Basem Saleh (bsaleh2@jhu.edu)

Keyboard Keystroke Audio Prediction Analysis

Abstract

This project seeks to experiment with the paper titled “A Practical Deep Learning-Based Acoustic Side Channel Attack on Keyboards.” The goal of the paper was to show how keystrokes of a keyboard can be identified through a neural network using the audio information grabbed. Members of Durham University detailed a process of extracting keystrokes from a laptop keyboard after using a phone microphone only a few inches away to record audio. For the project within this paper, the focus was placed on recreating this same experiment with a similar data collection process, but to focus on diversifying the dataset to include different keyboard switches, such as tactile and linear switches. In addition to training on different mechanical switches, the CoAtNet model used was modified to become a fusion model, so it can accept additional keyboard information that is non-auditory to help further training. Results shown indicate reliable performance for single key prediction, but poor across multiclass predictions and with a limited dataset collected. The overall dataset used had a dashboard created on Tableau Public to illustrate properties of the collected data.

Introduction

Cybersecurity remains a very important field within the technology field. Opportunities consistently arise for new methods that hackers of various backgrounds employ in order to steal data from unsuspecting victims. These methods normally include phishing emails or nefarious links that users will open or click that allows access to a user’s personal information. In the form of access to their computer or, if the hacker wanted to take information, they would create forms or emails that told users their personal information was required[1].

While these are good for general security, an old method of finding passwords is still s risk: key-logging. This type of program will consistently record all keystrokes performed by a user once the file reaches the computer[3]. That data is then fed back to a source in real time over the internet or if the malicious user knows of this computer, can grab the recorded data off of it. Recent developments have occurred that now been discovered that allows keystroke prediction without the need getting into the computer itself but rather utilize just an external microphone.

This project will focus on predicting keystrokes using recorded data from a MacBook Scissor keyboard, a linear mechanical keyboard, and tactile mechanical keyboard to attempt to prove the alternate hypothesis to accurately predict correct key presses using a machine learning model trained on different key switch keystroke data with an 80% accuracy or greater. Otherwise, reject the null hypothesis of training a model that is no better than random guess of keyboard strokes. This will check to see if the model requires further understanding of the keystroke audio, such as keyboard material and keyboard sizes to better predict keys. In addition, experiments will vary by limiting the scope to only one linear keyboard and completing binary

classification as well rather than on multiple keys on the keyboard. A fusion model will ingest both audio data, through Mel-spectrograms passed through the CoAtNet, combined with tabular data of keyboard sizes and switch types.

Background

The main background for this paper falls upon the paper titled *Deep Learning on Acoustic Side Channels.* The paper[4] this project focuses on accomplished the ability of predicting keyboard strokes using a trained neural network. The paper focused on data that was collected from a single MacBook Pro keyboard while recording keystrokes from a phone microphone only a few inches away to the side. In an effort to better preserve the controlling nature of the project, the phone was kept on a thick cloth and each desired key was pressed 25 times at varying pressures. This allowed them to create many data points for all of the keys required.

In addition to recording right next to the laptop, the model also predicted keystrokes over a Zoom call after recording data from keys in the same Zoom environment. Accuracy was very high, around 92% over zoom and 95% without zoom.

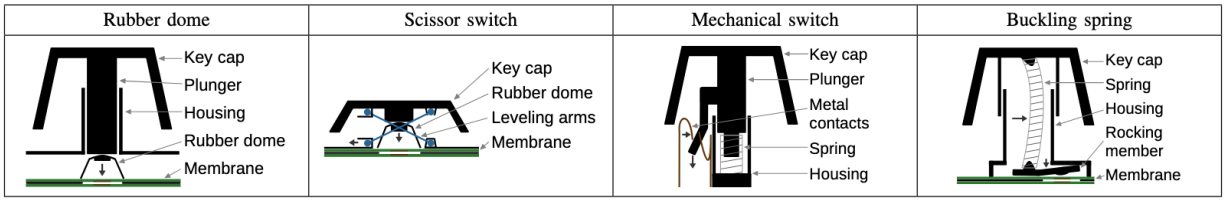

Fig. 1. Common Key switches[5]

One part of this paper that stuck was that this was only done on a MacBook Pro keyboard. This keyboard has a specific scissor design that is different than traditional membrane keyboards found on laptops or desktops, and also mechanical keyboards. While this paper proved useful in predicting keystrokes on a scissor keyboard, the chance to see if mechanical keyboards can have keystrokes predicted in a similar fashion would prove a bigger challenge.

Two additional papers, *SoK: Keylogging Side Channels*[5] and *Robust Keystroke Transcription from the Acoustic Side-Channel*[6] also perform similar model trainings and demonstrate results that show how easily predictions can be made. The three papers train various models to help with keystroke prediction. There is emphasis on tackling work across different types of keyboards, microphones, and environments. So, a focus of project is developing a recording setup that can be easily distributed and used to record data from different users who are enthusiasts of different keyboard types. Since recording different keyboards from different users includes different environments, this also tackles the issue of different microphones being utilized within different rooms as well.

Regarding keyboards themselves, there are a wide variety of switch types, such as mechanical switches with clicky, tactile, and linear type switches from different brands[7]. Clicky switches are bumpy and are very loud, making a clicking sound on the press and release

of a button. This includes switches similar to a Cherry MX Blue or Razor Green switches. Tactiles are slightly less noisy, depending on the chosen brand of switch. They can emit a loud sound on the press, but may not emit a sound on the release. And lastly, linear switches boast no sound as there is no clicky feeling when pressing the switch. In addition to mechanical keyboards, laptops also have membrane keyboards where all keys are connected and are set apart from the keyboard using a plastic membrane that presses to the board to make a connection[8]. Most laptops and office peripherals would have this keyboard. In addition, MacBook's have the Scissor keyboards that, while not fully membrane, have a unique leveling arms that press into a lighter membrane but return a light, mechanical feel.

There are many additional features to consider how similar certain keys may sound, sucher as the material the switch is made of, the material of the keyboard base, and many other factors. However, when it comes to sound, those are the most important features.

Approach

Keyboard Recorder Dataset

To start, the keyboard_recorder repository was created to start the actual recordings of the data. This repository houses the data collection scripts and a link to the recorded data within a JHU OneDrive folder. This repository aimed to repeat the same recording experiment from the DeepKeyAttack dataset by prompting a user to record specific keys and press them a certain number of times to be later preprocessed to extract those pulses.

Fig. 2. Recording setup of Tactile portion of Keyboard Recorder dataset.

The recording scripts instruct the user to place the microphone next to the keyboard, and then use two scripts. The first one to record general information, such as the type of key switch used, keyboard size, and the brand name of the keyboard, which was used to determine what the material of the keyboard was. This will allow the model to learn additional, tabular features that are separate from the audio and could prove useful for training. The second script allows user to select the key they want to record, and then record said data. After completing a key, users are prompted to move to the next key to continue recording. In total, there are three users worth of recordings that are comprised of a linear keyboard, a tactile keyboard, and scissor switches from a Macbook Pro keyboard(grabbed from the Deep Key Attack dataset). There are 48 total classes,

with keys ranging 0-9, a-z, brackets, semicolon, tilde, quote, forward slash, backslash, period, comma, plus, minus, and the space bar. Importing the Deep Key Attack dataset only provided 0-9 and a-z which becomes 36. The linear and tactile portions have all 48 classes.

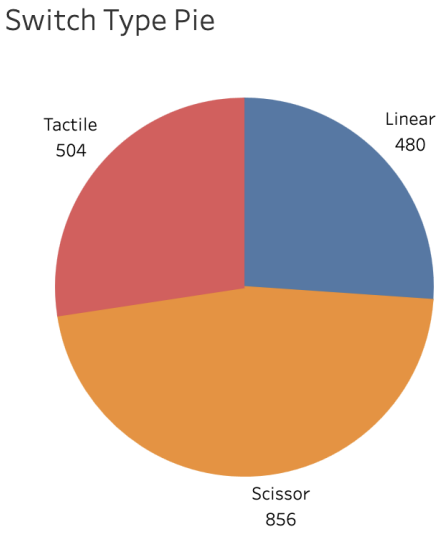

Fig. 3. Pie chart detailing split between the different recorded keyboards

Data Processing and Experimentation

In addition to the recording, proper audio file preprocessing was required to take place. Each audio file was a recording of twenty-five keystrokes. To better isolate the strokes to automatically, each file had its energy calculated, and then normalized across all of the data using a sliding window of size 50 samples. Then, two thresholds were placed on to the average sample: a lower threshold and an upper threshold. The lower threshold is set to .01%(0.001) to ignore noise that may enter the signal instead of a key. Then, an upper threshold is set to 1%(0.01). These were tested across various keys to get a threshold that can best get all of the keys from an audio file without accidentally grabbing unknown noise artifacts. And values within a continuous range of the signal where it exceeds the noise and upper threshold is considered a pulse. Sample locations are calculated from the original signal and cut as separate files for processing. Since this method doesn't grab everything properly and the keystrokes are given at varying pressures, not all 25 samples are collected per key using this method.

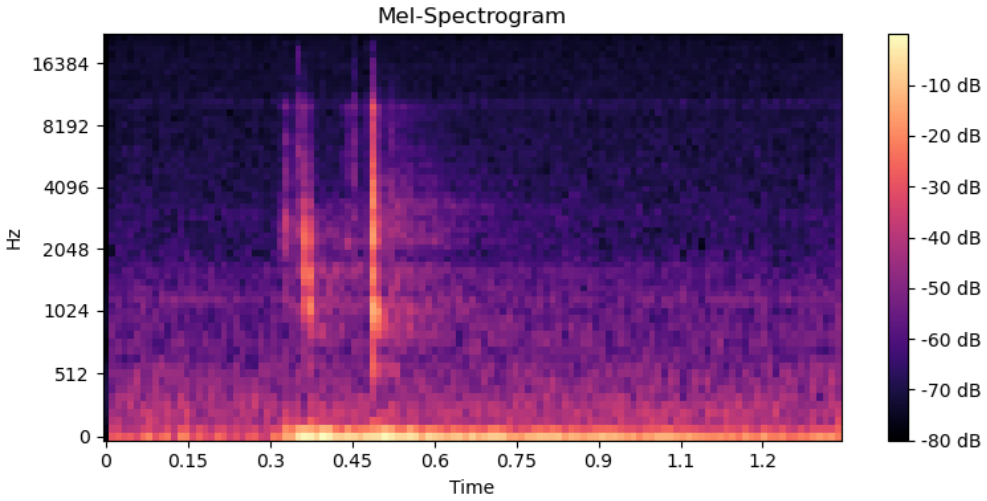

Fig. 4. Mel-Spectrogram of the Z key from the Tactile keyboard recording.

Audio was then passed through the Mel-spectrogram[14]. A Mel-spectrogram is a version of a general spectrogram[15]. A spectrogram, in digital signal processing, is a visual view of the signal wave representing all frequencies present within the signal. Depending on how the axis is oriented X time and Y represents the frequencies within the signal. When plotting, the more red a spectrogram is, the bigger the amplitude is of the captured frequency.

This concept is then built-upon to form Mel-spectrograms, in which the unit of frequency is adjusted to Mels: a logarithmic scale more representative of how humans hear sound. After pulse extractions from the recorded keystrokes are complete, they are placed through a function to produce Mel-spectrograms with the number of Mel coefficients at 64, a window size of 1024, and a hop length of 500. This was done to mimic the results of DeepKeyAttack paper preprocessing.

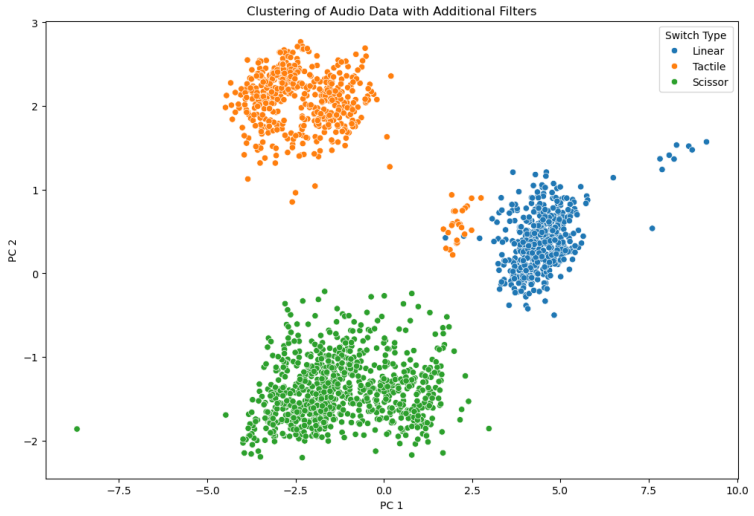

Fig. 5. Clustering of audio features, reduced to two PCA components for plotting. Hue is the type of switch from the Keyboard Recorder dataset.

To get a better understanding of the recorded data, each extracted pulse was placed through processing to extract thirteen Mel-frequency cepstral coefficients. This was chosen to extract features of the pulses while reducing dimensionality[10]. This is a commonly used

feature in automatic speech recognition. After returning the coefficients, they were put through a PCA algorithm to extract two components[11], which were then grouped into 3 clusters through Kmeans clustering[12]. The figure above shows clear distinction between the different audio pulses.

However while the pulses look separated from each other due to different switch types, this does not account for the different microphones used for the various recordings of each dataset[13]. So, instead of the pulse itself, there is potential that this is just the microphones being separated into different clusters and the clusters also just happen to have the different switches for each recording. A more substantial recording dataset could help further get around this.

In addition to the audio, each audio file is given addition data of the encoded keyboard size, encoded switch type, and keyboard material. This data is initially created in a tabular form and once completely encoded, will be used as additional input to the model, along with the Mel-spectrogram information.

The model used in this experiment is the Convolution and Self Attention Neural Network(CoAtNet)[16]. CoAtNet is a neural network architecture that combines convolutional layers and transformer encoders, aiming to leverage the strengths of both CNNs and transformers. The architecture is designed to improve efficiency and performance on various computer vision tasks by integrating local feature extraction capabilities of CNNs[17] with the global context modeling of transformers[18]. It's the same model as was done in the reference paper and will also be utilized this paper.

In addition to the basic architecture that is provided by this model, it was also converted to a fusion format[19]. After the base model performs processing on the audio data and learns it, it's merged with results of another, smaller multi-layer perceptron[20] that only ingests the encoded features. Once the output from both models are done, they are concatenated to a final output linear layer before moving to a classifier for all of the given keys.

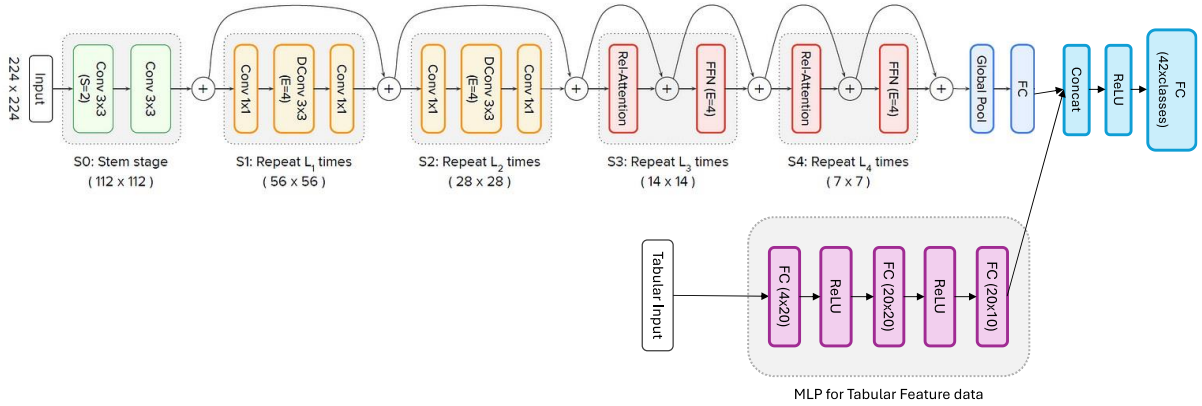

Fig. 6. Architecture of the CoATNet Fusion Modal. Top section is the CoAtNet before the fusion portion. Bottom is the Multi-layer Perceptron that processes feature data before concatenation. After concatenation, a final linear layer combines both outputs.

By combining both of these models together to format one output, this accomplishes the goal of a network both learning the audio distinction of individual keys on a keyboard based on location and at the same time, develops filters between different types of encountered keyboards, based upon what is in the given dataset.

In running these experiments, the most important piece is to check if the model is able to correctly identify a single label. By doing this, it will show the model is capable of predicting the correct letter from all keys. So, accuracy was chosen as the metric. Previously mentioned papers have used the same metric. Experiments to measure accuracy on this self-collected dataset involve the following:

1. Multiclass prediction of all 48 classes across tactile, linear, and scissor keys
    a. Proves the model can generalize well and won't be susceptable to different microphone recording environments.
2. Multiclass prediction all 48 classes on only the tactile dataset.
    a. Proves successful predictions on tactile keys and ensures the microphone is not a differing features in the dataset.
3. Binary prediction of 'h' versus all other classes using all keyboards
    a. Proves the aibility to identify a single character among all given keyboard types
4. Binary prediction of 'h' versus all other classes using only the tactile dataset.
    a. Proves a tactile keyboard is capable of still predicting the key pressed by reducing the class space.
5. Binary prediction of the spacebar versus all other classes using only the tactile dataset.
    a. Extra to determine if the spacebar, given it's distinct sound, is easier to predict than other data keys.

Training Method

All data was preprocessed in one jupyter notebook by creating one dataframe that featured encoded audio file data, alongwith a path to that dataset. During training, a custom PyTorch dataset performs the mel-spectrogram transformation when being loaded into the model. Since there weren't as many keyboards as desired, the only features, outside of the mel-spectrogram of audio, was the switch type used(tactile, linear, scissor) and the size of the keyboards. Experiments had modified epoch ranges since convergence happened much earlier for some experiments than others. All were trained on an NVIDIA 3080 GPU. Data was split 80/20 for training and testing using a total of 1840 pulses.

Results

TABLE I. KEYBOARD RECORDER EXPERIMENT ACCURACY METRICS

Experiments	Classes	Keyboard Recorder Dataset	Accuracy
DeepKeyAttack	0-9, a-z	MBPWavs	95.00%
Exp. #1	0-9, a-z, punctuation	Full	23.64%
Exp. #2	0-9, a-z, punctuation	Tactile	1.04%
Exp. #3	'h'	Full	98.64%
Exp. #4	'h'	Tactile	96.87%
Exp. #5	'space'	Tactile	94.79%

The table here highlights the accuracy of each experiment performed, the classes used during the experiment, and which portions of the Keyboard Recorder dataset used. The

experiment from *A Practical Deep Learning-Based Acoustic Side Channel Attack on Keyboards* is included here for comparison purposes.

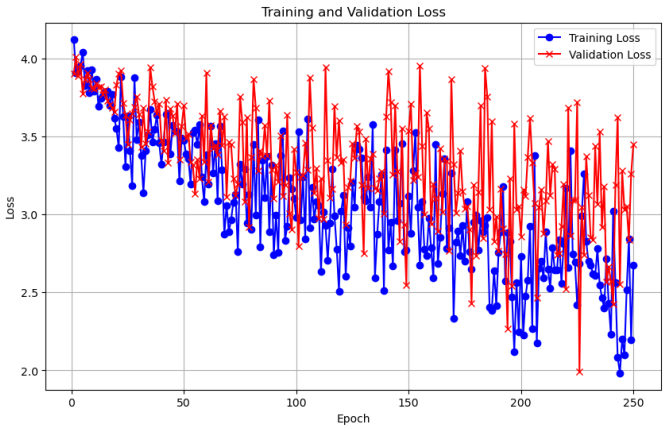

Fig. 7. Experiment #1 training and validation curves on the CoAtNet Fusion model. Using all of the Keyboard Recorder dataset

While convergence was continuing on the training and validation curves for Experiment #1, after 250 epochs, the overfitting was then present as the validation curve began to increase in error while training decreased. Table I indicates the accuracy after training to be close to 24%. While more training epochs show it fitting better on the data, total accuracy never exceeded beyond what's shown on Table I. Similar training curves were present for the rest of the experiments.

Conclusion

The experiments aimed to demonstrate that keystrokes could be differentiated based on keyboard switch types, but the results led to accepting the null hypothesis instead. The fusion model, which combined tabular data with audio features, showed promise in classifying certain key presses, like the letter 'h', but wasn't as effective for classifying all 48 switch types at once. Its utility was limited when only one keyboard type was used, as it didn't offer new information.

The poor performance in the experiments could be due to several factors: First, an insufficient amount of data was the most paramount. Had the dataset been more diverse with more sample recordings, then accuracy may have improved for a few experiments. Second, microphone variability may have heavily influenced these results. The tabular features may have encoded the background environment for which the three keyboards were recorded so instead of learning filters between the keyboards, the model may have only learned to differentiate the microphones used. Future experiments should consider recording all keyboards under the same. And lastly, there were more scissor keys present in the data than the other two switch types, causing a slight imbalance that may have affected performance of certain experiments. Overall, suggestions of better data collection and potential data augmentation could improve keystroke classification efforts.

References

[1] J. Wayburn, “Two-Step Phishing Campaign Exploits Microsoft Office Forms,” *Perception Point*, Jul. 25, 2024. https://perception-point.io/blog/two-step-phishing-campaign-exploits-microsoft-office-forms/ (accessed Aug. 11, 2024).
[2] Karim Toubba, “Security Incident December 2022 Update - LastPass - The LastPass Blog,” *Lastpass.com*, 2022. https://blog.lastpass.com/posts/2022/12/notice-of-security-incident
[3] “What is a Keylogger? | How to Detect Keyloggers,” *Malwarebytes*. https://www.malwarebytes.com/keylogger#:~:text=Keyloggers%20are%20a%20particularly%20insidious
[4] J. Harrison, E. Toreini, and M. Mehrnezhad, “A Practical Deep Learning-Based Acoustic Side Channel Attack on Keyboards,” 2023. Available: https://arxiv.org/pdf/2308.01074
[5] John V. Monaco (2018). SoK: Keylogging Side Channels. *2018 IEEE Symposium on Security and Privacy (SP)*, 211-228.
[6] Slater, D., Novotney, S., Moore, J., Morgan, S., & Tenaglia, S. (2019). Robust keystroke transcription from the acoustic side-channel. In *Proceedings of the 35th Annual Computer Security Applications Conference* (pp. 776–787). Association for Computing Machinery.
[7] “Switch Types - Mechanical Keyboard,” *Mechanical-Keyboard.org*, Nov. 26, 2015. https://www.mechanical-keyboard.org/switch-types/
[8] “Membrane Keyboards: Types, Uses, Features and Benefits,” www.iqsdirectory.com. https://www.iqsdirectory.com/articles/membrane-switch/membrane-keyboards.html
[9] B. Saleh, “Keyboard Recorder,” Github, Aug. 11, 2024. https://github.com/bsaleh524/keyboard_recorder (accessed Aug. 11, 2024).
[10] U. Kiran, “MFCC technique for speech recognition,” Analytics Vidhya, Jun. 13, 2021. https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/
[11] Z. Jaadi, “A Step by Step Explanation of Principal Component Analysis,” Built In, Feb. 23, 2024. https://builtin.com/data-science/step-step-explanation-principal-component-analysis
[12] P. Sharma, “The Most Comprehensive Guide to K-Means Clustering You’ll Ever Need,” Analytics Vidhya, Aug. 19, 2019. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
[13] P. Sharma, “The Most Comprehensive Guide to K-Means Clustering You’ll Ever Need,” Analytics Vidhya, Aug. 19, 2019. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
[14] “Mel spectrogram - MATLAB melSpectrogram,” www.mathworks.com. https://www.mathworks.com/help/audio/ref/melspectrogram.html
[15] “What is a Spectrogram?,” Pacific Northwest Seismic Network. https://www.pnsn.org/spectrograms/what-is-a-spectrogram

[16] Z. Dai, H. Liu, Q. Le, and M. Tan, “CoAtNet: Marrying Convolution and Attention for All Data Sizes.” Accessed: Aug. 11, 2024. [Online]. Available: https://arxiv.org/pdf/2106.04803
[17] IBM, “What are Convolutional Neural Networks? | IBM,” www.ibm.com. https://www.ibm.com/topics/convolutional-neural-networks
[18] A. Vaswani et al., “Attention Is All You Need,” Jun. 2017. Available: https://arxiv.org/pdf/1706.03762
[19] W. Li, Y. Peng, M. Zhang, L. Ding, H. Hu, and L. Shen, “Deep Model Fusion: A Survey.” Accessed: Aug. 11, 2024. [Online]. Available: https://arxiv.org/pdf/2309.15698
[20] S. Abirami and P. Chitra, “Multilayer Perceptron - an overview | ScienceDirect Topics,” www.sciencedirect.com, 2020. https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron
[21] “MX2A BLUE,” Cherry-world.com, 2024. https://www.cherry-world.com/mx2a-blue#:~:text=CHERRY%20MX%20BLUE- (accessed Aug. 11, 2024).