHITVA PATEL, and SANWAL HUSSAIN. "DATA SECURITY IN CLOUD: A REVIEW." Asian Journal of Advances in Research (2022): 76-83.

[44] Roy, Agniswar, Abhik Banerjee, and Navneet Bhardwaj. "A Study on Google Cloud Platform (GCP) and Its Security." Machine Learning Techniques and Analytics for Cloud Security (2021): 313-338.

[45] Mailewa, Akalanka, and Jayantha Herath. "Operating Systems Learning Environment with VMware" In The Midwest Instruction and Computing Symposium. Retrieved from http://www.micsymposium.org/mics2014/ProceedingsMICS_2014/mics2014_submission_14.pdf. 2014.

[46] Dageville, Benoit, Thierry Cruanes, Marcin Zukowski, Vadim Antonov, Artin Avanes, Jon Bock, Jonathan Claybaugh et al. "The snowflake elastic data warehouse." In Proceedings of the 2016 International Conference on Management of Data, pp. 215-226. 2016.

[47] Mbae, Oscar, David Mwathi, and Edna Too. "Secure Cloud Based Approach for Mobile Devices User Data." Open Access Library Journal 9, no. 9 (2022): 1-20.

[48] Shetty, Roshan Ramprasad, Akalanka Mailewa Dissanayaka, Susan Mengel, Lisa Gittner, Ravi Vadapalli, and Hafiz Khan. "Secure NoSQL based medical data processing and retrieval: the exposome project." In Companion Proceedings of the10th International Conference on Utility and Cloud Computing, pp. 99-105. 2017.

[49] Simkhada, Emerald, Elisha Shrestha, Sujan Pandit, Upasana Sherchand, and Akalanka Mailewa Dissanayaka. "Security threats/attacks via botnets and botnet detection & prevention techniques in computer networks: a review." In The Midwest Instruction and Computing Symposium.(MICS), North Dakota State University, Fargo, ND. 2019.

[50] Akintaro, Mojolaoluwa, Teddy Pare, and Akalanka Mailewa Dissanayaka. "Darknet and black market activities against the cybersecurity: a survey." In The Midwest Instruction and Computing Symposium.(MICS), North Dakota State University, Fargo, ND. 2019.

[51] Chandramouli, Ramaswamy, Michaela Iorga, and Santosh Chokhani. "Cryptographic key management issues and challenges in cloud services." Secure Cloud Computing (2013): 1-30.

[52] Ruiz Díaz, Blanca. "Deployment of a lab environment to identify and protect sensitive data in the cloud." Bachelor's thesis, Universitat Politècnica de Catalunya, 2022.

[53] Khalil, Maad M., Sergey E. Adadurov, and M. Sh Mahmood. "Mastering Google cloud: building the platform that serves your needs." Models and Methods for Researching Information Systems in Transport 2020 (MMRIST 2020) 1 (2020): 41-46.

[54] Carvalho, Daniel, João Morais, João Almeida, Pedro Martins, Carlos Quental, and Filipe Caldeira. "A Technical Overview on the Usage of Cloud Encryption Services." In European Conference on Cyber Warfare and Security, pp. 733-XI. Academic Conferences International Limited, 2019.

Imaging With 2.4GHz

Richard Anderson[*] , Brendan Betterman[*], Baozhong Tian[†]
Department of Mathematics and Computer Science
Bemidji State University, Bemidji, MN 56601
Richande218@gmail.com Brendan.Betterman@gmail.com
Baozhong.Tian@BemidjiState.edu

Abstract

The use of Wi-Fi waves for non-destructive imaging and sensing has become a popular area of research in recent years. This is due to the fact that Wi-Fi waves have the ability to penetrate solid objects and provide information about the interior of a structure, making them ideal for detecting people through walls. This paper aims to test the feasibility of detecting individuals through walls using Wi-Fi waves. The system developed in this work uses a combination of signal-processing techniques and machine-learning algorithms to produce an image of a person behind a wall.

The proposed method first collects raw Wi-Fi signals from a target area and then processes these signals to extract relevant information. This information is then used to construct an image which is then used to detect the presence of the person behind the wall.

The proposed system has several advantages over existing techniques for detecting people through walls. Firstly, the use of Wi-Fi waves eliminates the need for any physical contact with the target, making the method non-intrusive. Secondly, the system is less expensive compared to other existing techniques, such as millimeter-wave radar, which are often complex and expensive to implement.

In conclusion, this paper presents a new approach to detecting individuals through walls using Wi-Fi waves. The proposed system has the potential to make a significant impact in various applications and provides a foundation for further research in this field. The results of this study demonstrate the feasibility of using Wi-Fi waves for this purpose and highlight the potential of Wi-Fi waves for use in non-destructive imaging and sensing.

[*] Main student contributor

[*] Main student contributor

[†] Faculty sponsor, final editing and formatting

1 Introduction

The detection of individuals through walls is a challenging problem with significant implications for various fields, including security, healthcare, and search and rescue operations. Traditional methods for detecting people through walls often rely on physical contact or are expensive and too complex to implement. In recent years, the use of Wi-Fi waves for non-destructive imaging and sensing has emerged as a promising alternative for detecting individuals through walls. Wi-Fi waves have the unique ability to penetrate solid objects and provide information about the interior of a structure, making them ideal for detecting people behind walls.

In this paper, we will be testing the feasibility of using off-the-shelf and easy-to-obtain hardware to accomplish this task. The solution will be evaluated based on accuracy, speed, practicality, and ethical/privacy concerns. Accuracy will be determined by comparing the output of the solution to the tag on tagged images. The tags being the state of whether a person is present or not. Speed will be the measure of the total completion time to determine occupancy. Practicality will be based on availability, cost, and ease of use of the system in comparison to existing systems. Lastly, ethical and privacy concerns of the proposed solution will be mentioned and evaluated on severity.

The proposed system uses a combination of signal processing techniques and machine learning algorithms to extract relevant information from raw Wi-Fi signals collected from a target area using a motorized directional antenna to scan received signal strength indicator values and construct an image of a person, or lack thereof, behind a wall using these values. The proposed system is non-intrusive, less expensive, and less complex compared to other existing techniques, making it a promising alternative for practical applications. The results of this study provide a foundation for further research in this field and demonstrate the potential impact of Wi-Fi waves in various applications.

2 Existing Technologies

2.1 Ground Penetration Radar

Ground penetrating radar (GPR) works by sending electromagnetic waves (in the 10 to 1000 Hz range) into the probed material and receiving the reflected pulses as they encounter discontinuities. [Karbhari, 2011]. GPR data is usually recorded from a number of spatial positions by dragging the antennas along the surface of the ground or walls. A transmit and receive occur at each of the observation positions. The recorded data are then combined to form an image [Saleem et al., 2014]. GPR is mostly used for geological applications, but it is now being used for forensic [Solla et al., 2012] and archaeological applications [Mazurkiewicz et al., 2016] as well. GPR has its advantages such as its speed, it's non-destructive, and its high resolution. it is also very expensive. However, GPR can also be affected by the moisture content of certain materials, causing it to reflect in an undesirable manner [GPRS, 2023].

2.2 Millimeter Radar

Millimeter wave radar is a technology that uses waves in high-frequency bands ranging from 30-300GHz which have a wavelength of 1-10mm. It is particularly useful in cases when detecting through things such as fog, smoke, and dust, but can also even be expanded to detection through walls [Guan et al., 2020]. A complete millimeter wave radar system includes transmit (TX) and receive (RX) radio frequency (RF) components; analog components such as clocking; and digital components such as analog-to-digital converters (ADCs), microcontrollers (MCUs) and digital signal processors (DSPs). Traditionally, these systems were implemented with discrete components, which increased power consumption and overall system cost. System design is challenging due to the complexity and high frequencies [Gonzalez-Partida et al., 2009].

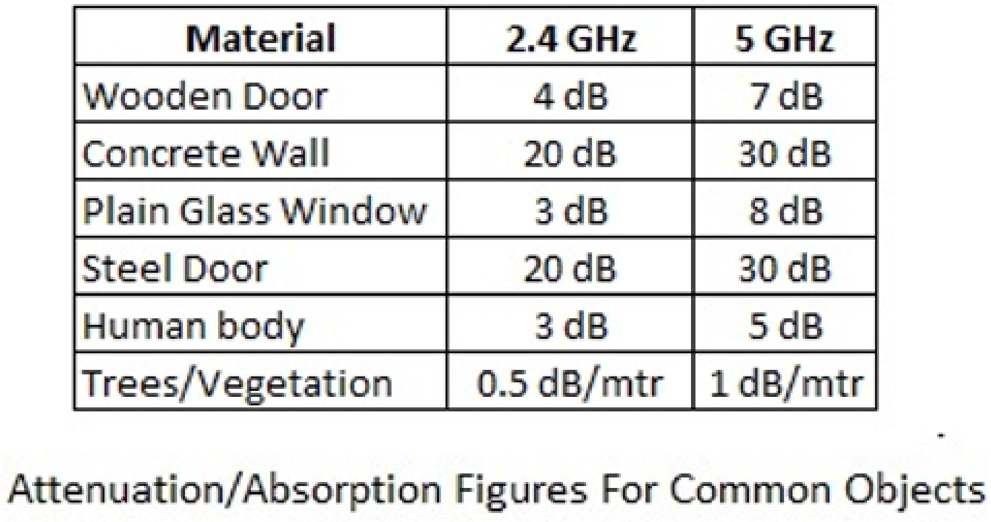
A common type of millimeter wave radar is called frequency-modulated continuous wave (FMCW) radar. FMCW radars transmit a frequency-modulated signal continuously in order to measure range as well as angle and velocity. This differs from traditional pulsed-radar systems, which transmit short pulses periodically [Gonzalez-Partida et al., 2009]. A common application for this system is in automobile proximity sensors [Guan et al., 2020], meteorological data collection on clouds [Bu et al., 2016], and even medical applications such as detecting heartbeats [Wu and Dahnoun, 2022]. The main advantages to these millimeter-wave radar systems are that they are relatively small in size, very accurate because of the wavelengths of the frequencies used, and resistant to interference. However, they are vulnerable to electrical towers and electromagnetic hotspots.

2.3 DensePose With WiFi

The most similar technology to the methods used in our paper is DensePose in combination with WiFi. This uses a very similar approach to the one proposed in this paper of using WiFi signals to generate input for a neural network. It then estimates the pose of a human. To accomplish this the raw channel state information (CSI) signals are collected by three antennas that are receiving information from another three antennas that are connected to a transmitter. The CSI is then cleaned by amplitude and phase sanitization. Then, a two-branch encoder-decoder network performs domain translation from sanitized CSI samples to 2D feature maps that resemble images. The 2D features are then fed to a modified DensePose architecture to estimate the UV map, a representation of the dense correspondence between 2D and 3D humans. This method is capable of estimating poses with reasonable accuracy, but struggles for poses that are not, or are rarely within the dataset [Geng et al., 2022].

3 Methodology

This project required the creation of a device that could generate an image using received signal strength indicator (RSSI) levels. In order to do this, custom hardware and software needed to be sourced and or created. The theory on why humans would be detectable is due to the fact that humans are 70% water which absorbs RF [Accolate, 2016]. As listed in Fig 1, a person should generate a 3dB low spot if they are in between the access point and the device. Since concrete and steel absorb more than wood concrete, brick, and metal-sheathed builds would be harder to image through.



Material	2.4 GHz	5 GHz
Wooden Door	4 dB	7 dB
Concrete Wall	20 dB	30 dB
Plain Glass Window	3 dB	8 dB
Steel Door	20 dB	30 dB
Human body	3 dB	5 dB
Trees/Vegetation	0.5 dB/mtr	1 dB/mtr

Attenuation/Absorption Figures For Common Objects

Figure 1: Attenuation Table from Accolade Wireless [Saleem et al., 2014].

3.1 Hardware

3.1.1 Antenna

Modern cameras have active-pixel sensors so they can simultaneously detect light values in a large grid. Due to scope and budget, this project needed to come up with a cheaper alternative. The use of a focused unidirectional antenna would grant us one pixel. One pixel isn't very useful as an image but the method of getting more pixels will be explained later in this paper. There are many different types of antennas, but the two common designs that fit the requirements are the Yagi style and the Cantenna. Yagi antennas are antennas that have multiple elements in parallel and are insulated from each other. These parallel rods can be used to amplify or absorb frequencies. Cantenna antennas are antennas that use a metal cylinder to boost signal strength in a direction. They are typically made of cans hence the name. These antennas are directional and both designs are fairly light. However, the yagi style antenna is smaller in footprint, which leads to less weight.

3.1.2 Mechanism

Physically moving the antenna to different x and y locations is the compromise made to keep this solution affordable. There are a few different methods that could be used to move the antenna. Corexy [Chiffey, 2022] or cartesian could be used to move the antenna to these different positions. Corexy is a motion system that moves on a horizontal plane combining the x and y. The motor's power increases since neither motor has to move another motor. Cartesian motion systems have two linear axes but one is mounted on the other so one motor will move another motor [Narayanan, 2018]. However, this would generate an isometric image of the room. Linear x and y movement could give really fine detail but the device would need to be the same scale as the image. Since isometric has problems with scale and cost. We used a movement method that can generate a perspective image on the panoramic plane. To achieve a panoramic x y plane we used a pan and tilt design. This had the benefit of being fairly compact, easy to design, and cost-effective since the mechanisms are centralized.

3.1.3 Radio Transceiver

In this application, an off-the-shelf motherboard equipped with a Wi-Fi transceiver was used. This hardware was chosen because of its availability and relatively cheap price. Note that any type of motherboard equipped with a Wi-Fi transceiver should be capable of accomplishing the scans performed in this application, it will only affect the speed at which the scans can be executed.

3.2 Software

3.2.1 Firmware

The device used to drive the stepper motors was an Arduino nano. Arduinos are microcontrollers that allow the user to connect software to hardware. This enabled the driving of the motors needed to move the antenna to different x and y locations. The requirement of this firmware was that it needed to listen to serial inputs and drive the motors accordingly. Serial inputs are used to connect the computer program to the Arduino via USB. This is useful since syncing the movement and the collection of data gives the ability to get RSSI in specified areas.

3.2.2 RSSI Collection

The RSSI collection can be done through multiple different libraries their names being Network-Manager, IW, and Aircrack-ng. All of these are Linux-only programs and all have pros and cons. The speed differentials and data output streams are the major differences. The IW library was the slowest since it generated the most amount of data that needed to be parsed through and it also had a hard-coded two-second cooldown. Extra data could prove to be useful since it allows the ability to exclude routers that cause noise. Network-Manager or Nmcli was quicker taking less than a tenth of a second to get data. However, the data does not update every time Nmcli is called. Nmcli stores the data and only refreshes every two seconds. Image banding was a result prior to this discovery.

The last RSSI collection library used was a modified version of Aircrack-ng. Aircrack-ng is a Linux tool used to do security checks on routers. The command Airodump-ng gives a continuous stream of live updating data of RSSI values and other information. This is useful when monitoring wireless networks but not useful when you need discrete results. The modification was one line of code that broke the program after the first cycle. Airodump-ng takes about half a second per call. It occasionally needs to be called multiple times if it does not find a device in the first cycle.

3.2.3 Image Generation

The images were generated using the RSSI value and then a heatmap was applied according to strength. After the scans were completed a CSV (comma-separated values) file was generated, each value was in between the range -30 dBm(decibel-milliwatt) and -120 dBm. For the image generation, the values were normalized into a range from 0 to 255 for the green channel and inverted for the red channel. The greener the pixel the better the signal strength and vice-versa for redness. When reading the CSV every other line was inverted and the image was rotated due to the scan method. The antenna scanned up a column, right one step, and down a column, in a zig-zag fashion. In Figure 2 the hypothesis is that the person reflected the signal. Figure 3 shows the room setup and how this is a possible explanation.

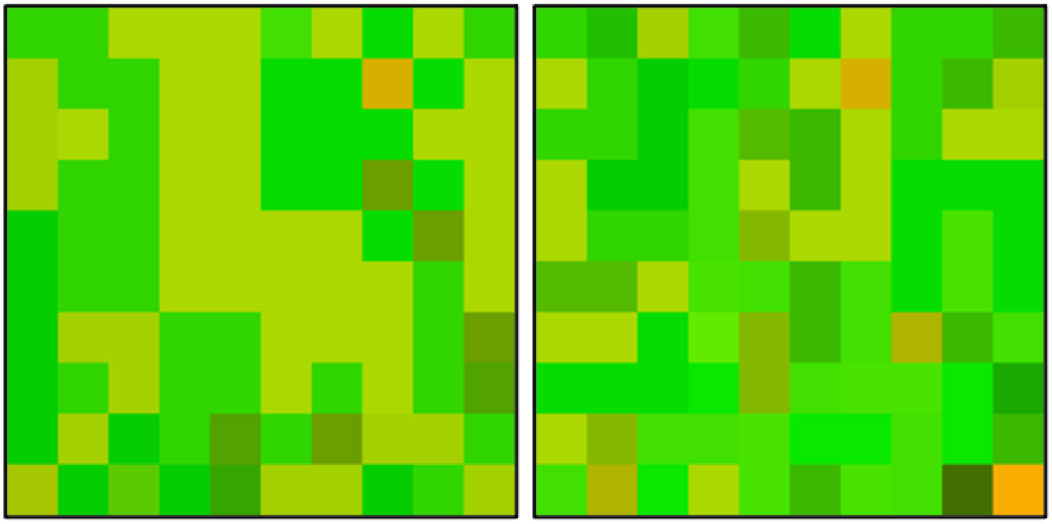


Figure 2: No Human (left) and Human (right).

3.2.4 Neural Network

There was no clear/consistent indicator of human presence to the naked eye when looking at the images produced by this solution. In order to see if there was any

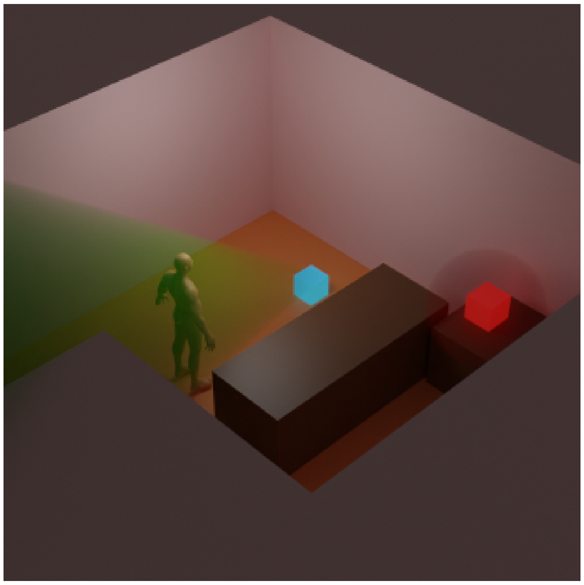


Figure 3: Render of Room the green volume is the area scanned and the red is a router in the room (not to scale).

pattern associated with this, an artificial neural network (ANN) was used to process the resulting data produced by the scan to create a pre-trained model. This model was used to predict the presence of a human between the antenna and the Wi-Fi source.

The layout of the neural network was as follows: 100 input nodes, 1 hidden layer of 50 nodes plus one bias node, a sigmoid activation function, and one output. All weights were given random initial weights. The 100 inputs were given the 100 pixels generated by the previously mentioned systems. The network was trained on the 46 training images and RSSI data that was generated for 1000 epochs at a learning rate of 0.001.

The network's weights were then saved and used to predict the presence of a human in 20 additional photos. The training data used a 0 to indicate that there was no human present and a 1 to indicate a human presence. To accomplish this, every output of the model was rounded to the nearest integer. If the value was 0 or below it was classified as a prediction of no human, and a value of 1 or higher was a human.

4 Results

4.1 Accuracy

Accuracy was hard to measure with insufficient data and a non-controlled environment. The ANN was able to reach an RMSE of 0.38 on the training data, but once applied to the testing data the predictions were not consistent or accurate. The majority of results ranged between 45-55% accuracy. In some cases, the model was able to reach up to 75% accuracy, but this is likely due to the randomization of initial weights in the ANN, and with such a limited size of test cases, it may have just been an anomaly. Another concern is that the ANN may have been overfitting the training data, leading to inaccuracies. All of this is likely due to a very small number of training points in relation to the degrees of freedom in the network. Professor Yaser Abu-Mostafa from Caltech states that as a rule of thumb, you need roughly 10 times as many examples as there are degrees of freedom in your model [Narayanan, 2018], which would mean in the case of our 50 nodes in the hidden layer we would require thousands of samples.

4.2 Speed

4.2.1 Wi-Fi Data Collection

The speed of the proposed solution is highly dependent on the hardware being used. Many motherboard's report rate on their Wi-Fi transceivers or the software associated with getting the required information is not high enough. This can cause each image to take upwards of 5 minutes for something as simple as a 10-pixel by 10-pixel scan. In testing, the proposed solution was able to get an image in anywhere from 67 seconds with no slowdowns to about 5 minutes with many slowdowns. This large range is caused by the device waiting to receive enough signal to collect the RSSI data. If there are many failures to receive the data, the routers rate-limit the number of times the data can be requested, the time increases significantly.

Another issue with the collection is the speed of the motors used to manipulate the direction of the antenna. Due to the weight of the antenna, and the relatively weak motors used to turn it, a gear reduction had to be used. This slowed the movement of the antenna as well as slowed the scans slightly because of this movement. However, this could be overcome by simply adding stronger motors, but that would in turn increase the price.

4.2.2 Neural Network Prediction

After the image is obtained, a pre-trained neural network is used to determine the human occupancy status is near instantaneous. With this network we were able to achieve a prediction from the complete Wi-Fi data in less than a millisecond, so a majority of the time taken with this implementation is due to the slow scan speed of the device. With this in mind, there is likely still room for optimization in the speed of the neural network.

4.3 Practicality

One of the main concerns with the proposed solution is its speed. With a 10-pixel by 10-pixel image taking on average 2-3 minutes, it would take an unreasonably long time to generate enough data to get a detailed image. Also, a moving person may not be detected at all, because they are not in the path of the signal for long enough.

This speed issue also leads into the problem of not being able to generate enough training data. At the current speed, generating the thousands of images required to train an ANN properly would take days of scan time, with a person in the focus of the antenna for some of these.

Additionally, the physical size of the antenna can limit its range and sensitivity, making it impractical for certain applications. The size of the antenna can affect the distance over which it can detect RF signals. A larger antenna may have a longer range, but it can be impractical to use in certain situations due to its size. On the other hand, a smaller antenna may not have a long-range or be sensitive enough to detect weak signals.

In an ideal scenario, a software-defined radio (SDR) would be used to get signal strength values, since it can directly control the gain applied to the antenna signal. It would also allow for faster communication between devices because it could directly interface with the custom software more easily.

While our solution may not have been the most accurate, it was cheap, with a material cost of less than $100. In comparison to things like GPR, Millimeter Wave Radar, and DensePose with WiFi, this is a great alternative in regards to price. GPR starts around $14,000 [GPRS, 2023] while millimeter wave radar [Lin and Hu, 2017] and DensePose fall in the $100-200 range depending on the equipment used [Geng et al., 2022].

4.4 Ethical Concerns

The ethical concerns for the method proposed are similar to that of current technologies. The capability to detect whether a person is present through walls can be an invasion of privacy. At a more affordable price point such as the proposed method, it is even more of a concern because it could be widely available to anyone.

The main issue with this device is that it is usable without the consent of those it's being used on, as well as without them knowing. If an accurate enough image could be obtained, whether it's through a long scan or through some other breakthrough, information could be captured about individuals that was not intended to be seen or monitored.

Another concern is that if a user of this implementation was able to speed up the collection of Wi-Fi data it may cause a denial of service to people in the area because of the constant polling of network information caused by the proposed solution. Even in the case that it does not cause complete outages it may degrade the experience for others in the area.

Lastly, because this is using information that is produced via radio waves, someone could potentially broadcast waves to lead to inaccuracy, since there is no way to validate which waves or lack of waves are authentic. This could lead to biased images, which could result in unfair treatment of individuals.

5 Conclusion

This project had a few shortcomings like slow imaging, not enough data for the neural network, poor design, and ineffective planning. These problems will need to be addressed for future iterations. The lack of resources played a role, using a wireless network card was a cheap alternative to a real SDR but it proved to be slow and hard to work with. Writing software for the network card was more of a hack rather than elegant code. Without properly rewriting firmware the network card could send out too much data causing a denial of service attack on local routers.

There were many unforeseen challenges that came up in this project. A lot of time was spent on the design of the robot that moves the antenna. There were countless redesigns due to the strength of the motors and tolerances. But the majority of time was devoted to finding a solution to the slow speed of scan times. Three different libraries gave results but there were many that either never worked or did not work in this application. Recompiling Aircrack-ng was a challenge but in the end, it did increase the speed by three times.

Rushing to implement ideas prior to proper research cause many dead ends and wasted time. Prior to the use of the wireless network card, a USB CrazyRadio-PA was used. There were many GitHub repositories that were using the device to get RSSI values in the 2.4ghz band. This was promising until the device arrive and we found out the device itself doesn't calculate the RSSI it gets it from the drones this device is typically used in combination with.

While the proposed implementation was not a complete success, it was able to generate images using WiFi waves. The images had slight differences between ambient and a person being in front but it was unclear if it was random noise.

Overall, in its current state, this implementation is not feasible. A few more things will be required to make it feasible. These are things such as increasing the speed of scanning for RSSI data, such as through an SDR, training the ANN on larger datasets, and collecting data in a more isolated environment. The method does however show promise for future research and may be capable of human detection through walls with further improvement.

6 Future Research Areas

6.1 Frequencies

Making a device to scan 5GHz could yield different results. Many routers output 5GHz alongside 2.4GHz. 5GHz has less of a penetration distance compared to 2.4GHz [Accolate, 2016]. The downside to 5GHz for users is a benefit when it comes to imaging. The waves are more likely to be absorbed by a human's presence and RSSI can be measured faster. Measurements occurring faster is due to the wave periods being faster so the amplitude can be measured more often. 5GHz antennas are about half the size of 2.4GHz antennas so the scanning device can be smaller. The con to 5GHz antennas is that their manufacturing tolerance is more precise. Making a Yagi antenna would require printing circuit boards.

6.2 Design

Redesigning the XY movement system would reduce backlash and make the device sturdier. The entire build was 3d printed including the antenna and some parts were reused from previous prototypes. A total redesign would make the movement less of a worry so more images could be taken automatically generating more data.

6.3 RSSI Collection methods

The major limiting factor of this device is the scan times. Reducing the scan times would require better-measuring equipment. The use of a proper SDR would reduce this time per pixel. It might be possible to reduce the time on the wireless network card method we used but that would require modifying the firmware or more workarounds. A system that used multiple antennas could generate speed up the collection process as well.

6.4 Controlled Environment

Controlling the environment was a challenge with this project. All of the scans took place in an apartment complex with many wireless access points and unknown whereabouts of people. Having a location with only one access point would be a better control scenario. Additionally, the device always had a person in the same room since it need to be monitored. A person in the room could throw off the RSSI values since wifi bounces off walls like a room of mirrors.

6.5 Different OS

The operating system used was a non-GUI(Graphical User Interface) version of Archlinux. The computer used was used for other projects where a GUI was not important and the speed gains were. Since this project's outputs were images a webserver was required to view the images. This made the computer less mobile since it needed to be connected to an ethernet cable since the wifi was disabled when scanning. Limited mobility caused issues when trying to image through different walls. Also since it had to be directly wired to the access point it was in the same room as the router. The 2.4 Ghz was turned off for some scans to get RSSI values from outside the room.

References

[Accolate, 2016] Accolate (2016). Why wifi is complicated: Wifi signal issues. https://www.accoladewireless.com/wlan-wifi-signal-issues/.

[Bu et al., 2016] Bu, L., Pan, H., Kumar, R., Huang, X., Gao, H., Qin, Y., Liu, X., and Kim, D. (2016). Lidar and millimeter-wave cloud radar (mwcr) techniques for joint observations of cirrus in shouxian (32.56 °n, 116.78 °e), china. *Journal of Atmospheric and Solar-Terrestrial Physics*, 148:64–73.

[Chiffey, 2022] Chiffey, J. (2022). Corexy vs cartesian 3d printers – which is best? `https://total3dprinting.org/corexy-vs-cartesian/`.

[Geng et al., 2022] Geng, J., Huang, D., and la Torre, F. D. (2022). Densepose from wifi. `https://doi.org/10.48550/arXiv.2301.00250`.

[Gonzalez-Partida et al., 2009] Gonzalez-Partida, J.-T., Almorox-Gonzalez, P., Burgos-Garcia, M., Dorta-Naranjo, B.-P., and Alonso, J. I. (2009). Through-the-wall surveillance with millimeter-wave lfmcw radars. *IEEE Transactions on Geoscience and Remote Sensing*, 47(6):1796–1805.

[GPRS, 2023] GPRS (2023). Ground penetrating radar (gpr) – an electromagnetic investigation method. `https://www.gp-radar.com/article/ground-penetrating-radar-gpr-an-electromagnetic-investigation-method`.

[Guan et al., 2020] Guan, J., Madani, S., Jog, S., Gupta, S., and Hassanieh, H. (2020). Through fog high resolution imaging using millimeter wave radar. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11470.

[Karbhari, 2011] Karbhari, V. M. (2011). Service life estimation and extension of civil engineering structures. *doi: 10.1016/B978-1-84569-398-5.50014-6.*, pages 291–301.

[Lin and Hu, 2017] Lin, J. and Hu, H. (2017). 79ghz to replace 24ghz for automotive mm-wave radar sensors. `https://www.digitimes.com/news/a20170906PD208.html`.

[Mazurkiewicz et al., 2016] Mazurkiewicz, E., Tadeusiewicz, R., and Tomecka-Suchon, S. (2016). Application of neural network enhanced ground-penetrating radar to localization of burial sites. *Applied Artificial Intelligence*, 30(9):844–860.

[Narayanan, 2018] Narayanan, K. (2018). Do you have enough data for machine learning? `https://www.forbes.com/sites/forbestechcouncil/2018/11/19/do-you-have-enough-data-for-machine-learning/?sh=4b84980d52e2`.

[Saleem et al., 2014] Saleem, H., Tyne, C. J. V., Batalha, G. F., and Yilbas, B., editors (2014). *Comprehensive Materials Processing*. Elsevier Ltd.

[Solla et al., 2012] Solla, M., Riveiro, B., MX, M. Á., and Arias, P. (2012). Experimental forensic scenes for the characterization of ground-penetrating radar wave response. *Forensic Science International*, 220(1-3):50–8.

[Wu and Dahnoun, 2022] Wu, J. and Dahnoun, N. (2022). A health monitoring system with posture estimation and heart rate detection based on millimeter-wave radar. *Microprocessors and Microsystems*, 94(10467).

Monocular Vision and Sensor Coupling for Indoor Localization

Houlin Chen, Lu Liang & Lei Wang
Computer Science Department
University of Wisconsin- Lacrosse
Lacrosse, Wisconsin, 54601
chen3653@uwlax.edu
lwang@uwlax.edu

Abstract

Although long used for positioning mobile devices, GPS has limitations in indoor environments due to blocked high-frequency signals and attenuation effects. This can lead to false readings. In this study, a low-cost, GPS-independent model of a neural network-assisted visual tracking system is proposed. The ArUco code is used to provide reference coordinates for the system and a neural network is used to optimize the location detection rate. To produce smooth tracking results, an inertial measurement unit (IMU) and vision-based position estimation are integrated. The proposed system significantly improves the accuracy, interference immunity, and real-time performance of the indoor tracking system. The instantiation of this system on a smartphone platform will likely enable a new cost-effective approach to indoor tracking. In the test phase, the proposed system obtained a tracking accuracy of 0.5 m without any help from expensive depth cameras or 3D LIDAR.

1. INTRODUCTION

To cope with the deviation of GPS readings for indoor positioning, a range of smart terminal-based indoor positioning technologies such as Wi-Fi, Bluetooth, Radio Frequency Identification (RFID) and Ultra-Wide Wave (UWB) technologies have emerged. wi-fi positioning systems (WPS) use the characteristics of nearby Wi-Fi hotspots and other wireless access points to discover the location of devices [1] [2].

Existing indoor localization and tracking systems either require prior knowledge of the environment, such as building floor plans, locations of Wi-Fi access points, Bluetooth beacons, and pre-established RF fingerprint databases, or expensive on-board equipment, such as 3D LiDAR, depth cameras, or omnidirectional cameras [3]. For example, Google Indoor Maps [4] can triangulate the approximate location of a user in an indoor mall (where the user is standing) using nearby WIFI points, user device Bluetooth, and the user device's built-in GPS. However, in practice, it is challenging to obtain comprehensive infrastructure information without infringing on the private rights of the user. Moreover, for existing computer vision techniques, the vision-only localization approach has extensive image processing, resulting in poor real-time performance [5] [6]. This approach cannot work in relatively poor lighting conditions. To address the problems of traditional visual localization algorithms such as poor anti-interference capability and limited real-time performance, we propose a deep learning-based visual localization method, leading to the design of an indoor localization system based on neural networks and sensor fusion. We improve the accuracy of visual localization by using neural network-based object detection and make the system efficient with real-time feedback results.

Vision-based indoor localization is used to extract information about the three-dimensional (3D) world from the two-dimensional (2D) images captured by a camera [7]. Discovering the correspondence between 3D points in the real-world environment and their 2D image projections is the most critical and complex step in this process. In our project, we use the ArUco code to help in the localization. The advantage of this marker is that a single marker provides enough correspondence information to calculate the camera pose. In addition, the internal binary encoding of the ArUco code allows the marker to maintain specific stability in terms of checks and corrections. We use neural networks to improve the accuracy of detecting markers under complex conditions such as fast motion and light changes.

In the process of detecting markers, we may encounter situations where the marker cannot be captured due to the limited view of the camera. The cell phone sensor has a relatively high acquisition frequency, and it can be used to fill this gap. On top of this, we also introduce Kalman filtering to correct the accumulation of sensor errors. The whole localization process starts by detecting the markers that appear in each frame of the video and the information obtained is the ID of the marker and the 2D coordinates of each of its corners. The 3D coordinates of the camera are obtained by solving the PnP (Perspective-n-Point) problem. At the same time, the sensors are working. The fusion of the accelerometer and gyroscope also allows for obtaining the current position. Therefore, this result will be used to fill the gaps where no markers are detected. Finally, based on the computed 3D coordinates of the device, the trajectory of the device can be plotted and compared with other systems.

2. METHODOLOGY

2.1. Overview

The goal of this project is to locate and track a mobile device and plot the 3D trajectory of the mobile device. The inputs to the system are video frames taken by the mobile device, measurements from inertial sensors, gyroscopes and magnetometers, and the outputs are its 3D motion trajectory. To accomplish the goal of this project, we divided the system into five modules. The whole process of this system is shown in Figure 1. The whole system is divided into six modules. The first module is the camera calibration. The results obtained from this module are the intrinsic and extrinsic parameters of the camera. This result will be used as input for the second module. The second module is the detection of the marker using YOLO. The result of the detection is the two-dimensional coordinates of the four corners of the marker and its unique ID. when the system does not detect the marker, it goes to the third module, which is the IMU-based localization. This module takes the values obtained from the IMU and fuses them to calculate the displacement of the device. The results obtained from this step are used as input to the Kalman filter in the fourth module to correct the results of the sensor fusion calculation. When the system clearly detects the marker, it proceeds to the fifth module, which uses the PnP algorithm to estimate the pose of the device. The main task of the last module is to present the results of the previous calculations and to plot the trajectory for comparison.

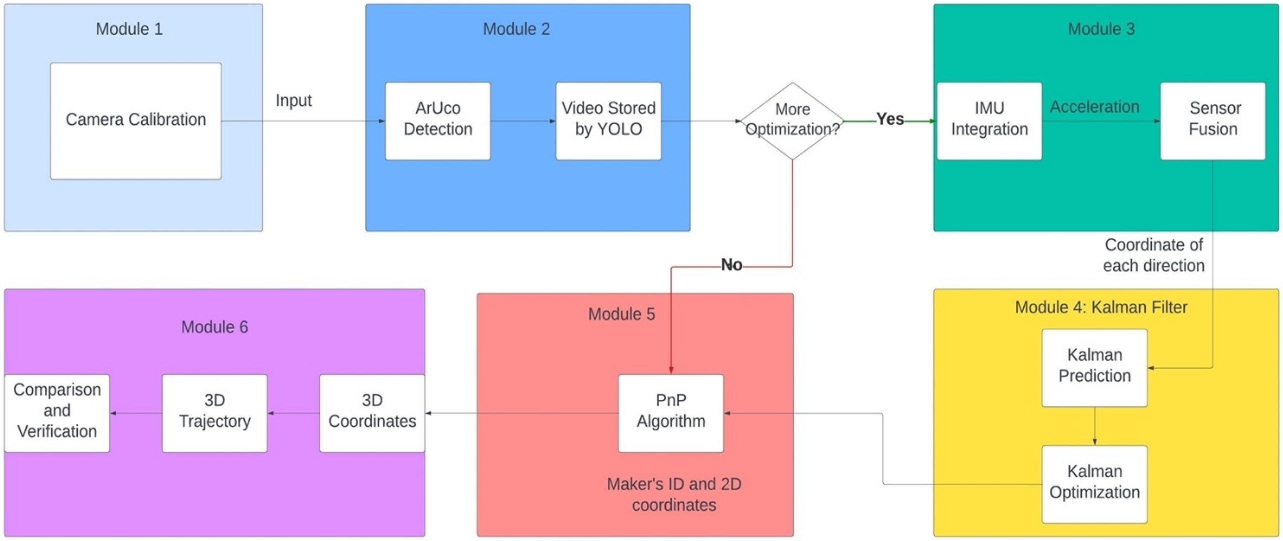


Fig. 1. The entire process of the system

2.2. Camera Calibration

The first module is the camera calibration. In machine vision applications, a geometric model of camera imaging is required to determine the 3D geometric position of a point on the surface of a spatial object and its corresponding point in the image. These geometric model parameters are the camera parameters. In most conditions, these parameters must be obtained through experiments and calculations, and this process of solving the parameters (intrinsic, extrinsic, and distortion parameters) is called camera calibration [8].

The first step in camera calibration requires converting the world coordinate system to the camera coordinate system. The world coordinate system (X_W, Y_W, Z_W), also known as the measurement coordinate system, is a three-dimensional right-angle coordinate system in which the spatial position of the camera and the object to be measured can be described. The position of the world coordinate system can be freely determined according to the actual situation. The camera coordinate system (X_C, Y_C, Z_C), also a three-dimensional right-angle coordinate system, the origin is located at the optical center of the lens, x and y axes are parallel to the two sides of the phase plane, the z-axis for the lens optical axis, and perpendicular to the image plane. The conversion process is shown in Equation (1).

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} \quad (1)$$

where R is a 3×3 rotation matrix, t is a 3×1 translation vector, $(X_C, Y_C, Z_C)^T$ and $(X_W, Y_W, Z_W)^T$ are the homogeneous coordinates of the camera coordinate system and world coordinate system, respectively.

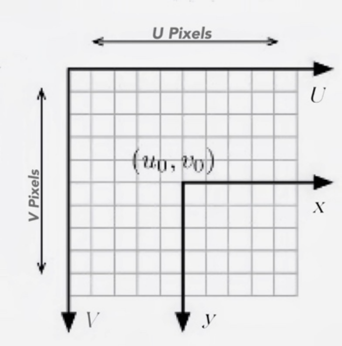


Fig. 2. Pixel coordinate and image coordinate

The next step is the conversion of pixel coordinates and image coordinates. As shown in Figure 2, the pixel coordinate system *uov* is a two-dimensional right-angle coordinate system that reflects pixel arrangement in the camera chip. The origin o is in the upper left corner of the image, and the *u* and *v* axes are parallel to the two sides of the image plane, respectively. The units of the axes in the pixel coordinate system are pixels. The pixel coordinate system is not conducive to coordinate transformation, so it is necessary to establish the image coordinate system XOY. The unit of its coordinate axis is usually millimeters (mm). The origin is the intersection of the camera's optical axis and the phase plane (called the principal point), which is the center of the image. X-axis and Y-axis are parallel to the u-axis and v-axis, respectively. Therefore, the two coordinate systems are translational, i.e., they can be obtained by translation. This conversion can be done by Equation (2).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} 1/dX & 0 & u_0 \\ 0 & 1/dY & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \quad (2)$$

where, dX, dY are the physical dimensions of the pixel in the X and Y-axis directions, respectively. u_0 and v_0 are the coordinates of the principal point.

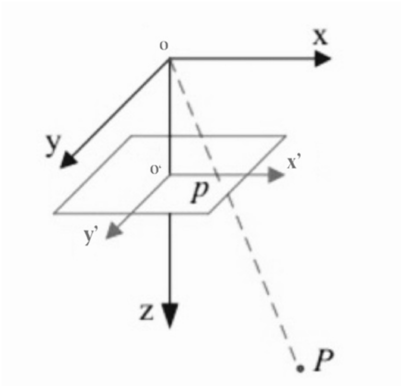


Fig. 3. Pinhole imaging principle

Figure 3 illustrates the relationship between any point P in space and its image point p. The line between P and the camera optical center o is oP, and the intersection of oP and the image plane p is the projection of the point P in space on the image plane. This process is perspective projection, as represented by the following matrix:

$$s\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3}$$

where s is the scale factor (s is not zero), and f is the effective focal length (the distance from the optical center to the image plane). $(x, y, z, 1)^T$ is the homogeneous coordinates of the spatial point P in the camera coordinate system xoy, and $(X, Y, 1)^T$ is the homogeneous coordinates of the image point p in the image coordinate system XOY. Combining Equations (1) to (3) we can get the intrinsic and extrinsic parameters of the camera.

$$s\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = \begin{bmatrix} 1/dX & 0 & u_0 \\ 0 & 1/dY & v_0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} = \begin{bmatrix} a_x & 0 & u_0 & 0 \\ 0 & a_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} = \boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{X}_w \tag{4}$$

where, $a_x = f/dX$, $a_y = f/dY$, are called the scale factors of the u and v axes. M_1 and M_2 are the intrinsic and extrinsic parameters of the camera, respectively.

2.3. YOLO Detection

The second module uses YOLO to detect markers. As we mentioned earlier, YOLO is a single-stage detector, which is fast and ideal for applications in real-time systems. The following will describe how YOLOV3 is applied to this project.

When a frame is passed into YOLOV3, this image is first resized to 416x416 grids, and a gray bar is added around the image to prevent distortion. YOLOV3 then splits the images into 13x13, 26x26, and 52x52 grids, which are used to detect large, medium, and small objects, respectively. Each grid point is responsible for detecting its lower right corner area, and if the object's center point falls in a grid, then the object's position will be determined by that grid point. In the example given in Figure 4, an image containing the object to be detected, i.e., the ArUco marker, is input into the YOLOV3 neural network and then surrounded by gray bars. In this image, the ArUco belongs to a large object, so a 13×13 grid image will detect the result. For the training part of YOLO, we choose to train on Google Colab.

2.4. Camera Pose Estimation

2.4.1. Principle&Conditions

The two-dimensional coordinates obtained from the detection of the markers will be used for the estimation of the camera pose. The camera poses estimation is mainly based on the PnP (Perspective-n-point) algorithm. General conditions for the PnP problem [9]:

- Coordinates of the n 3D reference points in the world coordinate system.
- Corresponding to these n 3D points, the coordinates of the 2D reference point are projected on the image.
- The intrinsic parameters of the Camera are denoted by M_1.

Based on our experimental results (for details, see Section IV), the EPnP algorithm is of the highest accuracy among the existing PnP algorithms.

Most non-iterative PnP algorithms will first solve for the depth of the feature point to obtain its 3D coordinates of it in the camera coordinate system. The EPnP algorithm, on the other hand, represents the 3D coordinates in the world coordinate system as a weighted sum of a set of virtual control points. For the general case, the EPnP algorithm requires the number of control points to be four, and these four control points cannot be coplanar. Because the camera's extrinsic parameters are unknown, the coordinates of these four control points under the camera reference coordinate system are unknown. Furthermore, if we can solve the coordinates of these four control points under the camera reference coordinate system, we can calculate the camera's pose [10].

2.4.2. Control Points and Barycentric Coordinates

Using the EPnP algorithm as a basis, the coordinates of the control points and are represented. This can help us to estimate homogeneous barycentric coordinates. Further, we can calculate barycentric coordinates from isometric coordinates. The positions of these points can help us to derive the camera angle and thus further estimate the camera pose.

In this paper, the superscripts ^{w} and ^{c} are used to denote coordinates in the world and camera coordinate systems, respectively. Then, the coordinates of the 3D reference points in the real-world frame are $p^w{}_i, i = 1,...,n$, the coordinates in the camera frame are $p^c{}_{i,} i = 1,...,n$. The four control points in the world coordinate system are $c^w{}_j, j = 1,...,4$, the coordinates in the camera reference coordinate system are $c^c{}_{j,} j = 1,...,4$.

$$\begin{bmatrix} \boldsymbol{p}_i^w \\ 1 \end{bmatrix} = \boldsymbol{C} \begin{bmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ a_{i4} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_1^w & \boldsymbol{c}_2^w & \boldsymbol{c}_3^w & \boldsymbol{c}_4^w \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ a_{i4} \end{bmatrix} \quad (5)$$

$[{p^W}_i{}^T 1]^T$ and $[{c^W}_j{}^T 1]^T$ are both isometric coordinates. Thus, we also get barycentric coordinates computed as follows: (See Appendix for detailed derivation)

$$\begin{bmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ a_{i4} \end{bmatrix} = \boldsymbol{C}^{-1} \begin{bmatrix} p_i^w \\ 1 \end{bmatrix} \qquad (6)$$

2.4.3. Selection of Control Points

A specific method for determining the control points is given in here. The set of 3D reference points is $P_i^W, i = 1,\ldots,n$ and the barycentric coordinates of the 3D reference points are chosen as the first control point:

$$\boldsymbol{c}_1^w = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{p}_i^w \qquad (7) \qquad\qquad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{p}_1^{w^T} - \boldsymbol{c}_1^{w^T} \\ \ldots \\ \boldsymbol{p}_n^{w^T} - \boldsymbol{c}_1^{w^T} \end{bmatrix} \qquad (8)$$

Donating the characteristic value of A^TA as $\lambda_{C,i,i=1,2,3}$, the corresponding feature vector is $v_{c,i,i=1,2,3}$. Thus, the remaining three control points can then be determined by the following formula:

$$\boldsymbol{c}_j^w = \boldsymbol{c}_1^w + \lambda_{c,j-1}^{\frac{1}{2}} v_{c,j-1}, j = 2, 3, 4 \qquad (9)$$

2.4.4. Solve for Coordinates of the Control Point in Camera Coordinates:

$u_i, i = 1,\ldots,n$ is the 2D projection of the reference point $p_i, i = 1,\ldots,n$, then,

$$\forall i, w_i \begin{bmatrix} \boldsymbol{u}_i \\ 1 \end{bmatrix} = K\boldsymbol{p}_i^C = K\sum_{j=1}^{4} a_{ij}\boldsymbol{c}_j^C \qquad (10)$$

Substitute $c^C{}_j = [x^C{}_j, y_j{}^C, z_j{}^C]^T$ into the above equation and write K in the form of focal length f_u, f_v and optical center (u_c, v_c), then,

$$\forall i, w_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} f_u & 0 & u_c \\ 0 & f_v & v_c \\ 0 & 0 & 1 \end{bmatrix} \sum_{j=1}^{4} a_{ij} \begin{bmatrix} x_j^C \\ y_j^C \\ z_j^C \end{bmatrix} \qquad (11)$$

Two linear equations can be obtained from Equation 11:

$$\sum_{j=1}^{4} a_{ij} f_u x_j^C + a_{ij}(u_c - u_i) z_j^C = 0 \qquad (12)$$

$$\sum_{j=1}^{4} a_{ij} f_v y_j^C + a_{ij}(v_c - v_j) z_j^C = 0 \qquad (13)$$

Concatenating all n reference points, we can obtain a linear system of equations:

$$Mx = 0 \qquad (14)$$

where M is a $2n \times 12$ matrix, x = $[\boldsymbol{c}_1^{c^T}, \boldsymbol{c}_2^{c^T}, \boldsymbol{c}_3^{c^T}, \boldsymbol{c}_4^{c^T}]$, x is the coordinate of the control point in the camera coordinate system, which is a 12×1 vector and x is in the right null space of M, or $x \in ker(M)$. Hence,

$$\boldsymbol{x} = \sum_{i=1}^{N} \beta_i \boldsymbol{v}_i \qquad (15)$$

In the equation above, v_i is the N eigenvector corresponding to the N null eigenvalues of M. For the i-th control point:

$$\boldsymbol{c}_j^C = \sum_{k=1}^{N} \beta_k \boldsymbol{v}_K^{[i]} \qquad (16)$$

where v_k is i-th 3×1 sub-vector of eigenvector v_k. Then we can obtain v_i by computing the eigenvectors of $M^T M$.

The next step is to calculate $\beta_{i,i=1,\ldots,N}$. Because the extrinsic parameters of the camera describe only coordinate transformations and do not change the distance between control points, thus:

$$\left\|\boldsymbol{c}_i^C - \boldsymbol{c}_j^C\right\|^2 \qquad (17) \qquad\qquad \left\|\sum_{k=1}^{N} \beta_k \boldsymbol{v}_K^{[i]} - \sum_{k=1}^{N} \beta_k \boldsymbol{v}_K^{[j]}\right\|^2 = \left\|\boldsymbol{c}_i^C - \boldsymbol{c}_j^C\right\|^2 \qquad (18)$$

This is a linear equation for $\beta_{ij,i,j=1,\ldots,N}$. In EPnP algorithm
[14], four cases $N = 1,2,3,4$ is discussed. When N takes different values, the number of unknowns of the linear equation is:

- N = 1, the unknown number is 1 • N = 2, the unknown number is 3
- N = 3, the unknown number is 6
- N = 4, the unknown number is 10

When $N = 4$, the number of equations is 6 and the number of unknowns is more than the number of equations. By commutativity of the multiplication, we have

$$\beta ab\beta cd = \beta a\beta b\beta c\beta d = \beta a0b0\beta c0d0 \qquad (19)$$

where $\{a^0,b^0,c^0,d^0\}$ represents any permutation of the integers $\{a,b,c,d\}$. Then we can reduce the number of unknowns. For example, if we solve for $\beta_{11},\beta_{12},\beta_{13}$, then we get $\beta_{23} = \frac{\beta_{12}\beta_{13}}{\beta_{11}}$.

2.4.5. Gauss-Newton Optimization

The objective function of the optimization is:

$$Error(\beta) = \sum_{(i,js.t.i<j)} \left(\left\|\boldsymbol{c}_i^C - \boldsymbol{c}_j^C\right\|^2 - \left\|\boldsymbol{c}_i^W - \boldsymbol{c}_j^W\right\|^2\right)^2 \qquad (20)$$

2.4.6. Calculating the Camera's Pose

The calculation of the pose of the camera in the EPnP algorithm is as follows.

1) Calculate the coordinates of the control point in the camera reference coordinate system.

$$c_i^c = \sum^{N} \beta_k v_k^{[i]}, i = 1,2,3,4 \qquad (21)$$

j=1

2) Calculate the coordinates of the 3D reference point in the camera reference coordinate system.

$$\boldsymbol{p}_i^c = \sum_{j=1}^{4} a_{ij} \boldsymbol{c}_j^c, i = 1, ..., n \qquad (22)$$

3) Calculate the barycentric coordinates $p^{w}{}_{0}$ of $p^{w}{}_{i}$,i = 1,...,n and matrix A:

$$\boldsymbol{p}_0^w = \frac{1}{n} \sum_{i=1}^{n} p_i^w \qquad (23) ,$$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{p}_1^{w^T} - \boldsymbol{p}_0^{w^T} \\ \cdots \\ \boldsymbol{p}_n^{w^T} - \boldsymbol{p}_0^{w^T} \end{bmatrix} \qquad (24)$$

4) Calculate the barycentric coordinates $p^{c}{}_{0}$ of $p^{c}{}_{i}$,i = 1,...,n and matrix B:

$$\boldsymbol{p}_0^c = \frac{1}{n} \sum_{i=1}^{n} p_i^c \qquad (25) ,$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{p}_1^{c^T} - \boldsymbol{p}_0^{c^T} \\ \cdots \\ \boldsymbol{p}_n^{c^T} - \boldsymbol{p}_0^{c^T} \end{bmatrix} \qquad (26)$$

5) Calculate H:

$$H = B^T A \qquad (27)$$

6) Calculating the singular value decomposition (SVD) of H:

$$H = U^{X} V^{T} \qquad (28)$$

7) Calculate the rotation R in the pose:

$$R = UV^{T} \qquad (29)$$

8) Calculate the translation t in the pose:

$$t = \boldsymbol{p}_0^c - Rp^{w}{}_{\mathbf{0}} \qquad (30)$$

l represents the camera's position in the real-world coordinate system:

$$l = -R^{-1} t \qquad (31)$$

2.5. Sensor Fusion

2.5.1. Role of IMU in Our System

The third module is the IMU-based localization method. This method is used as a replacement when visual localization is not available. As can be seen in Figure 1, when the marker is not detected, the system uses the IMU measurements to calculate the 3D coordinates of the mobile device. The sensors used in this system are mainly an accelerometer, gyroscope, and magnetometer, all of which are currently equipped in smartphones. Before we can use the sensor for localization, we need to convert the phone's coordinate system.

2.5.2. Coordinate System Conversion

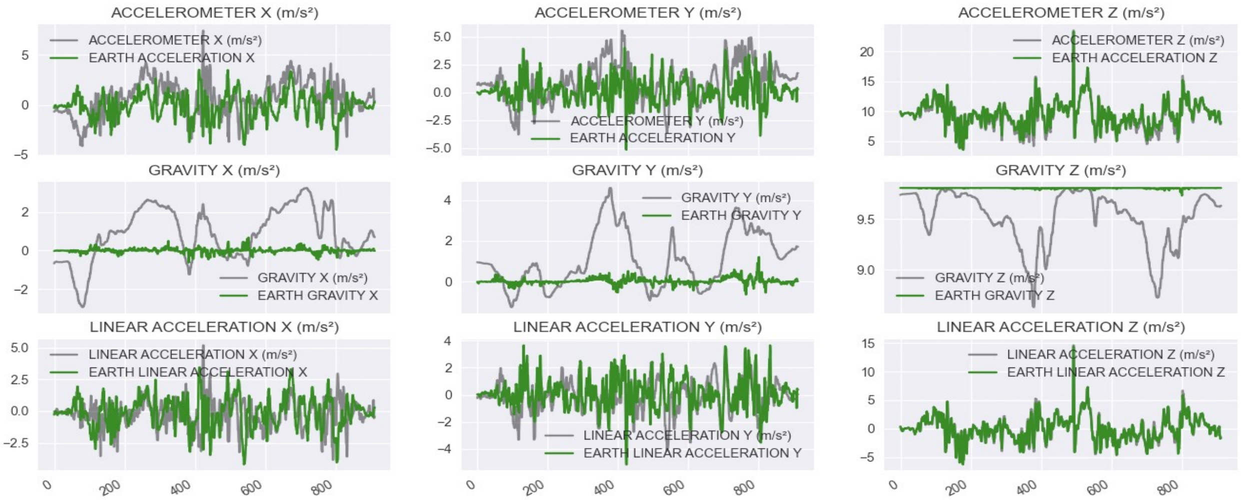


Fig. 4. Change in sensor values before(orange) and after(green) coordinate system conversion

The acceleration sensor of an Android phone refers to its coordinate system when measuring acceleration. Therefore, we need to convert the phone's coordinate system to an inertial, non-rotating coordinate system, which is the Earth coordinate system. This conversion will make it possible to hold the Android phone in any orientation and measure the correct acceleration vectors to calculate the phone's trajectory in the earth coordinate system. The transformation from the Phone coordinate to Earth's transformation [11] is done with formula (32) as shown below

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = R_z(\psi)R_y(\theta)R_x(\phi)\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} cos\psi & -sin\psi & 0 \\ sin\psi & cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} cos\theta & 0 & sin\theta \\ 0 & 1 & 0 \\ -sin\theta & 0 & cos\theta \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\phi & -sin\phi \\ 0 & sin\phi & cos\phi \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (32)$$

Where [X, Y, Z] are the phone's coordinate system linear accelerations, and R_z,R_y,R_x are the rotation matrices for each axis in order to rotate the [X, Y, Z] over to earth's [x,y,x] axes. The Euler angles (ψ,θ,φ) correspond to the angles about the pitch, roll, and yaw. The difference between the acceleration in the earth coordinate system (The green line) and the Android phone coordinate system (The gray line) can be seen in Figure 4. The vertical coordinate in the figure represents the value of the sensor, and the horizontal coordinate represents the sampling times (we used a sampling rate of 100 Hz). After the conversion, the Z-axis's gravity stays near 9.8 meters per second, while the gravity in the X-axis and Y-axis stays near 0 meters per second. This result is the same as what we know as common sense.

2.5.3. Displacement Estimation

The last step of the sensor module is to use the acceleration to calculate the displacement and thus estimate the distance. Acceleration is the rate of change of an object's velocity. At the same time, velocity is the rate of change in the position of the same object. In other words, velocity is the derivative of position, and acceleration is the derivative of velocity. Therefore, the following equation is available, (See Appendix for derivation)

$$\vec{s_x} = \int \left(\int (\vec{a_x})\, \mathrm{d}t \right) \mathrm{d}t \tag{33}$$

A similar expression for displacement in the y-axis and z-axis.

2.6. Kalman Filter Correction

After the initial value is given by the IMU, the error of the sensor itself will accumulate in the continuous optimization. The role of Kalman filtering here is to correct this error using linear iterations. Kalman filtering [12] is mainly divided into two steps, prediction, and correction. Prediction is the estimation of the current state based on the state of the previous moment, and correction is the integrated analysis based on the observation of the current state and the estimation of the previous moment to estimate the system's optimal state value. Then the process is repeated the next moment. The Kalman filter iterates continuously, it does not require many-particle state inputs, only process quantities, so it is fast and well-suited for state estimation of linear systems. Applying Kalman filtering to this project uses the position information obtained from vision-based localization to update the position information obtained from sensor-based localization.

3. SYSTEM TESTING EXPERIMENT

3.1. Design and Preparation

To demonstrate that our proposed method is superior to both sensor-based localization only and vision-based localization only, we first tested the effect of applying only one of these methods for localization and tracking. Lastly, we tested our complete system, i.e., applying the IMU to support visual localization.

Our experimental setup is as follows:

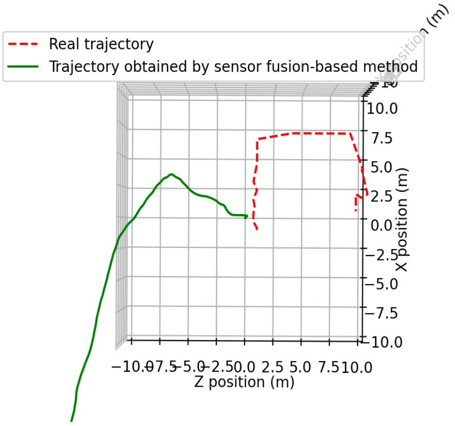
- Location: The laboratory in the Prairie Springs Science Center
- Device: Nokia 7.2
- Measuring tool: Tape measure
- Conditions: Sufficient and insufficient light; shade and no shade on markers

We choose a point in the lab as the origin of the world coordinates, and then affix ArUco markers at different heights

and on different surfaces. Each marker's location information was recorded to observe the difference between the system test results and the actual data results. We take the center point of the first marker as the origin of the world coordinate system.

In our indoor localization and tracking experiment, fifteen trails are conducted. Figure 5 is the actual 3D moving trajectory. The actual path shows the route of our experiment. The true path contains two corners, as well as a movement in the vertical direction. The maximum distance of motion in the x-axis direction reached 7 meters, and the maximum length of motion in the z-axis direction was 10 meters.

3.2. Testing of IMU-based Indoor Tracking & Vision-based Tracking



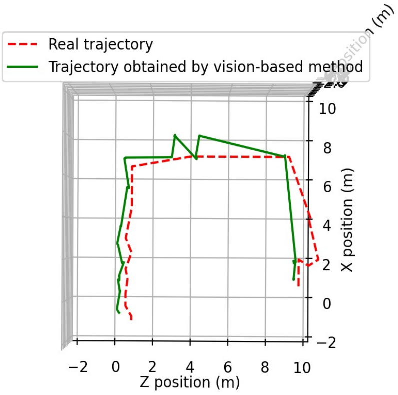


Fig. 5(left). Comparison of real trajectory and trajectory obtained by sensor fusion-based method
Fig. 6(right). Comparison of real trajectory and trajectory obtained by vision fusion-based method

We first conducted 15 experiments on the IMU-based indoor localization method to evaluate its effectiveness of the method. Figure 5 shows the results of the sensor fusion-based method through 15 experiments. As can be seen from the figure, the sensor-based method, which is also known as the IMU-based method, has a significant drift in obtaining motion trajectories. The reason why the IMU-based method has such a large error is that the interference of the gravitational acceleration in the vertical direction cannot be eliminated and there are residuals. In addition, the double integration of acceleration leads to the accumulation of errors. Another reason is the drift of the sensor during the measurement. Specifically, the reading of the inertial sensor is not zero when a moving device goes from rest to motion and back to rest. The second part of the experiment is for the vision-based localization method. Again, this part of the experiment was repeated fifteen times. Figure 6 shows the comparison between the trajectory obtained by the vision-based localization method and the real trajectory. It can be seen from the figure that the vision-based approach provides trajectories that are not smooth because it calculates the absolute position of the mobile device. This is caused by the fact that the vision-based localization method relies on markers as reference positions. When no marker is captured in FoV, the pose, and position of the camera cannot be estimated. In this case, there will be a gap in the trajectory, and the localization will continue once at least one marker is detected. As a result, the estimated motion trajectory will suddenly move from the previous position to the current one.

3.3. Testing of Our Optimized Method

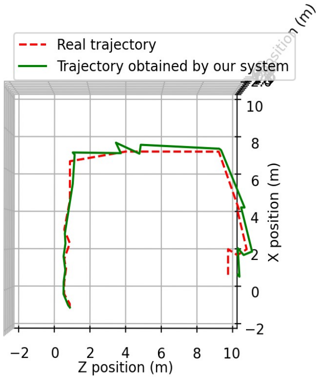


Fig. 7. Comparison of real trajectory and trajectory obtained by our system

The third part of the experiment is to evaluate our proposed method and test the reliability and effectiveness of our system. The experiment has also been repeated 15 times. Figure 7 points out that the obtained trajectories have no big jumps, which means that sensor-based localization successfully fills this gap when vision-based localization cannot be used. This largely indicates that this method combines the advantages of the two methods mentioned above and achieves optimization. The new method achieves improvements of 93.6%, 80%, and 84.4% in the x-axis, y-axis, and z-axis, respectively, with respect to the sensor-based method.

3.4. Experiment Results in Comparison

The experimental results show that the IMU-based indoor localization method has the worst performance. The average error of this method is about 8 meters in our 15 iterations of experiments. The vision-based localization method performs better, with an average error of less than 1 meter. Our system performed the best, with the highest accuracy of localization, with an average error of less than 0.5 m. Compared with the other two axes, the motion trajectory of the phone obtained by our system has a relatively large error in the Y-axis. The reason for this phenomenon may be that the markers are closer together in the vertical direction, causing repeated detection. Overall, this greatly enhances the accuracy of indoor positioning.

4. CONCLUSION AND FUTURE WORKS

Our proposed neural network-based indoor localization method improves the stability of the indoor tracking system. It is improved by introducing inertial sensors to assist vision-based localization. Compared to vision-based localization without the assistance of inertial sensors, our system avoids the inability to localize due to missing markers. The proposed method does not rely on any expensive depth camera and can be easily planted on a mobile device. We evaluate and validate our method with the prototype implementation on the smartphone platform. The experimental results show that the system has strong robustness to the complex indoor environment, strong anti-interference ability, high accuracy, and fast processing speed, which meets the demand for indoor localization. The neural network model we trained explicitly for detecting ArUco code has a breakneck detection speed, taking only 0.164 seconds to detect a single frame. Overall, the proposed method demonstrates a tracking accuracy of under 0.5 meters.

In terms of future work, we plan to increase further the number and diversity of datasets, which will significantly improve the accuracy of the neural network in detecting markers. This improves the accuracy of vision-based location detection of markers, which in turn improves the accuracy of our system's location. We also plan to include hardware devices like depth cameras in the approach. This will allow our visual localization method to no longer rely on markers and use objects already present in the indoor environment for localization.

5. APPENDIX

5.1. Derivation for barycentric coordinates

The EPnP algorithm expresses the coordinates of the reference point as a weighted sum of the coordinates of the control point:

$$\boldsymbol{p}_i^w = \sum_{j=1}^{4} a_{ij}\boldsymbol{c}_j^w, with \sum_{j=1}^{4} a_{ij} = 1 \tag{1}$$

where the a_{ij} are homogeneous barycentric coordinates. They are unique and can easily be estimated. In the camera coordinate system, the same relationship exists:

$$\boldsymbol{p}_i^c = \sum_{j=1}^{4} a_{ij}\boldsymbol{c}_j^c \tag{2}$$

Assuming that the extrinsic parameters (rotation matrix *R* and translation vector *t*) of the camera are [*R*,*t*] then a relationship exists between the virtual control points $c^w{}_j$ and $c^c{}_j$:

$$\boldsymbol{c}_j^c = \begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\end{bmatrix}\begin{bmatrix}\boldsymbol{c}_j^w \\ 1\end{bmatrix} \tag{3}$$

Considering that the EPnP algorithm expresses the reference point coordinates as a weighted sum of the control point coordinates, then can get:

$$\boldsymbol{p}_i^c = [\boldsymbol{R} \quad \boldsymbol{t}] \begin{bmatrix} p_i^w \\ 1 \end{bmatrix} = [\boldsymbol{R} \quad \boldsymbol{t}] \begin{bmatrix} \sum_{j=1}^4 a_{ij} c_j^w \\ 1 \end{bmatrix} \qquad (4)$$

$$\begin{aligned} \boldsymbol{p}_i^c &= [\boldsymbol{R} \quad \boldsymbol{t}] \begin{bmatrix} \sum_{j=1}^4 a_{ij} \boldsymbol{c}_j^w \\ 1 \end{bmatrix} \\ &= \sum_{j=1}^{4} a_{ij} \, [\boldsymbol{R} \quad \boldsymbol{t}] \begin{bmatrix} \boldsymbol{c}_j^w \\ 1 \end{bmatrix} = \sum_{j=1}^{4} a_{ij} \boldsymbol{c}_j^c \end{aligned} \qquad (5)$$

In the above derivation, the important constraint $\sum_{j=1}^4 a_{ij} = 1$of EPnP on the weight a_{ij} is used. Without this constraint, the above derivation will not hold. Putting the four control point constraints together yields the following equation:

$$\begin{bmatrix} \boldsymbol{p}_i^w \\ 1 \end{bmatrix} = \boldsymbol{C} \begin{bmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ a_{i4} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_1^w & \boldsymbol{c}_2^w & \boldsymbol{c}_3^w & \boldsymbol{c}_4^w \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ a_{i4} \end{bmatrix} \qquad (6)$$

Obviously, $[p^{W}{}_{i}{}^{T}1]^{T}$ and $[c^{W}{}_{j}{}^{T}1]^{T}$ are both isometric coordinates. The equation (1) in the reference paper, however, is essentially a linear combination of the isometric coordinates of 3D reference points with the isometric coordinates of the control points. Thus, we also get barycentric coordinates computed as follows:

$$\begin{bmatrix} a_{i1} \\ a_{i2} \\ a_{i3} \\ a_{i4} \end{bmatrix} = \boldsymbol{C}^{-1} \begin{bmatrix} p_i^w \\ 1 \end{bmatrix} \qquad (7)$$

5.2. Derivation for Displacement Expression

The relation between acceleration/ displacement and velocity: $\vec{a} = \dfrac{d\vec{v}}{dt}$, $\vec{v} = \dfrac{d\vec{s}}{dt}$

The relation between displacement and velocity: $\vec{a} = \dfrac{d(d\vec{s})}{dt^2}$

The integral is the opposite of the derivative. If the acceleration of an object is known, then we can obtain the position of the object using the double integral. Assuming that the initial condition is 0, then there is the following equation.

$$\vec{v} = \int (\vec{a}) \, \mathrm{dt}$$

$$\vec{s} = \int (\vec{v}) \, \mathrm{dt}$$

6. REFERENCES

[1] T. Lindner, L. Fritsch, K. Plank, and K. Rannenberg, "Exploitation of public and private wifi coverage for new business models," in *Building the E-Service Society*. Springer, pp. 131–148, 2004.

[2] G. E. Violettas, T. L. Theodorou, and C. K. Georgiadis, "Netargus: A snmp monitor & wi-fi positioning, 3-tier application suite," in *2009 Fifth International Conference on Wireless and Mobile Communications*. IEEE, pp. 346–351, 2009.

[3] J. Kunhoth, A. Karkar, S. Al-Maadeed, and A. Al-Ali, "Indoor positioning and wayfinding systems: a survey," *Human-centric Computing and Information Sciences*, vol. 10, pp. 1–41, 2020.

[4] G. MAPS, "Google indoor maps," Website, [Online]. Available from: https://www.google.com/maps/about/partners/indoormaps/, 2021(Accessed 22-December-2022).

[5] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.

[6] P. Roy and C. Chowdhury, "A survey of machine learning techniques for indoor localization and navigation systems," *Journal of Intelligent & Robotic Systems*, vol. 101, no. 3, pp. 1–34, 2021.

[7] N. Piasco, D. Sidibe, C. Demonceaux, and V. Gouet-Brunet, "A survey´ on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, 2018.

[8] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[9] A. A. B. Pritsker, *Introduction to Simulation and SLAM II*. Halsted Press, 1984.

[10] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[11] Wikipedia contributors, "Conversion between quaternions and Euler angles — Wikipedia, the free encyclopedia," [Online]. Available from: https://en.wikipedia.org/w/index.php?title=Conversion between quaternions and Eulerangles&oldid=998442809, 2021(Accessed 12-January-2023).

[12] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

Computing for Data Science Course

Mark Fienup

Computer Science Department

University of Northern Iowa

Cedar Falls, Iowa 50614

mark.fienup@uni.edu

Abstract

This paper describes the Computing for Data Science course offered at the University of Northern Iowa (UNI) as part of the Data Science minor. This course acts as a bridge for non-Computer Science majors between UNI's CS1 course, Introduction to Computing, and our upper-level Computer Science course: Database Systems. The course's goals and organization are described including layout of topics, labs, and programming assignments.

1 Introduction

During the last couple years, Data Science programs in higher education have exploded. At the University of Northern Iowa (UNI) our initial response was to create an interdisciplinary Data Science minor using mostly existing courses, but with a few new courses tailored for the minor. This paper describes the Computing for Data Science course offered by the Computer Science department at the University of Northern Iowa (UNI) as part of the Data Science minor -- see Appendix A. This new course acts as a bridge for non-Computer Science majors between UNI's CS1 course, Introduction to Computing, and our upper-level Computer Science course: Database Systems.

The prerequisite course for Computing for Data Science is the Introduction to Computing (CS 1510). The Introduction to Computing course is taught using the Python programming language and expects no prior programming experience by students. Its course description is: "CS 1510. Introduction to Computing — 4 hrs. Introduction to software development through algorithmic problem solving and procedural abstraction. Programming in the small. Fundamental control structures, data modeling, and file processing. Significant emphasis on program design and style."

Traditionally, the prerequisites for Database Systems are both Data Structures and Discrete Structures where the Discrete Structures course is a discrete mathematics course taught by the Computer Science department to have a stronger CS focus. The Data Structures is also taught using Python and has a course description of "CS 1520. Data Structures — 4 hrs. Introduction to use and implementation of data structures such as sets, hash tables, stacks, trees, queues, heaps, and graphs. Additional topics include searching algorithms, sorting algorithms, and algorithmic time and space complexity analysis. Design and implementation of programs using functional decomposition."

The two high-level goals of the Computing for Data Science course are:

Goal 1. Prepare students to succeed in the Database Systems course by providing the crucial knowledge from the traditional prerequisite courses of Data Structures and Discrete Structures.

Goal 2. Advance students understanding and skills in Data Science.

2 Organization of the Course

Python is the programming language used in the course so we can build upon student programming skills learned in the Introduction to Computing. Plus, Python is frequently used by Data Scientists in the real world since many "third-party" Data Science tools/libraries (e.g., NumPy, Pandas, SciPy, TensorFlow, etc.) are available through Python. To avoid the complication of students installing these Data Science tools/libraries, the course uses Google Colab or "Colaboratory" [1] which allows access to all of these tools with a just a web-browser. No configuration is required and allows for easy sharing of files via Google drive and GitHub. In addition, it allows access to GPUs for computationally intensive Data Science tools free of charge.

My teaching philosophy is that students learn best if they are actively engaged [2]. Traditional lecturing to students is not very effective for student learning. Research has shown that 10 minutes is about the maximum attention span of students during lecture. However, this attention span clock can be reset by having students trying to apply what they just heard about. That is why I like to structure my classes into mini-lectures followed by small group activities -- typically a series of questions. After each small group activity, I like to discuss "correct" answers to the questions. To keep discussion focused and on track, I hand-out paper copies of the mini-lecture material (e.g., code, diagrams, timings, etc.) and corresponding questions with space for them to record their answers and the correct answers. Students find these material useful when studying for exams.

Since Computing for Data Science is intended to teach programming skills, I find "tightly coupled" laboratory activities useful for student learning. While the course does not have a separate formal “closed-laboratory” time, the course is taught on Tuesdays and Thursdays in an 90-minute periods. Typically, Tuesday’s period introduces new material/topics in the mini-lecture fashion described above, and Thursday’s period is a laboratory activity applied extensions to what was discussed in the previous Tuesday period. In laboratory activities are typically timing code, writing code segments, and answering related analytical questions.

The current textbook Foundational Python for Data Science [3] is more of a minimalistic Python review and Data Science reference.

2.1 Three Instructional Units

To achieve the two goals for the course, it is split into three roughly equal units:

Unit 1. Python Programming for Data Science: This unit mainly addresses Goal 1 by covering select topics from Data Structures and Discrete Structures courses. Details of unit 1 are listed in Table 1 below.

Unit 2. Data Science Libraries: This unit mainly addresses Goal 2 by covering select Data Science tools: NumPy arrays, SciPy library, Pandas Series and DataFrames, and data visualization using matplotlib, Seaborn, Plotly and Bokeh. Details of unit 2 are listed in Table 2 below.

Unit 3. Advanced Python and Data Science Tools: This unit mainly addresses Goal 2 since the “advanced” Python coverage is not directly needed by the Database Systems course. The “advanced” Python topics are OOP (object-oriented programming) and the `re` (regular expression) module. The OOP coverage introduces a new programming paradigm frequently used in Python programs which the students were not exposed to in their Introduction to Computing course. The advanced Data Science tools currently covered are the `nltk` (Natural Language Toolkit) package and a brief introduction to machine learning libraries: TensorFlow and Scikit-learn. Details of unit 3 are listed in Table 3 below.

Week Number	Tuesday Period Topics	Thursday Period Lab Activity
1	Introduction to Google Colab and signed integer representation of data	Python Review: arithmetic expressions and "control statements:" if-elif, while loops, for loops, nested loops
2	IEEE 754 floating point representation and review of functions in Python	Python Review: strings, lists, dictionaries, list-of-lists, dictionary of string keys with list values
3	Introduction of big-oh analysis and big-oh of list and dictionary methods; general idea of hashing	Practice determining big-oh notations and timings of lists and dictionary methods
4	Text-file usage in Python and .csv processing	Practice .csv processing and writing of output text-file
5	Sets in Python and their relationship to databases	Practice with Python sets and frozensets
6	Review for Test 1	Test 1

Table 1: Python Programming for Data Science Unit Details.

Week Number	Tuesday Period Topics	Thursday Period Lab Activity
7	NumPy scientific computing package	Practice with NumPy 1-D and 2-D arrays: views vs. copy, filtering values, and array methods
8	SciPy scientific computing package	Practice SciPy modules to do image processing and graph processing with NumPy arrays
9	Pandas Series and DataFrame data structures	Practice using Pandas by processing .csv data file
10	Visualization of data with matplotlib, Seaborn, Plotly and Bokeh	Practice data visualization
11	Review for Test 2	Test 2

Table 2: Data Science Libraries Unit Details.

Week Number	Tuesday Period Topics	Thursday Period Lab Activity
12	OOP programming in Python	Practice writing Python classes, inheritance and creating objects
13	Python `re` (regular expression) module	Practice using `re` module
14	Introduction to Machine Learning	Practice using Scikit-learn and NLTK text classifier
15	Work Day	Review for Final/Test 3

Table 3: Advanced Python and Data Science Tools Unit Details.

2.2 Programming Projects

In addition to the weekly laboratory assignments, larger programming projects are assigned to allow students to practice designing and writing larger programs in Python. Details of the programming projects for Spring 2023 are listed in Table 4 below.

Project Number	Brief Description of Programming Project	Desired Learning Outcome
1	Rock, Paper, Scissors program	Review of Python basics and practice functional-decomposition design and using functions
2	Math Tutor program	Practice functional-decomposition design, using functions, and user-input validation
3	JUMBLE puzzle solver	Practice functional-decomposition design, using functions, selection of efficient data structures, and text-files
4	Steganography – Embedding secret message into an image and decoding the message	Practice functional-decomposition design, using functions, SciPy modules and NumPy arrays to do image processing, and binary bit manipulation
5	Dice game program – (TBD)	Practice OOD and OOP in Python

Table 4: Programming Projects Details for Spring 2023.

3 Conclusions

The Computing for Data Science course is currently being offered for only the second time this Spring 2023 semester. Because the Data Science minor is relatively new at UNI and the Computing for Data Science is only taken by non-Computer Science majors, it only has six students enrolled for Spring 2023. All 6 students are juniors or seniors with a Mathematics-Statistics/Actuarial Science major, and all are doing well in the course.

The first offering of the Computing for Data Science course during the Spring 2022 semester had only seven students with only three of these students having a declared Data Science minor. The remaining four students were using this course as a substitution on their Interactive Digital Studies (IDS) major from the Department of Communication and Media department. While all of the students had taken the prerequisite Introduction to Computing course, a couple of the IDS majors struggled with the programming aspects of the course. However, one IDS major really liked the course and switched their major to Computer Science.

Hopefully, enrollment in the Computing for Data Science course will grow as we get more Data Science minors outside of Computer Science. Starting Fall 2023 UNI's new general education program, UNI Foundational Inquiry (UNIFI), will include a Data Science certificate which contains a non-major CS1 type course with a Data Science focus that can help populate the Data Minor with more non-CS majors.

References

[1] Google Colab URL: https://colab.research.google.com/

[2] J. Philip East, and Mark Fienup, "Questions to Enhance Active Learning in Computer Science Instruction," Proceedings of the 35th Annual Midwest Instruction and Computing Symposium, (CDROM) April 2002.

[3] Foundational Python for Data Science, Kennedy Behrman. Pearson Education. ISBN: 978-0-13-662435-6

Appendix A – Data Science Minor Requirements (2022 – 2023 University of Northern Iowa Catalog)

Data Science Minor

The Data Science minor is an interdisciplinary program that integrates computer programming, machine learning, statistics, predictive modeling and visualization to provide students with broad based skills for extracting gainful information from data that originate from a variety of sources. A final project (ideally with corporate or non-profit partnerships) will ensure that students employ their skills to solve a real-world problem.

Statistics:		
STAT 1772	Introduction to Statistical Methods	3
STAT 4784/5784	Introduction to Machine Learning	3
Computer Science:		
CS 1510	Introduction to Computing	4
CS 2150	Computing for Data Science	3-7
or		
CS 1520 & CS 1800	Data Structures and Discrete Structures	
CS 3140/5140	Database Systems	3
Physics:		
PHYSICS 4160/5160	Data Visualization, Modeling and Simulation	3
Required Data Science Project		2-3
CS 4800	Undergraduate Research in Computer Science	
or MATH 4990	Undergraduate Research in Mathematics	
or PHYSICS 3000	Undergraduate Research in Physics	
Total Hours		**21-26**

Catapult Launch for Python Data Science Libraries

Leon Hannah Tabak

Department of Computer Science

Cornell College

Mount Vernon, Iowa 52314

l.tabak@ieee.org

Abstract

The author shares a description of a Jupyter notebook that introduces students to features of the Python programming language that may be unfamiliar to beginning programmers and to three libraries of software that they will use in their study of machine learning. The author's course, CSC316 Machine Learning, is open to students who have completed only a single programming course. The libraries (Matplotlib, NumPy, and Pandas) have wide application in data science and machine learning. Students learn by studying and experimenting with examples in the Jupyter notebook. Students experiment by modifying parameters in function calls. In several places, the notebook shows students two ways of computing the same result and invites students to try both.

1 Challenges for our students

Students of machine learning must learn how to work with large tables of data. Skills learned in a first course in computer science are not enough to meet this challenge.

Students at Cornell College take one course at a time. Classes meet five days per week for three and a half weeks (eighteen days). Classes typically meet three or four hours per day. The compressed schedule increases the importance of finding ways to bring students quickly up to speed on essential skills so that they can engage with the most important themes of the course.

In their first course in computer science, students learned how to use assignment statements, **if** statements, and **for** loops. They learned the difference between integers and floating point numbers. The first course introduced them to lists or arrays. It gave them just a little practice importing modules (e.g., math or turtle) and using functions found in those modules.

To work efficiently with large datasets, students will need to develop their skills with Python further. They will need to acquaint themselves with additional data types and data structures. They will need to learn how to identify and use library functions.

A greater knowledge of the Python programming language and a familiarity with powerful and popular libraries enables students to express themselves more concisely. Concise expression, in turn, allows quicker experimentation, a deeper understanding of the problems that they are trying to solve, and easier communication with teammates and clients.

The author has composed a Jupyter notebook and given it to his students to use as a means of quickly gaining the skills they need for their study of machine learning.

2 Python programming skills

The author's Jupyter notebook shows students how to assign one tuple to another. By letting Python unpack a tuple, students can replace two or more assignment statements with a single assignment statement.

The notebook shows students how to use slicing to extract a subset of elements from a list and how to use list comprehensions to build lists. The notebook introduces students to functional programming with examples that use Python's **lambda**, **map()**, and **filter()** functions. These features of the programming language enable students in many cases to replace loops that require several lines of code with a single line of code.

3 Powerful and popular libraries

3.1 NumPy

The notebook shows students how to create and populate NumPy arrays. NumPy arrays can model vectors. Students get practice with combining vectors arithmetically. Students use NumPy's functions to find the minimum, maximum, and mean values in an array. Students learn the definitions and significance of vector norms, standard deviation, and root mean square error. In these exercises, students see how a function call in a single line of code can substitute for a loop that spans several lines. They get faster computation and more readable source code.

3.2 Pandas

Pandas provides DataFrames. These are two-dimensional data structures. Unlike NumPy arrays, these are heterogeneous data structures—different columns in a table can hold different types of data. Students learn how to label, add, and delete columns and rows. They learn how to select rows and columns by indices, labels, and logical conditions. Students learn how to generate statistical summaries of the contents of a DataFrame.

3.3 Matplotlib

The notebook invites students to experiment with code that draws curves, points (scatter plots), and histograms (bar charts). Exercises within the notebook ask students to change foreground and background colors, the widths of lines, labels on axes, and titles on figures. Students learn how to specify the size of a figure and how to include several plots within a single figure.

4 Practice problems

The Jupyter notebook that the author has created for his students produces:

- A list of all prime numbers less than a given integer. The code identifies these prime numbers using the Sieve of Eratosthenes algorithm. This algorithm makes an array of non-negative integers. It marks zero and one to indicate that they are not primes, then examines the other integers in turn from smallest to largest. If it finds an integer that is not yet marked, it calls it prime and marks all of its multiples to indicate that they are not prime. (2 is prime, and so 4, 6, 8, and so on cannot be prime.)
- The number of primes less than a given integer and the number of primes within a specified interval (a density function).
- The distances between successive primes. This exercise uses slices and Python's **zip()** and **map()** functions.

- A list of all twin primes (pairs of prime numbers whose difference is two) less than a given integer. This exercise uses Python’s **filter()** function.
- A histogram that contains counts of the number of primes in an array whose least significant digit is 1, 3, 7, and 9. (There is only one prime number whose least significant digit is 2 and only one whose least significant digit is 5.)
- The totients of all positive integers less than a given integer. The totient of a positive integer N is the number of positive integers less than N that have no factor in common with N except for one. The algorithm for computing totients computes a product of differences. The function’s arguments are the integer N and an array of prime numbers less than or equal to N. This example shows students how to connect functions that produce and consume NumPy arrays.

The notebook reads data from an image file into a NumPy array. The image is a gray scale image. Using vector arithmetic, the code produces a negative image, images with fewer shades of gray (“posterized” images), and images with false colors. The same methods work for visualizing other kinds of two dimensional data with heat maps and contour maps.

The notebook generates and plots points on a line in the plane. It also generates points that lie near the line to simulate observations that contain noise, and then uses the method of least squares to find the line that best bits the noisy data. This final exercise prepares students for later experiments. They will write code to create small, synthetic datasets for the purpose of testing machine learning functions before applying those functions to larger, real datasets.

5 The value of Jupyter notebooks

A Jupyter notebook can contain formatted text, mathematical notation, images, hyperlinks, and executable code. An author of a notebook can describe a method for solving a problem with words, equations, and a computer program. Multiple views of a method for solving a problem enable readers to more easily understand the method. Readers can execute the code with different parameters. They can annotate an author’s explanations. Jupyter notebooks invite active learning.

The author did not attempt to teach his students all features of the Matplotlib, NumPy, and Pandas libraries (an impossibility, of course!). The author did not even attempt to identify the must important features. Instead, the author created a notebook that shows his students a sampling of features. It shows them a style of programming that will be unfamiliar to most. It invites them to try different ways of solving problems, and to adapt their own habits. Working with the notebook, students learn by doing. They learn how to build projects by altering a template or example in small steps until the product is their own.

References

View and download the author’s Jupyter notebook at bitbucket.org/leontabak/catapult.

Cyberbullying Classification Using Three Deep Learning models: GPT, BERT, and RoBERTa

Muhammad Abusaqer and Charles Fofie Jr

Department of Math and Computer Science

Minot State University

Minot, ND, USA

muhammad.abusaqer@minotstateu.edu; charles.fofiejr@minotstateu.edu

Abstract

This research paper presents a study on the classification of cyberbullying on social media feeds using three deep learning algorithms of GPT -3, BERT, and RoBERTa. Cyberbullying is a growing concern in social media, so it is crucial to develop systems for detecting and preventing it. Cyberbullying involves using technology to harass, threaten, embarrass, or target individuals based on age, gender, religion, etc. This paper proposes a system that leverages both machine learning and deep learning algorithms to detect cyberbullying and reduce its impact, particularly on teen suicides. The study trains the deep learning models on a dataset of 46,692 tweets.

Additionally, the study compares the performance of these deep learning models to traditional machine learning algorithms, including Support Vector Machines (SVM), Naive Bayes, and Decision Trees. The study results demonstrate that the deep learning models outperform the traditional machine learning algorithms in detecting cyberbullying. This study makes two contributions to the field. Firstly, it is one of the first studies to use the newly released deep learning models of GPT 3.0 from Open AI, BERT from Google, and RoBERTa from Facebook AI. Secondly, it supplies a performance comparison between these state-of-the-art deep learning models and traditional machine learning algorithms. The results of this study could also help develop tools to assist in monitoring social media for cyberbullying feeds and immediately deleting them.

Survey of Application of Machine Learning Methods in The Development of Network Intrusion Detection and Prevention Systems

Juliana Nkafu

Juliana.nkafu@gmail.com

Jun Liu

jun.liu@und.edu

School of Electrical Engineering and Computer Science
College of Engineering and Mines
University of North Dakota
Grand Forks, 58202

Abstract

Attacks targeting networks are increasing over time as Internet technology is widely adopted to foster communication in a plethora of professional and personal tasks. Attacks seek to damage and disrupt the integrity and confidentiality of connections and information exchanges. The ever-growing threats of cyber-attacks demand the urgency of developing robust security defense systems to protect business and client data. The primary goals of network defense systems are to identify, defend, and recover from network assaults. The core of network defense systems is the collective techniques for detecting and mitigating network intrusion. Network defense systems can be categorized into network intrusion detection systems (NIDS) and network intrusion prevention systems (NIPS). Network intrusion detection and prevention techniques can be categorized based on the approaches of detect network threats, the approaches of mitigating the threats, or a combination of both. The research and development on network intrusion detection and prevention techniques highly relies on the availability of representative security-related network datasets. Benchmark datasets are a good basis to evaluate and compare the quality of different network intrusion detection systems. Benchmark datasets with labeled data points of “normal” or “attach” serves as the important input to evaluate the quality of intrusion detection techniques to distinguish correctly detected attacks from false alarms. ML and DL have been used to improve IDS detection accuracy and reduce false positives. Our paper gives a survey of the application of machine learning and deep learning techniques in network intrusion detection, together with a list of available cybersecurity datasets used for model training. Network datasets labeled with information about malicious events are.

Keywords: *Network defense system, Intrusion Detection Systems, Intrusion Prevention Systems, Security-related network datasets, Machine Learning, Deep Learning*

1 Introduction

In recent year, cybercriminals launched a wave of cyberattacks that were not only highly coordinated, but also far more sophisticated than ever before. The emerging cloud evolution technologies have brought remarkable evolutions in network technology where different applications, services, and computing and storage resources are offered on demand to many users via the internet. Such an exponential growth in network technologies has offered many advantages and has improved communications. Although the Internet facilitates connection and communication, the integrity and confidentiality of these connections and information exchanges can be violated and compromised by attackers seeking to damage and disrupt network connections and network security. Each emerging network technology presents new security challenges and triggers the need for the development of detection tools and countermeasures to meet new demands.

Network attacks have become more sophisticated, and the foremost challenge is to identify unknown and obfuscated attacks as these authors use different evasion techniques for information concealing to prevent detection by an IDS. In the past, cybercriminals primarily focused on bank customers, robbing bank accounts, or stealing credit cards (Symantec, 2017). The new generation of attackers has become more ambitious and is targeting the banks themselves, sometimes trying to take millions of dollars in a single attack (Symantec, 2017). According to the purplesec.us report; on average, a malware attack costs a company over $2.5 million (including the time needed to resolve the attack). Individuals of phishing scams lost $225 on average. High profile incidents of cybercrime have demonstrated the ease with which cyber threats can spread internationally, as a simple compromise can disrupt a business' essential service or facilities. Many cybercriminals around the world are motivated to steal information, illegitimately receive revenues, and find new targets. For this reason, the detection of zero-day attacks has become the highest priority.

Several techniques for handling and classifying network traffic attacks have been proposed over the years. One approach is port-based, which involves identifying port numbers among those registered with the Internet Assign Number Authority (IANA). However, as the number of applications has grown, so has the number of unpredictable ports, and this technique has proven to be ineffective. This technique excludes account applications that do not register their ports with the IANA and use dynamic port numbers. Another technique proposed is the payload-based technique, also known as deep packet inspection (DPI), in which the contents of network packets are observed and compared to an existing set of signatures stored in a database. This method is more accurate than the port-based technique, but it does not work with network applications that use encrypted data. Behavioral classification techniques examine all network traffic received by the host to determine the type of application. The Network traffic patterns can be analyzed graphically as well as by looking at heuristic data such as transport layer protocols and the number of distinct ports contacted. Although behavioral techniques produce good results by detecting unknown threats, they are resource intensive and prone to false positives. Another technique, known as the rationale-based or statistical technique, looks at the statistical characteristics of traffic flow, such as the number of packets and the maximum, mean, and minimum packet size. Because these measurements are unique to each application, these statistical characteristics are used to identify different applications. However, there is an increasing need to combine this approach with techniques that can improve accuracy and speed up the classification of statistical patterns. Correlation-based classification groups packets into flows, or groups data packets with the same source and destination IP, port, and protocol. These are classified based on the correlation of network flows. Multiple flows are typically combined into a Bag of Flows (BoF). Although this technique outperforms statistical techniques because it eliminates feature redundancy, it has a high computational overhead for feature matching. As a result, the need to develop techniques to overcome the rising challenges persists.

The concepts of intelligent techniques, namely machine learning (ML) and deep learning (DL), became popular at the beginning of the twenty-first century. Researchers widely agreed that these techniques, which focus on using statistical methods and data to make computers think like humans, could increase the calculation potential. To address the limitations of non-intelligent techniques, computer scientists began to use intelligent techniques in network security. In network security, ML or DL algorithms can be trained on network data to distinguish between normal and malicious traffic and thus protect the network from intruders. Furthermore, if the network traffic is malicious, algorithms can be trained to identify the type of attack and take appropriate action to prevent the attack. The model can be taught to prepare individual defensive reactions by analyzing previous cyber-attacks.

This paper focuses on surveying the intelligent methods in network security can be useful in large businesses, organizations, law enforcement agencies, and banks that store sensitive information, as well as in personal networks. There are three significant contributions made by this article. (i) We did a systematic analysis to choose recent journal publications on various ML- and DL-based NIDS published during the last five years (2019-April 2022). (ii) We conducted a comprehensive examination of each publication and discussed its distinct characteristics, including its proposed methodology, evaluation criteria, network attack types, and datasets used. (iii) We did a review on the Network Intrusion and Prevention systems. (iv) On the basis of these observations, we presented recent trends in the use of AI approaches for NIDS, emphasized significant problems in ML-/DL-based NIDS, and outlined a variety of future directions in this crucial sector.

2 Network Attacks

For decades, networking technologies have been used to improve data transfer and circulation. Their continuous improvement has facilitated a wide range of new services. The utilization of mobiles is turning out to be an important component in our everyday life. An ever-increasing number of clients across the world rely upon their mobiles to trade messages, manage their personal documents, browse their emails. Additionally, mobiles facilitate online shopping which provokes clients to type their credit card numbers, security codes, usernames, and passwords. The goal of computer network defense (CND) is to prevent network intrusions that could lead to service/network denial, degradation, or disruptions by employing a variety of processes and defensive mechanisms that rely on computers and the internet. The smart city is one of the fastest growing fields. The fundamental goal of any smart city is improving the citizen's quality of life by offering a direct association to the administering body and providing better management to traffic, water, energy, air, waste, and more. Because of the variety of components in smart cities, security issues have become a significant concern.

A network attack is an approach to hurt, reveal, change, destroy, steal, or obtain illegal access to a network system resource. The attack could come from inside (internal attack) or from outside (external attack).

Existing review articles e.g., such as (Buczak & Guven, 2016; Axelsson, 2000; Ahmed et al., 2016; Lunt, 1988; Agrawal & Agrawal, 2015)) focus on intrusion detection techniques or dataset issue or type of computer attack and IDS evasion. The highly cited survey by Debar et al. (Debar et al., 2000) surveyed detection methods based on the behavior and knowledge profiles of the attacks. The major types of attacks can be categorized in the following list.

- **DDoS Attacks**

  These are attacks that attempt to disrupt the availability of service. Since distributed denial of service is easy to launch but not easy to detect, as in most cases the attacks traffic is very similar to legitimate traffic. DDoS attacks often originate from multiple sources, making them difficult to mitigate. To flood the target with traffic, the attacker uses "zombies"—compromised computers. HTTP requests, fake packets, and junk data can be this traffic. (Baek et al, 2019) conducted a study providing a model for assessing and identifying DDoS assaults on the network-level and service-levels of the bitcoin ecosystem. The dataset comprised of authentic DDoS attacks and included the impacted service, attack date, service type, number of postings, etc. The researchers collected statistical data such as maximum, minimum, total, and standard deviation from the Bitcoin block data. The researchers utilized PCA to extract features. MLP was used to identify DDoS, and the training set, validation set, and testing set were split 6:2:2 respectively.

- **Insider Threats**

  The term "insider threat" refers to a security risk posed to an organization by insiders having access to sensitive information, systems, or assets, such as employees, contractors, or business partners. This type of danger can manifest in a variety of ways, including inadvertent data breaches, malicious attacks, the theft of sensitive information, and the introduction of malware into the organization's systems. Because insider threats frequently include persons with high degrees of access and trust, it is easier for them to circumvent security measures and do damage. (Yuan et al, 2018). [12] utilized LSTM and CNN approaches to develop a model for detecting insider threats. They used the model to the CERT insider threat v4.2 dataset [13], which consisted of 32 M log lines, of which 7323 represented unusual activity. This edition of the CERT dataset contains a greater number of examples of insider threats than previous versions. The train-to-test ratio was 70% to 30%. The researchers initially extracted user behavior using LSTM, then extracted temporal characteristics and generated feature vectors. The researchers then turned the feature vectors into matrices of fixed size.

- **Phishing Attacks**

  Phishing assaults are a type of social engineering attack that seeks to acquire sensitive information, such as login credentials or financial information, by convincing victims they are talking with a reputable source, such as a bank or online service provider. These assaults are frequently delivered via email, text message, or phone call and may look to originate from a reputable entity, such as a bank or online business, requesting sensitive information or login credentials. The attacker may also include a link in the message that links to a phishing website that appears legitimate and requests sensitive information. (Mohammad et al,) built a self-structuring neural network based on ANN to recognize phishing website attacks. Phishing-related traits are essential for detecting highly dynamic web sites; hence, the network's architecture must be continuously enhanced. The suggested method solves this issue by automating the process of network architecture and displaying a high tolerance for noisy input, fault tolerance, and significant prediction accuracy. This was accomplished by accelerating the learning rate and adding neurons to the hidden layer. The objective of the constructed model was to achieve generalization ability, which necessitates that the classification accuracy throughout training and testing be as comparable as possible.

- **Malware**

  Malware is software that harms or exploits a computer system or network. Viruses, worms, trojan horses, ransomware, spyware, and adware are malware. Viruses spread by attaching themselves to emails or other files. Unlike viruses, worms replicate without attaching to files. Trojan horses are malware that masquerade as harmless software and fool users into downloading and installing them. Ransomware encrypts files and demands payment to decrypt them. Spyware steals passwords and login credentials from victims' computers. Adware shows unwelcome ads on victims' computers. Using the RF technique and the Kyoto 2006+ (Song et al, 2011) dataset, (Park et al, 2018) examined the recognition performance of several forms of attacks, including IDS, malware, and shellcode (total size 19.8 GB). The dataset contained three types of class: attack, shellcode, and normal.

- **Zero-Day Attacks**

  A zero-day attacks is a sort of cyberattack that exploits a previously unknown software application or operating system vulnerability. The phrase "zero-day" refers to the fact that the vulnerability has not been identified or revealed to the public; hence, the producer of the program has had zero days to patch the weakness and prevent it from being exploited. Zero-day attacks are especially perilous because they exploit unpatched vulnerabilities, allowing attackers to infiltrate a target's systems and steal sensitive data or install malware. These attacks can be launched by nation-states, criminal groups, or individual hackers for several goals, including cyber espionage, sensitive data theft, and financial gain. Several researchers have, surprisingly, focused on discovering zero-day attacks. (Beaver et al, 2013) conducted one such investigation using machine learning techniques that can distinguish between normal and malicious communications.

3 Network Intrusion Detection and Prevention Systems

Network security has recently received an enormous attention due to the mounting security concerns in today’s networks. computer security has become essential as the use of information technology has become part of our daily lives. Undoubtedly, IoT devices are vulnerable to various security attacks. There is a serious requirement for IDSs to secure IoT gadgets against security vulnerabilities.

3.1 Intrusion Detection System (IDS)

An IDS intensely monitors malicious network activities and notifies officials if an attack is detected with no prevention abilities. Signature-based and anomaly-based detection are the two most prevalent approaches used by IDS to identify threats. Typically, IDS works in three steps: monitoring, detecting, and warning. Firstly, it monitors the network traffic or the system. Secondly, it analyzes and identifies the pattern of connections and intrusion behaviors accordingly to the characteristics of the used algorithm. Finally, it generates an alarm immediately when detecting a suspicious activity for investigation. On the other hand, anomaly-based procedures attempt to differentiate malicious traffic from real traffic based on a change in the network traffic; thus, they can detect unknown threats. On the other hand, anomaly-based procedures attempt to differentiate malicious traffic from real traffic based on a change in the network traffic; thus, they can detect unknown threats.

- **Signature-based intrusion detection systems (SIDS)**
  Signature intrusion detection systems (SIDS) are based on pattern matching techniques to find a known attack; these are also known as Knowledge-based Detection or Misuse Detection. matching methods are used to find a previous intrusion. In other words, when an intrusion signature matches with the signature of a previous intrusion that already exists in the signature database, an alarm signal is triggered. SIDS usually gives an excellent detection accuracy for previously known intrusions. However, SIDS has difficulty in detecting zero-day attacks because no matching signature exists in the database until the signature of the new attack is extracted and stored. As a result of having to establish a new signature for each alteration, the efficiency of signature-based systems is drastically reduced. Moreso, the increasing rate of zero-day attacks has rendered SIDS techniques progressively less effective because no prior signature exists for any such attacks (Symantec, 2017).

- **Anomaly-based intrusion detection system (AIDS)**
  In AIDS, a normal model of the behavior of a computer system is created using machine learning, statistical-based or knowledge-based methods. Any significant deviation between the observed behavior and the model is regarded as an anomaly, which can be interpreted as an intrusion. The classification is based on heuristics or rules, rather than patterns or signatures. This category of strategies assumes harmful activity is different from user behavior; intrusions are anomalous user behavior. AIDS development involves training and testing. The training phase uses the typical traffic profile to create a model of normal behavior, and the testing phase uses a new data set to assess the system's ability to generalize new intrusions. AIDS can detect zero-day attacks without a signature database by analyzing abnormal user behavior (Alazab et al., 2012). Hence, AIDS has advantages. They can initially detect organizational malfeasances; An alarm is triggered whenever an intruder makes suspicious transactions in a compromised account. Second, because the system uses individualized profiles, cybercriminals can't detect routine behavior without triggering an alert.

- **Host-based IDS (HIDS)**
  A host-based intrusion detection system (HIDS) is a security solution that monitors and analyzes activity on individual computer systems or hosts in search of indicators of malicious behavior or unauthorized access. Unlike network-based intrusion detection systems (NIDS), which monitor

network traffic, host-based intrusion detection systems (HIDS) operate at the host level and are capable of detecting both internal and external threats. Comparing the present state of a host to its baseline or expected state and searching for deviations or anomalies that may signal an intrusion or breach is how HIDS software normally operates. Unauthorized changes to system files, attempts to access privileged resources or data, and the installation or execution of malicious software are examples of activities that may generate a HIDS warning. HIDS can detect insider attacks that do not involve network traffic. HIDS can provide an additional layer of defense against a wide variety of cyber threats, hence enhancing the security posture of individual hosts or systems.

- **Network-based IDS (NIDS)**
  A network-based intrusion detection system (NIDS) is a security solution that analyzes network data for indications of malicious or suspicious behavior. Unlike host-based intrusion detection systems (HIDS), which focus on specific hosts or systems, NIDS functions at the network level, examining network traffic in search of patterns or behaviors that may indicate a security issue. NIDS typically operate by recording network data in real-time and evaluating it for indicators of malicious activity or policy violations. This may involve recognizing known attack signatures, examining traffic patterns for anomalies, or employing machine learning techniques to discover patterns that may signal an attack. NIDS can be deployed as a standalone device or as a component of a larger network security architecture. Certain NIDS solutions are intended for integration with other security technologies, such as firewalls and intrusion prevention systems, to provide a more comprehensive protection against cyber-attacks.

3.2 Intrusion Prevention System (IPS)

The Intrusion Prevention System, often known as intrusion detection and prevention systems, is abbreviated as "IPS" (IDPS). It does a continual search across the network to identify any unauthorized or rogue control points that may be present. These points are identified based on changes in behavior. The system will automatically take preventative actions to deal with the dangers and protect itself from further damage. The protection of a network against harmful or unwanted packets and assaults is the primary purpose of an intrusion detection and prevention system (IDPS). IDPS is more effective than IDS because in addition to detecting risks, it is also able to respond appropriately to such threats. There are two different kinds of intrusion detection and prevention systems (IDPS): network-based intrusion detection and prevention systems (NIDPS), which examine the network protocol Sensors 2021, 21, 7070 6 of 43 to identify any suspicious activities; and host-based intrusion detection and prevention systems (HIDPS), which are utilized to monitor host activities for any suspicious events that may occur within the host. ML or DL based intelligent techniques ML and DL Can be adopted for effectively and efficiently identify network attacks.

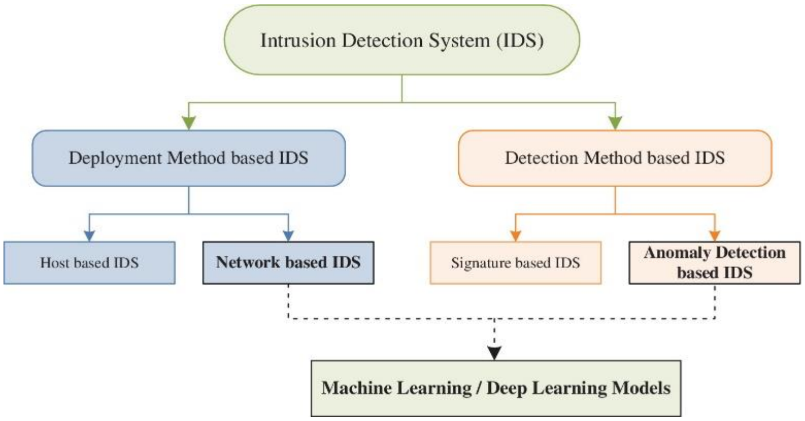

Intrusion Detection System (IDS)
Deployment Method based IDS
Detection Method based IDS
Host based IDS
Network based IDS
Signature based IDS
Anomaly Detection based IDS
Machine Learning / Deep Learning Models

Figure 1: Network Intrusion Detection Systems Classification Taxonomy.

4 Applying ML Techniques in the Design of NDS and IDS

People are always looking for methods, tools, or techniques that reduce the amount of effort required to perform a task efficiently. In Machine Learning, algorithms are programmed to attempt to self-learn based on their past experiences. After gaining knowledge from previous experiences, algorithms become quite capable of reacting and responding to conditions for which they were not explicitly programmed. Consequently, Machine Learning contributes significantly to intrusion detection. It attempts to recognize previously unrecognized hidden patterns that aid in intrusion detection. Machine Learning approaches for intrusion detection are so widespread and popular.

- **Improved Accuracy**
  The algorithms that comprise machine learning can examine huge volumes of data, recognize patterns, and learn from these observations. This can lead to more accurate intrusion detection, with fewer false positives and false negatives because of the process. It quickly examines and processes data to extract new patterns from it. For humans to analyze the data will require a substantial amount of time, which will increase proportionally to the quantity of data. Rule-based intrusion detection systems rely on established rules to determine which types of actions are regarded safe and which should raise a red flag. Rule-based method is inefficient due to the time-consuming nature of writing these rules for various instances. Machine Learning-based Intrusion Detection algorithms succeed in learning from existing patterns and can automatically recognize new patterns. And it accomplishes all of this in a fraction of the time required by rule-based systems.

- **Adaptability**
  Algorithms that are learned through machine learning can adjust to new kinds of threats and changing settings. The ability of machine learning models to identify new threats can progress in tandem with the development of novel intrusion techniques by cybercriminals. In addition, machine learning models are continually updated and retrained to accommodate the emergence of new types of assaults and the evolution of threat landscapes.

- **Scalability**

  Because machine learning models can process massive amounts of data, these models are scalable and therefore suitable for use in enterprise-level intrusion detection systems. This may be of particular benefit to companies that have a significant number of endpoints and network devices. The goal of intrusion detection in network security is to identify and respond to potential security breaches in real-time. Their ability to quickly and efficiently process and analyze large amounts of data makes them well-suited to handling the high volume of network traffic that is typical in modern networks.

4.1 Intrusion Data Sources

Intrusion data sources can be categorized by the methodologies used to detect intrusions (signature-based or anomaly-based) or by the input data sources used to detect anomalous behaviors (host-based or network-based). These methods use software, hardware, or both.

Datasets can test new methods. The dataset's amount and quality affect an IDS's performance. Packet-based, flow-based, or other formats can collect network traffic. Common features as evaluation bases assist researchers find suitable data sets for their evaluation scenarios. Network and transport protocols evaluate packet-based data. TCP, UDP, ICMP, and IP are the main protocols. Pcaps of these protocols include the payload. TCP, a reliable transport protocol, uses metadata including sequence numbers, acknowledgement numbers, TCP flags, and checksum values. Flow-based data have no payload and are created by collecting all packets that arrive within a defined time frame that share specific properties into a single flow. It contains the first view date, duration, and transport protocols and is used to match flow-based data attributes. Flows may be unidirectional or bidirectional. Netflow, IP-FIX, sFlow, and OpenFlow are flow-based formats (Mckeown et al, 2008). Other data includes packet- and flow-free data collections. This category includes flow-based data sets supplemented with packet-based or host-based log files.

Table 1: Overview of cybersecurity datasets

Dataset Name	Year	Attack types	Observation
DARPA	1998/99	Dos, R2L, U2R, Probe	Include irregular distribution of attack data instances Do not represent real network traffic
KDD Cup 99	1999	Dos, R2L, U2R, Probe	Suffer from redundant records and duplicate data samples.
Kyoto 2006+	2006/2008	DoS, Probe, U2R, R2L	While the Kyoto 2006+ dataset includes a diverse range of attack types, there may be other types of attacks that are not represented in the dataset, such as advanced persistent threats (APTs) or insider attacks.
NSL-KDD	2009	Dos, R2L, U2R, Probe	Lack of redundant records Limited number of attack types

UNSW-NB15	2015	Backdoors, portscans, DoS, Exploits, Spam, Reconnaissance, fuzzers, generic, Shellcode,Worms	Have list of new attacks and updated continuously.
CIC-IDS2017	2017	Brute force, Dos, DDoS, Portscan, Web, Botnet, Infiltration	contain some redundant data records Include attacks that resembles the real-world data
CSE-CIC-IDS 2018	2018	Brute force, Dos, DDoS, Portscan, Web, Botnet, Infiltration	Generate the dataset with the help of network profiles List a new scope of attacks produced from real network traffic
DARPA AI Next	2019	Malware infections, C2, Data exfiltration	The dataset was created with the specific goal of testing the performance of AI systems, and as such, it may not include all types of cyber threats or operations.
IoT-23	2020	Benign and malicious traffic	Have list of new attacks and updated continuously.

4.2 Challenges of Applying Machine Learning methods in Designing IDS

Since the accuracy of ML methods depends on the quality of data, it is crucial to provide appropriate datasets for training and testing phases. There's no doubt that ML methods have significantly improved the IDS landscape by rapidly identifying and frustrating attacks. However, Because of the continuous growing sophistication of the threats, ML remains incapable to keep up with this flow of threats due to some reasons outlined below

- **Lack of sufficient data**

  It is necessary to offer suitable datasets for the training and testing phases, as the precision of machine learning techniques is directly proportional to the quality of the data. Due to a lack of relevant datasets, it is impossible to conduct a security threat assessment or develop an efficient defense plan. Older training datasets are a significant challenge for any approach because they result in a detection performance that is only moderate when applied to fresh dangers. In addition to this, one of the challenges that machine learning specialists need to overcome is an imbalanced dataset. When a dataset has an uneven distribution of classes, for example when the ratio of harmful to benign samples is 2 to 40, we say that the dataset is imbalanced.

- **Labeled Sample Shortages**

  Because of their low cost and the ease with which they can be trained and put into practice; supervised learning methods are the most common type of machine learning approach utilized for intrusion detection systems (IDS). Yet, the most significant drawback of this machine learning category is that it cannot be trained without labeled examples. Unfortunately, there are not a lot of datasets that have labels, and the manual development of these labels is a time-consuming and expensive operation. It is widespread practice to use external whitelists and blacklists when labeling products, although the accuracy of these lists cannot be guaranteed, and they may also be of inferior quality.

- **Approaches of Attacks**
  Due to the strategic nature of the attacks and the fact that they are always adjusting their techniques, the application of machine learning is made more difficult. The defender makes repeated efforts to guard his holdings using all the resources at his disposal. So, in most situations, he sets up numerous layers of defensive systems and waits for some indication of an attack before acting. While it is true that most of the time, the attacker is aware of his objective as well as the type of defense that needs to be breached. He also has the advantage of knowing the exact time as well as the measures that will be taken during the assault. In addition, those who carry out attacks are continually developing new methods and making use of the most recent technological advancements. This includes both artificial intelligence and machine learning. To emerge victorious from the conflict, the defender needs to go above and beyond the simple duty of warding off attacks and instead focus on rendering them impossible

4.3 Different ML Techniques Applied in the Design of IDS

Machine learning is the process of extracting knowledge from large quantities of data. Machine learning models comprise of a set of rules, methods, or complex "transfer functions" that can be applied to find interesting data patterns, or to recognize or predict behavior (Dua & Du, 2016). This learning could either be supervised, unsupervised, semi-supervised, ensembled or hybrid. The goal of using machine learning techniques is to create IDS with improved accuracy and less requirement for human knowledge, which are some of the reasons why machine learning is popular these days.

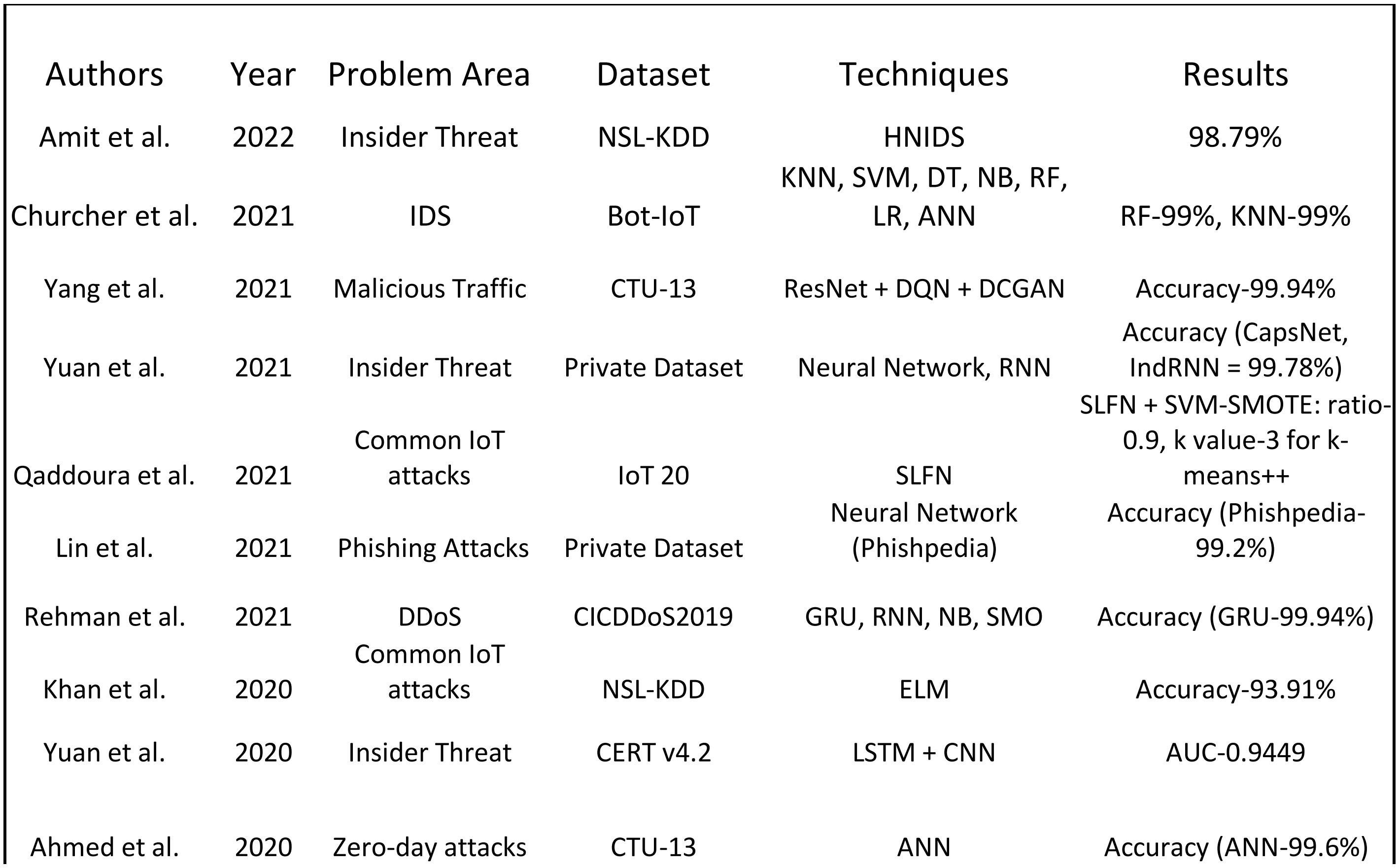
Supervised learning-based IDS techniques detect intrusions by using labeled training data. This approach usually consists of two stages, namely training and testing. In the training stage, relevant features and classes are identified and then the algorithm learns from these data samples. Each record is a pair, containing a network or host data source and an associated output value (i.e., label), namely intrusion or normal. Next, feature selection can be applied for eliminating unnecessary features. Using the training data for selected features, a supervised learning technique is then used to train a classifier to learn the inherent relationship that exists between the input data and the labelled output value. In the testing stage, the trained model is used to classify the unknown data into intrusion or normal class. The resultant classifier then becomes a model which, given a set of feature values, predicts the class to which the input data might belong. The performance of a classifier in its ability to predict the correct class is measured in terms of several metrics.
There are many classification methods such as decision trees, rule-based systems, neural networks, support vector machines, naïve Bayes, and nearest-neighbor. Each technique uses a learning method to build a classification model.

Unsupervised learning is a form of machine learning technique used to obtain interesting information from input datasets without class labels. The input data points are normally treated as a set of random variables. A joint density model is then created for the data set. In supervised learning, the output labels are given and used to train the machine to get the required results for an unseen data point, while in unsupervised learning, no labels are given, and instead the data is grouped automatically into various classes through the learning process. In the context of developing an IDS, unsupervised learning means, use of a mechanism to identify intrusions by using unlabeled data to a train the model.

Semi-supervised learning falls between supervised learning (with totally labelled training data) and unsupervised learning (without any categorized training data). Researchers have shown that semi-supervised learning could be used in conjunction with a small amount of labelled data classifier's performance for the IDSs with less time and costs needed. This is valuable as for many IDS issues, labelled data can be rare or occasional (Ashfaq et al., 2017). Several different techniques for semi-supervised learning have been proposed, such as the Expectation Maximization (EM) based algorithms (Goldstein, 2012), self-training (Blount et al., 2011; Lyngdoh et al., 2018), co-training (Rath et al., 2017), Semi-Supervised SVM (Ashfaq et al., 2017).

Combining machine learning techniques improves predicted performance. Boosting, Bagging, and Stacking are ensemble approaches. Conventional IDSs cannot be changed, cannot recognize new malicious threats, have low accuracy, and high false alarms. AIDS's false-positive rate. SIDS and AIDS form hybrid IDS. Hybrid IDS overcomes SIDS and AIDS. (Farid et al, 2010) suggested a hybrid IDS employing Naive Bayes and decision trees to detect 99.63% of KDD'99 datasets.

Table 2. Summaries of reviewed papers.



Authors	Year	Problem Area	Dataset	Techniques	Results
Amit et al.	2022	Insider Threat	NSL-KDD	HNIDS	98.79%
Churcher et al.	2021	IDS	Bot-IoT	KNN, SVM, DT, NB, RF, LR, ANN	RF-99%, KNN-99%
Yang et al.	2021	Malicious Traffic	CTU-13	ResNet + DQN + DCGAN	Accuracy-99.94%
Yuan et al.	2021	Insider Threat	Private Dataset	Neural Network, RNN	Accuracy (CapsNet, IndRNN = 99.78%)
Qaddoura et al.	2021	Common IoT attacks	IoT 20	SLFN	SLFN + SVM-SMOTE: ratio-0.9, k value-3 for k-means++
Lin et al.	2021	Phishing Attacks	Private Dataset	Neural Network (Phishpedia)	Accuracy (Phishpedia-99.2%)
Rehman et al.	2021	DDoS	CICDDoS2019	GRU, RNN, NB, SMO	Accuracy (GRU-99.94%)
Khan et al.	2020	Common IoT attacks	NSL-KDD	ELM	Accuracy-93.91%
Yuan et al.	2020	Insider Threat	CERT v4.2	LSTM + CNN	AUC-0.9449
Ahmed et al.	2020	Zero-day attacks	CTU-13	ANN	Accuracy (ANN-99.6%)

Letteri et al.	2020	Malware Attack	MTA KDD 19	MLP using AE optimization or RRw optimization	Accuracy (MLP with RRw opt.-99.60%)
Kim et al.	2020	DDoS	KDD-99, CICIDS2018	CNN, RNN	Accuracy (CNN-99% or more)
Alrashdi et al.	2019	Common IoT attacks	UNSW-NB15	RF	Accuracy (ML-99.34%)
Zhang et al.	2019	IDS	NSL-KDD	AE	F-Score-76.47% Recall-79.47%
Hu et al	2019	Insider Threat	Private Dataset	CNN	FAR-2.94% FRR-2.28%
Pektas et al.	2019	Botnet Attacks	ISOT HTTP, CTU-13	MLP + LSTM	ISOT: F score-98.8% CTU: F score-99.1%
Nguyen et al.	2018	IDS	UNSW-NB15, KDD-99, NSL-KDD	NNET	Accuracy (KDD-99-97.11%)

5. Conclusion

Network security is a major concern for individuals, businesses, and governments. With the current digital explosion, network security is essential to secure society's acceptance of the tens of thousands of services that rely on the network, the backbone of digital life. Hence, network security is essential. This study reviews IDS machine learning classification techniques. Researchers have used SVM, Nave Bayes, Neural Network, Gradient Boosted Tree, Decision Tree, k-nearest neighbors, multinomial randomness, forest classifier, stochastic gradient descent, and ensemble classifiers. SVM, Random Forest, and CNN can identify with high accuracy. We examined the most popular public datasets for IDS research, their data gathering methods, evaluation findings, and constraints. Newer and more complete malware activity datasets are needed because normal activities vary frequently and may lose effectiveness over time. DARPA/KDD99 does not include new malware operations. As these 1999 datasets are publicly available and no other appropriate datasets exist, testing is limited to them. These benchmarks no longer represent modern zero-day attacks. Combination techniques will be tested on the same dataset to improve detection and reduce false positives.

References

[Symantec, 2017] Symantec, "Internet security threat report 2017," April, 7017 2017, vol. 22.

[Goli et al., 2018] Y. D. Goli, R. Ambika, Network Traffic Classification Techniques-A Review. In Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018, Belgaum, India, 21–22 December 2018; pp. 219–222.

[Buczak et al., 2016] A. Buczak, E.Guven, (2016) A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials 18(2):1153–1176.

[Park et al, 2018] K. Park, Y. Song, Y. G. Cheong, Classification of attack types for intrusion detection systems using a machine learning algorithm. In Proceedings of the 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), Bamberg, Germany, 26–29 March 2018.

[Song et al, 2011] J. Song, H. Takakura, Y. Okabe, M. Eto, Inoue, K. Nakao, Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In Proceedings of the 1st Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, BADGERS 2011, Salzburg, Austria, 10 April 2011.

[Beaver et al, 2013] J. M. Beaver, C. T. Siymons, R. E. Gillen, A learning system for discriminating variants of malicious network traffic. In Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop, Oak Ridge, TN, USA, 8–10 January 2013.

[Baek et al, 2019] J. U. Baek, S. H. Ji, J. T. Park, M. S. Kim, DDoS Attack Detection on Bitcoin Ecosystem using Deep-Learning. In Proceedings of the 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS), Matsue, Japan, 18–20 September 2019.

[Mohammad et al, 2014] R. M. Mohammad, F. Thabtah, L. McCluskey, Predicting phishing websites based on self-structuring neural network. Neural Comput. Appl. 2014, 25, 443–458.

[Yuan et al, 2018] F. Yuan, Y. Cao, Y. Shang, Y. Liu, J. Tan, B. Fang, Threat Detection with Deep Neural Network. In Computational Science—ICCS 2018; Springer: Cham, Switzerland, 2018.

[Debar et al, 2000] Dacier, M Dacier Deber, and A. Wespi, "A revised taxonomy for intrusiondetection systems," in Annales des télécommunications, 2000, vol. 55, no. 7–8, pp. 361–378: Springer.

[Alazab et al, 2012] Hobbs M. Alazab, J. Abawajy, and M. Alazab, "Using feature selection for intrusion detection system," in 2012 international symposium on communications and information technologies (ISCIT), 2012, pp. 296–301.

[Claise, 2012] Cisco Systems NetFlow Services Export Version 9. Internet Engineering Task Force 2004. doi:10.17487/RFC3954.

[Claise, 2012] Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. Internet Engineering Task Force 2008.doi:10.17487/RFC5101.

[McKeown et al, 2016] Anderson N, McKeown, T. Balakrishnan H.Parulkar G. Peterson L. Rexford J. Shenker S. Turner J, OpenFlow: enabling innovation in campus networks. ACM SIGCOMM Comput Commun Rev 2008;38(2):69–74. doi:10.1145/1355734.1355746.

Dua and X. Du, Data mining and machine learning in cybersecurity. CRC press, 2016

[Ashfaq et al, 2017] R. Ashfaq, X-Z. Wang, JZ. Huang, H. Abbas, Y-L. (2017) Fuzziness based semisupervised learning approach for intrusion detection system. Inf Sci 378:484–497.

[Goldstein, 2012] Goldstein, "FastLOF: an expectation-maximization based local outlier detection algorithm," in Pattern recognition (ICPR), 2012 21st international conference on, 2012, pp. 2282–2285: IEEE.

[Lyngdoh et al, 2018] M.Lyndoh, I. Hussain, S. Majaw, and H. K. Kalita, "An intrusion detection method using artificial immune system approach," in international conference on advanced informatics for computing research, 2018, pp. 379–387: Springer.

[Rath et al, 2017] PS. Rath, NK. Barpanda, R. Singh, S. Panda (2017) A prototype Multiview approach for reduction of false alarm rate in network intrusion detection system. Int J Comput Netw Commun Secur 5(3):49.

[Farid et al, 2010] Harbi N.Farid, M. Z. Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection," arXiv preprint arXiv:1005.4496, 2010.

[Amit et al, 2022] Kumar B. Amit, Sachin Ahunja, Kumar L. Umesh, Sanjeev K. Sharma, Poongodi Manoharan., Abeer D. Algarni, Hela Elmannai, Kaamran Raahemifar, A hybrid intrusion detection model using EGA-PSO and improved random forest method. Sensor 2022, 22(16), 5986.

[Churcher et al, 2021] A Churcher, R. Ullah, J. Ahmad, Ur S. Rehman, F Masood, M. Gogate, F. Alqahtani, B. Nour, W.J. Buchanan, an experimental analysis of attack classification using machine learning in IoT networks. Sensors 2021, 21, 446.

[Yang et al, 2021] J. Yang, G. Liang, B. Li, G. Wen, T. Gao, A deep-learning- and reinforcement-learning-based system for encrypted network malicious traffic detection. Electron. Lett. 2021, 57, 363–365.

[Yuan et al, 2021] J. Yuan, G. Chen, S. Tian, X. Pei, Malicious URL detection based on a parallel neural joint model. IEEE Access 2021, 9, 9464–9472.

[Qaddoura et al, 2021] R. Qaddoura, A.M. Al-Zoubi, I. Almomani, H. Faris, A multi-stage classification approach for iot intrusion detection based on clustering with oversampling. Appl. Sci. 2021, 11, 3022.

[Qaddoura et al, 2021] R. Qaddoura, A.M. Al-Zoubi, H. Faris, I. Almomani, A multi-layer classification approach for intrusion detection in iot networks based on deep learning. Sensors 2021, 21, 2987.
[Lin et al, 2021] Y. Lin, R. Liu, M. Divakaran, J. Y. Ng, Q.Z Chan, Y. Lu, Y. Si, F. Zhang, J.S. D. Phishpedia, A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. In Proceedings of the 30th {USENIX} Security Symposium ({USENIX} Security 21, Online, 11–13 August 2021.
[Rehman et al, 2021] S. Rehman, M. Khaliq, S.I. Imtiaz, A. Rasool, M. Shafiq, A.R. Javed, Z. Jalil, A.K. B. Diddos, An approach for detection and identification of Distributed Denial of Service (DDoS) cyberattacks using Gated Recurrent Units (GRU). Futur. Gener. Comput. Syst. 2021, 118, 453–466.
[Yang et al, 2020] C.T. Yang, J.C. Liu, E. Kristiani, M.L. Liu, I. You, G. Pau, NetFlow Monitoring and Cyberattack Detection Using Deep Learning with Ceph. IEEE Access 2020, 8, 7842–7850.
[Zhang et al, 2019] C. Zhang, F. Ruan, L. Yin, X. Chen, L. Zhai, F.A. Liu, A Deep Learning Approach for Network Intrusion Detection Based on NSL-KDD Dataset. In Proceedings of 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID), Xiamen, China, 25–27 October 2019; pp. 41–45.
[Letteri et al, 2020] I. Letteri, G. Penna, L. D. Vita, M.T Grifa, MTA-KDD’19: A Dataset for Malware Traffic Detection. 2020.
[Kim et al, 2020] Kim J. Kim, H. Kim, M. Shim, E. Choi, CNN-Based Network Intrusion Detection against Denial-of-Service Attacks. Electronics 2020, 9, 916.
[Alrashdi et al, 2019] I. Alrashdi, A. Alqazzaz, E. Aloufi, R. Alharthi, M. Zohdy, H. Ming, AD-IoT: Anomaly detection of IoT cyberattacks in smart city using machine learning. In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference, CCWC 2019, Las Vegas, NV, USA, 7–9 January 2019.
27. Zhang, C.; Ruan, F.; Yin, L.; Chen, X.; Zhai, L.; Liu, F. A Deep Learning Approach for Network Intrusion Detection Based on NSL-KDD Dataset. In Proceedings of 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID), Xiamen, China, 25–27 October 2019; pp. 41–45.
[Hu et al, 2019] T. Hu, W. Niu, X. Zhang, X. Liu, J. Lu, Y. Liu, An Insider Threat Detection Approach Based on Mouse Dynamics and Deep Learning. Secur. Comm. Netw. 2019, 2019, 12.
[Pekta et al, 2019] A. S. Pekta, T. Acarman, Deep learning to detect botnet via network flow summaries. Neural Comput. Appl. 2019, 31, 8021–8033.
[Creech et al, 2014] G. Creech, J. Hu (2014a) A semantic approach to host-based intrusion detection systems using Contiguousand Discontiguous system call patterns. IEEE Trans Comput 63(4):807–819
[NSL-KDD, 2018] University of New Brunswick. NSL-KDD Data Set for Network-Based Intrusion Detection Systems. NSL-KDD Dataset. 2018.
[Song et al, 2006] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, K. Nakao, Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In Proceedings of the 1st Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, BADGERS 2011, Salzburg, Austria, 10 April 2011.
[Nguyen et al, 2018] K.K. Nguyen, D.T. Hoang, D. Niyato, P. Wang, D. Nguyen, E. DutkiewicCyberattack detection in mobile cloud computing: A deep learning approach. IEEE Wirel. Commun. Netw.Conf. WCNC 2018, 2018, 8376973.

Quantum Computing: An Assessment into the Impacts of Post-Quantum Cryptography

Roger G. Massmann
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
roger.massmann@go.stcloudstate.edu

Nick M. Grantham
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
nmgrantham@stcloudstate.edu

Akalanka B. Mailewa
Department of CSIT
Saint Cloud State University
St. Cloud, Minnesota, 56301
amailewa@stcloudstate.edu

Abstract

Quantum Computing continues to expand and rapidly approach large scale commercial usage. The advancement of quantum computing poses a threat to current cryptographic techniques and solutions are being researched rapidly to determine the correct course of action, culminating in a field known as Post-Quantum Cryptography, or PQC. This research gathers sufficient information and evidence to prove that quantum computing can be formally introduced into society where individuals can feel a sense of assurance that this technology is used for good rather than evil. For quantum computing, we weighed up whether the capabilities outweigh the costs, and if we can truly imagine a world where quantum computers can be a commercialized product. Quantum key distribution facilitates key exchange so that users can safely transmit messages over a quantum channel where the receiver will have access a key that will be the baseline for communication. Along with this comes the theorems that compensate for the possibility of photon leakage, or the possibility of an eavesdropper. With the evaluation of strength for a quantum computer comes the posed threat as to whether quantum computers can be used for the wrong reasons, therefore, post-quantum cryptography acts as the quantum proof measure, which may defend against quantum attacks and although PQC is currently limited to companies investing billions of dollars in research, the impact on the end-user is imminent. The objective of this review is to compile the current research and data in order to assess the impacts of Post Quantum Cryptography on organizations and agencies, and to approximate when and how these impacts will arrive at the end-user.

Keywords: Attacks; Quantum-Computing; Security; Vulnerabilities; Risks; Threats; Cryptography; Post-Quantum-Cryptography

1 INTRODUCTION

A technologically driven society stems from the structures and foundations of those who came well before us. From Stephen Wiesner's development of conjugate coding in the late 60's [1], onto Alexander Holevo who inspired the introduction of "Holevo's Bound" theorem, these individuals were just very few of the scientists that paved the way for Isaac Chung, Neil Gershenfeld and Mark Kubinec to successfully administer the first representation of two, three, five, and seven quantum bit quantum computers in 1998. The group's initial computation allowed for a two-qubit input/output scheme, and while recognized as a minor result mathematically, it can alternatively be viewed as a spark in opportunity for the future of quantum computing. Some may ask, why has quantum computing resurfaced as a 'new' generational phenomenon when it has been studied and implemented for over twenty years? The short answer to that question would be that it is based on interpretation and perspective as to how you understand the present versus the past of quantum computing. For years, quantum computing has been vastly a concept, but now, organizations are beginning to speak out about the properties, potential, and risk that quantum computing obtains. Now that quantum computing has transitioned from theoretical to realistic implementation, government agencies must now think of ways that will carry out or mirror the cryptographic algorithms that a standard computer would demonstrate in asymmetric or symmetric encryption onto a quantum-based computer. This leads us to the topic of quantum cryptography, better known as quantum key distribution (QKD), which was invented by Bennet and Brassard [2] as well as Ekert [3]. This essentially acts as the secure transmission of communication through a quantum channel that ensures both speed and security. The opportunities that arise through QKD are limitless, and it tends to be the driving factor to most organizations that are reassuring their customers that they should not be afraid of a quantum inspired future. See, Quantum keys are deemed unbreakable, which make them beneficial for secure message transmission, however on the other hand, message transmission might not always be used for good [4][2].

On the other end of the quantum spectrum comes the risk evaluation and assessment through post-quantum cryptography. As mentioned earlier, quantum computing has raised some very valid questions around how data is going to be protected, considering the power of message transmission and retrieval. Post-quantum cryptography aims to counteract and mitigate the risk factor around quantum computing, as well as reassure organizations and customers that their data is being handled sufficiently. U.S. Secretary of Homeland Security, Alejandro Mayorkas is adamant on embracing the transition to a quantum environment [5]. According to Mayorkas, the transition to post-quantum encryption algorithms is as much dependent on the development of such algorithms as it is on their adoption [5]. His statement suggests that there is a push for strong encryption practices within quantum computing, and that there is a priority around confidentiality of data now, and in the future. However, it isn't just Homeland Security and the US Government making statements about the potential and dangers of quantum computers, it's basically every major tech company. IBM, Google, Intel, Microsoft, just to name a few, are both embracing and preparing quantum computing, and have standards and roadmaps in place to ensure that when the day comes where quantum computing because natural practice, they will have all bases covered, whether it be through encryption to CVSS. With encryption in mind, there are various algorithms that are already being implemented in the IBM quantum lab, which can visualize what a quantum computer is capable of. From Shor's, to Grovers Algorithm, these methods will be analyzed and evaluated through the virtual quantum compiler, which is a great resource to enable us to get a gauge on how useful quantum systems can be in solving bulk data.

2 BACKGROUND

For this review, the referenced works had to meet several criteria to be considered. To have a full spectrum of perspectives, references were taken from the private sector publications, public sector publications, and from well-reviewed field experts. For the public sector publications, the references were only taken from well-known sources containing bodies of work in both classical computing as well as quantum computing to ensure the sources were unbiased and thoroughly researched. The sources from field experts all contained several references to other reviewed published works, as well as some sources that contained cited expert opinion. One of our first citations came from the Stanford Encyclopedia [6], a publication maintained by Stanford University, a private research institution with dedicated facilities for Computer Science, Theoretical Physics, and Quantum Information. Another source with a rich history in the field in computing is IBM [7], a corporation dedicated to advancing technology for business, which also was one of the first entities to begin research in quantum computing in the 1980's. In the public sector, sources used include the United States Department of Homeland Security [27], a long-standing contributor to the field of cryptography and cybersecurity that is publicly funded by the United States government. Each source was thoroughly vetted to ensure they were peer reviewed and approved by experts and contained well documented, accepted and most importantly accurate theory and information.

3 QUANTUM COMPUTERS

Quantum refers to the smallest particle in a physical state. It inherits the same properties of atomic or subatomic matter, involving electrons, neutrinos, and photons [8]. The reason why quantum computers are so relevant in society today, is due to the potential that they carry, as well as the threats that could arise. The characteristics of a quantum computer can essentially change the way technology operates as a whole. "Quantum supremacy" has been a term used to describe quantum computers. This is because there are abilities that a quantum computer inherits, that cannot be achieved by a classical computer. For example, Google's latest quantum computer claims to be able to complete a computation in around 200 seconds, whereas a classical computer would take 10,000 years [9][10]. Tech giants have publicly committed millions, if not billions of dollars toward large-scale quantum computers, and we could see fully functioning, publicly implemented quantum computers in as little as 5 years from now [10].

Classical computers function through binary states, this means that message transmission utilizes 0's and 1's to communicate and determine results [11][12]. Quantum computers on the other hand function on a qubit system, which factors in various possible states, expanding the functionality of message transmission. Quantum computers require temperature levels just shy of absolute zero to operate. Another drawback to quantum computers is that the price of processing has been a concern for many, where it can costs 10's of millions of dollars to function a commercialized quantum computer. This pulls into question, whether it is worth the investment, and to balance out our wants and needs as a society. The low scale quantum computers can function from as little as $5,000, with limited problem-solving capabilities. On the other hand, there is IBM, who acquires a 127 qubit quantum computer that would cost an exponentially large sum, exceeding the $15 million dollar mark.

IBM [13] mentions that they use cooled super fluids to avoid overheating. But with this extremely low temperature comes other perks of transmission for electrons in particular. Electrons attain a quantum mechanical effect where they travel with ease through a channel, making them superconductors. A term known as "Cooper pairs" describes the way that the electrons meet up and travel through the superconductors, carrying a charge over insulators that facilitate the transmission. These electrons are being passed through what is known as a quantum tunnel. IBM in particular, use a structure called Josephson junctions as their superconducting qubits. Therefore, when microwave photons are directed to the qubits in transmission, they can dictate their state, and note results.

Say there are two individuals living together. Both individuals need to decide what shirt they are going to wear for the day. Person one is in their room and has a selection of two shirts to wear for the day. The color of the shirts for person one is black and the other is white. For person two, they have many shirts to select from. Not only do they have many shirts to choose from, but they have shirts that range from black, to a grey, all the way onto a plain white. The bottom line is, person two has the option to wear a shirt that acquires both white and black, with various shades involved as well. Person one is the classical computer. This individual can only choose from two options, the black shirt representing a 0, and the white representing a 1. Individual two has an overflowing wardrobe, this person can choose from a huge selection of shirt colors. This is our quantum computer. This analogy represents superposition, where the states of a quantum particle can represent anything within the possibility of a binary digit [8].

The correlation between the photons that flow throughthe quantum channel is known as entanglement. To reach the state of entanglement, quantum particles follow a process where they flow through a laser lightinto crystal, by which they are then converted by crystal into entangled pairs of photons. Based on various QKD protocols, the entanglement process canbe perceived differently in each process.

4 QUANTUM KEY DISTRIBUTION

When looking at standard key distribution, there is always an element of whether Alice and Bob will be able to communicate safely without Eve intercepting the message in the middle. Many cryptographic practices will attempt to protect the key, however there can never be full protection guaranteed on classical methods. Classical cryptography bases encryption around the CIA triad, with the aim for messages to be confidential in nature, transmission being impenetrable or difficult to hinder through integrity, and finally available for both parties to access. Additionally, the CIA triad hasn't been the only thing referred to when verifying a secure method of communication, rather the objectives of authentication, digital signatures, and non-repudiation all mound into the structures of a successful cryptographic system [14][15][16].

The basic approach to cryptography is that plaintext message will be encrypted into ciphertext, which then becomes decrypted by the time it reaches the receiver. The properties of cryptography involve symmetric key, (also deemed as private key cryptography), which only has one key exchanged. Symmetric encryption is great for fast transfer and bulk encryption [17][18]. The image 1 below demonstrates how symmetric encryption works.

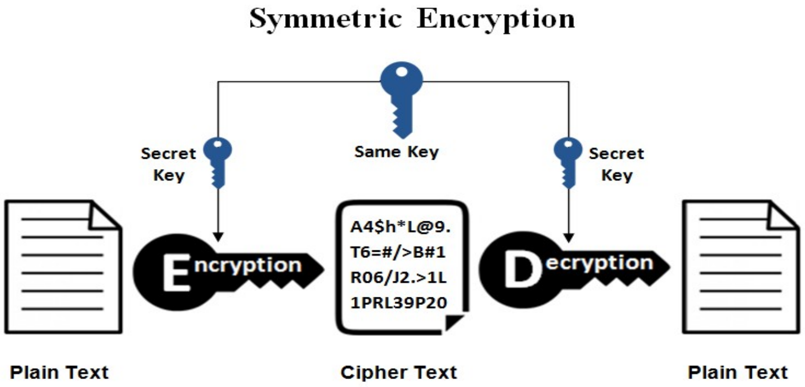


Figure 1: Symmetric Encryption [12]

Having one main key facilitating the communication. On the other hand, Asymmetric, or public key encryption utilizes a public channel that functions as a means of encryption, and the private channel functions as a decryption utility [19][20]. So, in the case of a user attempting to connect to a HTTPS server for instance, Alice (web server) attempts to communicate with Bob (browser). When Bob receives the public key from Alice, the message encrypts as a one-time symmetric key. Then the symmetric, private key acts as a baseline for the rest of the communication for both encryption and decryption [21].

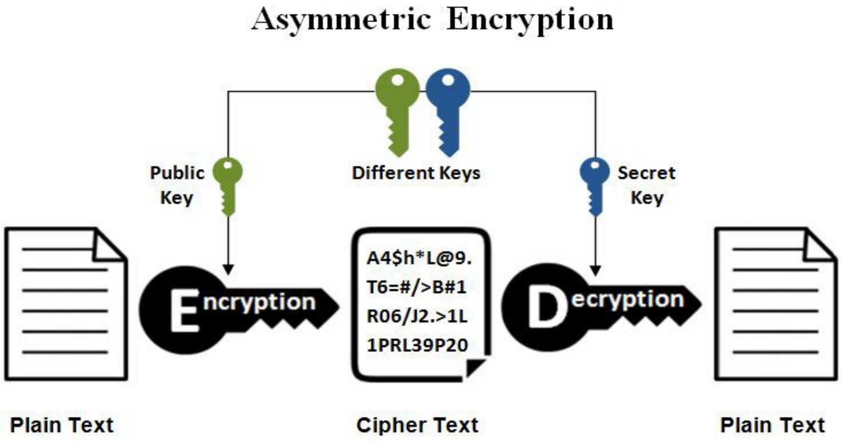


Figure 2: Asymmetric Encryption [17]

QKD is essentially the first application of quantum information science, and through years of study, analysis and finally, implementation, there have been products now available that acquire properties used through quantum based commercialized products. Although quite limited through key distribution over 100km, the progress of QKD is astonishing [22]. Just like classical key distribution, quantum key distribution uses an exchanged, shared key to communicate, however this method is facilitated through the quantum channel. The principle that separates quantum key distribution from classical.

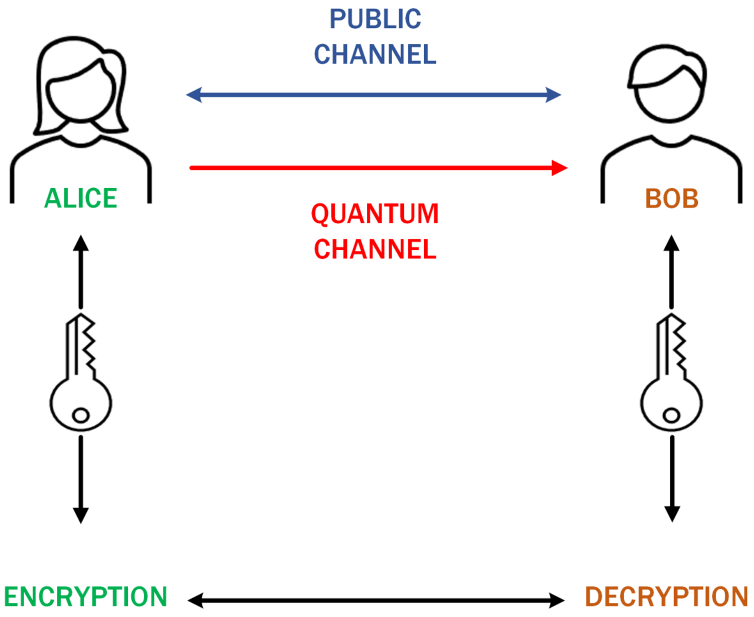
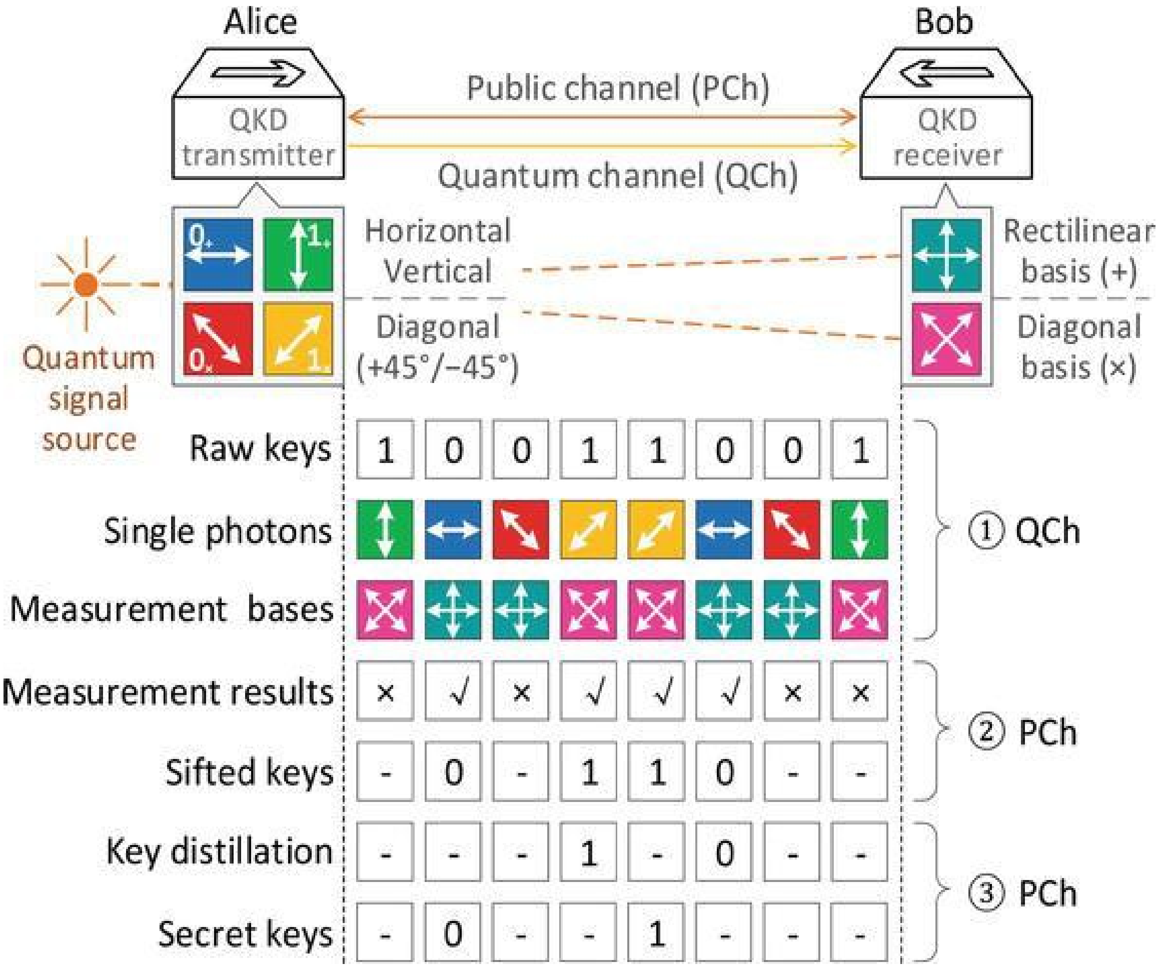


Figure 3: Quantum Key Distribution Diagram

Key distribution is the method of distribution that include the laws of physics to send and receive the qubits sent through the quantum stream. The distribution, capturing and development of the residual key is why quantum cryptography is deemed unbreakable, making it very controversial from all levels of the technological industry. To explain how QKD works with reference to the BB84 protocol by C.H Bennet and G. Brassard, we will use the standard Alice and Bob concept to demonstrate how message transmission successfully reaches the end user. During communication between Alice and Bob, Alice initiates the message through the production of a stream of photons. The photons flow through a polarizer, which then determines the quantum state of the cryptographic bits that will eventually reach Bob. However, the state of these quantum bits will be characterized as not only vertical and horizontal, but they can also carry a -45- degree or +45-degree angle that will travel towards Bob. [23].

- H (horizontal) codes for 0+
- V (vertical) codes for 1+
- +45 codes for 0×
- +45 codes for 1×

As a means of more advanced key determination, Bob will meet the quantum photon states with a randomizedphoton splitter, that will establish whether the photons. Will either pass through or be dropped in the case of an unmatched state? This is where the unbreakable nature comes into the picture. When Alice and Bob match up their sending and receiving keys, there will be inconsistencies as to whether Alice's photon state matches up with Bob's photon splitter state. The keys will correlate, and a residual/sifted key of the matched states will determine the final quantum key.



Raw keys	1	0	0	1	1	0	0	1
Single photons	↕	↔	↘	↗	↗	↔	↘	↕
Measurement bases	×	+	+	×	×	+	+	×
Measurement results	×	√	×	√	√	√	×	×
Sifted keys	-	0	-	1	1	0	-	-
Key distillation	-	-	-	1	-	0	-	-
Secret keys	-	0	-	-	1	-	-	-

Figure 4: QKD Based on BB84 [24]

The characteristics of QKD assume correctness and secrecy [22] over the channel that are near impossible to ensure. There are many factors that must be considered in quantum cryptography that can alter the state of the initial key or be intercepted by an eavesdropper. While we might believe that an ultimate secure method has been arranged by Alice and Bob through quantum key distribution, we must still factor in the possibility of an eavesdropper (Eve) intercepting or hindering transmission from Alice to Bob. For the QKD protocol to be deemed secure, it is important to understand that everything that can go wrong, very well will go wrong in the transmission. We have an equaling Alice and B equaling Bob. Ultimately, because QKD bases itself off randomness and estimation, if A = B, this is a successful key exchange. But the probability of this occurring without some type of interference is very low, therefore, it is paramount to account for the interference. In average instances of a long key distribution, Eve will gather about 50% of the raw key through eavesdropping over the fully exposed quantum channel (IE). With an error rate of about 25% (Q), we must factor in how successful the communication will be under the assumption of an eavesdropper within message transmission. With reference to Shannon information theory around parallelism as well as the Korner-Csiszar-Marton theorems [25], we will unpack the demonstration of how Alice maps to Bob with knowledge of a potential eavesdropper interference. For:

$$\mathbf{r = \max\{I(A : B) - IE , 0\}}$$ **[18]**

Where I(A : B) = H(A) + H(B) − H(AB) is the mutual information between Alice's and Bob's raw keys [15]. Variable H factors in Shannon entropy, meaning the amount of information within the variable is dependent on user input [26]. Having said this, if

$H(A) =(B) = 1$, one has $I(A : B) = 1 - H(Q)$, with a resend attack of $I(A : B) < IE$ [19]

Eve will then have more information as opposed to Bob through her manipulation of the quantum channel, and this will cause Bob to abort the communication.

This is an example of how quantum cryptography drops communication over any suspicion or inconsistency over transmitted messages. In another example, asymptotic and finite-key bound draws upon the corrections [27] important in the efficiency and robustness [28] of quantum key distribution. Additionally, these regimes determine whether through fault tolerance and error rate, can a quantum key still be developed? For these methods, Alice will devote as many signals toward Bob in attempts for Bob to receive as much data as possible, which outweighs the leakage to Eve. Similarly, to the previous example, if a substantial amount of data is exposed to Eve, the transmission will be aborted [29].

The asymptotic limit displays r as the running exchange of the key, N as the total number of signals exchanged by Alice towards Bob, and L being the length of the secret key.

$$\mathbf{r \infty = \lim (\mathit{L}/N) \rightarrow\infty \; N = \min H (A | E) - H (A | B)}$$ **[11]**

On the other hand, the finite key bound factors in the element of uncertainty within the communication, analyzing the level of leakage that can be withstood in a key distribution. [27]

$$\mathbf{r\, N = N = n\, N \min E | V \pm \Delta V\, H (A | E) - \Delta(n) - leak}$$ **[11]**

These theoretical distribution methods test the fault tolerance factor of quantum key distribution, and while it is great to suggest that Shannon theorem and parallelism will be secure enough to account for an absent eavesdropper, the fault tolerant based methods tend to be more practical

5 POST QUANTUM CRYPTOGRAPHY

As Quantum Cryptography grows and evolves, issues in security and vulnerabilities grow with it. While many companies seek to advance knowledge on the subject and improve the technology, other organizations seek to find the issues and risks with using Quantum Cryptography, and the risks associated with refusing to adopt these advancements as common business practice [30] [31].

The private and public sector have worked closely in the past to develop classical computing and security, and to create rules, regulations, and protocols. This is no different in quantum computing. While Google, IBM, and Intel work to expand the capabilities and usage of this technology, NIST and other agencies are quick to respond with common standardizations [30]. As of November of 2022, NIST is completing a third round of evaluating and selecting algorithms to be standardized in the field for PQC. The most recent status report updated in September of 2022 details the candidates and evaluation process, as well as algorithms no longer being considered. The report and operation as a whole set the precedent for the treatment of standardizing the field. The research, analysis, and results will impact the development of current and future algorithms to be standardized, while ensuring that the advancement of PQC is still under the control of the public.

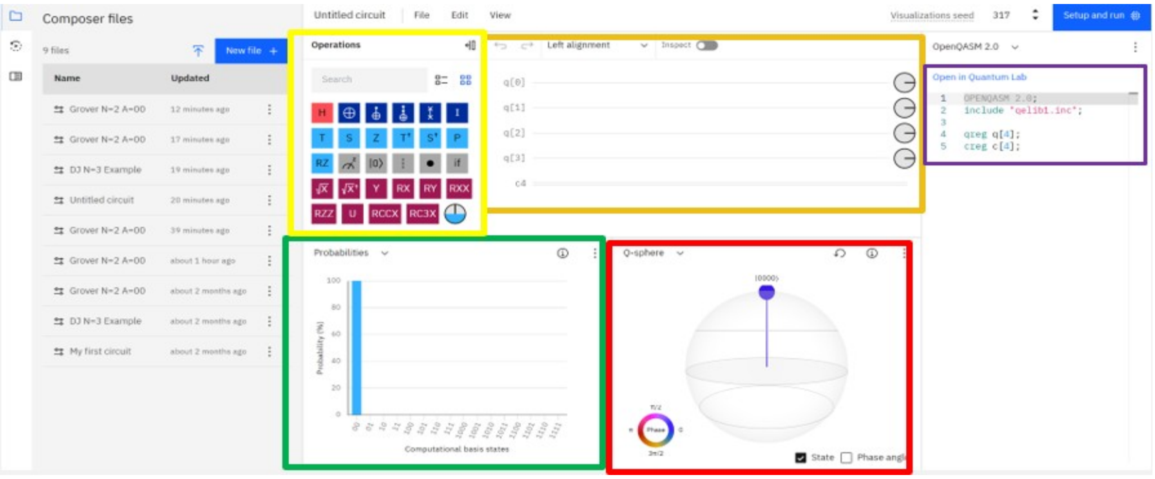
While standards are being set, private corporations continue to make strides to be the leader in the field. While classical computing continues to grow, it will eventually run into physical limitations [7] that can be effortlessly surpassed by quantum computing. With these advancements, however, the obsolescence of classical cryptography is also approaching. While quantum machines will be able to complete computations in minutes that would take classical computing decades, current cryptography simply won’t hold up to PQC. The most important issue for private corporations and government agencies is determining how to respond to PQC, when to respond, and when will all these factors make more sense financially for the companies. Recently, companies have even begun live testing of PQC algorithms. As of 2016, google announced a live experiment using their own post quantum key exchange algorithm on a small fraction of connections to Chrome servers [32]. This experiment demonstrated that practical applications of PQC in the real world are indeed feasible and are closer than most users are aware. More analysis into other private entities and public agencies are instrumental in determining exactly when PQC will overtake classical computing. With the private sector taking steps to revolutionize their own PQC power, the end user can already access an abundance of education material. Currently, IBM offers QISkit [33], a software that interprets high level languages such as Python and applies them to IBM’s quantum machine algorithms. This allows users to experiment and learn about quantum computing through the curriculum designed by IBM, or by experimenting individually in the quantum lab. The focus of this technology is not to allow the end user to execute complex quantum algorithms, the simulation has a timeout limit of roughly 10000 seconds, but to allow the user to gain a baseline understanding of quantum computing to educate the public. Ultimately, QISkit is a tool that is currently incapable of posing any form of a security threat. In fact, it is extremely unreasonable to consider any entity without access to a physical quantum machine be a true threat in the real of PQC, thus for the foreseeable future, quantum computing will remain virtually unaffected by script kiddies.

To understand what corporations are essentially racing towards, a quantifiable goal allows for a more concrete analysis. In quantum computing the term Quantum Advantage describes a machine that would be able to compute problems that no classical computer can [34]. As of right now, this goal is not immediately commercially viable, with no tangible outlook on when it might be. However, a more attainable benchmark is the 50-qubit quantum computer. In 2018, IBM announced the functional IBM Q, a 50-qubit quantum computer. This machine displayed the ability to compute problems previously considered unsolvable by machines [35]. Almost immediately after, Google responded with a processor that contained 72 qubits. Although these computers are far from being commercially available, they demonstrate that the technology currently exists to allow quantum computing to surpass classical computing, making PQC innovation a necessary field.

5.1 IBM QUANTUM COMPOSER

In our data findings, we will be using the IBM Quantum Composer to analyze how various algorithms work in quantum computing [36]. This will demonstrate the power of a quantum processor, and give a gauge around why quantum computing can be seen as both an advantage and disadvantage for the present and future of technology. As we open the Quantum Composer shown in figure 6, there are an overwhelming number of options as to how we, as the user, can input and output the data. Starting with the operations section; there are various inputs that can dictate the output of the quantum computation. The function, whether it's a classical gate, phase, non-unitary/modifier, or quantum operation, can be a building block to support a quantum algorithm or outcome. This is why it is so revolutionary for IBM to have this platform as an open-source space for individuals to be inspired by the potential of quantum algorithms. Next, the section to the right of the operations tab displays a staging type- environment, where the operations can be dragged and dropped to fulfil the users' algorithm choice.

In Figure 5, a Hadamard Gate has been placed in the staging environment. A Hadamard gate in particular convert |0⟩ and |1⟩ to |+⟩ and |-⟩ and linked to the super positioning states mentioned earlier [37].



Name	Updated
Grover N=2 A=00	12 minutes ago
Grover N=2 A=00	17 minutes ago
DJ N=3 Example	19 minutes ago
Untitled circuit	20 minutes ago
Grover N=2 A=00	39 minutes ago
Grover N=2 A=00	about 1 hour ago
Grover N=2 A=00	about 2 months ago
DJ N=3 Example	about 2 months ago
My first circuit	about 2 months ago

```
OPENQASM 2.0;
include "qelib1.inc";

qreg q[4];
creg c[4];
```

Figure 5: IBM Quantum Composer: Individuals to create own Quantum Algorithms

Next, there is a command line interface that will carry out the code that links to the staging environment. Therefore, the more code, the more changes to the CLI. When we added the ‘h’ gate, the CLI will then display an ‘h’ gate located at index 1, or qubit 1. The probabilities section displays the percentage of the outcome occurring based on the state, and the Q-sphere, gives a visual idea of the state, and phase angle, as well as the probability of the state occurring.

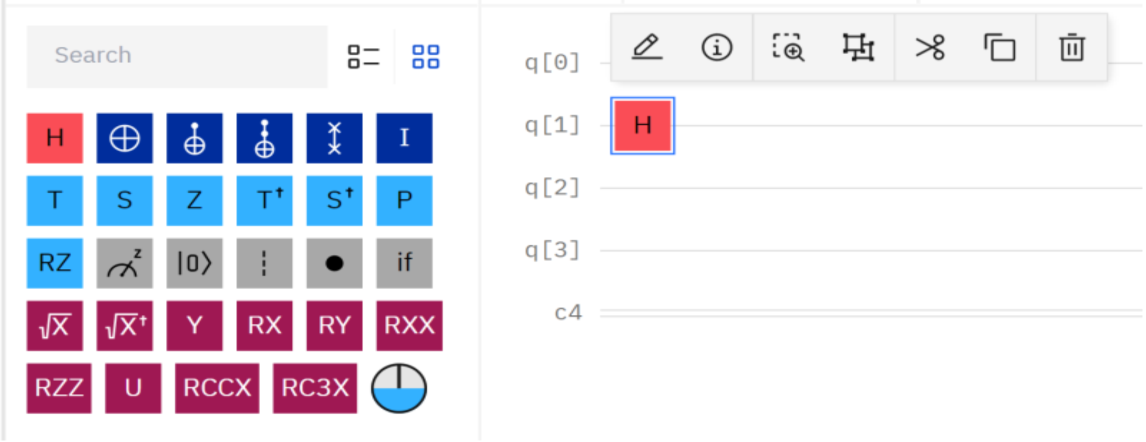


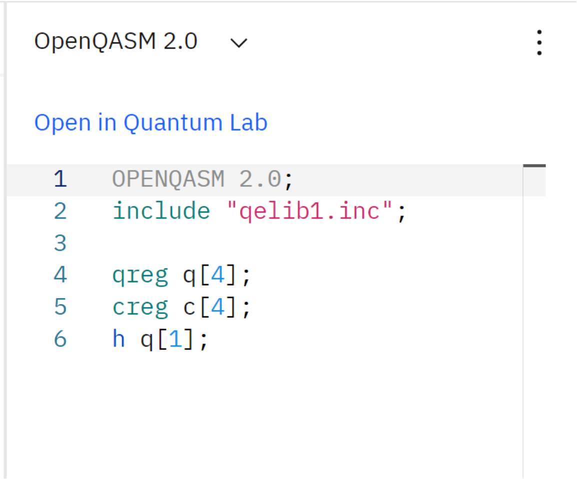
Figure 6: Operational structure with Hardamand Gate

5.2 SHOR’S ALGORITHM

In number theory, the former most efficient algorithm for finding prime factors of an integer was the general number field sieve or GNFS. That was until 1994, when mathematician Peter Shor introduced Shor’s algorithm, a polynomial-time quantum algorithm [38]. Shor’s algorithm allows for a near exponential decrease in the amount of time it takes to factor integers with digits greater than 1000.

The implications of Shor’s algorithm on PQC are immense, as future quantum computers with greater processing power might be able to apply the algorithm towards decrypting keys that would otherwise take hundreds of years to decrypt with classical computing. RSA encryption is reliant on the assumption that classical computing cannot efficiently factor large prime and semi-prime numbers. Because of this, in future applications of quantum computing, Shor’s algorithm could render RSA and other classical encryption techniques useless. To visualize this, it is known that the GNFS can be reasonably expected to be limited to roughly 200 digits. According to IBM’s QISkit’s current data [36], factoring a polynomial N with d decimal digits, GNFS would take exp(const * d1/3), almost exponentially longer than Shor’s which can be displayed as const *d3. While the classical algorithm’s record might take 1030 operations, Shor’s might take 10^7. With the IBM quantum composer, Shor’s algorithm can be displayed using IBM’s circuits and operations.

First, the reset operation is used to return the qubit to the state |0⟩. Then, an H gate, followed by a T gate and another H gate are applied to rotate the state and alter the angle. These figures demonstrate the application of Shor’s algorithm, as well as the written code in the quantum lab.



```
OPENQASM 2.0;
include "qelib1.inc";

qreg q[4];
creg c[4];
h q[1];
```

Figure 7: circuit code in IBM’s quantum lab

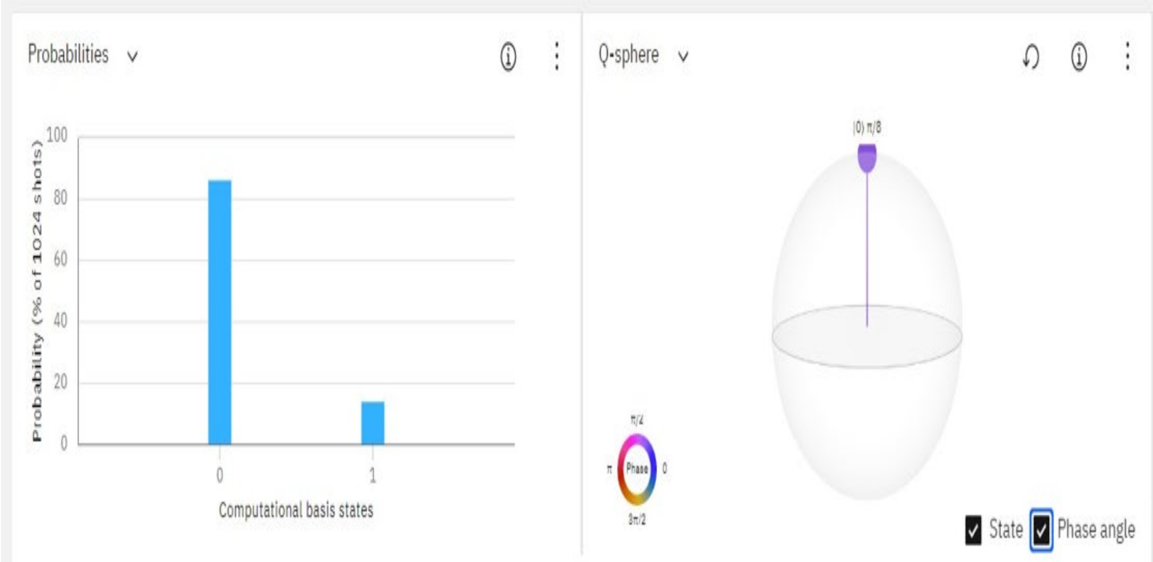


Figure 8: Shor’s Algorithm probabilities and Q-Sphere

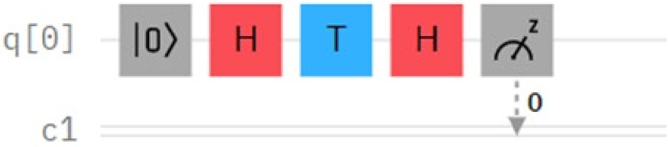


Figure 9: operational structure for Shor’s Algorithm

5.3 GROVERS ALGORITHM

In 1996, Lov Grover introduced quantum search function now known as Grover’s algorithm. The database search algorithm has the ability to analyse large amounts of data and narrow down the results to find the desired product [38]. To give a greater understanding around how powerful Grover’s Algorithm can be, think about a brute force attack, and how efficiently Grover’s algorithm could be implemented to gain access to a password. Grover’s algorithm is known to create a quadratic speedup, meaning it will drastically narrow down the tries it attempts before it gains a result. For example, if we have four cups, and one of these cups had a rock under it, the average amount of times it would take a classical computer to guess the right cup would be 2 and ¼ attempts. On the contrary, if a classical computer was to guess the correct cup, it would take 1.

Now, we will move into the IBM Quantum Composer to demonstrate how Grover’s Algorithm works. Using a template created from the IBM Quantum Composer, we can display the structure of Grover’s Algorithm, where the outcome that is sought after is in state 00. As shown below, we have both of our |0⟩|0⟩ states in qubit 1 and qubit 2. This is known as the reset operation.

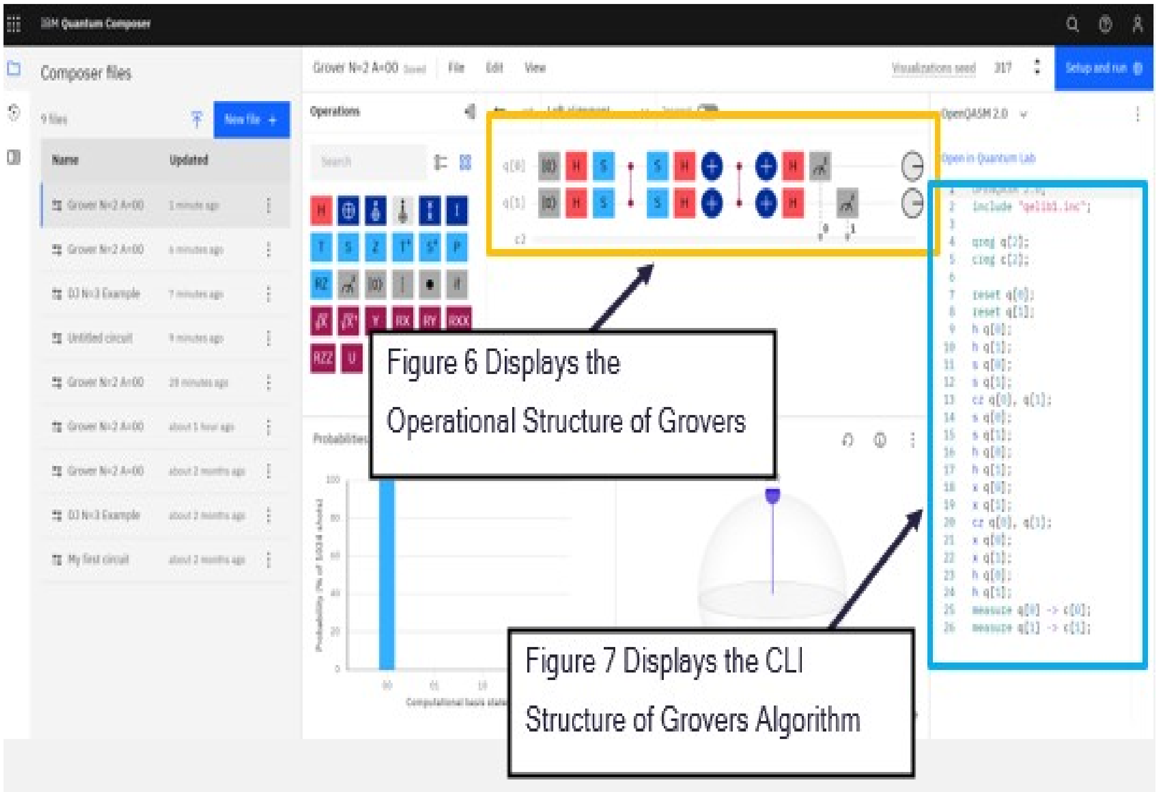


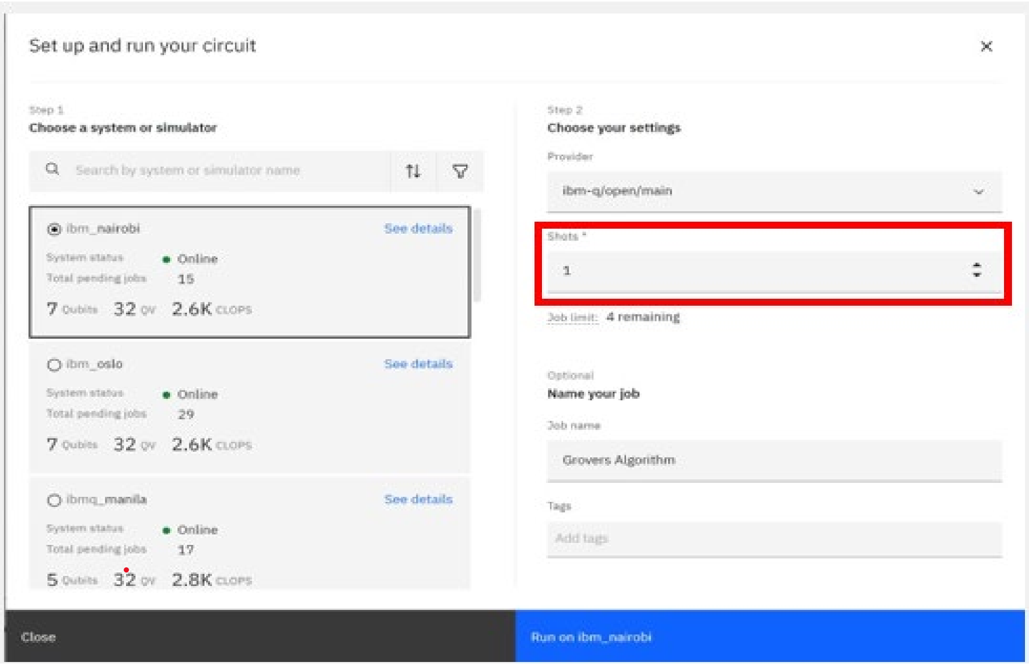
Figure 6 Displays the Operational Structure of Grovers

Figure 7 Displays the CLI Structure of Grovers Algorithm

```
OPENQASM 2.0;
include "qelib1.inc";

qreg q[2];
creg c[2];

reset q[0];
reset q[1];
h q[0];
h q[1];
s q[0];
s q[1];
cz q[0], q[1];
s q[0];
s q[1];
h q[0];
h q[1];
x q[0];
x q[1];
cz q[0], q[1];
x q[0];
x q[1];
h q[0];
h q[1];
measure q[0] -> c[0];
measure q[1] -> c[1];
```

Figure 10: IBM Quantum Composer with Grover’s Algorithm

5.3.1 Data Extraction of GROVER’S algorithm



Set up and run your circuit ×

Step 1
Choose a system or simulator

Search by system or simulator name

◉ ibm_nairobi — See details
System status ● Online
Total pending jobs 15
7 Qubits 32 QV 2.6K CLOPS

○ ibm_oslo — See details
System status ● Online
Total pending jobs 29
7 Qubits 32 QV 2.6K CLOPS

○ ibmq_manila — See details
System status ● Online
Total pending jobs 17
5 Qubits 32 QV 2.8K CLOPS

Step 2
Choose your settings

Provider: ibm-q/open/main

Shots *: 1

Job limit: 4 remaining

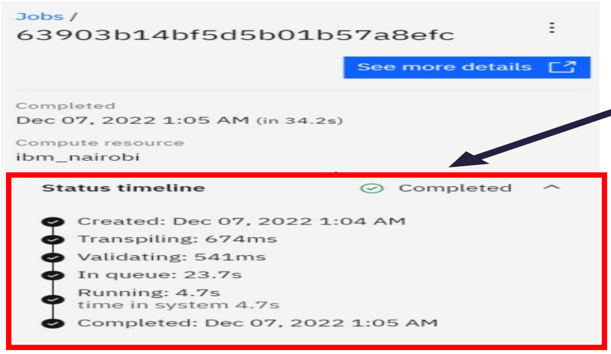
Optional
Name your job

Job name: Grovers Algorithm

Tags: Add tags

Close | Run on ibm_nairobi

Figure 11: Run with 1 Shot



Jobs /
63903b14bf5d5b01b57a8efc

See more details

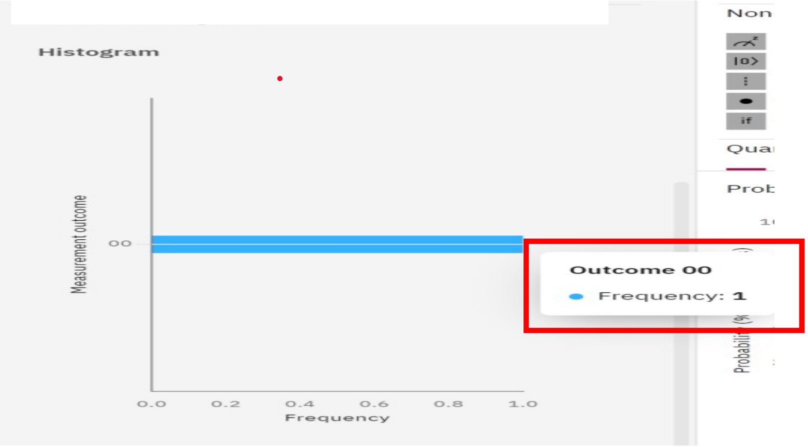
Completed
Dec 07, 2022 1:05 AM (in 34.2s)

Compute resource
ibm_nairobi

Status timeline — Completed

Created: Dec 07, 2022 1:04 AM
Transpiling: 674ms
Validating: 541ms
In queue: 23.7s
Running: 4.7s
time in system 4.7s
Completed: Dec 07, 2022 1:05 AM

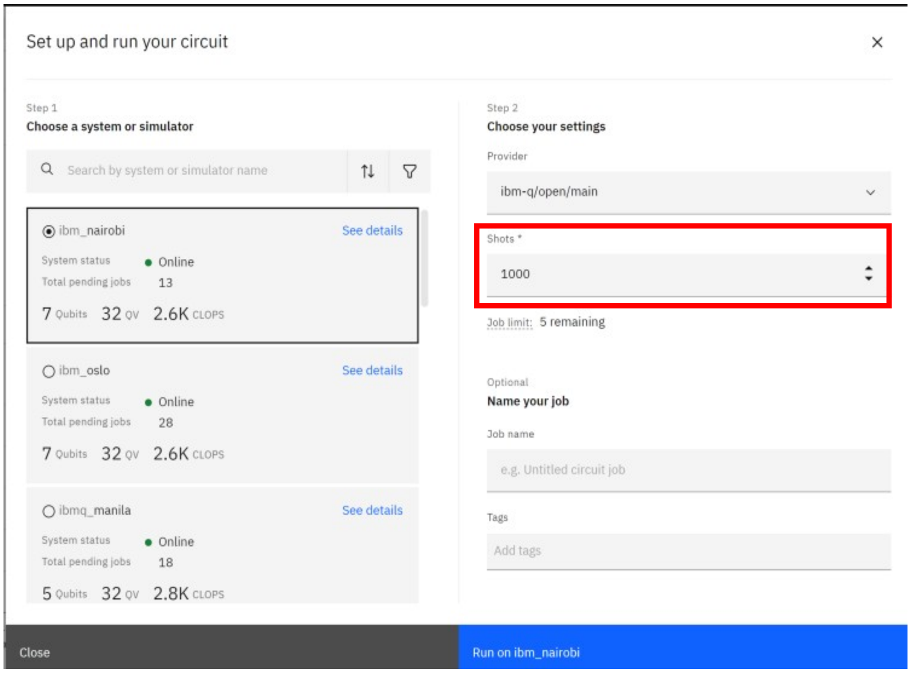
The job on the left displays a must faster result time, due to the size of the results that aresought out



Outcome 00
● Frequency: 1

Figure 12: Shows the test 1 histogram

As we would like to receive the result of 00, and only have 1 attempt to do so, this proves the accuracy of the algorithm.



Set up and run your circuit

Step 1
Choose a system or simulator

Search by system or simulator name

○ ibm_nairobi — See details
System status ● Online
Total pending jobs 13
7 Qubits 32 QV 2.6K CLOPS

○ ibm_oslo — See details
System status ● Online
Total pending jobs 28
7 Qubits 32 QV 2.6K CLOPS

○ ibmq_manila — See details
System status ● Online
Total pending jobs 18
5 Qubits 32 QV 2.8K CLOPS

Step 2
Choose your settings

Provider: ibm-q/open/main

Shots *: 1000

Job limit: 5 remaining

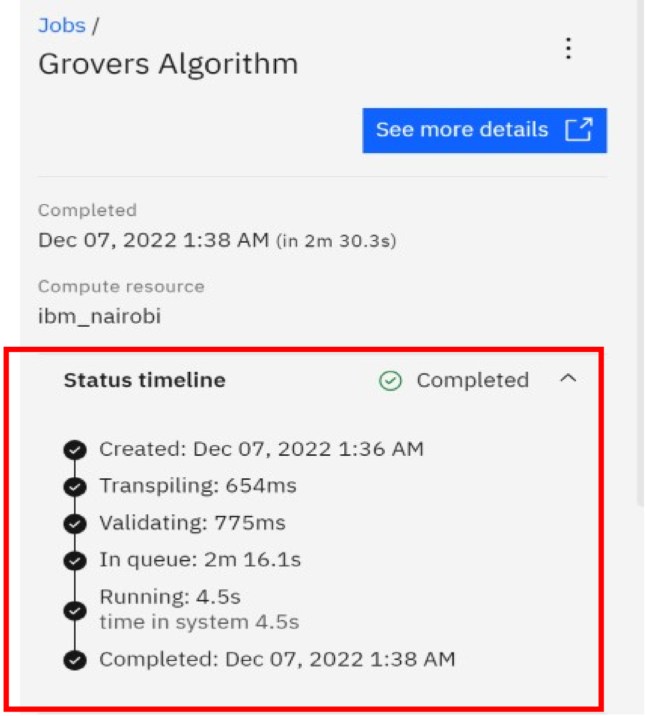
Optional
Name your job

Job name: e.g. Untitled circuit job

Tags: Add tags

Figure 13: Run with 1000 shots

Alternatively, the size of the algorithm affects the queue time of the results to occur



Jobs /
Grovers Algorithm

Completed
Dec 07, 2022 1:38 AM (in 2m 30.3s)

Compute resource
ibm_nairobi

Status timeline — Completed

- Created: Dec 07, 2022 1:36 AM
- Transpiling: 654ms
- Validating: 775ms
- In queue: 2m 16.1s
- Running: 4.5s
  time in system 4.5s
- Completed: Dec 07, 2022 1:38 AM

Figure 14: Details of status timeline 2

To receive the result from a bulk set of attempts, we can see that the result shows a 96.8% success rate for 00

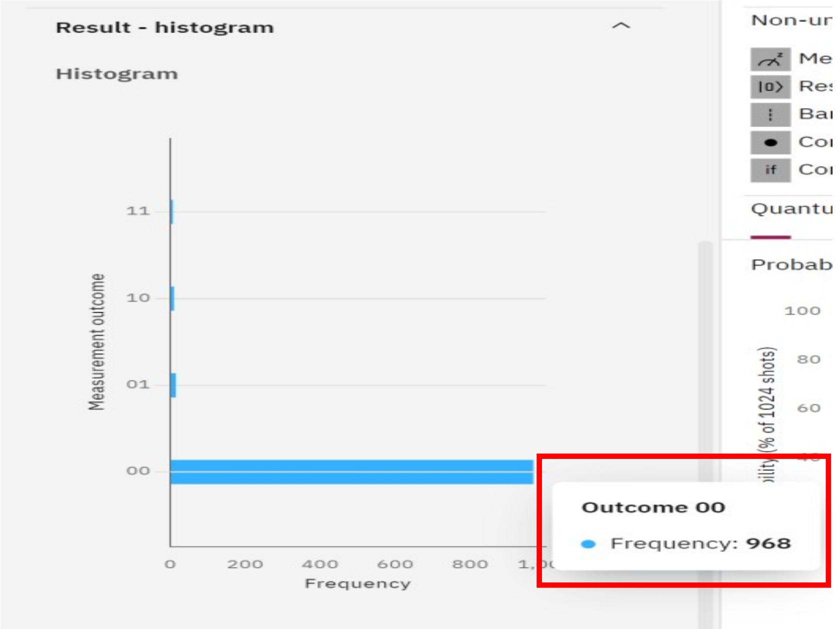


Figure 15: Shows the test 2 histogram

5.4 ALGORITHM ANALYSIS

In the analysis of Grovers Algorithm using the IBM Quantum Composer, the program is very impressive in how it simulates quantum processes. The functionality within the program is great from an educational point of view, and clearly outlines how Grovers algorithm works. The findings expressed that Grovers algorithm is very accurate in searching for items within an array. Working on a theoretical basis is a great way to demonstrate processes, however from further research, it is not necessarily certain that Grovers algorithm can be exploited through quantum computation. [1] Grovers algorithm proves to be a great searching utility, however from a cryptographic tool standpoint, there may be implications that any quantum computer may face, such as noise or other physical interferences. On the other hand, Shor's algorithm provides not only a strong display of the advantages of quantum cryptography, but also an insight into how the field of cryptography will need to adapt to changes that were not previously predicted. The ability of Shor's algorithm to factor large prime numbers poses a new challenge for researchers everywhere. Although quantum computers do not currently have the capabilities to execute the algorithm in a way that could impact current encryption standards, the implication that it could is enough to help advance the research of post-quantum cryptography to search for a solution when it is eventually needed. It also alerts researchers to the idea that many classical cryptography techniques will simply not be viable in the near future, and the progress of cryptography is benefiting from this push for new technology.

6 CONCLUSION

In this paper, our aim was to gather sufficient information and evidence to prove that quantum computing can be formally introduced into society where individuals can feel a sense of assurance that this technology is used for good rather than evil. For quantum computing, we weighed up whether the capabilities outweigh the costs, and if we can truly imagine a world where quantum computers can be a commercialized product. Quantum key distribution facilitates key exchange so that users can safely transmit messages over a quantum channel where the receiver will have access a key that will be the baseline for communication. Along with this comes the theorems that compensate for the possibility of photon leakage, or the possibility of an eavesdropper. With the evaluation of strength for a quantum computer comes the posed threat as to whether quantum computers can be used for the wrong reasons, therefore, post-quantum cryptography acts as the quantum proof measure, which may defend against quantum attacks. And although PQC is currently limited to companies investing billions of dollars in research, the impact on the end-user is imminent. Regardless of the commercial viability, or lack thereof, of quantum computing, the current innovations clearly demonstrate a commercial need for corporations to pursue Quantum Advantage. This means Post- Quantum Cryptography, if not now, will be a pressing concern for companies in the near future. The first organization or entity to attain quantum advantage will acquire an insurmountable lead in the field and cause a shift in the entire landscape of computing, not just quantum. Ultimately, however, the end-user currently only feels the impact in a theoretical sense. Access to education material on quantum computing is as close to PQC as any individual not affiliated with a large agency will attain within this decade, if not century. The current most pressing issue to the end-user is data confidentiality, which likely will not be breached with quantum computers for decades. When it is, companies and government entities will be well prepared, having already been working towards post-quantum cryptography for years. And with various algorithms just like Grover's and Shor's, we can see that extraordinary research and results will be carried on throughout the years and built upon at an advanced rate that we cannot even comprehend.

References

[1] SGate at qiskit.org. Available at: https://qiskit.org/documentation/stubs/qiskit.circuit.library.SGate.html (Accessed: December 6, 2022).

[2] Bennett, C. H. & Brassard, G. in Proc. IEEE Int. Conf. on Comp. Sys. and Signal Processing 175–179 (Bangalore, India, 1984)

[3] Ekert, A. K. Quantum cryptography based on Bell's theorem. Phys. Rev. Lett. 67, 661–663 (1991).

[4] Gamnis, Steven, Matthew VanderLinden, and Akalanka Mailewa. "Analyzing Data Encryption Efficiencies for Secure Cloud Storages: A Case Study of Pcloud vs OneDrive vs Dropbox." Advances in Technology (2022): 79-98. (DOI:10.31357/ait.v2i1.5526)

[5] Post-Quantum Cryptography | Homeland Security. Available at: https://www.dhs.gov/quantum#:~:text=%E2%80%9 CThe%20transition%20to%20post%2Dquantum,lat ter%20remains%20in%20its%20infancy. (Accessed: December 5, 2022).

[6] Hagar, A. and Cuffaro, M. (2019) Quantum computing, Stanford Encyclopedia of Philosophy. Stanford University. Available at: https://plato.stanford.edu/entries/qt-quantcomp/ (Accessed: December 5, 2022). Homeland sec

[7] Barde, Nilesh, et al. "Consequences and Limitations of Conventional Computers and Their Solutions through Quantum Computers." Issue, vol. 19, 2011, p. 161, lejpt.academicdirect.org/A19/161_171.pdf.

[8] What is Quantum Computing: Microsoft Azure, What is Quantum Computing | Microsoft Azure. Available at: https://azure.microsoft.com/en- us/resources/cloud-computing-dictionary/what-is- quantum-computing/#introduction (Accessed: December 5, 2022). IBM

[9] Gisin, N. et al. (2001) Quantum cryptography, arXiv.org. Available at: https://arxiv.org/abs/quant- ph/0101098v2 (Accessed: December 5, 2022).

[10] Gillis, A.S. (2022) What is quantum cryptography?, Security. TechTarget. Available at: https://www.techtarget.com/searchsecurity/definitio n/quantum-cryptography (Accessed: December 5, 2022).

[11] Khan, Muhammad Maaz Ali, Enow Nkongho Ehabe, and Akalanka B. Mailewa. "Discovering the Need for Information Assurance to Assure the End Users: Methodologies and Best Practices." In 2022 IEEE International Conference on Electro Information Technology (eIT), pp. 131-138. IEEE, May 2022. (DOI:10.1109/eIT53891.2022.9813791)

[12] Mailewa, Akalanka, and Jayantha Herath. "Operating Systems Learning Environment with VMware" In The Midwest Instruction and Computing Symposium. Retrieved from http://www.micsymposium.org/mics2014/ProceedingsMICS_2014/mics2014_submission_14 .pdf. 2014.

[13] Quantum encryption vs. Post-Quantum Cryptography (with infographic) (2022) QuantumXC. Available at: https://quantumxc.com/blog/quantum-encryption- vs-post-quantum-cryptography-infographic/ (Accessed: December 5, 2022).

[14] Mailewa, Akalanka, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Mechanisms and techniques to enhance the security of big data analytic framework with mongodb and Linux containers." Array 15 (2022): 100236. (DOI:10.1016/j.array.2022.100236)

[15] Sanyal, A. (2021) Symmetric, asymmetric and quantum encryption- an introduction to quantum cryptography, LinkedIn. Available at: https://www.linkedin.com/pulse/symmetric-asymmetric-quantum-encryption-introduction- sanyal/ (Accessed: December 5, 2022).

[16] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Security assurance of MongoDB in singularity LXCs: an elastic and convenient testbed using Linux containers to explore vulnerabilities." Cluster Computing 23 (2020): 1955-1971.

[17] Sanyal, A. (2021) Symmetric, asymmetric and quantum encryption- an introduction to quantum cryptography, LinkedIn. Available at: https://www.linkedin.com/pulse/symmetric-asymmetric-quantum-encryption-introduction- sanyal/ (Accessed: December 5, 2022).

[18] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Dynamic & portable vulnerability assessment testbed with Linux containers to ensure the security of MongoDB in Singularity LXCs." In Companion Conference of the Supercomputing-2018 (SC18). 2018.

[19] Rathore, A. (2022) Quantum key distribution: The Future of Secure Communication, Electronics For You. Available at: https://www.electronicsforu.com/technology-trends/quantum-key-distribution-future-secure- communication (Accessed: December 5, 2022).

[20] Mailewa Dissanayaka, Akalanka, Roshan Ramprasad Shetty, Samip Kothari, Susan Mengel, Lisa Gittner, and Ravi Vadapalli. "A review of MongoDB and singularity container security in regards to hipaa regulations." In Companion Proceedings of the10th International Conference on Utility and Cloud Computing, pp. 91-97. 2017.

[21] Shetty, Roshan Ramprasad, Akalanka Mailewa Dissanayaka, Susan Mengel, Lisa Gittner, Ravi Vadapalli, and Hafiz Khan. "Secure NoSQL based medical data processing and retrieval: the exposome project." In Companion Proceedings of the10th International Conference on Utility and Cloud Computing, pp. 99-105. 2017.

[22] Tomamichel, M., Lim, C., Gisin, N. et al. Tight finite-key analysis for quantum cryptography. Nat Commun 3, 634 (2012). https://doi.org/10.1038/ncomms1631

[23] Rathore, A. (2022) Quantum key distribution: The Future of Secure Communication, Electronics For You. Available at: https://www.electronicsforu.com/technology-trends/quantum-key-distribution-future-secure- communication (Accessed: December 5, 2022). www-nature

[24] Zhao, Y. et al. (2018) Quantum key distribution (QKD) over software-defined optical networks, IntechOpen. IntechOpen. Available at: https://www.intechopen.com/chapters/63116 (Accessed: December 8, 2022).

[25] Muchnik, A. A. (2002). Conditional complexity and codes. Theoretical Computer Science, 271(1), 97–109. https://doi.org/10.1016/S0304- 3975(01)00033-0

[26] J. Lin, "Divergence measures based on the Shannon entropy," in IEEE Transactions on Information Theory, vol. 37, no. 1, pp. 145-151, Jan. 1991, doi: 10.1109/18.61115.

[27] Quantum Cryptography and Computing: Theory and Implementation, edited by R. Horodecki, et al., IOS Press, Incorporated, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/stcloud- ebooks/detail.action?docID=3014999.

[28] Meyer, T. (2007) Finite key analysis in quantum cryptography. Available at: https://www.osti.gov/etdeweb/servlets/purl/2106051 5 (Accessed: December 5, 2022).

[29] Singh, Nicholas, Kevin Bui, and Akalanka Mailewa. "Robust Efficiency Evaluation of NextCloud and GoogleCloud." Advances in Technology (2021): 536-545. (DOI:10.31357/ait.v1i2.5392)

[30] Moody, D. (2022). Status Report on the Third Round of the NIST Post-Quantum Cryptography Standardization Process. https://doi.org/10.6028/nist.ir.8413-upd1

[31] Dissanayaka, Akalanka Mailewa, Susan Mengel, Lisa Gittner, and Hafiz Khan. "Vulnerability prioritization, root cause analysis, and mitigation of secure data analytic framework implemented with mongodb on singularity linux containers." In Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis, pp. 58-66. 2020.

[32] Venables, Phil. "How Google Is Preparing for a Post-Quantum World." Google Cloud Blog, 6 July 2022, cloud.google.com/blog/products/identity- security/how-google-is-preparing-for-a-post- quantum-world.

[33] "QISKit -- Quantum Information Software Kit for Quantum Computation." IBM Research Blog, 20 Feb. 2018, www.ibm.com/blogs/research/2018/02/qiskit- index

[34] Preskill, John. "Quantum Computing in the NISQ Era and Beyond." Quantum, vol. 2, 6 Aug. 2018, p. 79, 10.22331/q-2018-08-06-79.

[35] Pierce, Alan. "The IBM Q Is a Working 50 Qubits Quantum Computer - ProQuest." Www.proquest.com, May 2018, www.proquest.com/openview/67be7836dbb91b17e9d118359f1a02ca/1.pdf?pq- origsite=gscholar&cbl=182.

[36] Shor's algorithm. (n.d.). IBM Quantum.
https://quantumcomputing.ibm.com/composer/docs/iqx/guide/shors-algorithm

[37] Jozsa, R. (1999) Searching in grover's algorithm, arXiv.org. Available at: https://arxiv.org/abs/quant-ph/9901021 (Accessed: December 6, 2022).

[38] IBM Quantum Experience - Dashboard. (n.d.). IBM Quantum Experience. https://quantum- computing.ibm.com/

Automation in the Food Service Industry, and its Wide-Reaching Effects

Sieger Canney

Department of Computer Science

Buena Vista University

610 W 4th St, Storm Lake, IA 50588

cannsie@bvu.edu

Abstract

Automation in the food service industry is rapidly advancing and most are either unaware of the developments or consider them interesting but relegate them to nothing more than "just another flashy tech demonstration". Automation is quickly moving to the forefront of the industry and these changes forecast wide-reaching effects not only across the food service industry but also society as a whole.

The purpose of this paper is to research and educate on the recent developments in automation within the food service industry and its potential industrial, ethical, and economical effects.

1 Introduction

When it comes to automation, the food service industry is no stranger. It is common knowledge that most large food service chains have their own dedicated factories where ingredients are automatically processed, packaged, and shipped out to franchise locations with minimal human interaction. This process has been automated for a long time due to the consistent and repetitive nature of ingredient preparation. Alternatively, at franchise locations there is a lot more manual labor with employees taking orders, preparing the meals, and delivering the food to customers. In the past, it was technologically challenging and highly expensive to try to implement automation in local franchises due to the complexity and nuance needed to handle local tasks. But with recent technological advances in AI and robotics, it is both cheaper and easier to implement automated systems into local franchises and restaurant chains are starting to chase this trend. If this pace is kept, it could have wide-reached industrial, ethical, and economical effects.

1.1 Number of Restaurant Workers

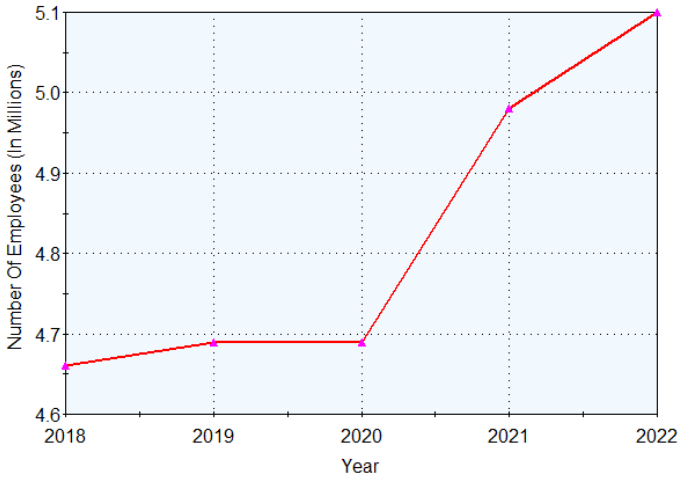


Table 1: Number of Fast-Food Employees Working in the United States [1]

As of 2021, there were 11 million people working in the restaurant industry [1], with almost five million working specifically in fast food, with this number continuously growing [2]. With an industry that affects this many lives, any huge shift could have ripples across all of society. With automation being known to drastically change how industries function, the recent trend of automating more and more restaurant processes forecasts potential drastic societal and economic repercussions. It is imperative that the general public be informed and well-educated on these potential outcomes so that they may be prepared for when they occur.

2 Emerging Examples

The automating of the food service industry is emerging in many different ways. One example is robotics companies such as Miso Robotics, a company based on creating robots that automate kitchen tasks. Another example is the food service giant, McDonald’s, which is experimenting with automating all aspects of customer interaction including taking orders and food delivery. There are even examples of some smaller restaurant chains in the midst of developing fully automated “ghost kitchens” that will be able to take orders, produce, and serve meals with minimal interaction needed. An example of an establishment that has already fully implemented automation, Mezli, is quoted to be the first fully automatic restaurant.

2.1 Miso Robotics

Miso Robotics is one of the most prominent robotics companies in the world of food service automation. They are best known for their robot named Flippy. Flippy first debuted in a restaurant in 2018 solely to flip patties on a grill when an AI deemed it the proper time. Since then, Flippy has evolved into being able to track and manage multiple deep-frying stations while using AI and visual machine learning to determine when food is properly cooked [3]. Many businesses have already started implementing Miso robots in stores with White Castle being a prominent example using specifically a Flippy model. Other examples are Panera trying out a robot called Sippy to moderate coffee temperature and volume and Chipotle experimenting with a Miso robot named Chippy in an attempt to automate the process of cooking tortilla chips in stores [4].

2.2 McDonald’s

In contrast to Miso robotics, McDonald’s seems to be experimenting with automating all of the tasks outside of the kitchen. Recently in Fort Worth, Texas, a mostly automated McDonald’s was opened specifically designed to minimize human interaction. When placing an order, a customer must either access the digital kiosks located in the small interior or go online and use the McDonald’s app. The drive-through window utilizes a conveyor belt system, promoting zero contact between customer and employee. This unique McDonald’s still contains a restaurant crew, though they mainly work in the kitchen. McDonald’s states this model of restaurant is designed for people who want to pick up their food and go [5].

2.3 Fully Automated “Ghost Kitchen”

Ghost kitchens are typically places that cook and prepare food, but then have it delivered somewhere for off-site consumption. They are a relatively new type of restaurant that lives and dies based on the nature of apps such as Grubhub and DoorDash [6]. Typically, ghost kitchens are only manned by a staff dedicated to cooking orders while leaving the handling and serving of orders to delivery apps. Companies such as Pizza HQ, Cala, Hyper-robotics, and Nommi are currently attempting to automate the kitchen side of the process [7].

The pizza franchise 800 Degrees is launching a new spinoff called 800 Go. It is set to be a robotic ghost kitchen hybrid almost like a pizza vending machine. They have partnered with the robotics company Piestro in order to design and produce the machines. They plan on opening 3,600 locations from 2021-2026. The more compact design and nature of these “ghost kitchens” are projected to have double the profit margin of one of their traditional venues [8].

2.4 Mezli

Mezli had its grand opening on August 28, 2022, and claims to be the world’s first fully autonomous restaurant. It was founded in 2021 by three Stanford engineers and is located in San Francisco. The restaurant features a highly customizable Mediterranean-themed menu with a large variety of grain and protein bowls. Ingredients are prepped every morning in an off-site kitchen and are then brought and inserted into the food truck-sized restaurant. The on-site restaurant has a digital kiosk where customers order and pay.

After receiving an order, the machine uses a high-tech oven to reheat the meat and then utilizes a conveyer belt to combine it with other ingredients within a bowl. It is then possible to pick up your order in small pickup windows located at the far side of the restaurant [9].

3 Potential Effects

The food service industry contains millions of jobs so understandably any potential shift would affect millions of people. When McDonald’s new mostly automated franchise hit the mainstream media, there was a lot of public backlash and criticism claiming that instead of raising wages, they simply replaced workers with autonomous machines [5]. While this may not be the exact reason, if the trend of replacing food service jobs with automated machines continues by expanding not only to other McDonald’s locations but also to other chains, it is highly likely that we will see increased job loss. On the contrary, ever since the 2019 COVID-19 outbreak, restaurants have been reporting labor shortages. They state that they received a drop in labor during the pandemic and that it has been a struggle to return to a number of employees that makes it comfortable to run their businesses. This is also a reason that a lot of fast-food cashiers have been replaced with electronic kiosks [10].

Both 800 Go and Mezli have stated interest and intent in their ability to rapidly expand using their new automated systems [8,9]. The ability to rapidly expand could change the food service industry as we know it and could quickly rise to be the next big food service trend. The convenience of fast food is only expanded upon by allowing for smaller, more autonomous experiences more akin to the convenience of a vending machine with the quality of a restaurant’s level of food.

References

[1] IBISWorld, "Fast Food Restaurants in the US - Employment Statistics 2005–2029," IBISWorld, [Online]. Available: https://www.ibisworld.com/industry-statistics/employment/fast-food-restaurants-united-states/. [Accessed 17 March 2023].

[2] S. R. Department, "Number of employees in the restaurant industry in the United States from 2010 to 2021," 1 August 2022. [Online]. Available: https://www.statista.com/statistics/203365/projected-restaurant-industry-employment-in-the-us/.

[3] Miso Robotics, "sec.gov," 31 12 2020. [Online]. Available: https://www.sec.gov/Archives/edgar/data/1710670/000110465921059344/tm2114854d1_partii.htm. [Accessed 18 2 2023].

[4] A. Lucas, "Why restaurant chains are investing in robots and what it means for workers," CNBC, 27 12 2022. [Online]. Available: https://www.cnbc.com/2022/12/27/restaurant-chains-are-investing-in-robots-bringing-change-for-workers.html. [Accessed 13 2 2023].

[5] P. Aitken, "McDonald's unveils first automated location, social media worried it will cut 'millions' of jobs," Fox Buisness, 24 12 2022. [Online]. Available: https://www.foxbusiness.com/technology/mcdonalds-unveils-first-automated-location-social-media-worried-will-cut-millions-jobs. [Accessed 20 2 2023].

[6] J. Miller, "What’s a ghost kitchen? A food industry expert explains," The Conversation, 1 6 2021. [Online]. Available: https://theconversation.com/whats-a-ghost-kitchen-a-food-industry-expert-explains-163151. [Accessed 3 3 2023].

[7] M. Wolf, "Our Ghost Kitchen Future Will Be Automated," The Spoon, 17 11 2021. [Online]. Available: https://thespoon.tech/our-ghost-kitchen-future-will-be-automated/. [Accessed 25 2 2023].

[8] J. Guszkowski, "800 Degrees unveils robotic pizza spinoff, 800 Go by Piestro," Restaurant Business, 12 11 2021. [Online]. Available: https://www.restaurantbusinessonline.com/technology/800-degrees-unveils-robotic-pizza-spinoff-800-go-piestro. [Accessed 25 1 2023].

[9] T. Huddleston, "These Stanford engineers built a fully autonomous restaurant in San Francisco that could make your lunch cheaper," CNBC, 26 8 2022. [Online]. Available: https://www.cnbc.com/2022/08/26/mezli-stanford-engineers-built-fully-autonomous-restaurant-in-sf.html. [Accessed 1 3 2023].

[10] J. Dorer, "Labor Shortages for Restaurants: A Look at A Long-Term Solution," QSR Magazine, 1 12 2022. [Online]. Available: https://www.qsrmagazine.com/outside-insights/labor-shortages-restaurants-look-long-term-solution. [Accessed 20 2 2023].

Investigating Curiosity in Student Text Data

Paul Meisner, Mitchell Hanson, Naeem Seliya,
Benjamin Fine, Rushit Dave*, and Mounika Vanamala

Computer Science Department
University of Wisconsin – Eau Claire,
Eau Claire, WI 54701
seliyana@uwec.edu

Abstract

We present a unique question-based student text responses analysis that can help instructors better identify what drives students to be more engaged in their learning. To determine the level of inquisitiveness among students, data is collected utilizing the Question Formation Technique. Data collection involves presenting students with thought-provoking QFocus statements, prompting them to formulate their responses in form of questions. The data is analyzed through Natural Language Processing, which is then analyzed using the WEKA machine learning tool. Feature selection is performed using filter-based feature rankers and wrapper-based feature subset algorithms. The course subject instructors determined that the extracted features provide meaningful insight into the "Propensity for Exploration" within the student text responses as a measure of their curiosity. Through an empirical mining of words/sentences that prove a curious disposition in text data produced by students in response to thought-provoking and critical thinking analysis, we obtained promising results, including an interesting distribution of results among the different applied feature ranker and subset algorithms.

* Department of Computer Information Science, Minnesota State University, Mankato.

1. Introduction

The skill of how to learn and apply new knowledge is a vital skill students need to develop. A student's curiosity in exploring a topic supports learning that knowledge [1], building upon what is taught in the given course. Curiosity has been associated with workplace learning and job performance [3]. Curiosity supports lifelong learning, a desirable outcome of students' education [2]. Given the benefits curiosity can have on self-directed learning and job performance, it is important to be able to identify whether students are exhibiting curiosity in the assignments and lab work.

Text mining has seen increasing focus on the investigation of sentiment [4], behavior analytics [5], linguistic understanding [6] improving product marketing [7], and pedagogical improvements [8]. Our project focuses on a relatively novel area, i.e., curiosity detection in text. This paper presents preliminary, yet promising, results of empirically mining words that demonstrate a curious disposition (of the students) in text data produced by students in response to thought-provoking and critical-thinking exercises. The success of our project could positively impact efforts to assess both curiosity and its impact on educational outcomes.

Grossnickle [9] has provided a framework for understanding facets, factors, and dimensions of the construct of curiosity that are relevant to the education audience. The key dimensions identified in the framework for curiosity include focus of curiosity (physical, perceptual, social, and epistemic), scope of curiosity (breadth vs. depth), cause of curiosity (diversive vs. specific or interest vs. deprivation), and consistency of curiosity across situational contexts (state vs. trait) [9]. Curiosity is positively linked to inquiry-based learning [10]. Questions are artifacts of curiosity. People consider children as curious when they ask many questions about a variety of topics, and particularly when they creatively combine ideas. Questions are posed to bring to light that which is unknown or not fully formed within the mind of the individual posing the question.

It seems reasonable to consider question data sets (especially from students) as a starting point for detecting a curious disposition. The datasets for our data mining approach to curiosity detection in students' text come from applying the Question Formulation Technique (QFT) [11], originally developed by the Right Question Institute [12], in an upper-level Artificial Intelligence (AI) course in an undergraduate computer science program. Previously, we performed relatively similar work for a lower-level Electric Circuits course in an undergraduate Electrical Engineering program. A portion of five class sessions in the AI course was utilized to obtain the QFT data, to improve students' ability to formulate questions and to support their curiosity on course topics. Compared to an expert examining student text data to determine curiosity levels, we envision our data mining solutions could provide substantial aid to experts. The solutions developed will be useful to detect whether curiosity is demonstrated in the results of the QFT exercises, provide analysis on key dimensions of curiosity, and potentially predict associated behaviors of students'.

The metric used in our study to assess the curiosity level of each student's question in the QFT data is "Propensity for Exploration". This metric is chosen because the dimension of curiosity that is most relevant to self-directed learning is the desire to identify knowledge gaps and seek out knowledge to close those gaps [14]. Propensity for exploration (PE) attempts to capture the identification of knowledge gaps and demonstration of some understanding of the landscape of the topic, which supports curiosity and the desire to seek out the knowledge [14]. Specifically, PE considers the degree to which the question identifies characteristics of, or layers within, the subject of the question, the degree to which relationships between the primary subject and other topics is identified, how relevant those characteristics and relationships are, and how well the question directs the attention of the audience within the landscape of the topic. Each question in the dataset is labeled as belonging to one of two categories for PE: 1 (Low) and 2 (High).

In a given dataset (student text responses set as per the QFT process), all unique words (tokens) are considered as features or attributes, after removing general stop-words typically observed in text data. Feature Selection (FS) methods have been applied to reduce the high dimensionality of the obtained datasets. We investigate five different FS and they include: two wrapper-based feature subset selection methods and three filter-based feature ranker techniques. The wrappers involve the C4.5 decision tree classifier with the BestFirst and GreedyStepwise search algorithms, while the filters consist of the ChiSquared, ReliefF, and GainRatio algorithms. The algorithm and associated parameter details for the C4.5 classifier and the five feature selection methods considered in our study is provided in [15]. Each FS method provides a reduced set of features for domain experts to examine to determine whether the selected features are indeed correlated with the different PE levels.

The important conclusions determined from our case study are that the two wrapper-based algorithms tend to yield the same feature subsets, and the three filters provide relatively less similarity in general. Among all five feature selection methods examined, GainRatio and ChiSquared are determined as the best approach for our case study, because it identifies words relevant to the subject that highly correlate to a particular level (class) of PE even if they are sparsely represented in the dataset. We note that like most machine learning-based studies, the case study results are determined on the underlying dataset and the algorithms investigated. Our proposed approach, however, can be applied to other curiosity exercise datasets as well, and provide the relevant experts a better insight into the student data.

The rest of the paper is structured as follows: Section 2 details the case study methodologies including QFT, feature selection, modeling approach, and data preparation and processing; Section 3 presents and discusses the various results obtained from our case study; Section 4 concludes our paper with a brief summary of the work done and some directions of future work.

2. Methodology

2.1 Question Formulation Technique

A student's ability to formulate insightful questions is a critical life skill that enables the student to engage with the content for a deeper understanding and learning [21]. Questions serve the purpose of making clear and concrete that which is unknown or misunderstood by the student. By making the unknown concrete, a pathway for exploration, engagement and learning is opened to the student. As the student engages with the resources needed to answer the question, inevitably more questions are formed and new connections between topics are discovered. This process of questions driving deeper inquiry and learning is the premise of question-driven learning [1], sometimes referred to as inquiry-based learning [16].

Question-driven learning is hypothesized to stimulate curiosity and supports problem solving [1]. It has also been combined with Problem-Based Learning (PBL) in [17] to examine the role students' questions played in driving their learning. Students actively contribute to the development of a biology course through the questions they pose in [18]. Beatty et al. [19] present an audience response system combined with question-driven instruction to engage students in active knowledge building, as the instructor uses real-time formative feedback to tailor the classroom experience to student inquiry. An adaptive, question-driven intelligent tutoring system is developed and discussed in [20].

One framework that allows students to engage in question formulation as an exercise is the Question Formulation Technique (QFT) [11]. The QFT has been developed by the Right Question Institute [12] to empower students with the ability to formulate relevant and specific questions. The QFT involves a combination of (1) divergent thinking, (2) convergent thinking, and (3) metacognition, and is designed to be a collaborative exercise, ideally with groups of four students. One student should be selected as the recorder to record the generated questions.

The first stage of the QFT is called question-storming, in which the students generate as many questions as possible on a topic in a specified amount of time. The mode of thinking utilized during this stage is divergent thinking, as the students spontaneously form questions based on a prompt as soon as the question comes to mind. The prompt that introduces the topic to the students is called the question focus (or QFocus). The QFocus can be a statement, quote, set of images, video, audio clip, or any other type of prompt that sets the students on the path of generating questions on the desired topic. Typically, the QFocus is a provocative or outrageous statement, such as "Torture can be justified" [11]. Sometimes, it may be selected to emphasize a conceptual conflict, such as "For an RC circuit, forever is just five time constants away" [13]. There are four essential rules that govern the question-storming process to motivate the students stay on task of generating many questions while also encouraging a safe, inclusive space [11]. The first rule is to produce as many questions as possible in the allotted time. The second rule is to write the questions exactly as stated (including grammatical errors). The third rule is to not discuss or judge the quality of the questions during the question-storming process. The final rule is to try to formulate everything as a question. The instructor's primary function during the question-storming process is to encourage the students to adhere to the rules and cajole groups that are slow to generate questions. The instructor should not judge the quality of questions, neither with constructive feedback nor praise, as it undermines the divergent thinking process.

At the end of the question-storming process, the group should have many questions, some of which may be similar or complementary. The second stage is question refinement in which students eliminate equivalent questions, combine complementary questions to formulate multifaceted questions, eliminate grammatical errors, and generally improve the questions. This process involves convergent thinking, as students must analyze the questions to see how to improve the set of questions. The third stage is question prioritization. The instructor should provide some criterion or set of criteria on which to prioritize the questions. Some options include propensity for exploration, relevance to the topic, importance to the topic, question complexity, or level of student interest. The criteria selected by the instructor should be related to the desired purpose for which the questions will be used, e.g., a research paper, design project, or topic motivation [13].

2.2 Feature Selection Techniques

In machine learning, the typical task is to model a learner with the given dataset to predict a target feature (related to the given domain) based on a given number of predictor features. In the case of a dataset with a very large number of features and/or with the presence of data noise (especially feature noise), a feature selection (FS) process is performed prior to building the final predictive model. The former is applicable to our study where a token-based feature importance approach is taken, as explained in the Section 2.3 of this paper. We investigate five different FS approaches commonly used in the data mining domain, and they include [15]: two wrapper-based feature subset selection methods and three filter-based feature ranker techniques.

The wrapper-based approaches work by using a search algorithm (e.g., BestFirst and GreedyStepwise) to find a subset of features that collectively defines the performance of the classification model. A classifier is built using a given feature subset and evaluated using a performance metric. The classifier used in our study is the C4.5 decision tree, and the performance metric used is the Area Under the Receiver Operating Characteristic curve (AUROC). The ROC curve plots the true positive rate versus the false positive rate, for a given class. The wrappers yield a feature subset that collectively provide the best classification performance. Therefore, the size of the subset can vary for each dataset and no elements of the subset may be removed when building the classifier and presenting the results.

The filter-based feature ranker techniques consist of the ChiSquared, ReliefF, and GainRatio algorithms [15]. Rather than providing a subset of features as in the case of the wrapper approaches, the filters provide an ordered rank list of all the features from the best to the worst, based on a given performance metric. The ChiSquared attribute evaluator in Weka evaluates the worth of an attribute by computing the value of the chi-square statistic with respect to the class attribute. GainRatio is a modification of Information Gain by reducing its bias on highly branching features. It considers the number and size of branches when choosing a feature. This is done by normalizing information gain by the Intrinsic Information, which is defined as the information needed determine the branch to which an instance belong (the class label). The ReliefF algorithm computes a feature score by using the identification of feature value differences between nearest neighbor instance pairs. A "hit" occurs when a feature value difference is observed in a neighbor instance belonging to the same class, yielding a reduction in the feature score. Conversely, a "miss" occurs when a feature value difference is noted in a neighbor instance belonging to a different class, yielding an increase in the feature score. The distance function and the number of nearest neighbors is the key variants for ReliefF.

The open-source WEKA data mining and machine learning tool is used to implement our case study experiments, including the training of the classifiers and implementing the five feature selection algorithms [15]. In our study, all parameters other than the specific feature selection algorithms used and C4.5 classifier for the wrapper-based approaches in the Weka tool are set to default.

2.3 Modeling Methodology

2.3.1 Data Collection

A 400-level course in Artificial Intelligence (AI) was considered for our data collection purposes. The QFT methodology was applied to obtain the question-based responses to five QFocus statements (labeled in the form of, Qx), and they are:

Q1. AI did my homework. I did not cheat.
Q2. AI is the worst thing to happen to law enforcement.
Q3. AI has a singular moral code.
Q4. AI creates a more equitable job market.
Q5. AI algorithms should discriminate.

The data collection from the three stages of the QFT methodology were obtained from students of the course. To get a relatively decent size of dataset we focus our analysis and case study experiments on the questions obtained from stage two (question refinement) of the QFT methodology. To our knowledge there is no direct measure to evaluate a student's curiosity degree, thus, we use an associated concept, Propensity of Exploration (PE), as a measure to provide insight into a student's degree of curiosity. A panel of three domain experts (two faculty and a senior-class student) evaluated the stage two questions for their PE potential and scored them as either: Low (1) and High (2). The scoring of Low and High of the target feature PE, is used during the feature selection process for finding the tokens (words) in students' questions that best reflect the different levels of PE, and thus, different degrees of curiosity. Thus, the PE scoring values are used as categories or classes to group the different question-based answers students developed in stage two.

2.3.2 Data Preparation and Processing

The question-based answers collected from participants was digitized into a Microsoft Word Document. We used the respective session number (each QFocus statement session) to label each document created. A final document was created in which we formatted the questions given by students by changing capital letters into lowercase letters and by removing any numbers, punctuation, and special characters. Subsequently, as mentioned earlier, after completing the data digitization process, each question is labeled as either Low or High according to their potential for Propensity of Exploration.

Every question in the dataset was analyzed to find the unique words it had. Toward this goal, we created a python script would open the Word documents using the docx Python library and find all the words in each question in the entire document. The QFocus prompt at the start of the document and the dashes to separate the prompt from the questions generated by the students are exempted from the word search. All the words were put into an array variable called "word_list" which was then looped through to find all the stop-words in the document. We used the stop-word list from the Natural Language Toolkit (NLTK) python package. This package can be installed with "pip install nltk.” We compare every word in "word_list" to the words in the stop-word text file and added the non-stop words into another array variable called "filtered_words.” This array variable still contained the PE ranking at the end of every question. To find all the unique words in the “filtered_words" array, we put every unique word into another array variable called "unique_words.”

Lastly, we created two variables to calculate the number of unique words used per question and an array to write the occurrence of unique words used in a single question to a CSV file. We created a dictionary in which the unique word was set as the key, which defined the number of occurrences per question (called "num_words_dict"). A second variable, “num_words,” was created with the intention of storing the occurrences of unique words from the "num_words_dict" in an array. To create and write into a CSV file, the CSV python library was used to insert the question number, occurrences of unique words in a question, and the PE label for the question for each QFocus session.

2.3.3 Modeling for Feature Selection

The WEKA data mining and machine learning too was used to conduct the feature selection experiments in our case study. As elaborated previously, the collected data was converted from text data into numerical form based on a text-to-tokens transformation and using a standard stop-words list. Each unique word that is not in the stop-words list is referred to as a token. Through feature selection, we obtained meaningful tokens (insightful for the PE metrics) and were able to reduce the sparsity of the high-dimensional sparse data set. The feature selection process was based on three filter-based feature rankers and two wrapper-based feature subset selection algorithms. The rankers included Chi-squared, GainRatio, and ReliefF, while the wrappers included BestFirst and GreedyStepwise search algorithms with the C4.5 decision tree as the classifier and AUROC as the performance evaluation metric.

3. Case Study Results

The frequencies of the Low-PE and High-PE question-based students text responses are shown in Figure 1, which shows the data for each of five QFocus statements, Q1, Q2, Q3, Q4, and Q5 (the QFocus statements are provided in Section 2.3.1). While there is not clear cut pattern across all the Qs, in general the High-PE questions are in higher numbers relative to the Low-PE questions. The exception being Q1, which could be reflective of students being new to the QFT process and in general responded with Low-PE questions.

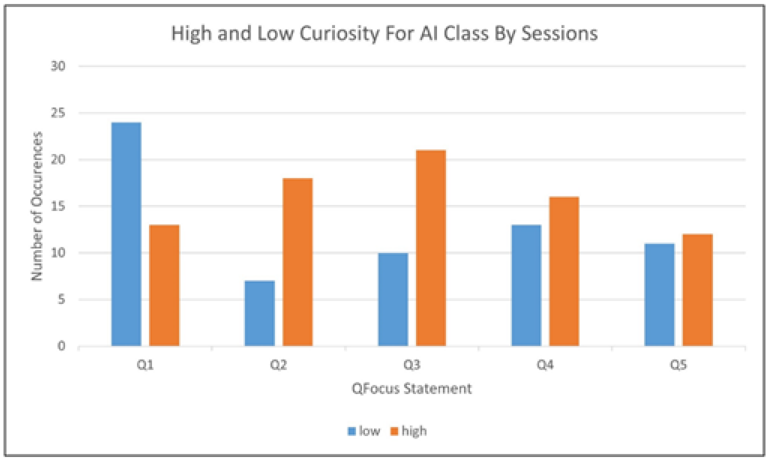


Figure 1: The PE (Curiosity Degree) Frequency of Low/High Questions

Best-First Search Based Wrapper FS				
Q1	**Q2**	**Q3**	**Q4**	**Q5**
use	ai	defines	used	discrimination
might	use	---	---	algorithms
build	---	---	---	good
homework	---	---	---	bias
GreedyStepwise Search Based Wrapper FS				
Q1	**Q2**	**Q3**	**Q4**	**Q4**
use	ai	defines	used	discrimination
might	use	---	---	algorithms
---	---	---	---	good

Table 1: Feature Subset Selection by the Wrappers.

Chi-Squared Ranker				
Q1	**Q2**	**Q3**	**Q4**	**Q5**
use	scenario	could	done	algorithms
might	ai	ai	remove	could
copyright	used	act	bids	people
cheating	maliciously	immoral	ai	ai
rules	times	way	specifically	begins
generated	breach	still	job	discriminate
considered	citizens	following	market	ethical
problem	privacy	code	certain	use
comes	ways	likely	sectors	oversees
person	people	scenario	biased	negative
Gain Ratio Ranker				
Q1	**Q2**	**Q3**	**Q4**	**Q5**
use	scenario	could	done	algorithms
might	ai	ai	remove	could
copyright	used	act	bids	people
cheating	maliciously	immoral	ai	ai
rules	times	way	specifically	begins
generated	breach	still	job	discriminate
considered	citizens	following	market	ethical
problem	privacy	code	certain	use
comes	ways	likely	sectors	oversees
person	people	scenario	biased	negative
ReliefF Ranker				
Q1	**Q2**	**Q3**	**Q4**	**Q5**
use	enforcement	define	might	discrimination
might	ai	gain	used	algorithms
past	worse	term	well	things
model	things	would	kinds	discriminate
less	cars	like	things	bias
programming	self	face	algorithms	thing
assignments	driving	follows	spell	good
give	scenario	scenario	different	avoid
teach	something	irobot	others	calls
study	hindering	us	make	train

Table 2: Features Selected by the Three Filter-based Rankers.

The feature subsets selected by the two search algorithms of the wrapper-based feature selection approach is shown in Table 1. The table shows the selected feature subset for each of the five QFocus statements, where the top half of the table represents the Best-First search algorithm's results while the bottom half of the table represents the GreedyStepwise search algorithm's results. Recall that a wrapper uses a classifier and a classification performance metric during its feature subset selection process. In our study we used the C4.5 decision tree classifier and the AUROC performance metric. Once a wrapper-based feature selection is done, any machine learner can be used with the selected feature subset to train and evaluate classifiers.

From Table 1, we observe that the two feature subset search algorithms generally yield similar or identical results. In the case of Q2, Q3, and Q4, the feature subsets are identical, and in the case of Q1 and Q5, the feature subsets are relatively similar. A deeper look at the features selected for each QFocus statement, we can notice interesting observations. For Q1 ("AI did my homework. I did not cheat."), the features are strongly reflective of the meaning of the statement in addition to extracting the "homework" token from the statement as an important feature. The tokens, "use" and "might," are semantically relevant to the Q1 statement, e.g., "I might have cheated." Similar observations can be interpreted from looking at the features selected for the other QFocus statements. For example, in the case of Q5, key tokens from the statement itself are observed as strong predictive features by the two wrapper-based feature selection methods. Finally, the tokens, "use" or "used", occur frequently in the table, which is intuitive given the different QFocus statements.

The feature selection results of the three filter-based rankers are shown in Table 2. After the word-to-vector tokenization process in our case study, the feature dimensionality was very large compared to the data point dimensionality. And since a filter-based ranker orders the different features from best to worst based on the predictive capability, we select the top 10 features to focus upon in this case study. This was done because in general the top 10 features, for a given ranker and QFocus statement, yielded the highest performance metric of the respective ranker. Among the features ranked by the three filters, as shown in Table 2, we observe that GainRatio or Chi-Squared provided identical token (for a given QFocus statement) both in features and their respective rankings. This was also observed in our previous study [21], where the QFT process was conducted for an Electric Circuits lower-level undergraduate course. The ReliefF ranker provided somewhat different top 10 features and their respective ranking for the different QFocus statements. With ReliefF, in general, the respective QFocus statements, both in terms of words and their semantics, yielded features that were intuitively or directly reflective of the respective statements. For example, with Q2, tokens such as "enforcement", "ai", and "worse" have a direct correlation with the QFocus statement, while tokens such as "hindering" and "scenario" provide a more semantic-based correlation with the QFocus statement.

4. Conclusion

The paper investigates data mining and machine learning techniques toward providing an insight into predicting the degree of curiosity a student has for a given course-related topic. To determine the level of curiosity among students engaging in an upper-level Artificial Intelligence course in an undergraduate computer science program, data is collected utilizing the Question Formation Technique. The latter collects text responses from students via a process with three stages, namely, divergent, convergent, and prioritization.

Data collection involves presenting students with thought-provoking five different QFocus statements, prompting them to formulate their responses in form of questions. The data is analyzed and interpreted through NLP for which Python-based scripts are developed toward an efficient organization of the student text responses, which is then analyzed using the WEKA data mining and machine learning tool. Feature selection is performed using three filter-based feature rankers and two wrapper-based feature subset algorithms. The course subject instructors determined that the extracted features provide meaningful insight into the “Propensity for Exploration” within the student text responses as a measure of their curiosity degree levels.

The five QFocus statements included: “AI did my homework. I did not cheat.”; “AI is the worst thing to happen to law enforcement.”; “AI has a singular moral code.”; “AI creates a more equitable job market.”; and “AI algorithms should discriminate.” For the case study presented the best results were obtained with the Gain Ratio and ChiSquared filter-based rankers. While providing important features, the wrapper-based feature subset selection process yielded fewer tokens for the domain experts to analyze and evaluate the degree of curiosity (PE) in student text responses. Our future work will include performing the QFT and machine learning based approach presented here for other courses, both computer science courses and non-computer science courses.

References

[1] D. L. Schwartz, J. M. Tsang, and K. P. Blair. The ABC’s of How We Learn: 26 Scientifically Proven Approaches, How They Work, and When to Use Them. W. W. Norton & Company, Inc., New York, NY, 2016.

[2] M. J. Kang, M. Hsu, I. M. Krajbich, G. Loewenstein, S. M. McClure, J. T.-Y. Wang, and C. F. Camerer, “The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory,” *Psychological Science*, 20(8), pp. 963-973, 2009.

[3] T. G. Reio Jr. and A. Wiswell, “Field investigation of the relationship among adult curiosity, workplace learning, and job performance,” *Human Resource Development Quarterly*, 11(1), pp. 5-30, 2000.

[4] J. Prusa, T. M. Khoshgoftaar and N. Seliya, “The Effect of Dataset Size on Training Tweet Sentiment Classifiers,” In *Proceedings of the 14th IEEE International Conference on Machine Learning and Applications* (ICMLA), pp. 96-102, Miami, FL, 2015.

[5] A Singh, J Ranjan, and M Mittal, “Big Data and Behavior Analytics”, Handbook of e-*Business Security*, Ch. 9, 2018. Taylor & Francis, CRC Group.

[6] P. Kendeou, P. Broek, A. Helder, and J. Karlsson J, “A Cognitive View of Reading Comprehension: Implications for Reading Difficulties,” *Learning Disabilities Research & Practice*, 29(1), pp. 10–16, 2014.

[7] L. McGarrity, “What Sentiment Analysis Can Do for Your Brand?” *Marketing Profs*, April 2016.

[8] A. E. Barth, S. Vaughn, P. Capin, E. Cho, S. Stillman-Spisak, L. Martinez, and H. Kincaid, “Effects of a Text-processing Comprehension Intervention on Struggling Middle School Readers,” *Topics in Language Disorders*, 36(4), pp. 368-389, 2016.

[9] E. M. Grossnickle, “Disentangling Curiosity: Dimensionality, Definitions, and Distinctions from Interest in Educational Contexts,” *Educational Psychology Review*, 28(1), pp. 23-60, 2016.

[10] T. J. van Schijndel, B. Jansen, and M. Raijmakers, “Do individual differences in children’s curiosity relate to their inquiry-based learning?” *International Journal of Science Education*, 40(9), pp. 996-1015, 2018.

[11] D. Rothstein and L. Santana. Make Just One Change: Teach Students to Ask Their Own Questions. Harvard Education Press, Cambridge, MA, 2015.

[12] The Right Question Institute, www.rightquestion.org.

[13] H. J. LeBlanc, K. Nepal, and G. S. Mowry, “Stimulating Curiosity and the Ability to Formulate Technical Questions in an Electric Circuits Course Using the Question Formulation Technique (QFT),” *IEEE Frontiers in Education Conference* (FIE). Indianapolis, IN. October 2017.

[14] G. Loewenstein, “The psychology of curiosity: A review and reinterpretation,” *Psychological Bulletin*, 116, pp. 75-98, 1994.

[15] E. Frank, M. A. Hall, and I. H. Witten, The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[16] M. Pedaste, M. Mäeots, L. A. Siiman, T. De Jong, S.A. Van Riesen, E.T. Kamp, C.C. Manoli, Z.C. Zacharia, and E. Tsourlidaki, “Phases of inquiry-based learning: Definitions and the inquiry cycle,” *Educational Research Review*, 14, pp. 47-61, 2015.

[17] C. Chin and L.-G. Chia, “Problem-based learning: Using students' questions to drive knowledge construction,” *Science Education*, 88(5), pp. 707-727, 2004.

[18] M. Shodell, “The question-driven classroom: student questions as course curriculum in biology,” *The American Biology Teacher*, 57(5), pp.278-281, 1995.

[19] I. D. Beatty, W. J. Leonard, W. J. Gerace, and R. J. Dufresne, “Question driven instruction: teaching science (well) with an audience response system,” In *Audience response systems in higher education: applications and cases*, pp. 96-115. IGI Global, 2006.

[20] P. J. Muñoz-Merino, M.F. Molina, M. Muñoz-Organero, and C.D. Kloos, “An adaptive and innovative question-driven competition-based intelligent tutoring system for learning,” *Expert Systems with Applications*, 39(8), pp. 6932-6948, 2012.

[21] N. Seliya, H. LeBlanc, B. Hylton, Z. Youssfi, and M. Schweinefuss. Examining Trends in Curiosity Levels of Student Text-Based Questions. In *Proceedings of the 2019 American Society for Engineering Education Conference*, Tampa, FL, June 2019.

Video Interpolation and Extrapolation using a Transformer with Relative Positional Embedding & Relative Global Attention

Autumn Beyer Mitchell Johnstone Sam Keyser Ryan Kruk
Tyler Schreiber Tillie Pasternak Michael Conner

Department of Electrical Engineering & Computer Science

Milwaukee School of Engineering

{beyera, johnstonem, keysers, krukr, schreibert, pasternakt, connerm}@msoe.edu

Abstract

Video frame interpolation is a popular technique used to increase the frame rate of a video sequence, resulting in smoother and more fluid playback. This process involves generating intermediate frames between existing ones, which fill in the gaps and produce a more natural and visually appealing video. The goal of this technique is to estimate motion information between frames and synthesize new pixels to fill in the gaps. This is typically achieved using convolutional neural networks (CNNs) that have been trained on large datasets of videos. In recent years, transformers have been used for video interpolation tasks, showing significant progress in this field. However, there is still limited knowledge on the use of relative positional embeddings to help capture more complex relationships between different sections of frames based on position. To address this gap, our research investigates the use of relative positional embeddings in video frame interpolation and extrapolation. Our goal is to capture complex spatial relationships between frames in a video sequence that can improve the accuracy and quality of interpolated frames. In addition to exploring the use of relative positional embeddings in video frame interpolation, we also investigate their effectiveness in video extrapolation. Video extrapolation involves generating new frames beyond the end of a given video sequence, which is a challenging task due to the lack of visual information available. By using relative positional embeddings, we aim to capture the spatial relationships between frames in both the forward and backward directions, which can lead to more accurate and realistic extrapolation results. To conduct our experiments, we use the Vimeo90K dataset, which is widely used in the field and allows for easy comparison with other models. Our research contributes to the growing body of knowledge on the use of transformers in video processing and

provides new insights into the potential benefits of relative positional embeddings. Video frame interpolation and extrapolation are essential techniques used in various applications, such as video compression, slow-motion effects, and video enhancement. Our research aims to improve the accuracy and quality of these techniques. In conclusion, we believe that our research will provide new insights into the use of relative positional embeddings in video processing and contribute to the development of more effective and accurate video frame interpolation and extrapolation methods.

1 Introduction

Video interpolation is the process of generating a frame in between two consecutive frames. This process increases the frame rate of a video, allowing users to create slow motion videos in the native frame rate or create a smoother video with more frames per second. Beyond the obvious applications to media, video interpolation also holds promise for better data compression. If high quality video interpolation can be achieved, videos can be stored as a set of several key frames, with the intermediate frames being recoverable via interpolation.

Video extrapolation, on the other hand, aims to generate a new frame after being given a sequence of some number of frames. The traditional application of video extrapolation is for video prediction, which is predicting the next image given a sequence of preceding images. Extrapolation can also be used for novel view synthesis (NVS), which is the general task of trying to generate a view of a scene given some number of complementary views [3]. An example might be trying to generate the side profile of a car, given a view of the car at 45 degrees.

Recently transformers have been applied to the problem of video interpolation. Zhihao et al. recently published a video interpolation model, the Video Frame Interpolation Transformer, built on top of the transformer architecture which achieved state of the art results. In this paper we will evaluate how their base model behaves for the task of video extrapolation, as well as adding relative global self-attention and relative positional encoding.

While video interpolation and extrapolation can produce impressive results, the quality of the output frames can be further improved with the use of advanced techniques such as relative global attention and relative positional embedding. Relative positional embedding helps the model establish an improved spatial relationship between relative location of pixels within a frame. By having this established relationship, the model’s output will generate a more accurate results when frames contain multiple dynamic objects.

In Summary, the integration of relative global attention and relative positional embedding can improve the quality of output frames in video interpolation and extrapolation by capturing long-range dependencies and establishing improved spatial relationships between pixels within a frame, leading to more accurate predictions and smoother, more natural-looking video sequence.

2 Background

2.1 Transformer Architecture

The transformer is a deep neural network architecture originally introduced for sequence transduction in the 2017 paper “Attention is all you Need” by Vaswani et al [4]. The transformer uses the same encoder-decoder scheme used by previous sequence transducers but introduces the concept of “self-attention”. Self-attention allows the model to weigh individual parts of the input when generating its prediction. This mechanism is expanded to multi-head attention, which computes the attention several times in parallel. This allows the network to “attend” to, or pay attention to, several parts of the sequence at once.

Position embeddings are a way of incorporating positional information into the input representation, by adding a fixed-length vector to each token's embedding that encodes its position within the sequence. These embeddings are typically learned during training and are often represented as sinusoidal functions of different frequencies and offsets.

Beyond basic self-attention, there are several variations on attention which have been introduced. We describe a few which we used in our own architecture.

Global self-attention is a variant of self-attention that computes attention weights between all pairs of positions in the input sequence, instead of just between adjacent positions. This enables the model to capture long-range dependencies between different parts of the sequence, which can be particularly important in tasks like language translation or summarization.

In contrast, cross-attention computes attention weights between positions in different sequences, allowing the model to attend to different parts of the key sequence based on the information in the query sequence. This mechanism is useful for tasks like machine translation or multimodal learning, where the model needs to attend to multiple sequences or modalities.

Causal attention is a variant of self-attention that only allows the model to attend to positions that come before the current position in the input sequence. This is important in tasks like language modeling or text generation, where the model must generate output one step at a time based on previous output and prevent the model from "cheating" by attending to positions that come after the current position.

2.2 Patches

Transformers are trained on a sequence of frames to create the next element. This makes sense in terms of image extrapolation, as we would be taking two images and predict the next.

However, evaluating the images without modification has some issues. Mostly, it’s that a computer has difficulty in maintaining context. For example, if the input images were evaluated wholly, it would be difficult to capture regional movements within an image. To improve this process, 16x16 pixel patches are extracted from the input images and fed to the transformer as the input sequence.

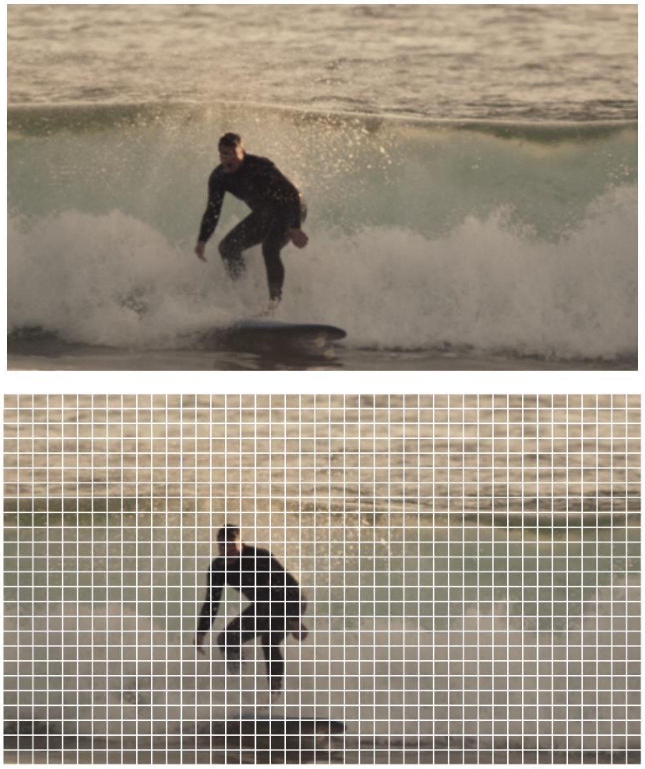


Figure 2: Sample input image before patch encoding, then after patch encoding

This allows the transformer model to capture regional movements more finely, as it can focus on individual components within the images.

2.3 Relative Positioning

Transformers by default use absolute positional embedding, which just encodes the absolute position of a token within the overall sequence. However, this approach loses the information encoded by the positions of each token relative to each other, as well as enforcing a maximum sequence length. Shaw et al. introduced a method for using the pairwise distances to create positional encodings in the paper “Self-Attention with Relative Position Representations” [2]. The pairwise distances get added to the keys during the computation of attention,

$$e_{ij} = \frac{x_i W^Q \left(x_j W^K + a_{ij}^K\right)^T}{\sqrt{d_z}} \ (1)$$

and then again as a subcomponent of the values.

$$z_i = \sum_{j=1}^{n} \alpha_{ij} \left(x_j W^V + a_{ij}^V\right) \ (2)$$

We don’t apply the second formula in our implementation as it was found to not have a significant effect on our results.

Computing the relative positional embeddings requires O(L^2D) memory where L is the length of the sequence and D is the hidden state size.

Huang et al. use a trick called “skewing” to efficiently compute the relative positional embedding with ever expanding a [1]. We will not replicate their full method in this section, but they were able to lower the memory footprint to O(LD), which makes it feasible to train with a much longer sequence length. We use the skewing method in our implementation.

3 Our Architectures

The first model explored was a baseline transformer architecture. Custom implementations were created for the cross-attention, global-attention, and causal self-attention. From there, the rest of the transformer architecture was written to take the image sequence as patches. This gave a baseline model to base our results on.

To try to modify the models from prior implementations of image extrapolation and interpolation using a transformer model, relative attention was implemented to attempt to see if the relative positioning of the patches in the input images were relevant to the output images. The logic behind this decision was that the positioning of patches within an image would be important to their context, so incorporating the position of the patch in the attention equation should capture correlating movement relations.

Our model was inspired by the Relative Global Attention discussed in the paper “Music Transformer” [1]. While the context is different, as that particular paper focused on music note sequences and our model used image patch sequences, the application is relevant as both music notes and image patches use neighboring components to provide context.

4 Results

After comparing the two models, our current evaluation has yielded inconclusive results with regards to the performance of the two models under consideration. Specifically, we have not been able to identify any significant differences between their results. The model incorporates relative global attention, as opposed to the base attention. This does not appear to exhibit any substantial divergence in terms of its efficacy or accuracy when compared to its counterpart.

Moreover, our experiments have led us to believe that both models generate similar results when presented with the same set of data. While further tests may be necessary to validate these initial findings, our current analysis indicates that there is no marked discrepancy in the performance or accuracy once relative global attention is used.

5 Further Work

This paper explored some basic interpolation and extrapolation applications using basic and relative encoding attention. One extension that could be done using extrapolation is to prolong the input sequence by using the output image as the next term in the sequence. In doing so, one could propagate the extrapolation to generate entire videos from a starting sequence of images, continually updating the sequence in a recurrence style.

Similar to the extrapolation to generate a new image, repeated interpolation could provide higher quality slow-motion videos. These may not match up to the real motions of the object due to some sampling issues, but for some motions it may be able to properly describe the real function of the real actions.

The relative positioning incorporated in the relative global attention is useful for sequences of data, such as strings of words. One issue in the current setup is that the patches generated from the image are strung out, meaning that there is no vertical association between the patches. A potential exploration is modifying the relative positioning to incorporate the vertical position as well, potentially using a metric such as the Manhattan Distance between patch locations.

Otherwise, giving improper images may create interesting results. It would be interesting to see the different effects of extrapolation and interpolation on the same set of images. For example, if the input images were of a house and a rainbow, does the extrapolation make something completely different? Does interpolation create a morphing house and rainbow? More work will have to be done to see the effects of these processes.

References

[1] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, Douglas Eck: "Music Transformer", 2018; arXiv:1809.04281.

[2] Peter Shaw, Jakob Uszkoreit, Ashish Vaswani: "Self-Attention with Relative Position Representations", 2018; arXiv:1803.02155.

[3] Yunzhi Zhang, Jiajun Wu: "Video Extrapolation in Space and Time", 2022; arXiv:2205.02084.

[4] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, Ming-Hsuan Yang: "Video Frame Interpolation Transformer", 2021; arXiv:2111.13817.

Constructing a UX Testing Platform using Embedded Computing Systems

Ariana Beeby Erik Steinmetz
Department of Math, Statistics, and Computer Science
Augsburg University
2211 Riverside Ave, Minneapolis, MN 55454
{beebya|steinmee}@augsburg.edu

Abstract

In this work we demonstrate how to create a platform for UI/UX experiments based on a Raspberry Pi computer configured with a touchscreen display.

User interface studies are often conducted by personal observation of a user engaging with a piece of software. Many kinds of experiments can be conducted without the need for human oversight at the time of the experiment. This work is intended to build a device to monitor user interactions in an unsupervised kiosk mode, providing a platform for these kinds of experiments.

The work presented in this paper explains the hardware configuration and software stack used to set up the kiosk. We present a detailed look at the software design, including which programs were chosen along with the configuration settings necessary for the software and hardware components to combine and create a versatile experimentation platform.

1 Introduction

The goal is to set up the Raspberry Pi as a kiosk allowing users unsupervised interactions with the computer in a controlled and limited environment. This environment enables UI/UX experiments to be conducted such as observing First Click placement and length of engagement. To allow entirely unsupervised user interactions, the machine will run software which automatically records and documents desired user events. It will also have a "default" display mode to which it returns after a set amount of time, reducing the effects of a previous user's interactions on future engagements. This eliminates the need for a human observer of the user, whose presence may interfere with the natural flow of the subject's interaction with the software. A machine may also be able to offer a more accurate and consistent record of the interactions than a human observer.

2 Background

User interface and experience testing can be broken down into a number of unique tests that focus on individual elements necessary for the user experience, including both performance and quality-focused tests. Some examples include task completion, clickstream analysis, and First Click testing [1][2]. With some of these tests, a moderator is required to have direct interaction with participants as they are guided through each task, requiring more time and resources to be put towards testing. The presence of a moderator could also have a undesired effect on the outcome of each test as well. Other tests only require observation of the participants and their interactions with the software. For First Click testing, the participant's first interaction with a given situation is recorded, with a desired outcome on the tester's part. This results in the tracking of correct and incorrect interactions based off the first click [1][2].

3 Development Process and Architecture

The software created for the kiosk setup of the Raspberry Pi is based on a tutorial guide outlined on the official Raspberry Pi website [3]. A bash script is created that blocks aspects of the desktop environment and cursor allowing for a clean display – kiosk mode – and opens a specified website on the Chromium browser. In order to run the data collection software created specifically for this kiosk, a command was added to the script which opens and runs a python script in the background. Completing the setup a service file is created that ensures the device will boot automatically in its kiosk mode by executing the bash script on power-up.

The focus of the data collection in the system described in this paper is purely to get the locations and times of the interactions with the kiosk system. Unlike First Click testing, data collection will continue after the first interaction, allowing for analysis of the overall experience, including engagement time.

For the creation of the data collection software, a library that could monitor input events from the screen on the site was required. A number of libraries were explored, including xdotool, a shell command that is used in the kiosk bash script. When initially implemented, xdotool's tracking of the mouse-click behavior did not detect any touches, but the mouse-enter behavior would detect when the cursor would briefly appear on screen on touch. This resulted in the logging of two duplicate events, as the appearing and disappearing of the cursor on touch counted as separate mouse-enter behaviors.

Ultimately, the Pynput [4] and logging Python libraries were utilized. Pynput is a high-level library that allows for the simulation and monitoring of keyboard and mouse events. Though not extensive in its functionality, pynput fits the needs of the data-collection this device requires, and has streamlined usability. Similar to xdotool's mouse-click behavior, pynput's click detection does not read a touch on the screen as a click. However, as the cursor only appears on touch, the event is detected as a movement of the mouse. Unlike xdotool, a touch event is only detected once and at the moment the initial contact is made, or the cursor is first moved. This eliminates the logging of redundant information. The Python logging library was used to record these events and the specific time of the events to a logging file. This file can be written and read while the monitoring program is active, allowing the device to remain in kiosk mode while observations are being made.

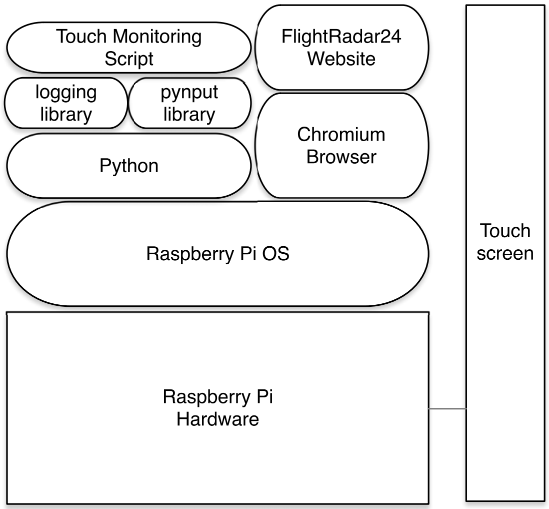


Figure 1: The Hardware-Software Stack

The overall architecture of the system is shown in Figure 1, including the Raspberry Pi hardware and the various software components as outlined above.

The current version of the kiosk runs on a seven inch touchscreen with the computer mounted on the back, so it can act as a portable experimentation system. This setup is seen in Figure 2.

4 Results

When powered up, the kiosk boots and after about thirty seconds ends up in its default state, showing the `flightradar24.com` website that the user should interact with, as shown in Figure 3.

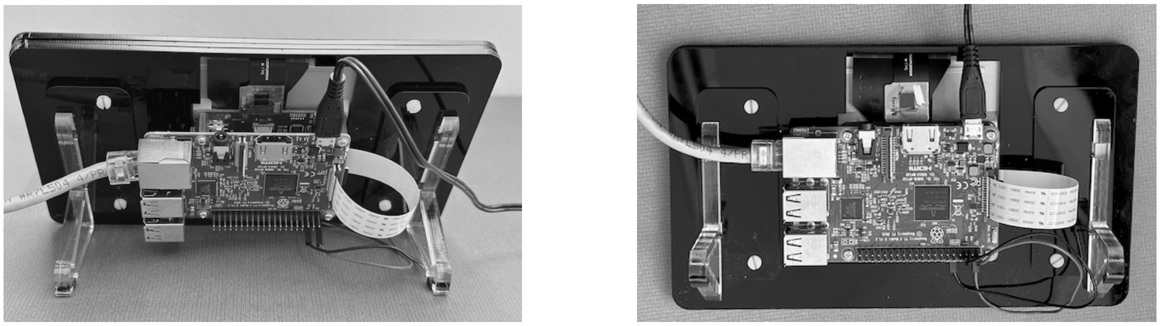


Figure 2: Kiosk Hardware with Touchscreen and Computer

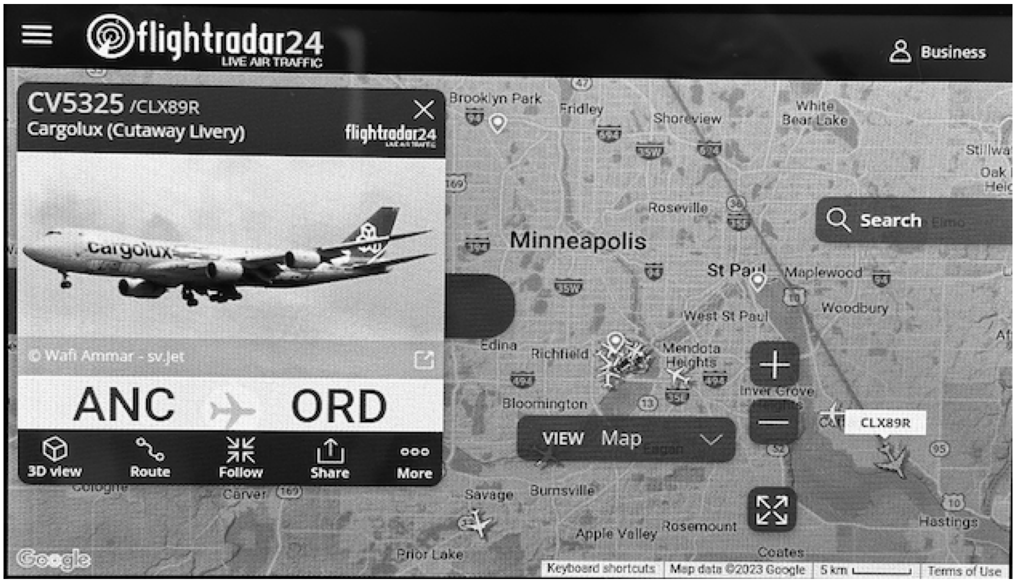


Figure 3: Flight Radar Screenshot

As the user interacts with the kiosk, each of their taps is recorded in a log for examination and analysis at a later time by the user interaction experimenter. The log tracks the time of the tap, as well as the location on the screen using the corresponding x and y coordinates. An example log file is shown in Figure 4.

The user can customize the kiosk.sh file to specify the desired site to start the kiosk in. They can also specify in the Python script the file path and desired name for their log. For the user's experimentation, the device can be left in a location where participants can interact with the kiosk. Each interaction, or touch, will be logged in the file, which can be accessed remotely by the user through an ssh connection with the device. As changes are made to the site for each test, or different sites are specified in kiosk.sh, all that is required of the user is to stop and restart kiosk.service, or restart the device,and specify a separate file in data-collection script to track the new interactions.

```
2023-03-17 14:27:14,108  move to (539, 248)
2023-03-17 14:27:15,118  move to (512, 245)
2023-03-17 14:27:16,216  move to (571, 229)
2023-03-17 14:28:13,617  move to (588, 304)
2023-03-17 14:28:14,545  move to (600, 309)
2023-03-17 14:28:15,505  move to (565, 362)
2023-03-17 14:28:16,489  move to (554, 432)
2023-03-17 14:28:17,425  move to (590, 354)
2023-03-17 14:28:18,869  move to (602, 313)
2023-03-17 14:28:19,417  move to (573, 355)
```

Figure 4: Sample Log Entries

5 Conclusions and Future Work

This first build is a success as it creates a clean kiosk display on a Raspberry Pi device, and monitors and records touch events on the device while in the kiosk mode. However, the touch event data is rudimentary, only recording the x and y coordinates of the touch. This data requires more work on the researches part to mirror the coordinates over the website display in order to see what elements were interacted with. It also does not account for touches that would result in the leaving of the page, which could result in data that is not applicable to the desired testing.

In order to more accurately monitor events, future versions of this kiosk will continue to use the Raspberry Pi, but attached to a larger screen. The data-collection software will also be expanded upon, allowing for finer-grained control. In order to expand the software, the exploration of a different, lower-level recording and monitoring library will be required.

References

[1] BAILEY, R. W., WOLFSON, C. A., NALL, J., AND KOYANI, S. Performance-based usability testing: Metrics that have the greatest impact for improving a system's usability. In *Human Centered Design* (Berlin, Heidelberg, 2009), M. Kurosu, Ed., Springer Berlin Heidelberg, pp. 3–12.

[2] DUMAS, J. S., AND FOX, J. E. Usability testing. In *The Human-Computer Interaction Handbook, Third Edition*. CRC Press, 2012, pp. 1222–1241.

[3] FOUNDATION, R. P. How to use a raspberry pi in kiosk mode. `https://www.raspberrypi.com/tutorials/how-to-use-a-raspberry-pi-in-kiosk-mode/`. Accessed in March 2023.

[4] PALMER, M. pynput 1.7.6 library. `https://pypi.org/project/pynput/`, 2022. Accessed in March 2023.

XprospeCT: CT Volume Generation from Paired X-Rays

Benjamin Paulson, Joshua Goldshteyn, Sydney Balboni, John Cisler, Andrew Crisler,
Natalia Bukowski, Julia Kalish, Theodore Colwell

Department of Electrical Engineering and Computer Science

Milwaukee School of Engineering

1025 N Broadway St, Milwaukee, WI 53202

(paulsonb, goldshteynj, balbonis, cislerj, crislera, bukowskin, kalishj, colwellt)
@msoe.edu

Abstract

Computed tomography (CT) is a beneficial imaging tool for diagnostic purposes. CT scans provide detailed information concerning the internal anatomic structures of a patient, but present higher radiation dose and costs compared to X-ray imaging. In this paper, we build on previous research to convert orthogonal X-ray images into simulated CT volumes by exploring larger datasets and various model structures. Significant model variations include UNet architectures, custom connections, activation functions, loss functions, optimizers, and a novel back projection approach.

1 Introduction

Computed tomography (CT) allows for detailed views of anatomic structures by reconstructing hundreds of X-ray images taken from a full rotation around a patient's body at various angles to produce three-dimensional volumes. The information acquired by a CT scan is beneficial to the diagnostic process as an exploratory analysis tool but is resource-demanding and requires high levels of ionizing radiation exposure with an average effective radiation dose of 7 mSv per chest CT [4]. In contrast, two-dimensional X-ray imaging is a widely used medical imaging modality as it is resource-efficient and results in a lower average radiation dose of 0.1 mSv per chest X-ray [4]. Although more efficient, X-rays lack the detailed spatial information obtained from a three-dimensional imaging modality.

Many healthcare protocols involve an X-ray scan for preliminary information gathering, allowing healthcare workers to determine the need for risking higher radiation exposure and expending more resources with additional imaging modalities. Although X-ray scans provide a clear view of bone structures, soft tissues are less defined. Machine learning methods have shown to be an effective way to extrapolate the two-dimensional information from X-rays into three-dimensional space, remedying the shortcomings of traditional X-rays while equipping professionals with additional information for making well informed decisions.

We propose to improve upon past research by making a model capable of generating simulated CT scans from X-ray image inputs with high detail. In the recent work, "X2CT-GAN: Reconstructing CT from Biplanar X-Rays with Generative Adversarial Networks", Ying *et al.* researched the simulation of CT scans from orthogonal X-ray views [1]. This paper's approach included a Generative Adversarial Network (GAN) framework with a specialized generator. We hypothesize that using a larger dataset of CT exams and testing several network architectures will allow for improved reconstruction results.

In this work, we develop deep-learning models for reconstructing CT scans from paired lateral and anterior-posterior X-rays to generate an accurate volume for structures in the chest region. These reconstructed CT scans may allow for the preliminary detection of medical abnormalities, enabling a patient's care team to determine the need for additional resource-intensive imaging methods.

2 Data

The models in this study depend on two distinct datasets: the Rad-ChestCT dataset [2] and the CheXpert dataset [3]. The former dataset was obtained from the Center for Virtual Imaging Trials at Duke University and consists of 35,747 chest CT scans acquired from 19,661 adult patients. Access was granted to 10% of the data, comprising 3,630 chest scans. This dataset was developed by Rachel Draelos, an MD/PhD student, with the objective of facilitating machine learning models focused on CT scans. The CheXpert dataset is an open-source dataset that comprises of 224,316 chest radiographs taken from 65,240 patients. For training the style transformation model, a total of 2,390 paired orthogonal X-rays were used. To be used as the ground truth in the reconstruction model, the CT images were normalized by scaling the volumes from -1000, 1000 HU to between 0 and 1. Next, these volumes were resized to either 128×128×128 or 64×64×64 to be inputs for the reconstruction model. The dataset was split, with 80% partitioned for training and 20% for validation. Furthermore, 30 test images were extracted and used to compare models as shown in the Experiments Table (Table 1).

3 Models

Several experiments were performed using simulated X-ray images as the input for the reconstruction model, which takes a set of paired X-rays and yields a simulated CT volume. The creation of simulated X-rays required style transfer models – one for each X-ray view – which created paired data from the Rad-chest CT [2] and CheXpert [3] datasets.

3.1 Style Transfer General Adversarial Network

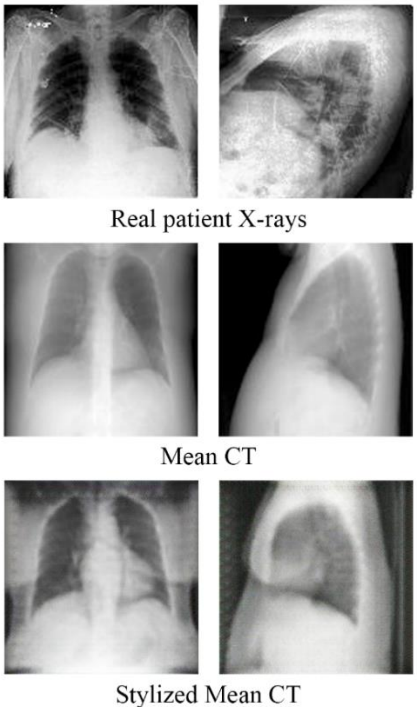
Obtaining paired data from the separate CheXpert and Rad-ChestCT datasets presented a challenge. Past research into a model capable of learning the mapping between two unpaired datasets resulted in CycleGAN [5]. We used CycleGAN as a style transformation model to obtain paired data for training. The style transformation model was trained on X-rays from the CheXpert dataset and mean CT scans that were originally a part of the Rad-ChestCT dataset. The style transformation model was employed to learn the style of a real X-ray that can then be transferred to mean CT scans in the process of generating paired data.

The CT scans of the Rad-ChestCT dataset were averaged along the coronal and sagittal axis to obtain two-dimensional CT scans in both the anterior-posterior and lateral views. The mean CT scans were inputs to the style transfer general adversarial network which then stylized the mean CT scans to have a closer resemblance to real X-rays. This process resulted in paired simulated X-rays and CT scans that were used to train the reconstruction model in some experiments.

Changes were made to the CycleGAN model to better process the style transformation that is required to go between mean CT scans and X-rays. The first major change included significantly decreasing the learning rate, as the level of the style transform amounted to a non-linear contrast transformation as opposed to transitioning between two different styles. This was paired with the use of an exponential moving average, which reset the weights to their moving average to increase the generalization of the model. This also allowed for finer level updates between epochs.

On top of these changes, the Nadam optimizer was used instead of Adam due to the better level of convergence that it achieved. The Nadam and Adam optimizers work similarly, except for incorporating Nesterov momentum which performs updates to the gradient based on the projected update, as opposed to the current value of each parameter. This allows for a lower level of overshooting once a local minimum has been found and has resulted in better styled X-rays than those obtained using pure Adam.

Another change was made to increase the Lambda value to 20. The Lambda variable describes the cycle consistency loss which affects how accurate the style reconstruction is by comparing the original input when going both forward and backward through the CycleGAN training. This was originally set to 10 for most uses, but it was discovered that the pixel alignment improved the style transfer when using a higher value by testing different values. However, if the value was too high (greater than 25), the style transfer ended up becoming very weak and the differences between the input and output became insufficient. With a lambda of 20, a style transfer occurred that included significant X-ray characteristics (such as less contrast of different depth regions as compared to mean CT scans) while maintaining a similar shape to that of the original.



Real patient X-rays

Mean CT

Stylized Mean CT

Figure 1: Examples of real Patient Coronal and Sagittal View X-rays from the CheXpert Dataset, Mean Coronal and Sagittal View X-rays, and Stylized Coronal and Sagittal View X-rays

3.2 Reconstruction Model

The reconstruction model is responsible for the transformation of two orthogonal X-rays to a simulated CT volume. Our reconstruction model is inspired by the "X2CT-CNN+B" reconstruction model proposed in the paper “X2CT-GAN: Reconstructing CT from Biplanar X-Rays with Generative Adversarial Networks” [1], comprised of two UNet encoder-decoder architectures for corresponding simulated X-ray inputs. Following the decoders, a fusion network with custom skip connections interweaved from the orthogonal decoder architectures iteratively extracts their convolved features into the yielded simulated CT volume. Modifications made to the initial “X2CT-CNN+B” reconstruction model – as well as associated result discussions – are compiled in “Section 4: Experiments & Results”. Featured experimental adjustments ranged from architectural to hyperparameter changes. Contrasting previous work, the discussed reconstruction models were trained using paired images of real, as opposed to averaged, orthogonal X-Ray images and associated CT scans in some experiments.

Stemming from challenges faced regarding bit limitations of Keras Tensors (see 6.3) [7], the initial architecture (Figure 2) was modified from the “X2CT-CNN+B” for 128×128-pixel images rather than 256×256-pixel images, resulting in significant memory savings due to CT labels consequently reduced to 128×128×128 voxel scans. The UNet architecture was chosen due to its ability to preserve features – via skip connections – from input images regardless of the convolutional changes that occur throughout the encoding portion of the network. This preservation through skip connections was important due to the emphasis on small details in medical-imaging applications.

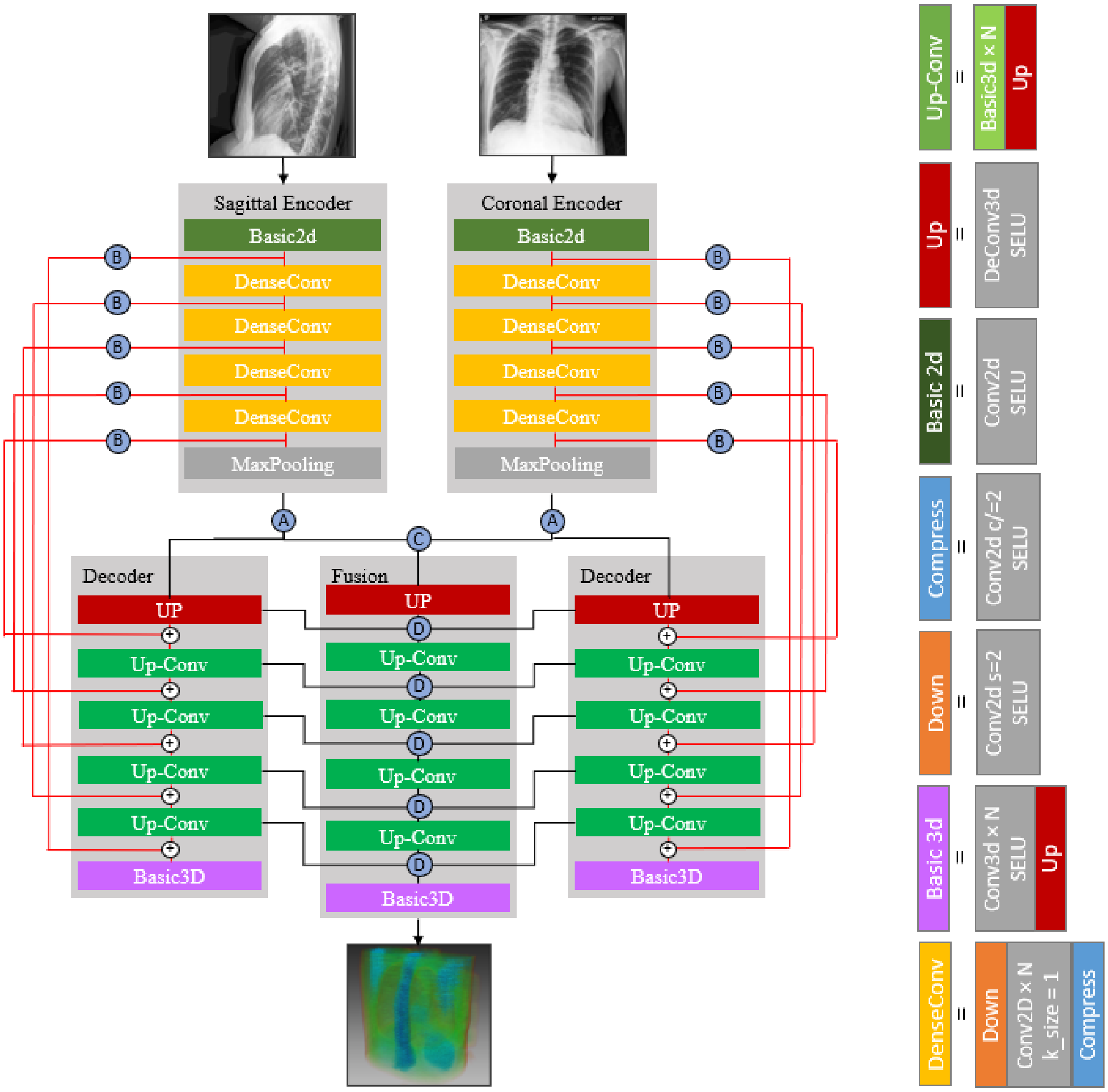


Figure 22: Dense-512 Reconstruction Model with Model-Layer Legend

3.3 Reconstruction Model Connections

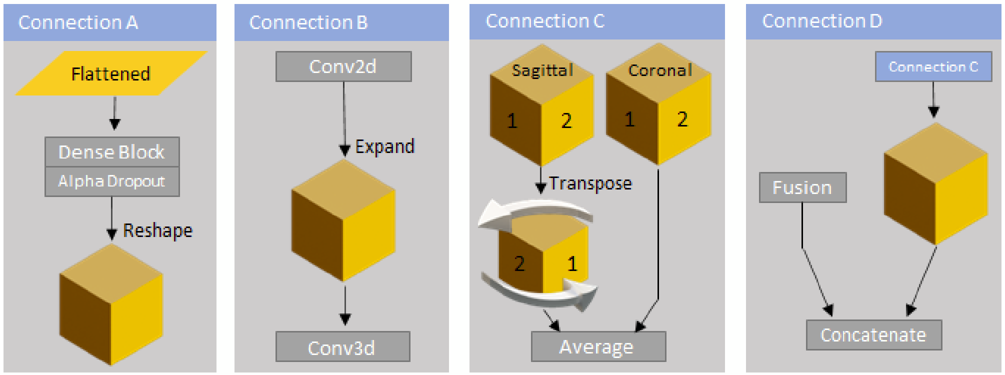


Figure 33: Architectural View of Connections

The connections that link the individual components of the reconstruction model are divided as Connection-A, Connection-B, Connection-C, and Connection-D. The placement of these connections within the overall architecture did not change across experiments; however, if the connection was altered, the internal definition was changed without changing the external I/O of each connection. This was done to maintain the capability of the overall structure while testing smaller model changes.

Connection-A reshapes the input tensor to a uniform three-dimensional output regardless of the input dimensionality. Following the convention for tensors in Keras, this is represented by a five-dimensional tensor formatted in order of following index: batch size, rows, columns, depth, and channels. Further discussion about the final Connection-A architecture is provided in section 4.3.

Connection-B augments the skip connections of each UNet architecture present throughout the reconstruction model. Because the decoding portion of the UNet consists of five-dimensional tensors – but the skip connections from correlating encoding layers are four-dimensional – the skip connections must also account for transforming two-dimensional image data to a three-dimensional representation to be concatenated with the decoded layers' outputs. The presence of Conv2D and Conv3D layers within the connection respectively match the number of channels and transform to an expanded five-dimensional tensor.

Connection-C transposes the sagittal and coronal voxel representations such that the sagittal view is appropriately orthogonal to the correlating front-facing coronal voxel representation. The result of this permutation is then averaged to produce a coherent voxel representation for input into the decoder layers of the fusion network.

Connection-D uses Connection-C yet accepts an additional input from the previous decoding layer of the fusion network to be concatenated with the averaged output of Connection-C.

4 Experiments & Results

Below is a brief, tabular overview of the experiments discussed in this section.

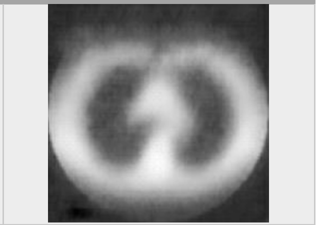
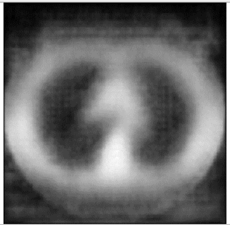
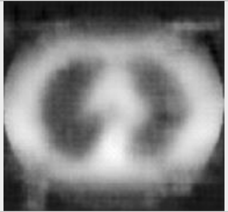
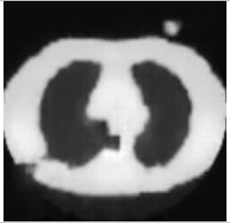
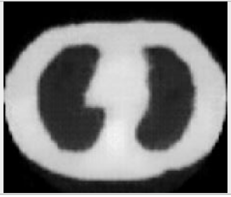
Model Name	Input X-ray Images	Training Loss function	Optimizer	Test Accuracy MSE (±0.001)	Example Axial Image
64-Dense	Mean CT	MSE	Adam	0.0834	
64-Dense-Styled-Inputs	Stylized CT	MSE	Adam	0.0751	
512-Dense	Mean CT	MSE	Adam	0.0819	
512-Dense MAE	Mean CT	MAE	Nadam	0.0970	
512-ProjInj	Mean CT	MAE	Nadam	0.0980	

Table 1: Model names, input x-ray images, training loss function, optimizer, test accuracy, and example axial images.

Note: In every experiment, the SELU activation function is used due to its self-normalizing properties. This allows for a faster and more accurate level of training than using ReLU layers followed by batch normalization.

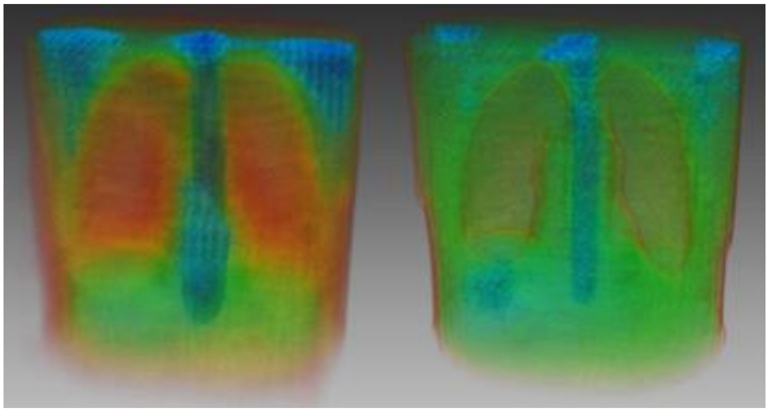


Figure 4: 3D Volume Renderings of predictions (64-Dense-Styled-Inputs [Left], 512-ProjInj [Right]). Features including the lungs and spine can be seen. All models failed to synthesize dense bones, as shown by the lack of ribs.

4.1 Input X-Ray Dimensions

Several experiments were performed with various input dimensions (not included in experiment table). An analysis was performed on image pixel sizes of 128×128 and 64×64 – outputs were kept at similar dimensionality matching the input dimension height and length.

The 256×256 input images were unable to train due to TensorFlow’s tensor size limit (Section 5.2). The 128×128 with a “64-Dense” architecture produced results comparable to “X2CT-CNN+B” [1]. Additional experiments were performed with a 64×64 input, further decreasing the required memory and training time for the stylized CheXpert dataset; however, the decreased amount of encoding/decoding layers resulted in significant artifacts for predicted CT-Scans due to the lack of model structure to recognize specific features. Furthermore, the reduced resolution of the input X-rays restricts the amount of data the model must recognize to identify key features and patterns from the mean CT scans.

4.2 Training Loss Function & Optimizer

Initial experiments with choice of training loss function and optimizer followed guidelines set by existing work [1]; these guidelines are Mean-Squared Error Loss and Adam Adaptive Optimizer [9]. Results with these guidelines produced CT scans with noise surrounding key X-Ray features: lungs, spine, and rib cage.

While noisy output may indicate a deep-learning model deficiency in adequate capacity for the application, variations around key features may indicate an inappropriate loss function. Mean-Squared Error as the training loss function lacks spatial awareness – a key factor in outputting noise-minimized CT scans. This resulted in multiple experiments with varying loss functions, resulting in Mean Absolute Error (MAE) being identified as the best candidate for effectively reducing the blurriness of the output CT scan data. While usage of MAE as a loss function reduces the precision of output features, this model is intended to be used for preliminary imaging purposes rather than full medical diagnoses.

Learning-rate optimization was initially performed with Adaptive Moment Estimation (Adam) to use the low memory requirements and capability to effectively optimize the training process of a large dataset [9]. However, Adam has significant difficulties escaping local minima whereas Nadam incorporates Nesterov-Accelerated momentum to minimize these convergence issues. Nadam often produces models with better generalization capabilities in less training time [10]. Experiments with Nadam as the optimizer, along with the benefits of MAE as the associated training loss function produced the clearest results throughout the reconstruction model experiments.

Ultimately, MAE and Nadam were chosen due to reduced blurriness within the CT scan output. However, this decreased blurriness does not mean the CT scan is less accurate due to the inherent quality of MAE and Nadam that tend to perform very similarly to the more commonly used MSE and Adam in other scenarios [10]. The MAE loss function allowed for predicting model losses linearly as opposed to squaring the error and treating each degree of error with the same amount of importance. Using Nadam allowed for the use of Nesterov acceleration with Adam, which updated the momentum weights in response to the future location as opposed to the current location (potentially preventing overshooting in local minima).

4.3 Connection-A Architecture

The first two experiments involved differing Connection-A architectures involving the number of dense-layer neurons at the apex of both the sagittal and coronal UNet architectures. “64-Dense” and “512-Dense” distinguish these experiments as the number of neurons within the dense block. Minimal visual and quantitative differences were detected between these changes; however, “512-Dense” produced CT-Scans more aware of features outside the immediate intrathoracic cavity such as shoulder blades and upper and lower vertebrae.

4.4 Back Projection Injection

Back projection injection is the largest architectural change to the previous alterations of “X2CT-CNN+B” mentioned above. This novel approach includes a simplified version of back projection with the orthogonal X-ray images. It produces CT scans with increased visibility of granular yet key features such as the vertebrae. This approach was considered because back projection is the result of a calculation, preserving spatial data, rather than requiring additional data collection and is similar to conventional CT reconstruction methods.

5 Challenges

5.1 Memory Limitations of Single-GPU Training

During the creation of the reconstruction model outlined in (Figure 2)**,** we came to the realization that the quantity of layers used in combination with the image size were consuming a lot of memory during run-time. The biggest cause of our memory was the final size of our output: 128×128×128 pixels. Many of the intermittent models tried to use high pixel inputs and outputs to increase accuracy of the result but failed due to reaching the memory limit. This restricted the maximum quality of the outputs produced. The final reconstruction model consumed around 25 GiB of memory and the Projection Injection consumed around 32 GiB of memory.

5.2 TensorFlow’s Max Limit on Parameters

Another issue encountered during testing of model architectures was the TensorFlow library limitation on max parameters in a layer. TensorFlow limits its parameters per layer at 2^{32} parameters. This also contributed to our max possible output size being 128×128×128 pixels, again restricting the max quality of our results. We ended up working within Keras’ layer constraint, but future work could look at solving this problem, potentially with PyTorch (Section 6.1) to create more detailed results.

6 Future Work

6.1 Switch to PyTorch to Solve Max Parameters Constraint

As mentioned in the Challenges, TensorFlow’s Keras library has a limited number of parameters allowed per layer. This resulted in our reconstruction model being limited to an output of 128×128×128 for the simulated CT scan. To achieve more defined simulated CT scans, the model output layer needs more parameters. The best alternative we identified was switching to PyTorch’s libraries [10], which would require a large code change.

6.2 Transformer Architecture

Currently, all models tested directly input the coronal and sagittal X-Ray images into the reconstruction model’s encoders, as shown in (Figure 2). However, past research [6] has shown that a multi-headed attention layer before the encoder could lead to increased quality in the final picture.

6.3 Multi-GPU Training

The implementation of the reconstruction model uses a single GPU to train. However, this was relatively slow, limiting the number of epochs that models could train for. Using multiple GPUs may reduce training time and allow for the model to get better results by training across more epochs.

6.4 Use Medical Label AI to Test Accuracy of Output

Other research [3] has attempted to label CT scans from the Rad-ChestCT data set. To verify the validity of the output from the models, it could be beneficial to match labels found in both the Rad-ChestCT and CheXpert datasets and then train the model on the Rad-ChestCT data and verify the accuracy using CheXpert as the validation set. This would ensure that the model is reconstructing the CT scans properly instead of losing information in the process.

6.5 Train on Larger-scale X-Ray Images

To speed up training times and reduce run-time memory consumption, it was necessary to reduce the input X-Rays to 128×128. However, this traded the maximum quality of the resulting CT scans. To improve the reconstruction model, scaling up the input images to at least 256×256 or potentially higher may be useful. This will increase the train time and memory consumption, but with other future directions mentioned above, these effects could be reduced.

6.6 Partner with Medical Experts to Evaluate Usability

In the future, it would be beneficial to validate this model with medical experts to determine its relevance in clinical settings. Experts could also identify areas of improvement, identifying possible future research directions.

7 Conclusion

This work explores various techniques to reconstruct CT scans from paired orthogonal X-rays. Usage of the reconstruction model was primarily used to facilitate the generation of simulated CT volumes, along with several experiments performed using simulated X-ray images created by the style transformer model. While the stylized model (table 1, row 2) has the best mean squared error value, we consider the back projection model (table 1, row 5) to provide more accurate results for granular features within the final CT scan compared to other models that were experimented with. We hope to further our research by using multiple GPUs, adding attention to our reconstruction model, improving input X-Ray image size, and look at evaluating our model, both by using CT labeling models as well as receiving feedback from medical experts. With further research, we hope to create a model that clinicians can use to improve the diagnostic process.

References

[1] X. Ying *et al.*, "X2CT-GAN: Reconstructing CT from Biplanar X-Rays with Generative Adversarial Networks." Available: https://openaccess.thecvf.com/content_CVPR_2019/papers/Ying_X2CT-GAN_Reconstructing_CT_From_Biplanar_X-Rays_With_Generative_Adversarial_Networks_CVPR_2019_paper.pdf

[2]"RAD-ChestCT Dataset - CVIT - Center for Virtual Imaging Trials," *CVIT - Center for Virtual Imaging Trials*, Oct. 11, 2022. https://cvit.duke.edu/resource/rad-chestct-dataset/ (accessed Feb. 10, 2023).

[3] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, Jul. 2019, doi: https://doi.org/10.1609/aaai.v33i01.3301590.

[4] "Radiation risk from medical imaging," *Harvard Health*, 30-Sep-2021. [Online]. Available: https://www.health.harvard.edu/cancer/radiation-risk-from-medical-imaging. [Accessed: 10-Feb-2023].

[5] J.-Y. Zhu, T. Park, P. Isola, A. Efros, and B. Research, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," 2017. Available: https://arxiv.org/pdf/1703.10593.pdf

[6] Y. Wang and Q. Xia, "TPG-rayGAN: CT reconstruction based on transformer and generative adversarial networks," *Third International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI 2022),* Jan. 2023, doi: https://doi.org/10.1117/12.2655901.

[7] Martín Abadi *et al.*, 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems'. 2015.

[8] A. Paszke *et al.*, 'PyTorch: An Imperative Style, High-Performance Deep Learning Library', in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.

[9] K. P. Diederik and J. Ba, "Adam: A Method for Stochastic Optimization," *arxiv*, vol. 1, no. 9, Dec. 2014.

[10] T. Dozat, "Incorporating Nesterov Momentum into Adam," 2016.

dptv: A New PipeTrace Viewer for Microarchitectural Analysis

Adam Grunwald
Department of Computer Science
University of Wisconsin-La Crosse
grunwald5629@uwlax.edu

Phuong Nguyen[†]
FPT Software
Hanoi, Vietnam

Elliott Forbes
Department of Computer Science
University of Wisconsin-La Crosse
eforbes@uwlax.edu

Abstract

In computer architecture research, it is common to test the effectiveness of new microarchitectural features by modeling hardware structures and logic in a software simulator. The benefit of simulation is that those features can be implemented with configurable parameters (sizes, widths, latencies, etc.) that an architect can vary to assess performance across a suite of benchmarks. Typically the simulator gives overall summary statistics for how each benchmark ran, which can be compared to the output of other configurations. But summary statistics are usually averages across entire benchmarks, which can hide key fine-grained details about a configuration's run-time behavior. Furthermore, summary statistics may make it difficult to find exactly where performance diverges when comparing the same code region of a benchmark across two processor configurations.

Pipeline tracing tools, for example gem5's O3 Pipeline Viewer [2], help to visualize instruction-level behavior. These tools require that a simulation not only provides summary statistics, but also a record of all processor events for all dynamic instructions for each cycle. Those events are written to a trace file, which can then be opened by a visualization tool that plots the trace, with each dynamic instruction of a benchmark along the negative y-axis (program order), and pipeline stage events for each cycle on the positive x-axis. Pipeline tracing tools greatly increase the visibility into the run-time behavior of a benchmark for a given processor configuration, but still lack the ability to compare multiple configurations.

This paper presents a new pipeline trace viewer, called Dual PipeTrace Viewer, or dptv for short. This program can visualize a single trace file, like previous trace viewers. But dptv can also take two trace files as input. dptv uses the SDL2 graphics library to provide a more flexible graphical interface, allowing for for easy panning and zooming on the diagram, with additional functionality to search the trace and to peek at specific stage information. When opening two trace files, dptv verifies both trace files represent the same dynamic instruction stream, and then overlays the pipeline stage events for both traces. In this way, dptv makes it more obvious where two processor configurations have performance that diverges.

This paper will outline the specifics of how dptv functions and provide examples of where analysis is made easier using this tool.

[†] Author contributed to this paper while at University of Wisconsin-La Crosse.

1. Introduction

The Dual PipeTrace Viewer (stylized as *dptv*) is a graphical tool for visualizing traces generated by a processor simulator. Traces are a record of the dynamic instruction stream of a benchmark program being executed by a processor simulator. Tools of this nature visualize the trace as a plot with the dynamic instructions along the y-axis and each cycle along the x-axis, making each row in the plot contain processor events for one dynamic instruction. This style of visualization greatly aids a computer architect – trends and patterns can more obviously attract our eye. The visual format of a pipetrace viewer comes naturally, as most undergraduate computer architecture courses use pipeline timing diagrams [11] to visualize the instruction-level parallelism achieved by a pipelined processor execution model.

As microarchitectural complexity has increased, and workloads have grown in scope, architects have relied more heavily on ever-more sophisticated simulators when looking for performance bottlenecks, and opportunities for improvements to systems. However, the effort required to analyze performance bottlenecks have become highly nuanced. Interactions between instructions and microarchitectural structures cause subtle performance differences that can potentially lure a computer architect to iteratively add debugging code to a simulator, re-execute a benchmark, use the insight gained to add more debugging code, re-execute a benchmark, and so on. This problem only increases when considering alternatives for microarchitectural features (sizes of hardware tables, policies decisions, widths, history depths, etc.). Simulators typically output summary statistics for how a given benchmark ran overall. However, summary statistics can potentially hide the subtle interactions and events that can affect overall performance, since these statistics are often averages across a large number of instructions.

The main motivation for a tool such as the one being presented is to ease these burdens. First there is a desire to visualize traces generated from a simulator in a way that is more easily navigable. Pipeline tracing tools like the *dptv* allow for the user to investigate more closely the run-time behavior of a particular execution than statistics alone would provide. Second is the desire to compare two traces of the same benchmark under different microarchitectural configurations more easily. It is common to run a benchmark multiple times to compare which configuration gives the best performance, however it could also be useful to tell where exactly in a benchmark performance diverges between configurations. If designed right, a pipeline tracing tool is a prime candidate for such a task. By superimposing the traces of multiple executions of a benchmark under different microarchitectural models, it becomes easier to both identify at what point performance diverges, but also the specific behavior that may have caused performance to differ.

dptv aims to make the process of visualizing and comparing traces as easy as possible. Figure 1 shows a basic screenshot from *dptv*. Its interface is graphical, using the SDL2 graphics library [3], which allows the user to both zoom and pan around the diagram, familiar to anyone having used photo editing or CAD tools. By hovering the mouse over a particular pipeline stage marker, identified by a character representing the type of stage, more information about the machine state at that point in time can be viewed (shown in the upper-right corner). The trace can be searched for occurrences of text in the pipeline stage information, the program counter, instruction text, and more. Users can zoom in/out on the diagram, enabling both visualization of the long-term performance of the trace, and also investigation of pipeline events over a short period of time.

Several pipetrace viewers have been proposed [2] [13] [14] in the literature. The main feature which differentiates *dptv* from these other tools is the ability to visualize two traces at once. When visualizing two traces, each trace will be interleaved, effectively showing traces atop each other, with each row alternating which trace its information is from. As previously mentioned, this greatly helps with comparing and contrasting the characteristics of the two executions. Since varying microarchitectural features potentially impacts the clock period, *dptv* allows the two traces to operate at different frequencies – scaling the visualization to accurately represent the actual run-time of the traces.

In addition to visualization goals, the development of *dptv* is also making integration into a wide variety of simulation infrastructure as a first class design constraint. As such, *dptv* is instruction set agnostic, simply treating instructions as plain text. Furthermore, pipeline stages can be annotated with arbitrarily many events in a key-value pair format. And *dptv* was written in standard C, with the goal of portability in mind – with only the reliance of the SDL2 library in addition to standard C libraries.

The remainder of this paper will go more in-depth on specific aspects of *dptv*. Section 2 provides more context on the current state of the art for tools of this nature, as well as related prior work. Section 3 will cover the various features provided by *dptv* in depth. Section 4 will give an example in which using *dptv* aids in quickly understanding the difference in performance between two traces. And finally Section 5 will conclude the paper with a brief discussion on the current status of the tool, what is finished and what still needs to be done, and ideas for future work.

2. Background

Architects have grappled with architectural complexities, and scale issues since the earliest simulators and benchmarks. Using graphical visualization tools to better understand computer performance has a well established history.

Gao, et al. [10] provide a survey paper that summarizes the current landscape of a wide variety of visualization tools, 21 tools in total, as of their publication date in 2011. They further derive a characterization of these tools, according to a taxonomy proposed by Card [9], borrowing insight from the Human-Computer Interaction research community. These 21 tools span the gamut from tools that more appropriately graph summary statistics, up to tools that provide visualizations of data synthesized across highly dimensional data sets.

Accelerator and GPU-style many core architectures present a class of bottlenecks that tend not to manifest in general purpose processor architectures. For instance, memory bandwidth is more highly constrained in many-core architectures. The lock-step execution of warps in GPUs, and control divergences present unique challenges. And to meet these challenges, visualization tools more appropriately highlight the needs of these architectures. In [5], Ariel et al. present a tool for displaying DRAM channel utilization, warp control divergences, histograms of static instruction usage, as well as mechanisms to show which instructions are the sources of bottlenecks for GPU and SIMT-style architectures. Similarly, Alsop et al. provide a GPU-specific visualization [4] for highlighting sources of memory stalls.

Not all architectures have performance as their main target metric. Mobile systems have energy efficiency and power as first class design constraints. In [6], the authors instrument a multithreaded processor simulator to track with activity counters that are fed to a tool that is able to display a processor floorplan with a 3D height map such that the height represents the relative power consumption of hardware units. This visualization can be scrubbed forward or backward in time to see how the height map (thus, power consumption) changes as the program behavior stresses different hardware units over time.

The closest relatives of *dptv* exist in four tools, TraceVis [13], PSE [12], GPV [14] and `o3` [2]. None of these tools allow for traces from multiple microarchitectural configurations to be visualized simultaneously.

TraceVis [13] uses the same pipeline timing diagram style of visualization as *dptv*. Furthermore, TraceVis uses the Qt windowing toolkit to allow for zooming in and out, like *dptv* allowing for visualizing both course and fine grained program behavior.

The Processor Simulation Elucidator (PSE) [12] uses the GTK windowing toolkit to similarly display Pipeline Event Diagrams (PEDs). These PEDs are simply pipeline timing diagrams, with pipeline event information associated with each instruction. PSE does have additional modes of visualization not found in *dptv*, used to show how performance metrics (those that will likely become summary statistics) vary over time.

The Graphical Pipeline Viewer (GPV) [14] is an early example among these pipeline timing diagram style viewers, written in Perl TK. GPV is integrated into the once popular SimpleScalar [8] simulation framework, and also incorporates a visualization of power consumption.

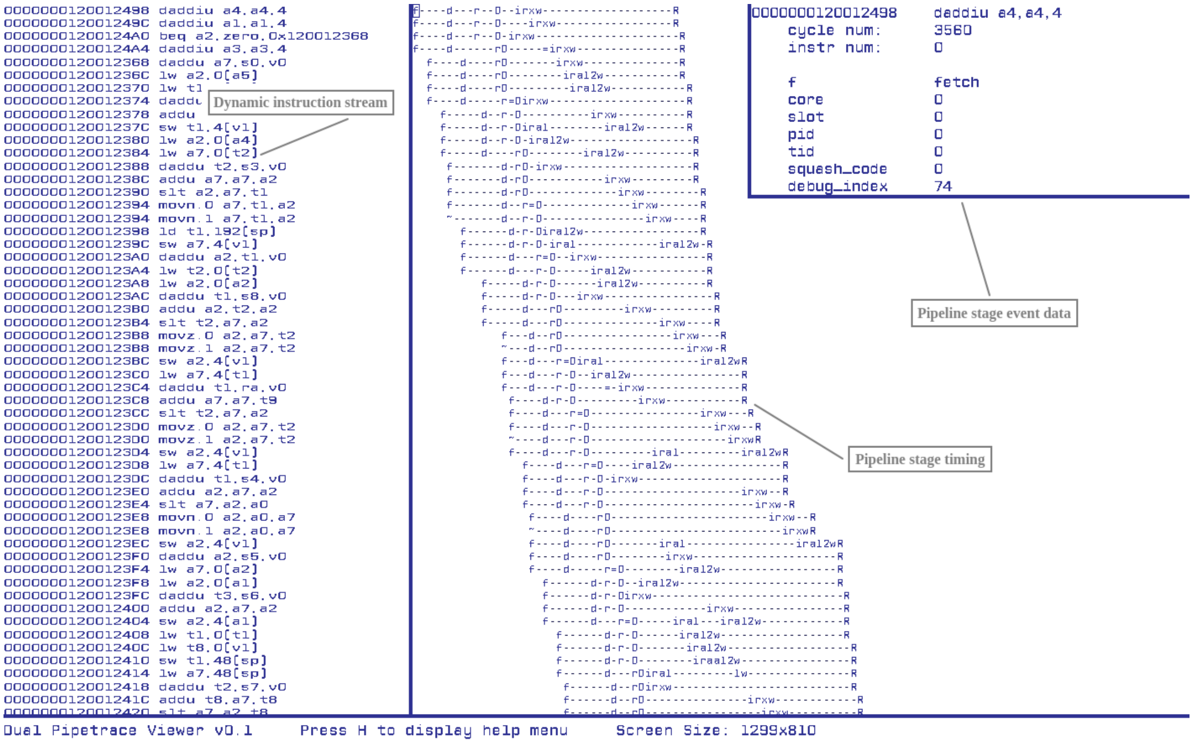
A somewhat rudimentary, but comparable, tool bundled with the popular `gem5` [7] simulator is the `o3 pipeline viewer` [2]. Again, `o3` shows a pipeline timing diagram style visualization. However, `o3` produces a non-interactive, text-only output file, formatted with a fixed column width. Each column represents one cycle of execution time. When an instruction stays in-flight for more cycles than there are columns of text, then the remaining cycles simply wrap to the next line. This limitation makes it extremely difficult to find performance bottlenecks and correlate those bottlenecks with pipeline events.

3. *dptv* Viewer Tool

This section will go in-depth on the usage and features of *dptv*, as well as the visualization provided.

dptv is currently setup to build in a Linux environment, with the only extenal library used being SDL2 [3] for rendering. *dptv* is invoked on the command-line with either one or two trace file names as input. If only one trace file name is provided, then the program will open in single-trace mode. If two traces are provided, the viewer will first verify that the trace data held within the files are the same instruction stream, and if so, the visualization will open in dual-trace mode. This is a key new feature, unique to *dptv*. Traces from two different instruction streams are not meaningful comparisons, and if the tool determines that the *committed* instructions are not the same, it will simply fail to start the visualization. Since only committed instructions are compared, if two traces differ by squashed instructions only, then *dptv* can still visualize the two trace.

It is possible that two processor configurations operate at different clock frequencies, depending on microarchitectural features. *dptv* handles this possibility. If the two traces should be displayed at different frequencies, a command-line argument can be passed when invoking *dptv* which will scale the traces based on the ratio of the frequencies. Additional command-line options exist to change the color palette used, as well as the position and size of the window.



```
0000000120012498 daddiu a4.a4.4
000000012001249C daddiu a1.a1.4
00000001200124A0 beq a2.zero.0x120012368
00000001200124A4 daddiu a3.a3.4
0000000120012368 daddu a7.s0.v0
000000012001236C lw a2.0[a5]
0000000120012370 lw t1
0000000120012374 daddu
0000000120012378 addu
000000012001237C sw t1.4[v1]
0000000120012380 lw a2.0[a4]
0000000120012384 lw a7.0[t2]
0000000120012388 daddu t2.s3.v0
000000012001238C addu a7.a7.a2
0000000120012390 slt a2.a7.t1
0000000120012394 movn.0 a7.t1.a2
0000000120012394 movn.1 a7.t1.a2
0000000120012398 ld t1.192[sp]
000000012001239C sw a7.4[v1]
00000001200123A0 daddu a2.t1.v0
00000001200123A4 lw t2.0[t2]
00000001200123A8 lw a2.0[a2]
00000001200123AC daddu t1.s8.v0
00000001200123B0 addu a2.t2.a2
00000001200123B4 slt t2.a7.a2
00000001200123B8 movz.0 a2.a7.t2
00000001200123B8 movz.1 a2.a7.t2
00000001200123BC sw a2.4[v1]
00000001200123C0 lw a7.4[t1]
00000001200123C4 daddu t1.ra.v0
00000001200123C8 addu a7.a7.t9
00000001200123CC slt t2.a7.a2
00000001200123D0 movz.0 a2.a7.t2
00000001200123D0 movz.1 a2.a7.t2
00000001200123D4 sw a2.4[v1]
00000001200123D8 lw a7.4[t1]
00000001200123DC daddu t1.s4.v0
00000001200123E0 addu a2.a7.a2
00000001200123E4 slt a7.a2.a0
00000001200123E8 movn.0 a2.a0.a7
00000001200123E8 movn.1 a2.a0.a7
00000001200123EC sw a2.4[v1]
00000001200123F0 daddu a2.s5.v0
00000001200123F4 lw a7.0[a2]
00000001200123F8 lw a2.0[a1]
00000001200123FC daddu t3.s6.v0
0000000120012400 addu a2.a7.a2
0000000120012404 sw a2.4[a1]
0000000120012408 lw t1.0[t1]
000000012001240C lw t8.0[v1]
0000000120012410 sw t1.48[sp]
0000000120012414 lw a7.48[sp]
0000000120012418 daddu t2.s7.v0
000000012001241C addu t8.a7.t8
0000000120012420 slt a7.a2.t8
```

```
0000000120012498    daddiu a4,a4,4
    cycle num:      3560
    instr num:      0

    f               fetch
    core            0
    slot            0
    pid             0
    tid             0
    squash_code     0
    debug_index     74
```

```
Dual Pipetrace Viewer v0.1     Press H to display help menu     Screen Size: 1299x810
```

Figure 1. *dptv* at default zoom in single-trace mode, annotations indicate different aspects of the visualization

3.1. Visualization

Figure 1 shows an instance of *dptv* which has just been opened visualizing an example trace in single-trace mode. Note that for the figures in this paper the background color has been changed to white to make them more gray scale printer-friendly. The pane on the right shows the trace displayed as previously described, with each pipeline stage plotted against time on the x-axis and dynamic instruction stream order on the y-axis. Each character column of the pipeline stage timing represents a cycle. Therefore, a column shows which pipeline stages were executed on the same cycle, and each row contains instructions from the execution of the execution of the benchmark. The left pane shows the list of dynamic instructions executed (and their instruction addresses) on the same row as their associated stages.

3.2. Stage Information

Pipeline stages are represented on the plot (right pane) by a single character which indicates which stage was executed. For example, 'f' represents Instruction Fetch, 'd' represents Instruction Decode, etc. The processor model shown in Figure 1 is an out-of-order execution pipeline, so instructions can potentially be buffered for several cycles before being selected for execution. The visualization uses dash '–' characters to indicate that an instruction has not executed any new pipeline stage for that cycle.

By hovering the mouse over one of the stages, additional information can be viewed. This additional information is the event data for the pipeline stage, for the given instruction. During simulation, an arbitrary amount of pipeline stage event data can be saved to the trace file. The event data is in a name/value pair. Figure 1 shows the information from a fetch stage being viewed (notice there is a mouse cursor highlighting the 'f' character for the very first instruction). Stage information is shown in the top-right corner and includes the full name of the stage, the instruction it's from, the cycle and instruction position, as well as stage-specific information such as the values of registers, which lane of the superscalar processor it is being executed on, etc. Stage information is included in the trace so what exact information is included is dependent on the program generating the trace.

3.3. Movement

The user can pan the diagram either by using the left mouse button and dragging the view, or by using key bindings. If the trace is off the window, a key binding can be pressed to snap the view horizontally back into the window. Another key binding can be pressed to reset the view to the default location as upon startup.

Zooming is done using the mouse scroll wheel. Scrolling down will zoom out, and scrolling up will zoom in. The zoom is centered on the current location of the mouse pointer. These zoom mechanics should be very familiar to any user having experience using CAD tools, photo editors, or (for example) Google Maps.

When zoomed out far enough, the letters representing each stage would become unreadable. Instead, *dptv* will switch the visualization such that the text disappears (both panes), and the pipeline stage timing panes switch to lines. These lines will connect like-pipeline stages from one instruction to the next – by default, a line to connect the first pipeline stage, Instruction Fetch and a second line for final commit/retirement of instructions. These pair of lines thus show the lifetime of all instructions, the space between these lines. Figure 2 shows an example trace, zoomed out far enough to show the line representation of the instruction stream. There are roughly 10 thousand dynamic instructions represented at the zoom level represented in this figure.

The pipeline stage that is connected by lines in the zoomed-out mode can be changed to a single stage, for any of the pipeline stages. For example, it might be helpful to see a line that connects the Execute pipeline stage for each instruction. This is changed with an additional command-line argument when invoking *dptv*.

3.4. Searching

Facilities to search through traces is also implemented in *dptv*. A key binding will initiate a search (the slash, similar to initiating a search in `vi/vim`), after which the search term can be typed and enter can be pressed to search. After entering in the search term additional key bindings allow the user to continue to the next match of the same search term. Both the current match and all other matches are highlighted. Unlike `vi/vim`, the search terms are not regular expressions.

dptv will search traces for occurrences of the search term; by default it will search in the instruction text, pc text, stage identifier (the single character on the plot), stage name, and the fields in the stage information. There is functionality to search in only one selected field, done by typing '/', then a character, then another '/', then the search string. For example, "/i/lw" will search only the instruction text for the string "lw". Additionally, the contents of a specific stage event data can be searched using "/v:(name)/(value)". A list of fields to search in and their associated search characters is included in the help text of the program.

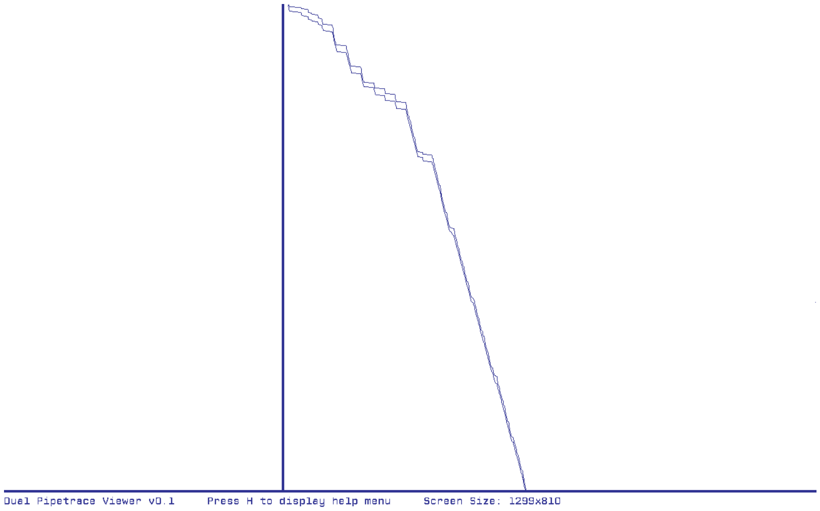


Figure 2. *dptv* at zoomed out in single-trace mode, such that lines connect the fetch stages and retire stages

There is also functionality to jump to a specific point in the trace, initiated by pressing ':', entering the instruction number to jump to, then pressing enter.

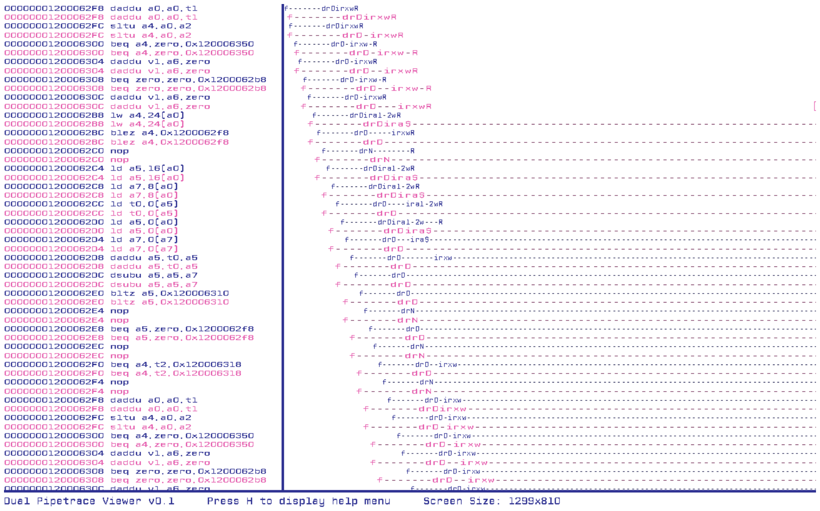
3.5. Dual-Trace Mode

When visualizing two traces, the plot will alternate between each trace for each row, as seen in Figure 3. Notice the right pane shows the same instruction twice, colored differently – these are the same instruction from each simulated processor configuration. Each traces instructions and stages are colored differently, including when zoomed out. These colors can also be changed at the command-line. As discussed previously, both traces are expected to be from the same benchmark, meaning all completed instructions (i.e. all instructions which retire) are the same in both traces. However, squashed instructions may differ, for example if one configuration predicts a taken branch while the other does not – whichever configuration mispredicted will have squashed instructions that will not retire. *dptv* will align all matching retired instructions, inserting dummy instructions to fill any necessary gaps. *dptv* will not attempt to visualize traces which do not have a common stream of committed instructions. If the two traces begin at different points in execution, by default *dptv* will remove the instructions before/after the common section.

Note again that the two traces do not need to be recorded at the same frequency. If they are not (also shown in Figure 3), then the trace with the slower frequency will be stretched horizontally to line the traces up correctly in time. This also means that two stages with the same cycle position but from different traces may not line-up in time.

The panning and zooming system works well with the dual-trace view to make comparing the traces easy. For example, the view can be zoomed far out to look at the trend lines of both traces, and then can be zoomed back in to points where the traces performance diverges, allowing the user analyze the specific details of what happens differently between them.

When comparing two traces, it is possible, in fact very likely, that the two traces have performance that progressively displays one trace off of the window. It is helpful to be able to re-establish a point in the



```
0000000120006 2F8 daddu a0.a0.t1
```

Figure 3. *dptv* displaying traces of instructions that were executed on two different processor configurations

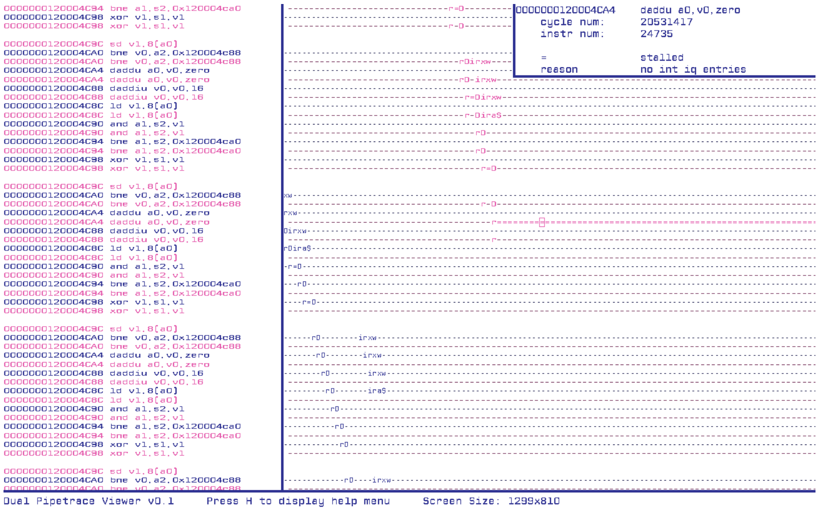
visualization where the traces are realigned in time. For example, suppose the two traces may have the same performance for a phase of the benchmark, but maybe one trace has a cache miss that did not occur in the opposing trace. This causes the lower performing trace to stall for a little bit, pushing the visualization off the right hand side of the window. Maybe later both traces return to the same level of performance.

A similar scenario is shown in Figure 3. At the beginning of time both traces start together, but after a while, both traces will continue to be spaced out (notice fetch ‘f’ for like instructions diverge) even if their performance reconverges. To remedy this, the right mouse button can be pressed to realign the traces to that point time. The traces will shift horizontally to converge at the point clicked. This realignment can be done at any zoom level.

4. Example Usage

For an example of performance comparison and the utility of *dptv*, we simulated a phase of the `462.libquantum` benchmark from the SPEC CPU2006 Benchmark Suite [1]. This benchmark is known to have performance that scales well with additional microarchitectural resources. We ran the simulation twice: once with a superscalar width of one and again with a superscalar width of two, and all other microarchitectural variables (hardware table sizes, predictor implementations, etc.) remaining constant. The summary statistics showed that the performance of the two executions were very close, however the one-wide simulation ran slightly faster than the two-wide simulation. This is a surprising result, and prompts the question of *why* this would be the case. We load the generated `462.libquantum` traces into *dptv* to elucidate the run-time behavior. Figures 4 and 5 shows a zoomed-in view and a zoomed-out view, respectively. The one-wide trace is colored in cyan and the two-wide trace is colored in magenta in both figures.

It becomes immediately evident that the phase of `462.libquantum` traces consists of a small loop body, seven static instructions. This same loop continues for the entirety of both traces. In the zoomed-out view we can see that both traces seem to delay execution (the horizontal lines on the zoomed-out graph)



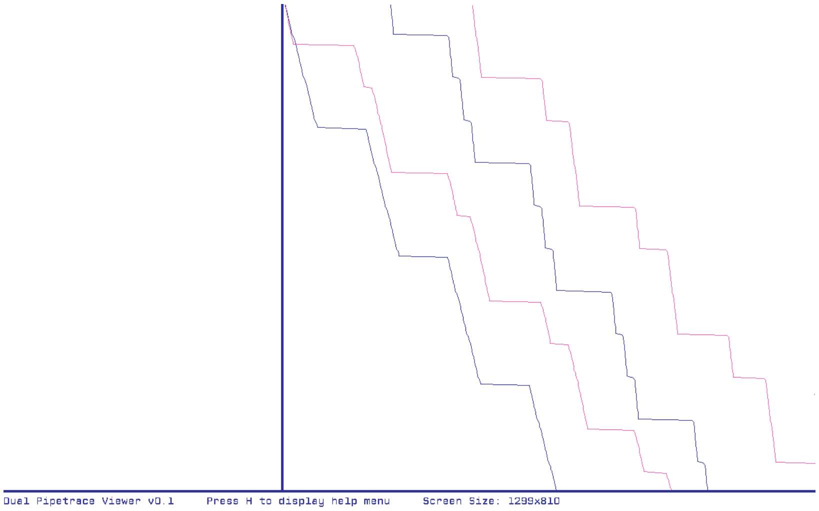
```
000000012000 4C94 bne a1,s2,0x120004ca0
000000012000 4C98 xor v1,s1,v1
000000012000 4C98 xor v1,s1,v1

000000012000 4C9C sd v1,8(a0)
000000012000 4CA0 bne v0,a2,0x120004c88
000000012000 4CA0 bne v0,a2,0x120004c88
000000012000 4CA4 daddu a0,v0,zero
000000012000 4CA4 daddu a0,v0,zero
000000012000 4C88 daddiu v0,v0,16
000000012000 4C88 daddiu v0,v0,16
000000012000 4C8C ld v1,8(a0)
000000012000 4C8C ld v1,8(a0)
000000012000 4C90 and a1,s2,v1
000000012000 4C90 and a1,s2,v1
000000012000 4C94 bne a1,s2,0x120004ca0
000000012000 4C94 bne a1,s2,0x120004ca0
000000012000 4C98 xor v1,s1,v1
000000012000 4C98 xor v1,s1,v1

000000012000 4C9C sd v1,8(a0)
000000012000 4CA0 bne v0,a2,0x120004c88
000000012000 4CA0 bne v0,a2,0x120004c88
000000012000 4CA4 daddu a0,v0,zero
000000012000 4CA4 daddu a0,v0,zero
000000012000 4C88 daddiu v0,v0,16
000000012000 4C88 daddiu v0,v0,16
000000012000 4C8C ld v1,8(a0)
000000012000 4C8C ld v1,8(a0)
000000012000 4C90 and a1,s2,v1
000000012000 4C90 and a1,s2,v1
000000012000 4C94 bne a1,s2,0x120004ca0
000000012000 4C94 bne a1,s2,0x120004ca0
000000012000 4C98 xor v1,s1,v1
000000012000 4C98 xor v1,s1,v1

000000012000 4C9C sd v1,8(a0)
000000012000 4CA0 bne v0,a2,0x120004c88
000000012000 4CA0 bne v0,a2,0x120004c88
000000012000 4CA4 daddu a0,v0,zero
000000012000 4CA4 daddu a0,v0,zero
000000012000 4C88 daddiu v0,v0,16
000000012000 4C88 daddiu v0,v0,16
000000012000 4C8C ld v1,8(a0)
000000012000 4C8C ld v1,8(a0)
000000012000 4C90 and a1,s2,v1
000000012000 4C90 and a1,s2,v1
000000012000 4C94 bne a1,s2,0x120004ca0
000000012000 4C94 bne a1,s2,0x120004ca0
000000012000 4C98 xor v1,s1,v1
000000012000 4C98 xor v1,s1,v1

000000012000 4C9C sd v1,8(a0)
000000012000 4CA0 bne v0,a2,0x120004c88
000000012000 4CA0 bne v0,a2,0x120004c88
```

```
000000012000 4CA4    daddu a0,v0,zero
    cycle num:       20531417
    instr num:       24735

    =                stalled
    reason           no int iq entries
```

```
Dual Pipetrace Viewer v0.1     Press H to display help menu     Screen Size: 1299x810
```

Figure 4. *dptv* showing fine grained instruction behavior of the `462.libquantum` benchmark for two processor configurations

on a semi-regular basis. The delays seem to always occur on the same instruction when zoomed-in, but not on every loop iteration. In-between the delays the two-wide trace has a steeper slope, meaning it executes more instructions over the same cycle time, as is expected. What makes it fall behind the one-wide seems to be both the number and duration of these delays, so our next goal is to find out what is causing those. The simulator used to generate these traces provides information about each stall in the instruction that caused it. Zooming into a section where the two-wide execution delays and scrolling up (going back in instruction order but staying at the same point in time), we find the result of the delays is an instruction stalling due to a full issue queue. Since the two-wide fills execution resources more quickly than the one-wide, these structures fill, eventually stalling the processor front-end. The simulator used to generate these traces uses typed lanes for the back-end, requiring at least a lane for arithmetic



```
Dual Pipetrace Viewer v0.1     Press H to display help menu     Screen Size: 1299x810
```

Figure 5. *dptv* showing broad instruction behavior of the `462.libquantum` benchmark for two processor configurations instructions, a lane for CTIs, and a lane for memory instructions. Thus the back-end width must be at least three-wide, greater than the frond-end width of the simulated one-wide and two-wide. Since the two-wide fills the issue queue at a higher rate than the one-wide, yet they both drain the issue queue at the same rate depending on the instruction mix, this causes the two-wide to exceed the issue queue capacity at times when the one-wide did not.

While the problem most certainly could have been identified without *dptv*, the use of this tool greatly increased the speed at which the issue could be identified and possible fixes considered.

5. Status and Future Work

dptv is still under active development, although as of this publication date, its code base is quite mature and stable. It is able to open large trace files, upwards of hundreds of thousands to millions of dynamic instructions. Although extensive testing has not been done, relatively modest hardware (8th generation Intel Core i5 laptop, 8GB memory, integrated graphics) can fluidly navigate traces. Testing on this viewer tool will continue, but it is not anticipated that any major changes will be necessary.

Development and testing has primarily been carried out on a Linux platform. Future work will ensure compatibility with Apple MacOS and Microsoft Windows environments.

In addition to the stand-alone viewer tool, it is also anticipated to release libraries written in various languages, intended to be easily integrated into existing processor simulations. These libraries will provide a straight-forward API to collect trace information while simulating benchmarks. As a simulation completes, the API can then automatically generate the trace file in the appropriate file format to be read into the viewer tool. Currently, there is only one library, written in C++, which has been integrated into an in-house developed processor simulator that is not publicly available. A possible change to the code base will use the same file format as that used by `o3`, to simplify integration with `gem5`. Currently, *dptv* uses a custom plain-text file format.

Up-to-date project status, and eventual code release will be maintained at https://cs.uwlax.edu/~eforbes/dptv/.

References

[1] "SPEC CPU2006 Benchmark Suite." The Standard Performance Evaluation Corporation, 2006. [Online]. Available: http://www.spec.org/cpu2006/

[2] "gem5: Visualization," 2023, Reference. [Online]. Available: https://www.gem5.org/documentation/general_docs/cpu_models/visualization/

[3] "Simple DirectMedia Layer," 2023, Reference. [Online]. Available: https://www.libsdl.org/

[4] J. Alsop, M. Sinclair, R. Komuravelli, and S. Adve, "GSI: A GPU Stall Inspector to Characterize the Sources of Memory Stalls for Tightly Coupled GPUs," in *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software*, April 2016, pp. 172–182.

[5] A. Ariel, W. Fung, A. Turner, and T. Aamodt, "Visualizing Complex Dynamics in Many-Core Accelerator Architectures," in *Proceedings of the 2010 IEEE International Symposium on Performance Analysis of Systems and Software*, March 2010, pp. 164–174.

[6] A. Baranov, P. Panfilov, and D. Ponomarev, “PowerVisor: A Toolset for Visualizing Energy Consumption and Heat Dissipation in Modern Processor Architectures,” in *Proceedings of the 12th International Conference on Parallel Computing Technologies*, September 2013, pp. 149–153.

[7] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, “The gem5 Simulator,” *ACM SIGARCH Computer Architecture News*, vol. 39, no. 2, pp. 1–7, May 2011.

[8] D. Burger, T. M. Austin, and S. Bennett, “Evaluating Future Microprocessors: the SimpleScalar Tool Set,” 1996.

[9] S. Card, “Information Visualization,” in *The Human-Computer Interaction Handbook: Fundamentals*, January 2002, pp. 209–543.

[10] Q. Gao, X. Zhang, P. Rau, A. Maciejewski, and H. Siegel, “Performance Visualization for Large-Scale Computing Systems: A Literature Review,” *Human-Computer Interaction. Design and Development Approaches: 14th International Conference, HCI International*, vol. 10, no. 1, pp. 450–460, July 2011.

[11] J. Hennessy and D. Patterson, *Computer Architecture: A Quantitative Approach, 5th ed.* Waltham, MA: Morgan Kaufmann, 2012.

[12] D. Koppelman and C. Michael, “Discovering Barriers to Efficient Execution, Both Obvious and Subtle, Using Instruction-Level Visualization,” in *Proceedings of the First Workshop on Visual Performance Analysis (held in conjunction with Supercomputing SC14)*, November 2014.

[13] J. Roberts and C. Zilles, “TraceVis: An Execution Trace Visualization Tool,” in *Workshop on Modeling, Benchmarking and Simulation*, June 2005.

[14] C. Weaver, K. Barr, E. Marsman, D. Ernst, and T. Austin, “Performance Analysis Using Pipeline Visualization,” in *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software*, November 2001, pp. 18–21.

Practical studying and conscious lifestyle

Thao Huy Vu
Computer Science
Maharishi International University
Fairfield, IA, 52557
thaovu@miu.edu

Asaad Saad
Computer Science
Maharishi International University
Fairfield, IA, 52557
asaad@miu.edu

Abstract

Nowadays, getting a job in the technology field is a fashion, especially becoming a part of the software industry. Many people from different backgrounds want to change their paths to become software developers. At the same time, there are many programs to educate a person to have a skillset for the technology industry. However, how can we effectively teach students who can not only do the job immediately when they finish academy or university but can survive well in this challenging industry livelong? One method I would want to propose in this paper is to provide education which makes students study happily and have a good lifestyle. In my proposal, the first thing we as educators need to ensure that students can see the future in their learning. The simple way to do that is by leading students to actual projects. What they learn is not just skeptical things on the air for their exams, but they can use what they learn immediately to apply to their projects similar to other actual projects or systems they can quickly check out in reality. That will show them the way how the projects in actual companies work. It helps them not be surprised when taking interviews, starting the job, or getting along on projects. The second thing is nourishing them with a good lifestyle which can improve their mind. It means that they should always be aware of all activities they have done daily. They need to acquire all knowledge daily, and they can improve themselves continuously. I call this a conscious continuous improvement which is one of the essential skills for modern life because our life is changing speedily day in and day out. For this second purpose, I want to recommend four things that contribute to a mindful lifestyle: healthy foods, sleeping, exercising, and

meditation. If all students can achieve this happy studying and the proposed lifestyle day-to-day, they can not only have a great job in the technology industry. Still, they can follow the industry deliberately in their lifetime.

Thao Huy Vu and Asaad Saad

Computer Science

Maharishi International University

Fairfield, IA, 52557

thaovu@miu.edu and asaad@miu.edu

1 Introduction

Based on my experience as a computer science student at three different universities in three other countries and more than ten years working as a software engineer, I recently found that conventional education needs to change to be a truly transformative environment for students. In this paper, I propose a method for modern education in computing that includes two essential aspects: practical studying and a conscious lifestyle. Why do we need this change?

The main goal of this education methodology is to educate students to emerge into the industry faster. Then they can improve themselves like how they learn at universities, which is conscious, continuous improvement. Students can see their future at universities by doing actual projects; they can discuss these projects with teachers or classmates who might become their future colleagues. Moreover, universities complete the purpose of transitioning students to industry. That is also the target of my first facet, practical studying. However, modern life is so stressful, and people can quickly get anxiety or depression. Therefore, it is very critical to maintain physical and mental health.

The second part of this proposal has a healthy lifestyle to deal with the demanding and intensive industry. Four things contributing to this lifestyle are healthy foods, sleep, exercise, and meditation. If students and educators keep this lifestyle, they will live with their full potential in reality.

This proposal brings a new way of teaching and learning to the conventional education systems. As a result, all computing educators can leverage the existing educational methodology and my proposal to create effective teaching and learning.

2 Practical studying

As I mentioned, practical studying is the learning method with a practice that is as close to reality as possible. If students learn this way, they can go to work quickly and efficiently. That completes one purpose of education. However, are students able to do that with the current education system?

Most universities nowadays still keep the traditional way of education. We teach the students a lot of knowledge for passing the exams. I do not blame the exams, but the conventional education style slows how students can create a happy life with their learning after graduating. Some say that teaching for exams is essential knowledge, but technology is changing too fast, and the studying time of students is limited. Hence, we must teach students only a few things they can use immediately in their work and quickly grasp new concepts based on mastery. Students can only apply a few concepts in their actual projects, not all.

But what if we select some fundamental theory to teach them, and they can use them in their practice? As a result, they can fully understand these concepts for the exams, but in reality. They can go to work immediately with their knowledge and expand their knowledge from their work. That means they store all their hypothesis in long-term storage because they fully understand and apply that in their practices.

It is a very sustainable way to do this in computing because many people worldwide add new knowledge daily. It is the same that many people working in this industry need to study new knowledge day-to-day. Practical studying is a suitable method in this case. Students learn the basic and latest technology that many companies want to employ. Therefore, they can easily apply for a job or go to work. Moreover, students with this studying can expand their knowledge like how they study practically at universities.

One good example of this education is the university of Waterloo, one of the top engineering universities in Canada, because they have a lot of practical training for students even though they are a new university [1]. But unfortunately, applying this way to other universities is hard because we can only have a few internships for all students.

However, we can create a hands-on experience at the university. I called that practical studying. This simple way leads students to work toward projects as close as actual products or real projects based on their understanding of fundamental hypotheses. They can have hands-on experience at the universities when performing in the same manner as the companies require. Educators must keep updated on the industry and then teach students to do something similar. They can study real projects and perform the tasks like they are in the company. Moreover, they can deeply understand the concept by knowing and applying it to the projects.

I want to mention two things in the practical studying here: fundamental concepts and application. Moreover, I also discuss the evaluation advantages if we apply this studying.

2.1 Fundamental concepts

The fundamental concepts are basic and profound. They are like blocks that students can use to accomplish more important things. Educators can introduce complex knowledge, but it should be optional. Educators must classify the ability to select the good ones for the industry and teach students. And students can study these basic blocks for their grades but for actual projects. Complex topics should be at will, it means students can explore if they are enthusiastic, but they know it is not mandatory. Educators must also create exams or homework based on the basic blocks. Educators might have difficult questions, but it is just for excellent students.

Moreover, educators discuss the basic knowledge necessary to pass the course and their future projects. Many people will wonder, if we only provide the basic concepts, how can students create something different? However, a new idea is just a permutation of basic knowledge. If students understand the foundation, they can make new ideas sometimes.

2.2 Application

Many universities teach students all academic knowledge. They expect that the more students have, the more they can accomplish their job. This conventional method works well for people who want to study higher education if they have intensive academic knowledge. However, after working with some new graduates in the industry and considering my case as a fresher again, I wondered why universities do not teach students more actual applications.

The traditional way does not work well for students wishing to have jobs and have to up to date the coming knowledge after graduation. In the computing industry, the quicker you can work on projects, the deeper you can understand these projects. Unfortunately, many graduates take months to join the projects, and some cannot finish their probation because of their poor performance; even though they know many things, they need to learn more about the basics. Even worse, they sometimes cannot use their knowledge because companies move to new and better ones. It wastes a lot of time and money from students, companies, and universities.

As I proposed here, educators should lead students to do actual projects or to build applications as close as the real ones they can refer to in their capacity. For example, we can ask students to create an online library, but we can ask them to go to the library to check the software and get the idea. After that, they can see the product and develop a similar outcome. They reinvent the wheels but study from the existing ones, not the old ones that can be obsolete. After that, these projects become their reference or building block; they can understand and apply their knowledge in the future.

2.3 Easy Evaluation

One problem in conventional education is that students think their professors are on a different side than them when they get low grades. If we see sports, all students like their coaches because they think coaches support them to improve.

What are the differences between coaches and teachers? They are the same. They all bring knowledge to students—however, the way they evaluate students differently. Coaches grade depending on the practical exercise that is clear and reasonable to students because they can feel and see their progress in their practices. Teachers do the same thing, but they grade students through exams. Grades or exams are not the problem, but it does not persuade students whey they get low grades.

What if we cooperate in sports and academics? What if we have grades, but it not only shows the exams but also the progress of actual projects in the course? That is what practical studying aims. If we apply this methodology, the grades are not a big problem between teachers and students because they can know their proceed in everything, not just see one facet of the exam. It helps to dissolve the barrier between teachers and students. Teachers can quickly evaluate students through their movement in homework, projects, and exams. Seeing the benefit of the course in their job later inspires the studying purpose of students. Moreover, all students want to gain hands-on experience to understand the hypothesis practically and fully. And they can pass exams naturally easier than memorizing the bunch of knowledge that they only use for these exams.

3 Conscious lifestyle

As a student, I saw a common phenomenon that most of my friends spent overnight studying for the exam. They got good grades, but they remembered nothing after that. At present, it is still happening at many universities. Many students learn wrongly because universities only focus on discipline rather than lifestyle. Moreover, many argue that we do not need to guide students since they are adults. It is a drawback of conventional education systems.

Since the computing industry is very stressful due to the fast-changing and demanding environment, new graduates will be shocked if the universities do not prepare well for them. If students only keep the current lifestyle in this industry, it is tough to keep them happy.

I propose a new lifestyle, which students can apply at universities and later, is the conscious lifestyle. This lifestyle includes healthy foods, exercise, sleep, and meditation. Universities must be transformative environments for students to practice this lifestyle before using it outside.

3.1 Healthy foods

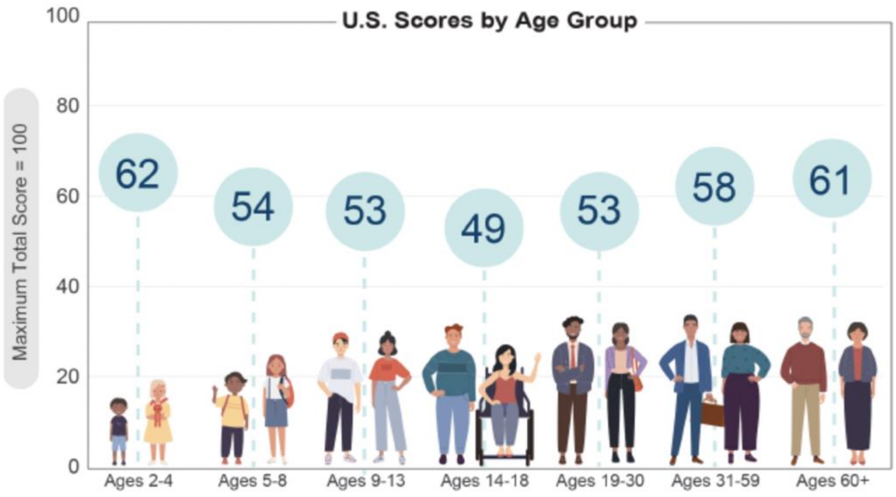


Figure 1: HEI scores for Americans.

As depicted in the above picture from USDA (United States Department of Agriculture), adults from 14 to 30 years old have the lowest grades on Healthy Eating Index [2]. It means they only care a little about healthy eating since they are in the golden age. I know that they are healthy, can absorb any food, and can easily overcome stress or illness. However, if they can eat well during this period, they can improve their long-term health. Thus, they can study many new things if they have excellent health.

There is an old saying you are what you eat. You know that foods are essential to building solid physical and mental health because foods provide all nutrients that the body needs to produce energy for any activity. Since life is too fast, many people prefer to eat quickly. It is the same for students. They need more time to prepare good meals, but universities should encourage them to have healthy food. They should eat more vegetables, fruits, and enough protein. If they can eat healthy foods, they can have good health to obtain knowledge quickly and release stress.

Moreover, many people drink too much alcohol, which can increase the harmful effects. According to NIH (National Institute for Alcohol abuse and alcoholism), an alcohol overdose can cause mental confusion, difficulty remaining conscious, vomiting, seizure, trouble breathing, slow heart rate, clammy skin, dulled responses (such as no gag reflex, which prevents choking), and shallow body temperature [3]. Students have a lot of parties, and many of them drink to a great extent. They know the side effects of alcohol, but they still want to drink because it is fun. If we remind or encourage them not to drink while they have a course since they will lose their knowledge due to some blocks from alcohol. When students do not use alcohol, their brain is conscious, and they can gain understanding quickly.

3.2 Exercise

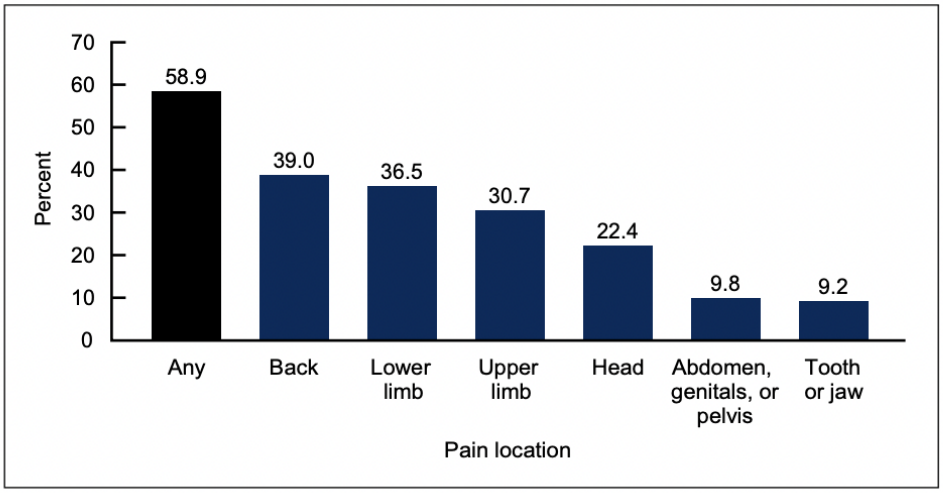


Figure 1: Back, Lower Limb, and Upper Limb Pain Among U.S. Adults, 2019.

As depicted in the above figure from CDC (Center for Disease Control and Prevention), 39 percent of adult Americans had back pains in 2019. Nowadays, people can have all services at home and even work from home with the internet. However, many people spend hours in front of computers without physical activities. Consequently, they have problems with their backs after a few years.

One of the best natural killers for this pain is exercise. With 30 minutes of walking, running, or any physical activity, bones, muscles, and other orgasms are more robust. Consequently, the body can resist many diseases. Moreover, exercise helps improve mental health, release stress, and sleep well. Some students might have many reasons to avoid exercise, but universities must have a way to stimulate them to practice daily for their long-term future. It can benefit students, universities, and society.

3.3 Sleep

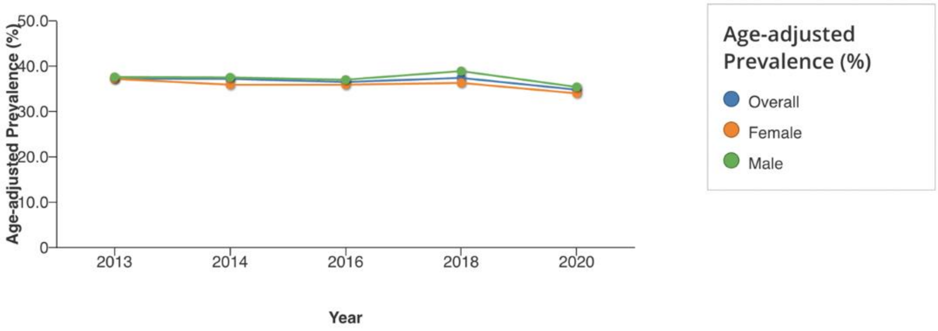


Figure 2: Age-adjusted Prevalence.

This figure reports the short sleep duration (less than 7 hours within 24 hours) is approximately unchanged from 2013 to 2020. According to CDC, one-third of adult Americans do not have enough sleep (7-9 hours per 24 hours) [4] during this period.

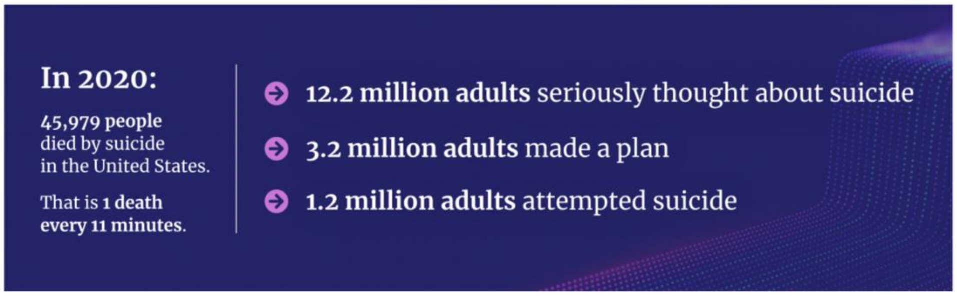
Sleep is essential to rest and rejuvenate the body for the following activities. If we lack sleep, we cannot have a fully awakened brain, and our actions cannot reach their full potential. Hence, it causes low performance and stresses out.

According to new research on the relationship between sleep and memory [8], long-term memory formation depends on sleep quality. There are two sleep stages: early sleep or slow wave sleep (SWS) and late sleep or rapid eye movement (REM) sleep. SWS reactivate recently encoded neuronal memory representations and transforms respective presentations for integration into long-term memory. And then, REM sleep may stabilize transformed memories. This process is significant in determining how students can keep their knowledge for the future.

Hence, students need to sleep enough to recharge their bodies, gain knowledge, and prepare for the next day. However, many students do not care about this tiny thing. They still study overnight. Libraries at many universities open very late for students. Students will be exhausted sooner or later, and when they have vacations, they want to sleep more or drink too much to release stress. This way damages their health in the long term. Universities and students should cooperate to resolve this simple problem.

Universities and educational institutions should encourage and remind educators and students to sleep well for a while later to have the full potential power. It will help both teaching and learning better. Moreover, students do not have stress and have excellent sleep habits in their daily life later.

3.4 Meditation



In 2020:

45,979 people died by suicide in the United States.

That is **1 death every 11 minutes.**

- **12.2 million adults** seriously thought about suicide
- **3.2 million adults** made a plan
- **1.2 million adults** attempted suicide

Figure 3: Suicide statistics.

As shown in the above figure from CDC, more than 45 thousand people died by suicide in the United States in 2020 [5]. As you know, the main reason is the mental problem these people cannot resolve since people have a lot of stress, depression, or anxiety daily. Since we cannot sort out all issues by relaxing or sleeping, the remaining accumulates in our bodies. The accumulation gets bigger like we pump air into a balloon every day. The balloon will expose itself for some days if we do not find a solution.

One of the best solutions to fix this problem is meditation. Many scientists have found that meditation helps to boost learning ability and creativity, improve brain functioning, and reduce stress and anxiety [5]. As their recommendation, if people practice meditation twice daily, their mental health will be much better.

Since the computing industry is changing too fast, anyone wanting to join it must improve their relief. Students have a lot of pressure because they are new to the environment without taking care of their parents. Universities are a transformative place before students can go into the industry. Therefore, it is essential that universities can lead students to practice meditation, which they can use this tool for dealing with stressful reality.

Conclusion

As you know, to live well in the modern world, people need to do their job well and have continuous self-studying capabilities to gain new knowledge. To do that, they need to have a crystal understanding of fundamental concepts and excellent mental and physical health to survive in this technology world which is moving extremely fast daily. And universities are vital in guiding students to have a long, happy life after graduation.

I have proposed a new method for modern computing education in this paper. The first thing in this method is practical learning which students will gain knowledge based on their projects. As a result, they can pass the academic requirements but can get a job quickly. To have good mental health at universities and in the future, the second part of this method is a conscious lifestyle so that students can have a healthy body and a great mind. To have this lifestyle, universities guide students to have good eating and drinking, exercise, sleep, and meditation habits. If students can apply this method daily, they can have a better life in the fast-changing computing world.

References

[1] Author Salman Khan. The One World School House. pages 236–237, September 2013.

[2] U.S. Department of Agriculture, Center for Nutrition Policy and Promotion. https://www.fns.usda.gov/hei-scores-americans

[3] National Institute on Alcohol abuse and alcoholism. Alcohol and brain. https://www.niaaa.nih.gov/publications/alcohol-and-brain-overview.

[4] Centers for Disease Control and Prevention. Data and Statistics: Short Sleep Duration Among US Adults. https://www.cdc.gov/sleep/data_statistics.html.

[5] Centers for Disease Control and Prevention. Data and Statistics: Short Sleep Duration Among US Adults. https://www.cdc.gov/suicide/suicide-data-statistics.html

[6] Maharishi International University. Transcendental Meditation: https://www.miu.edu/about-miu/transcendental-meditation-technique.

[7] Centers for Disease Control and Prevention. Data and Statistics: Back, Lower Limb, and Upper Limb Pain Among U.S. Adults, 2019. https://www.cdc.gov/nchs/data/databriefs/db415-H.pdf

[8] Bijorn Rasch and Jan Born. About Sleep's role in Memory. 2013 April. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3768102/

Darknet Traffic Using Deep Learning

Muhammad Abusaqer and Quinn Sullivan

Department of Math and Computer Science

Minot State University

Minot, ND, USA

muhammad.abusaqer@minotstateu.edu

quinn.sullivan@minotstateu.edu

Abstract

This research investigates the potential of deep learning for the analysis of darknet traffic, a network operating outside the traditional internet and frequently associated with illegal activities such as drug dealing, trafficking, and exploitative content. The darknet also provides a secure communication and information exchange platform for privacy-conscious individuals. The study aims to classify darknet traffic into 8 categories - P2P, Audio-Streaming, Browsing, Video-Streaming, Chat, Email, File-Transfer, and VOIP - for accurate categorization of real-time applications and to support law enforcement agencies in detecting and preventing malicious activities. A custom-designed Artificial Neural Network (ANN) model was trained using the CIC-Darknet2020 dataset to perform multi-class classification of darknet traffic. The ANN model's performance was compared to two established machine learning algorithms, XGBoost and RandomForest. The results demonstrated that although the ANN model showed promise, it was outperformed by both XGBoost and RandomForest models. This paper presents a contribution by applying deep learning to the CIC-Darknet2020 dataset and comparing its performance with traditional machine learning models. The findings highlight the potential capabilities of deep learning models in analyzing darknet traffic and suggest avenues for future improvements.

1 Introduction

The Darknet, a hidden and encrypted part of the internet, has become a significant challenge for law enforcement and cybersecurity professionals due to its association with criminal activities and illicit services [1]. Operating on overlay networks such as Tor and I2P, the Darknet provides a high level of anonymity and privacy for its users, making it an attractive platform for illegal activities such as drug trafficking, hacking, and financial fraud [2]. The rapid growth and increasing complexity of Darknet ecosystems necessitate advanced monitoring and analysis tools to identify and combat these malicious activities [3]. Recent research has focused on developing novel methods for Darknet traffic classification, utilizing machine learning and deep learning techniques to detect and analyze suspicious activities and enhance cybersecurity efforts [4] [5]. As the Darknet continues to evolve, researchers and practitioners must adapt their approaches and develop innovative strategies to stay ahead of emerging threats and protect the integrity of online systems and networks.

This research paper focuses on the classification of a darknet dataset using both traditional machine learning and deep learning models. The results from the Artificial Neural Network (ANN) model will be compared with those from traditional machine learning models, namely RandomForest and XGBoost. The motivation behind this research is to explore the potential of machine learning in darknet classification and to gain an understanding of how various algorithms perform in this context. The study is based on the CIC-Darknet2020 dataset [6] [7].

2 Related Work

One of the relevant studies in the field of darknet traffic classification is the work by Iliadis and Kaifas in [7]. The authors explored the application of various machine learning models to classify darknet traffic effectively. The primary motivation behind their research was the growing importance of identifying and classifying darknet traffic for cybersecurity purposes, as understanding the nature of such traffic can provide valuable insights into potential threats and vulnerabilities. In their study, Iliadis and Kaifas [7] experimented with a range of machine learning algorithms, including Decision Trees, Random Forests, k-Nearest Neighbors (k-NN), and others. They aimed to find the most accurate and efficient approach for classifying darknet traffic. Their research highlighted the need for advanced methods that can accurately distinguish between different types of darknet traffic, which can, in turn, contribute to improved cybersecurity measures and threat detection. The findings of [7] serve as a valuable reference for the current research, as they provide insights into the effectiveness of different machine learning techniques in the context of darknet traffic classification. The comparison of their results with the outcomes of the present study, which employs an Artificial Neural Network (ANN) model along with RandomForest and XGBoost classifiers, can shed light on the relative performance of deep learning approaches versus traditional machine learning methods in this domain.

In [8], DarknetSec, a novel self-attentive deep learning framework, has been proposed to improve darknet traffic classification and application identification. It employs a cascaded model combining a 1D Convolutional Neural Network (CNN) and a bidirectional Long Short-Term Memory (Bi-LSTM) network to capture local spatial-temporal features from packet payloads. The self-attention mechanism, integrated into the feature extraction network, uncovers hidden relationships among the extracted content features. DarknetSec also extracts side-channel features from payload statistics to enhance performance. Evaluated on the CICDarknet2020 dataset, DarknetSec outperforms state-of-the-art methods, achieving a multiclass accuracy of 92.22% and a macro-F1-score of 92.10%. It also maintains high accuracy in other encrypted traffic classification tasks.

Almomani proposed a novel darknet traffic analysis and classification system based on modified stacking ensemble learning algorithms in [9]. The study focused on utilizing stacking ensemble learning, a machine learning technique that combines multiple learning mechanisms to generate more accurate predictions. The system was evaluated on a dataset containing over 141,000 records from CIC-Darknet 2020, the same dataset used in this study. The experimental results showcased the classifiers' ability to distinguish between benign and malignant traffic, with accuracy rates exceeding 99% during the training phase and 97% in the testing phase. The study utilized a two-tiered learning stacking scheme that incorporated both individual and group learning, with three base learning methods, including neural networks, random forests, and support vector machines. The ensemble approach demonstrated better performance compared to single techniques, particularly when handling small historical windows, suggesting that the system becomes more robust and accurate as data grows. Despite limitations related to performance and privacy concerns, the proposed system offers a promising direction for future research in darknet traffic classification and analysis, exploring various ensemble schemes and methodologies to enhance its effectiveness against different types of attacks [9].

Sridhar and Sanagavarapu in [5] conducted a study on darknet traffic classification, aiming to enhance network security by detecting threats or risks. The authors used the standard CIC-Darknet2020 dataset, which contains instances of both benign and darknet traffic. They performed feature importance analysis using the Chi-Squared statistical score for feature selection and addressed the imbalance of classes by applying oversampling with Conditional Generative Adversarial Networks (Conditional GANs). The multi-class classification of traffic encryption types was carried out using the Random Forest classifier, achieving a 97.87 F1-Score for traffic encryption classification. In their conclusion, they suggested exploring feature extraction through Principal Component Analysis and employing Recurrent Neural Networks for detecting attacks over time as potential future work.

The paper [10] presents an approach to darknet traffic analysis using a weight agnostic neural network (WANN) framework for real-time detection of malicious intent. The authors propose a method that leverages big-data analysis techniques and network management practices to process and classify darknet traffic data. They aim to improve the efficiency and effectiveness of malicious intent detection in darknet traffic by using a WANN framework, which is capable of learning and generalizing from limited training data. This study contributes to the ongoing research on darknet traffic classification and detection of malicious activities. The proposed WANN framework offers a promising approach to enhance cybersecurity efforts by automating the process of detecting threats in real-time.

Al-Qatf et al. in [11] proposed a deep learning approach for network intrusion detection that combines a sparse autoencoder with a Support Vector Machine (SVM). The authors recognized the importance of effective network intrusion detection systems to counter the growing number of cyber threats. They introduced a deep learning method that leverages a sparse autoencoder to extract relevant features from network traffic data and an SVM classifier to categorize the traffic as normal or malicious. The proposed system was trained and tested on a dataset consisting of various network traffic instances. The results indicated that the combined deep learning approach outperformed traditional machine learning techniques in terms of detection accuracy and generalization performance. This research highlights the potential of hybrid deep learning methods in enhancing network intrusion detection and providing more effective solutions for cybersecurity professionals.

3 Proposed Methodology

3.1 Dataset

The dataset used in this research is the CIC-Darknet2020 dataset, obtained from the Canadian Institute for Cybersecurity [6]. The Darknet-2020 dataset was chosen over other available datasets due to its recency and relevance to the research objectives. The dataset encompasses a mix of data types, including numerical, categorical, and text features. The 'Label' column in the dataset contains eight distinct classes, which are P2P, Audio-Streaming, Browsing, Video-Streaming, Chat, Email, File-Transfer, and VOIP. These classes represent different types of darknet traffic that the models aim to classify in this study.

3.2 Dataset Preprocessing

The original dataset comprised 141,530 rows with 85 columns. However, given the computational resource constraints faced by the authors, a random subset of 8,000 rows was selected for the experiment. During the preprocessing stage, the authors encountered issues with certain columns that negatively impacted the model's performance. Consequently, these problematic columns were dropped, along with a few others that did not contribute significantly to the output. Additionally, several preprocessing steps were applied to handle missing values and encode categorical features. Large entries in the dataset were replaced with NaN, missing values in numeric columns were imputed using mean imputation, and non-numeric columns with missing values were imputed using the most frequent (mode) imputation method. All non-numeric features, excluding the 'Label' column, were encoded as integers. The target variable ('Label') was encoded as integers using the LabelEncoder from the scikit-learn library.

3.3 Machine Learning Models

In this research, three different models were employed: an Artificial Neural Network (ANN) model, a RandomForest classifier, and an Extreme Gradient Boosting (XGBoost) classifier. The comparison aimed to evaluate the performance of the deep learning approach, as represented by the ANN model, against the well-established machine learning techniques of RandomForest and XGBoost in the context of darknet traffic classification. For all models, the dataset was split into training (80%) and testing (20%) sets, with the features scaled using the StandardScaler from the scikit-learn library.

3.3.1 Artificial Neural Network Model

Artificial Neural Networks (ANNs) are a class of machine learning algorithms that mimic the structure and function of the human brain, allowing them to learn patterns from data [12]. The neural network was designed with three layers: an input layer with 64 nodes and a ReLU activation function, a hidden layer with 32 nodes and a ReLU activation function, and an output layer with a softmax activation function [13] [14] [15] [16]. The model was trained for 10 epochs with a batch size of 32, and the optimizer used was the Adam optimizer [17].

3.3.2 RandomForest Model

RandomForest is an ensemble learning method that constructs multiple decision trees and combines their output to improve overall model performance and reduce overfitting [18] [19].

The RandomForest classifier in this research was instantiated with 100 estimators and a random state of 42 to ensure reproducibility. The model was then trained on the training set and used to make predictions on the testing set.

3.3.3 XGBoost Model

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting machines that uses a combination of decision trees and optimization techniques to improve model accuracy and speed [20] [21] ,

The XGBoost classifier [19] [20] in this research was trained on the training set, with the 'use_label_encoder' parameter set to 'False' and the 'eval_metric' parameter set to 'mlogloss'. After training, the classifier was used to make predictions on the testing set.

For all three models, evaluation metrics, including accuracy, precision, recall, and F1-score, were calculated to assess their performance in the context of darknet traffic classification.

3.4 Evaluation Metrics

To compare the performance of the Artificial Neural Network (ANN) model with RandomForest and XGBoost models in the context of darknet traffic classification, the following evaluation metrics were employed: accuracy, precision, recall, and F1-score.

3.4.1 Accuracy

Accuracy is the proportion of correct predictions (both true positives and true negatives) made by the model out of the total number of instances in the dataset. It is a commonly used metric to measure the overall performance of a classifier [22].

Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)

However, accuracy alone may not be an appropriate measure when the data is imbalanced, as it can be misleading when the majority of the instances belong to one class [23].

3.4.2 Precision

Precision is the proportion of true positives out of the total number of instances predicted as positive by the model. In other words, it measures the ability of the classifier to correctly identify the positive instances among all the instances predicted as positive [24].

Precision = True Positives / (True Positives + False Positives)

Precision is a useful metric in the context of darknet traffic classification when the cost of false positives is high, such as in identifying malicious activities where incorrectly labeling benign traffic can lead to unnecessary investigations or countermeasures [4].

3.4.3 Recall

Recall, also known as sensitivity or true positive rate, is the proportion of true positives out of the total number of actual positive instances in the dataset. It measures the ability of the classifier to identify all the positive instances [24] [4].

Recall = True Positives / (True Positives + False Negatives)

3.4.4 F1-score

F1-score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall [25]. It is particularly useful when dealing with imbalanced datasets, as it takes into account both false positives and false negatives [26].

F1-score = 2 * (Precision * Recall) / (Precision + Recall)

An F1-score of 1 indicates perfect precision and recall, while an F1-score of 0 indicates that either precision or recall (or both) are zero.

4 Experiment

4.1 Results

In this experiment, the performance of a custom-designed Artificial Neural Network (ANN) was assessed and compared to two established machine learning models, XGBoost and RandomForest, with respect to their classification accuracy. To ensure a fair comparison, all models were initially trained on a smaller dataset and subsequently tested on a larger dataset containing 8,000 randomly selected rows. The performance metrics for each model are as follows:

	Accuracy	**Precision**	**Recall**	**F1-Score**
Neural Network	0.74	0.74	0.74	0.71
XGBoost	0.83	0.83	0.83	0.83
RandomForest	0.81	0.80	0.81	0.80

Table 01: Experiments Results

5 Discussion

The results demonstrate that both XGBoost and RandomForest models outperformed the custom-designed neural network in terms of accuracy, precision, recall, and F1-scores for classifying darknet traffic. Although deep learning models hold great potential, the neural network did not surpass the performance of the XGBoost and RandomForest models in this specific classification task.

6 Future Work

In future work, the authors plan to enhance the ANN model by adding more nodes and hidden layers and continue experimenting until satisfactory results are achieved compared to the XGBoost classifier. Additionally, the entire dataset will be utilized for a more comprehensive analysis.

7 Conclusion

In conclusion, this study investigated the performance of a custom-designed ANN model in comparison to established machine learning models, XGBoost and RandomForest, for darknet traffic classification. The results showed that the ANN model did not outperform the other two models in this specific task. Further experimentation and improvements to the ANN model are necessary to achieve better classification results.

References

[1] M. Chertoff and T. Simon, "The impact of the dark web on internet governance and cybersecurity," Global Commission on Internet Governance, Paper Series 6, 2015. [Online]. Available: https://www.cigionline.org/static/documents/gcig_paper_no6.pdf

[2] G. Owen and N. Savage, "The Tor dark net /," 2015, Accessed: Feb. 28, 2023. [Online]. Available: https://policycommons.net/artifacts/1223621/the-tor-dark-net/1776697/

[3] D. Moore and T. Rid, "Cryptopolitik and the Darknet," Survival, vol. 58, no. 1, pp. 7–38, Jan. 2016, doi: 10.1080/00396338.2016.1142085.

[4] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

[5] S. Sridhar and S. Sanagavarapu, "DarkNet Traffic Classification Pipeline with Feature Selection and Conditional GAN-based Class Balancing," in 2021 IEEE 20th International Symposium on Network Computing and Applications (NCA), Nov. 2021, pp. 1–4. doi: 10.1109/NCA53618.2021.9685743.

[6] "Darknet 2020 | Datasets | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/datasets/darknet2020.html (accessed Feb. 09, 2023).

[7] L. A. Iliadis and T. Kaifas, "Darknet Traffic Classification using Machine Learning Techniques," in 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST), Jul. 2021, pp. 1–4. doi: 10.1109/MOCAST52088.2021.9493386.

[8] J. Lan, X. Liu, B. Li, Y. Li, and T. Geng, "DarknetSec: A novel self-attentive deep learning method for darknet traffic classification and application identification," Computers & Security, vol. 116, p. 102663, May 2022, doi: 10.1016/j.cose.2022.102663.

[9] A. Almomani, "Darknet traffic analysis, and classification system based on modified stacking ensemble learning algorithms," Information Systems and e-Business Management, pp. 1–32, Feb. 2023, doi: 10.1007/s10257-023-00626-2.

[10] K. Demertzis, K. Tsiknas, D. Takezis, C. Skianis, and L. Iliadis, "Darknet Traffic Big-Data Analysis and Network Management for Real-Time Automating of the Malicious Intent Detection Process by a Weight Agnostic Neural Networks Framework," Electronics, vol. 10, no. 7, Art. no. 7, Jan. 2021, doi: 10.3390/electronics10070781.

[11] M. Al-Qatf, Y. Lasheng, M. Al-Habib, and K. Al-Sabahi, "Deep Learning Approach Combining Sparse Autoencoder With SVM for Network Intrusion Detection," IEEE Access, vol. 6, pp. 52843–52856, 2018, doi: 10.1109/ACCESS.2018.2869577.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[13] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," International Journal of Forecasting, vol. 14, no. 1, pp. 35–62, Mar. 1998, doi: 10.1016/S0169-2070(97)00044-7.

[14] R. Tadeusiewicz, "Neural networks: A comprehensive foundation: by Simon HAYKIN; Macmillan College Publishing, New York, USA; IEEE Press, New York, USA; IEEE Computer Society Press, Los Alamitos, CA, USA; 1994; 696 pp.; $69–95; ISBN: 0-02-352761-7." Pergamon, 1995.

[15] C. M. Bishop and P. of N. C. C. M. Bishop, Neural Networks for Pattern Recognition. Clarendon Press, 1995.

[16] Z. Hu, Z. Zhang, H. Yang, Q. Chen, and D. Zuo, "A deep learning approach for predicting the quality of online health expert question-answering services," Journal of Biomedical Informatics, vol. 71, pp. 241–253, Jul. 2017, doi: 10.1016/j.jbi.2017.06.012.

[17] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 29, 2017. doi: 10.48550/arXiv.1412.6980.

[18] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[19] A. Liaw and M. Wiener, "Classification and Regression by randomForest," vol. 2, 2002.

[20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[21] X. He et al., "Practical Lessons from Predicting Clicks on Ads at Facebook," in Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, in ADKDD'14. New York, NY, USA: Association for Computing Machinery, Aug. 2014, pp. 1–9. doi: 10.1145/2648584.2648589.

[22] J. D. Kelleher, B. M. Namee, and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies. MIT Press, 2020.

[23] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[25] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[26] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv, Oct. 10, 2020. doi: 10.48550/arXiv.2010.16061.

Cyberbullying Classification Using Three Deep Learning models: GPT, BERT, and RoBERTa

Muhammad Abusaqer and Charles Fofie Jr

Department of Math and Computer Science

Minot State University

Minot, ND, USA

muhammad.abusaqer@minotstateu.edu; charles.fofiejr@minotstateu.edu

Abstract

This research paper presents a study on the classification of cyberbullying on social media feeds using deep learning algorithms, including GPT-2, BERT, and RoBERTa Transformers. Cyberbullying is a growing concern in social media, so it is crucial to develop systems for detecting and preventing it. Cyberbullying involves using technology to harass, threaten, embarrass, or target individuals based on age, gender, religion, etc. This paper proposes a system that leverages both deep learning and traditional machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, and Random Forest, to detect cyberbullying and reduce its impact, particularly on teen suicides. The study trains the models on a dataset of 1,000 tweets selected randomly from a larger dataset of 46,692 tweets.

The study compares the performance of these deep learning models to traditional machine learning algorithms in terms of accuracy, precision, recall, and F1-score. The study results demonstrate that the RoBERTa Transformers model outperforms the other models, highlighting the effectiveness of leveraging large-scale pre-trained language models for cyberbullying detection. However, the results also reveal that traditional machine learning algorithms, such as SVM and Random Forest, can still offer competitive performance compared to some transformer-based models, particularly when computational resources are limited.

This study makes significant contributions to the field by providing a performance comparison between state-of-the-art deep learning models and traditional machine learning algorithms for cyberbullying detection. In addition, the results of this study could help develop tools to assist in monitoring social media for cyberbullying feeds and immediately deleting them, thereby ensuring the safety and well-being of online users.

1 Introduction

The advent of social media has revolutionized communication, allowing people from around the world to connect and share their lives. However, this unprecedented access to global communication has given rise to a troubling phenomenon: cyberbullying. Cyberbullying is defined as the use of technology to harass, threaten, embarrass, or target individuals based on factors such as age, gender, or religion. As a growing concern, particularly among adolescents, cyberbullying has been associated with severe psychological consequences, including depression, anxiety, and even suicide. Therefore, it is imperative to develop robust systems for detecting and preventing cyberbullying on social media platforms.

This research paper presents a comprehensive study on the classification of cyberbullying on social media feeds using three state-of-the-art deep learning algorithms: GPT-3 from OpenAI, BERT from Google, and RoBERTa from Facebook AI. These deep learning models are trained on a dataset of 46,692 tweets to detect instances of cyberbullying. Additionally, the study compares the performance of these deep learning models with traditional machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, and Decision Trees, which have been widely used in previous research on cyberbullying detection.

The primary contributions of this study are twofold. First, it is among the first studies to employ the newly released GPT-3, BERT, and RoBERTa deep learning models in the context of cyberbullying detection. Second, it provides a comprehensive performance comparison between these cutting-edge deep learning models and traditional machine learning algorithms. By demonstrating the superiority of deep learning models in detecting cyberbullying, the results of this study have the potential to inform the development of tools that can aid in monitoring social media for cyberbullying content and enable timely intervention, ultimately creating a safer online environment for users of all ages.

2 Related Work

Researchers studied cyberbullying using different machine learning and deep learning algorithms.

A recent study by Hani and his colleagues presents a supervised machine-learning method aimed at detecting and mitigating cyberbullying [1]. Authors employed various classifiers to train and find bullying behavior. Upon evaluation of the proposed technique using a cyberbullying dataset revealed that the Neural Network (NN) model outperformed other classifiers, achieving an accuracy rate of 92.8%. The Support Vector Machine (SVM) model followed closely with an accuracy of 90.3%. Moreover, the NN model proved superior performance compared to classifiers employed in similar research efforts when tested on the same dataset.

In a study focusing on automatic cyberbullying detection in a social media text, researchers explored the feasibility of identifying posts written by bullies, victims, and bystanders in online bullying situations [2]. They developed a fine-grained annotated cyberbullying corpus for both English and Dutch languages and employed linear support vector machines with a rich feature set to perform a series of binary classification experiments. The study aimed to find which information sources contribute the most to the task of automatic cyberbullying detection. The results showed promising outcomes for detecting cyberbullying-related posts, with the optimized classifier achieving F1 scores of 64% and 61% for English and Dutch, respectively, significantly outperforming baseline systems [2].

In a recent study examining the rise of cyberbullying in digital spaces, particularly social media, researchers aimed to detect cyberbullying comments automatically using machine learning and deep learning techniques [3]. They highlighted the various forms of cyberbullying, such as sexual remarks, threats, hate mail, and spreading false information about individuals, and noted the long-lasting impacts on victims, both physically and psychologically. The study revealed an increase in cyberbullying-related suicides in recent years, with India being one of the top four countries with the highest number of cases. To address this issue, the researchers employed metrics like accuracy, precision, recall, and F1-score to evaluate the performance of their models [3]. The study found that the Gated Recurrent Unit, a deep learning technique, outperformed all other techniques considered in the paper, achieving an impressive accuracy of 95.47%.

In a recent study examining the critical role of cybersecurity in safeguarding complex networks of client and organization data, researchers emphasized the importance of cybersecurity for individuals, families, corporations, agencies, and educational institutions [4]. The study highlights the potential of machine learning in advancing the cybersecurity landscape, particularly in light of the massive amounts of data collected by modern businesses and infrastructure systems. As data becomes increasingly central to various business-focused and infrastructure systems, the authors argue that machine learning and artificial intelligence are gaining traction across all domains of today's systems, whether on-premises or in the cloud [4]. By incorporating these advanced technologies, cybersecurity teams may be better equipped to protect sensitive data and maintain the integrity of mission-critical systems.

Trong and his colleagues focused on detecting cybersecurity events. The researchers emphasized the importance of event detection (ED) to identify event trigger words within the cybersecurity domain [5]. To facilitate future research, the authors introduced a new dataset for this problem, comprising manual annotations for 30 significant cybersecurity event types and a large dataset size suitable for developing deep learning models. Compared to previous datasets for this task, the new dataset includes more event types and supports the modeling of document-level information, potentially enhancing performance [5]. The researchers conducted extensive evaluations using current state-of-the-art methods for ED on the proposed dataset, revealing the challenges associated with cybersecurity ED and presenting numerous research opportunities in this area for future work.

3 Proposed Methodology

3.1 Dataset

The study used a dataset from Kaggle. The dataset has 47,693 tweets with cyberbullying labels. Each tweet is labeled to one of these six classes: "not_cyberbullying", "other_cyberbullying", "age", "ethnicity", "gender": 4, and "religion". Because of limited computing power, the authors randomly sampled 1000 rows from the dataset for running the experiments.

3.2 Dataset Cleaning

The tweet data has been cleaned from punctuation, stopwords, and nonalphanumeric text, as they do not contribute to the classification. Also, the text was transferred to lowercase.

3.3 Machine Learning

The study used three machine learning classifiers, Multinomial Naïve Base (MultinomialNB), Support Vector Machine (SVM), and RandomForest (RF).

3.4 Transformers for NLP

Transformers form the underlying architecture for many popular NLP models, such as BERT, RoBERTa, and GPT. They were proposed by a team of researchers from Google in 2017 in the paper, Attention Is All You Need [6]. The study used three transformers models, GPT – 2.0, RoBERTa, and BERT.

3.5 Evaluation Metrics

Accuracy, precision, recall, and F1-score are common evaluation metrics used to assess the performance of classification models. These metrics provide insights into the model's ability to identify and classify instances in the data correctly, and each metric focuses on a different aspect of the classification task.

3.5.1 Accuracy

Accuracy represents the ratio of correct predictions (encompassing both true positives and true negatives) made by the model to the entire number of instances within the dataset. This metric is frequently utilized to evaluate a classifier's overall performance [7].

Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives) [8]

Nonetheless, relying solely on accuracy might be unsuitable when dealing with imbalanced data, as it could produce misleading results when the majority of instances pertain to a single class [9].

3.5.2 Precision

Precision refers to the fraction of true positives out of all instances predicted as positive by the model [7]. In essence, it evaluates the classifier's capability to accurately identify positive instances among all predictions deemed positive.

Precision = True Positives / (True Positives + False Positives) [8]

Precision serves as a valuable metric in situations where the consequences of false positives are significant, such as spam detection. In this context, wrongly classifying a legitimate email as spam could result in the loss of crucial information [10].

3.5.3 Recall

Recall, also referred to as sensitivity or true positive rate, represents the fraction of true positives out of the entire count of actual positive instances in the dataset [7]. It assesses the classifier's capacity to identify all positive instances accurately.

Recall = True Positives / (True Positives + False Negatives) [8]

Recall emerges as a critical metric in scenarios where the repercussions of false negatives are substantial, such as in medical diagnoses where undetected diseases can lead to grave consequences [11].

3.5.4 F1-score

The F1-score serves as the harmonic mean of precision and recall, offering a unified metric that harmonizes both precision and recall [12]. It proves particularly valuable when working with imbalanced datasets since it considers both false positives and false negatives.

F1-score = 2 * (Precision * Recall) / (Precision + Recall) [8]

An F1-score of 1 signifies impeccable precision and recall, while an F1-score of 0 denotes that either precision or recall (or both) amounts to zero [13].

4 Results

In this study, the authors evaluated various machine learning models, including BERT Transformers, RoBERTa Transformers, GPT Transformers, Random Forest Classifier, Multinomial Naïve Bayes, and Support Vector Machine (SVM) to detect cyberbullying in a dataset of tweets. The models were trained and tested on a smaller subset of the dataset (1000 samples) for computational efficiency. The performance of each model was measured using accuracy, precision, recall, and F1-score. The results obtained are shown in Table 1 contains data.

	Accuracy	Precision	Recall	F1-Score
RoBERTa Transformers	0.83	0.82	083	0.82
BERT Transformers	0.78	0.76	078	0.73
GPT Transformers	0.18	0.03	0.18	0.05
RandomForest	0.75	0.79	0.75	0.76
MultinomialNB	0.71	0.73	0.71	0.68
SVM	0.75	0.80	0.75	0.76

Table 1: Evaluation results of ML and Transformers.

5 Discussion

The results indicate that among the evaluated models, the RoBERTa Transformers achieved the highest performance in terms of accuracy, precision, recall, and F-1 score. This demonstrates the potential of using advanced pre-trained transformer models for cyberbullying detection tasks. On the other hand, the GPT Transformers showed significantly lower performance compared to the other models, possibly due to the model's inherent design as a generative language model rather than a classification model.

The traditional machine learning algorithms, such as Random Forest, Multinomial Naïve Bayes, and SVM, showed competitive performance compared to the BERT Transformers, with SVM having similar performance in accuracy, recall, and F1-score. This suggests that, despite the advancements in deep learning and natural language processing, traditional machine learning algorithms still hold potential for cyberbullying detection tasks, especially when computational resources are limited.

In conclusion, the choice of the model for detecting cyberbullying in social media text data depends on the available computational resources and the desired performance metrics. While the RoBERTa Transformers model provides the best overall performance, traditional machine learning algorithms, such as SVM and Random Forest, can still offer competitive results with lower computational requirements.

6 Future work

Future research in this area could explore the use of other advanced transformer models, such as GPT-3, or investigate the benefits of combining multiple models through ensemble techniques to further improve classification performance. Additionally, examining the impact of different data preprocessing and feature engineering methods, as well as incorporating domain-specific knowledge, could provide valuable insights for enhancing the detection of cyberbullying in social media text data.

7 Conclusion

In this research, the authors investigated the performance of various machine learning models, including BERT Transformers, RoBERTa Transformers, GPT Transformers, Random Forest Classifier, Multinomial Naïve Bayes, and Support Vector Machine (SVM), for the task of cyberbullying detection in social media text data. The experiments demonstrated the potential of advanced pre-trained transformer models in achieving high performance for this challenging task.

The RoBERTa Transformers model outperformed the other models in terms of accuracy, precision, recall, and F1-score, highlighting the effectiveness of leveraging large-scale pre-trained language models for cyberbullying detection. Despite the relatively lower performance of the GPT Transformers model, the results emphasize the importance of model selection and fine-tuning strategies to match the characteristics of the specific classification task.

The study also revealed that traditional machine learning algorithms, such as SVM and Random Forest, can still offer competitive performance compared to some transformer-based models, particularly when computational resources are limited. These findings underscore the value of considering a diverse range of classification techniques for cyberbullying detection, depending on the available resources and desired performance metrics.

In conclusion, this research contributes to the growing body of literature on the application of machine learning for cyberbullying detection, offering valuable insights into the performance of various models and guiding the development of more effective and efficient detection systems. As social media continues to play an increasingly prominent role in current days, the ability to accurately and swiftly identify and address instances of cyberbullying is critical for ensuring the safety and well-being of online users.

References

[1] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, “Social Media Cyberbullying Detection using Machine Learning,” *International Journal of Advanced Computer Science and Applications*, vol. 10, pp. 703–707, Jan. 2019, doi: 10.14569/IJACSA.2019.0100587.

[2] C. V. Hee *et al.*, “Automatic detection of cyberbullying in social media text,” *PLOS ONE*, vol. 13, no. 10, p. e0203794, Oct. 2018, doi: 10.1371/journal.pone.0203794.

[3] A. K. G and D. Uma, “Detection of Cyberbullying Using Machine Learning and Deep Learning Algorithms,” in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, Aug. 2022, pp. 1–7. doi: 10.1109/ASIANCON55314.2022.9908898.

[4] R. Kumar and E. Al, "Detection of Cyberbullying using Machine Learning," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 9, Art. no. 9, Apr. 2021, doi: 10.17762/turcomat.v12i9.3131.

[5] H. Man Duc Trong, D. Trong Le, A. Pouran Ben Veyseh, T. Nguyen, and T. H. Nguyen, "Introducing a New Dataset for Event Detection in Cybersecurity Texts," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 5381–5390. doi: 10.18653/v1/2020.emnlp-main.433.

[6] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Mar. 30, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[7] S. Kotsiantis, I. Zaharakis, and P. Pintelas, "Machine learning: A review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, pp. 159–190, Nov. 2006, doi: 10.1007/s10462-007-9052-3.

[8] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[10] G. V. Cormack, "Email Spam Filtering: A Systematic Review," *INR*, vol. 1, no. 4, pp. 335–455, Jun. 2008, doi: 10.1561/1500000006.

[11] T. Fawcett and P. Flach, "A Response to Webb and Ting's On the Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions," *Machine Learning*, vol. 58, pp. 33–38, Jan. 2005, doi: 10.1007/s10994-005-5256-4.

[12] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley California USA: ACM, Aug. 1999, pp. 42–49. doi: 10.1145/312624.312647.

[13] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv, Oct. 10, 2020. doi: 10.48550/arXiv.2010.16061.

Automated Categorization of Cybersecurity News Articles through State-of-the-Art Text Transfer Deep Learning Models

Aden Scott, JT Snow, Muhammad Abusaqer

Department of Math and Computer Science

Minot State University

Minot, ND, USA

nathan.a.scott@minotstateu.edu, joshua.snow@minotstateu.edu,
muhammad.abusaqer@minotstateu.edu

Abstract

The rapid increase in cybersecurity events highlights the need for effective detection and organization of news articles that report such incidents. This research investigates deep learning techniques to categorize and organize a large dataset. The study uses a dataset of 3732 cybersecurity news articles classified into four categories: cyberattack, data breach, malware, and vulnerability. The time-consuming and error-prone task of manual categorization can be efficiently and accurately accomplished with deep learning methods. Moreover, the use of deep learning for the categorization and organization of news articles can provide insights into trends and the latest developments in the field.

This study applies the latest deep learning models for categorizing cybersecurity news articles automatically. Specifically, the research evaluates the performance of three state-of-the-art text transformation deep learning models on the cybersecurity news dataset, including BERT, GPT, and RoBERTa. The study reports the results of the categorization and compares the three models.

This paper's primary contribution is applying the latest deep learning models for categorizing and organizing cybersecurity news articles to show performance comparison of the models, highlighting the need for further research into deep learning for text classification in the cybersecurity domain. The results of this study could also help develop tools to assist cybersecurity professionals in keeping up to date with the latest developments in the field.

1 Introduction

The safety of data is a major concern for all internet users. As technology advances, more aspects of life move online, creating opportunities for cybercriminals to exploit. Consequently, the risks of cyberattacks and data breaches are increasing. Cybersecurity professionals face the challenge of effectively detecting and organizing news articles related to cybersecurity events. With thousands of news articles published daily, categorizing them by topic and relevance is time-consuming and error-prone. With all the news on cybersecurity emerging every day, it is difficult to always stay up to date on the latest trends because there is so much information to process.

To address this issue, the authors explore the use of deep learning techniques, specifically focusing on how text transfer deep learning models can be trained to automatically categorize and organize a large dataset of cybersecurity news articles. The goal is to improve the accuracy and efficiency of categorization and provide better insights into the latest developments in the field.

The authors use a dataset consisting of 3,732 cybersecurity news articles, pre-classified into four categories: cyberattack, data breach, malware, and vulnerability. Text preprocessing was applied to the dataset to prepare it for experimentation. The study's methodology encompasses experimentation on the dataset using text transformation deep learning models (BERT, GPT, and RoBERTa) for article classification in automatically categorizing these articles. The authors assess the performance of these models using accuracy, precision, recall, and F1 score as evaluation metrics

Previous research has employed different deep learning models, such as Convolutional Neural Networks (CNNs), Logistic Regression, and Long Short-Term Memory (LSTM) networks, for the automatic categorization of cybersecurity news articles.

Given the increasing importance of cybersecurity in today's world and the growing volume of news articles on cybersecurity events, there is a need for efficient and effective methods for organizing and categorizing these articles. Deep learning in text classification has proven to be effective, which is why authors want to implement it to automatically categorize cybersecurity news articles. The authors hope that this research contributes to the development of accurate and efficient methods for organizing and analyzing cybersecurity news and help cybersecurity professionals stay up to date on the latest threats and developments in the field.

2 Literature Review

Many studies have proven that deep learning is a powerful tool for text classification but is usually applied in a broader domain. Rather than specifically cybersecurity news, many are categories of all news.

A study by (Zhang, 2021) explored the use of deep learning models for classifying news articles into distinct categories, such as event news and ordinary news, addressing challenges related to lengthy text data length and feature extraction difficulties. Traditional approaches, which relied on single-word vectors, considered only the relationship between words while neglecting the crucial relationship between words and categories. Zhang developed a customized DCLSTM-MLP model, combining deep learning algorithms like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Multilayer Perceptron (MLP). This model processes word vector and word dispersion information simultaneously to capture the relationship between words and categories. The study achieved an accuracy of at least 90% using models like CNN, text-LSTM, CNN-MLP, CLSTM, and DCLSTM-MHP. Building upon Zhang's work, the authors of the present study focus on classifying cybersecurity news articles using deep learning text transformers. These transformers have demonstrated remarkable success across various natural language processing tasks, offering enhanced classification performance and adaptability to the rapidly evolving cybersecurity landscape.

Arora, et al. (2022) conducted a study on the performance comparison of different machine learning algorithms for Hindi news classification. Text classification is the process of categorizing text into predefined classes. Before applying machine learning algorithms to the extracted text, the authors implemented preprocessing and feature engineering techniques such as count vectorizer, TF-IDF, and word2vec. Preprocessing in this context is challenging due to the presence of multisensory words, conjunctions, punctuations, and special characters in the Hindi language. The model they developed accepts Hindi news headlines from predefined categories like Entertainment, Sports, Tech, and Lifestyle news. After preprocessing, the corpus size containing unique words was 54,44,997. Out of the different combinations tested, the multinomial Naïve Bayes algorithm with count vectorizer achieved the highest accuracy of 85.47%. This study demonstrates the potential of using machine learning algorithms for news classification in languages other than English, which can be informative for research on the classification of cybersecurity news articles using deep learning text transformers.

Another study by Deepak Singh (2021) used deep learning models to classify articles into five different categories: sports, business, politics, entertainment, and tech. This dataset contained 1490 entries. The top four models used were: Logistic Regression, Random Forest, Multinomial Naive Bayes, and Support Vector Classifier. All of which had an average accuracy of 97% as well as an average F1 score of 97%. The results showed that deep learning models could achieve high accuracy and F1 score while categorizing a smaller dataset.

González-Carvajal and Garrido-Merchán (2021) conducted a study comparing the performance of BERT, a state-of-the-art machine learning model, against traditional machine learning text classification approaches using TF-IDF vocabulary. BERT has gained popularity in recent years due to its ability to handle a wide range of NLP tasks, including supervised text classification, without human supervision. The authors aimed to provide empirical evidence supporting or refuting the use of BERT as a default method in NLP tasks. Their experiments demonstrated the superiority of BERT over traditional methods and its independence from features such as the language of the text, adding empirical evidence supporting the use of BERT as a default technique for NLP problems.

This finding highlights the potential of using advanced models like BERT for the classification of cybersecurity news articles using deep learning text transformers.

There are limitations in the existing research, like the lack of standardization in datasets and evaluation metrics. Most studies use different datasets and evaluation metrics, making it difficult to compare the results. Another limitation is the lack of transparency in deep learning models. Deep learning models are often seen as “black boxes,” and it is difficult to understand how they get their classifications.

Overall, the existing research proves the potential of text classification and organization deep learning can have in the cybersecurity domain. With that said, more research is needed to standardize datasets and evaluation metrics.

3 Methodology

In the research, the authors used The Hacker News Dataset from Mendeley Data (Ahmed et al., 2021). We planned to use this dataset to train and evaluate the three deep learning models on how well they classified the evaluation set.

3.1 Dataset and Preparation

The Hacker News Dataset that was used in this study consists of 3,732 cybersecurity news articles collected from thehackernews.com website. The dataset was preprocessed to remove irrelevant information. The articles are pre-classified into four categories: cyberattack, data breach, malware, and vulnerability. These categories were not completely balanced. The number of labels of Cyberattacks and Data breaches is less than Malware and Vulnerability. Having this unbalanced data set could lead to biased models, but authors will measure this using specific metrics. This is illustrated in Table 1 below.

Category	**Number of Articles**
Cyberattack	364
Data breach	699
Malware	1327
Vulnerability	1352

Table 1: The number of articles across the four categories in the dataset.

3.2 Dataset Preprocessing

Prior to inputting the text into the deep learning models, the authors carried out several preprocessing steps. They first removed stop words and special characters from the text and converted them to lowercase. Then tokenized the text.

3.3 Model Selection and Training

In terms of model selection and training, the authors evaluated three state-of-the-art text transfer deep learning models: BERT, RoBERTa, and GPT. They aimed to obtain a diverse set of results from these three popular deep learning algorithms for evaluation purposes.

3.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model introduced by Devlin et al. (2019) that revolutionized NLP. Its bidirectional context representation allows it to learn both the left and right context of a word, leading to significant performance improvements in various NLP tasks (Devlin et al., 2019).

3.3.2 RoBERTa

RoBERTa (Robustly Optimized BERT) is an adaptation of BERT by (2019) that refines BERT’s pre-training methodology and data processing. RoBERTa’s modifications result in improved performance on downstream tasks (Liu et al., 2019).

3.3.3 GPT

GPT (Generative Pre-trained Transformer) is a transformer-based model developed by OpenAI that focuses on learning the left context of a word. GPT is pre-trained on a large-scale unsupervised language modeling task and fine-tuned for specific NLP tasks (Radford et al., n.d.).

3.4 Software and Hardware

The authors implemented the models using the PyTorch deep learning framework and ran the experiments on a MacBook Pro. (2.8 GHz Quad-Core Intel Core i7). We used Python 3.10 and several Python libraries, such as Pandas, NumPy, and Scikit-learn, for data preprocessing and analysis.

3.5 Model Evaluation

The performance of each model on the testing set was evaluated using the accuracy, precision, recall, and F1-score metrics. We believed that the combinations of these metrics would give us a good evaluation of the performance of the models. Finally, the authors compared the performance of the models and discussed the results.

3.5.1 Accuracy

Accuracy measures the percentage of correctly predicted instances out of all instances. The mathematical representation for this calculation is accuracy = (true positives + true negatives) / (true positives + false Positives + true negatives + false negatives). This is the simplest metric and could be misleading with an imbalanced dataset (Kelleher et al., 2020).

3.5.2 Precision

Precision measures the percentage of correctly predicted positive instances out of all instances that were predicted as positive. The mathematical representation for this calculation is precision = true positives / (true positives + false positives) (Kelleher et al., 2020) (Sokolova & Lapalme, 2009).

3.5.3 Recall

Recall measures the percentage of correctly predicted positive instances out of all actual positive instances. The mathematical representation for this calculation is recall = 2 * (precision * recall) / (precision + recall) (Sokolova & Lapalme, 2009).

3.5.4 F-1 Score

The F-1 score combines precision and recall scores into a single number. The mathematical representation for this calculation is F-1 Score = true positives / (true positives + false negatives). The F-1 score is a good metric to use when the classes are imbalanced since it is the mean of precision and recall (Chicco & Jurman, 2020) (Powers, 2020).

4 Results

In their research, the authors evaluated the results of three deep learning models: BERT, RoBERTa, and GPT. The performance of each model was measured based on accuracy, precision, recall, and F1-score. Upon completion of the training for all three models, the authors reported the best results for each, as shown in Figure 1.

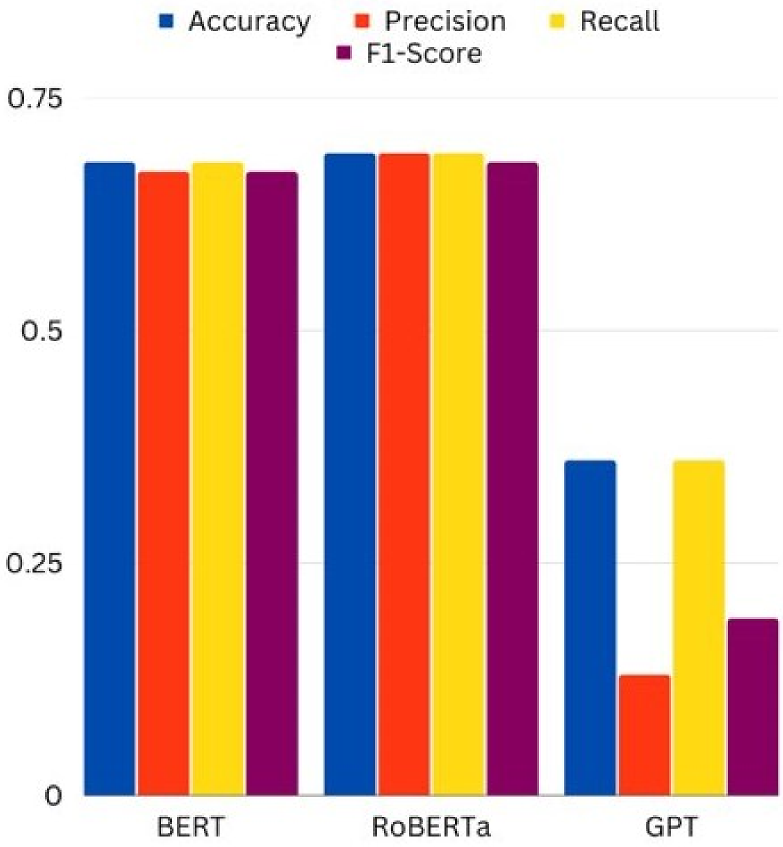


Figure 1: Evaluation results for deep learning models.

5 Discussion

The authors' experiment aimed to compare the performance of three different deep learning models when classifying articles from the Hacker News dataset.

The results point to RoBERTa being the best model for this specific task. RoBERTa performed slightly better than BERT with an accuracy of 0.69, precision of 0.69, recall of 0.69, and F1-Score of 0.68. *These results show the potential of using these large pre-trained language models to classify news articles in the cyber security domain*.

In contrast, the GPT model did not perform as well as the other two, with an accuracy of 0.36, precision of 0.13, recall of 0.36, and F1-Score of 0.19. *The results suggest that GPT struggled with this specific task and is better suited for other natural language tasks like text prediction or generating text*.

6 Future work

Future research in this area, as suggested by the authors, could enhance model accuracy by employing larger, more diverse datasets for training, as well as exploring various pre-trained language models to find the optimal architecture for this specific task. Additional experimentation with these algorithms may facilitate the development of tools for rapidly detecting trends in the cybersecurity domain, thereby enabling companies to bolster security and respond to emerging trends more promptly.

7 Conclusion

In this study, the authors compared the results of three text transformers' deep learning algorithms for classifying articles from the Hacker News dataset. The findings indicate that RoBERTa outperforms the other two models tested for this task. The experiment contributes to the understanding of the strengths and limitations of machine learning models in natural language processing, emphasizing the need for further research into these models, particularly within the cybersecurity domain. The authors also hope their work inspires additional research into predicting trends in the cybersecurity field.

References

Ahmed, M. F., Anwar, M. T., Tanvir, S., Saha, R., Shoumo, S. Z. H., Hossain, M. S., & Rasel, A. A. (2021). *Cybersecurity News Article Dataset*. *1*. https://doi.org/10.17632/n7ntwwrtn5.1

Arora, M., Dhingra, B., Gupta, D., & Singh, D. (2022). Performance Comparison of Different Machine Learning Algorithms on Hindi News Classification. In A. Khanna, D. Gupta, S. Bhattacharyya, A. E. Hassanien, S. Anand, & A. Jaiswal (Eds.), *International Conference on Innovative Computing and Communications* (pp. 323–333). Springer. https://doi.org/10.1007/978-981-16-2597-8_27

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

González-Carvajal, S., & Garrido-Merchán, E. C. (2021). *Comparing BERT against traditional machine learning text classification* (arXiv:2005.13012). arXiv. https://doi.org/10.48550/arXiv.2005.13012

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies*. MIT Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pre-training Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Powers, D. M. W. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation* (arXiv:2010.16061). arXiv. https://doi.org/10.48550/arXiv.2010.16061

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). *Improving Language Understanding by Generative Pre-Training*.

Singh, D. (2021, December 27). Text Classification of News Articles. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Zhang, M. (2021). Applications of Deep Learning in News Text Classification. *Scientific Programming*, *2021*, e6095354. https://doi.org/10.1155/2021/6095354