Sri Lanka Institute of Information Technology

Keystroke Dynamics Authentication Systems

Project ID – 24-25J-073

Individual Project Proposal Report

Submitted by:

Student Registration Number	**Student Name**
IT21340864	E M N Edirisinghe

Department of Computer System Engineering

Date of submission

Tuesday, August 22, 2024

Declaration

I declare that this is my own work, and this proposal does not incorporate without acknowledgement of any material previously submitted for a degree or diploma in any other. university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except were the acknowledgement is made in the text.

Name	Student ID	Signature
Edirisinghege E M N	IT21340864	

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

.... 22/8/24

Signature of the supervisor

Date

Abstract

This proposal explains on the development of a new keystroke dynamics-based authentication system for increasing the security and the verification of the user. The keystroke dynamics which can be categorized under the behavioural biometric systems, involves the use of key writings styles to deliver consistent and seamless form of identification. This system fills the gap of the existing advanced cybersecurity threats and various attack approaches in the modern world.

It is organised to consist of the general background of keystroke dynamics and a literature review to assess the state of research in this field and derive the research question. It involves objectives such as to enhance the mechanism of authentication, advanced technologies in addition to methodologies have to be incorporated. The steps used include acquisition and pre-processing of the keystroke data, feature selection, and the use of more new sophisticated models such as CNNs and RNNs for improving the authentication performance.

Some of the commonly used tools as for the data analysis, model training, visualization are Scikit-learn, TensorFlow, PyTorch, Matplotlib, Seaborn respectively. It will involve an evaluation of the technical possibility in terms of system design and computing power and resources as well as economic viability and timing schedule. Implicit and explicit costs include data acquisition, technology equipment and accessories, and application software.

The commercialization strategy looks at market prospects in cybersecurity and protection, health and smart living environments, and considering revenue models as licensing and subscription, direct sales. Thus, the growth strategy is based on pilot projects, the focus on certain markets, and general tips to provide the effective deployment and further scaling of the system.

In summary, the proposed keystroke dynamics authentication system made efforts of being a reliable, flexible, and easy-to-use solution for improving the current security framework and coping with the existing key issues of current authentication systems.

Table Of Contents

Table of Figures

1. Introduction

In today's digital world, securing sensitive data such as emails, contacts, organization confidentiality data, and more is more crucial. Traditional authentication mechanisms like passwords, pins, patterns are vulnerable to various attacks. leading to fishing attacks, unauthorized access, unauthorized modifications, and data breaches.

As much as the threats become more and more elaborated in recent years. The issue of having functional mechanisms for authenticating people or systems becomes more appropriate now than ever before. Passwords or PINs are the most common methods used by people to protect their computer and accounts in general from being hacked. However, password and PINs are not safe from different types of attacks, which is why the researchers feel the need to invent better ways of protection. The solution to this challenge, hence, is behavioural biometrics which refers to a practice that captures individual behavioural patterns.

This proposal targets on the creation of a system that will be used to authenticate users basing on the keystroke dynamics. Keystroke dynamics in turn implies the identification of users via typing behaviour or typing characteristics. In contrast to other biometric systems which are based biometric such as fingerprint, face recognition, keystroke dynamics provide real-time and non-intrusive security addition which does not interfere with the normal users.

The proposed system envisions using this set-up to capture and analyse the temporal and spatial features of typing behaviour using state-of-the-art machine learning techniques such as CNN-RNN hybrids. This way, through a strong data preprocessing stage and the use of the most advanced algorithms the system will be able to provide a rather high level of user identification accuracy and reliability.

The following table shows most of the methods of authentication along with their advantages and disadvantages. Security, usability, and compliance and other aspects indicate why keystroke dynamics is indeed a viable form of secure and even easily implementable form of authenticating systems.

Types	Vulnerability	User Friction	Regulatory	Accuracy
Passwords	**High** Vulnerable to keylogging, phishing, and brute-force attacks	**High** Requires to user remember the password and handle complex passwords.	**Low** Not fully comply with security regulatory standards	**Low** Depends on the password complexity
Multi-Factor Authentication	**Medium** vulnerable to social engineering attacks or loss of the device	**Medium** Requires additional steps to user login. Then reducing user convenience	**High** Often meets regulatory standards, especially in finance and healthcare sectors.	**High** Increase security by applying multiple authentications
Keystroke Dynamics	**Very Low** Compared to other methods, typing is very hard to spoof because of unique typing behaviour.	**Low** Because it operates in the background, the user has minimal impact	**High** Provides authentication that meets strict regulatory standards for secure access.	**High** Provides high accuracy by personalized behaviour analysis

The research problem, research objectives that have been set and research methodology that will be used in the development of a keystroke dynamics-based authentication system are the parts of this proposal. It also encompasses a technical plan and an economic and market plan of the project to cover all aspects of the project's implementation and market entry. [1]

2. Background and Literature Review

2.1 Overview of Keystroke Dynamics

Keystroke dynamics is a newly developed biometric modality that deals with the typing behaviour of a person for the purpose of identification and verification. Keystroke dynamics is different from other types of physiological biometric which includes fingerprints and facial recognition in that regards it as a behavioural biometric where it measures the way a person types on the keyboard. This method takes advantage of the truth that different people type in different manners, due to elements for example hand-to-eye coordination, impaired coordination, and methods of thinking.

Keystroke dynamics has attracted a lot of attention and has been applied in the field of cybersecurity because of the increasing sophistication and urgency of effective mechanisms for user identification. As the threats are becoming more elaborate and complex, measures as passwords and PIN codes are inadequate in protecting information and networks. Keystroke dynamics is once again an authentication technique that can be used continuously, without interfering with the identified traditional methods, and provide stronger security if used in parallel or instead of them.

The project will be mainly conducted researching flight time, dwell time and other features of keystroke dynamics. These features will be extracted as well as the designed aspects will be subjected to machine learning to establish the sound model that will be used in the authentication of the user. The proposed system will be evaluated in different scenarios and checker to see that is flexible for different users and the different attacks like mimic attack or replay attack. [2]

In this project, the capability of replacing keyboard as biometric system to provide evidence for potential of keystroke dynamics as a reliable and scalable biometric modality for the enhancement of authentication systems will be investigated. The overall result of this study could potentially lead to the extended use of keystroke dynamics in a more generic domain and for the personal-computer and business security solutions.

2.2 Existing Research

This research in keystroke dynamics has become significance in behavioural biometrics, since it may be used to improve user authentication systems. By using different patterns in each individual's typing behaviour, such as the length of key presses, the time between keys presses, typing speed, typing rhythm, and error rate, this biometric approach allows for authenticating users individually.

1.Convolutional Neural Networks (CNNs) in Keystroke Dynamics

Recent developments have automated the extraction of spatial features from keystroke dynamics data with research on keystroke dynamics. Finding patterns in the data that are tough to see with spatial feature extraction techniques is a specialty of CNNs. An increased accuracy rate and robustness in user authentication systems may be attributed. As shown by research, CNNs ability to recognize patterns in keypress durations and intervals, flight time, and dwell time. These systems effectively decrease the dimensionality of the keystroke data while maintaining important characteristics that characterize a user's typing behaviour by using many convolutional layers like pooling layers, and flattering layers. [3]

2.Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs)

RNN known as Long Short-Term Memory (LSTM) has received a lot of attention in terms of analysing temporal sequences which are evident in the keystroke dynamics. LSTMs are especially good at can modelling kind of typing where order, when keys are being pressed, is important. The studies show that it is possible to use LSTMs to capture temporal dependencies in keystroke data, allowing improving the recognition of legitimate users against the background of impostors. In the same regard, Bidirectional RNNs have helped bring to the analysis of keystroke sequences in forward plus reverse methods to provide a richer set of temporal patterns.

3.Hybrid CNN-RNN Models

There are integrated models that have attempted to make use of both CNNs for the spatial features and RNNs especially LSTMs and GRUs for the temporal sequences. These models have been proved to enhance the reliability of the systems which are used in keystroke-based biometrics authentication. For example, an application of CNN might initially get the raw keystroke inputs and possibly determine temporal features such as key press time and gap which can be input to an LSTM for temporal features. Such an approach has been found to improve the chances of the system to distinguish between users with close typing patterns.

4.Privacy and Security Concerns

Although keystroke dynamics can be used to uniquely identify a user without requiring any additional information from the user, issues to do with privacy and security still pose a major issue. Unlike other physical biometrics like fingerprint, face recognition, Keystroke dynamics can be regally hence causing privacy issues. However, the research has also looked into how to effectively safeguard keystroke data; ways such as encryption and anonymization. Furthermore, it is crucial to guarantee that the used system cannot become a victim of such types of attacks as, for example, replay attacks or spoofing.

5.Challenges and Future Directions

There are some issues that researchers still have to face while employing the CNNs, RNNs, and other hybrid models for keystroke dynamics. For example, the kind of typing behaviour may be variant due to the kind of device used, the context of typing, or even the exhaustion of the typist, these factors can influence the precision of these systems. Another open problem is to reach real-time authentication with very low latency whereas preserving high accuracy at the same time. For better performance, future research is expected to endeavour on how best to make such systems more flexible for different scenarios and how to make the systems more resilient to variations in users' performance.

6.Innovations in Data Collection and Preprocessing

Some of the applied recent studies also attempted to investigate new approaches towards the collection and initial data preprocessing of keystroke data. These freely available online datasets including the CMU Benchmark Dataset also allowed to train and test their models on various typing styles. Data preprocessing techniques such as normalization, cleaning and noise reduction are used to ensure the raw keystroke data quality and accuracy. These techniques assist in bringing the data to a common format, something which is useful when dealing with datasets obtained from various sources or devices.

2.3 Research Gap

Even though considerable studies on keystroke dynamics research and its usage in biometric authentication systems and effective, there remain certain notable research gaps in this discipline to accomplish its full benefits. These gaps therefore show further areas of research and enhancement of the keystroke dynamics-based systems by enhancing their reliability, validity, and applicability.

1.Handling Various Typing Patterns

Unlike other biometric modalities, the main problem associated with keystroke dynamics is the variation in typing behaviour. Several reasons may lead to typing behaviour variation including type of keyboard like physical or touch and the context of typing like typing formal document, typing a text message, chatting or even the state of the typist, be it emotional or physical. Existing models usually provide poor generalization from one of these contexts to another, and hence lower accuracy and reliability are observed. This research poses questions for future research on how such variations can be fixed so that the model becomes more robust to these variations can be learned by the models so as to normalize the differences which might come into play, perhaps by feeding contextual information into the model or by using machine learning techniques that are able to discover invariance. [1]

2.Scalability for Large-Scale Deployments

The approaches applied in the keystroke dynamics systems have proved to be effective in controlling limited environments but the large-scale implementation of these systems is very difficult. Concerns like processing latency, storage capacity and the fact that these systems require handling large number of user profiles become impractical to implement. Future studies in this area may extend on the improvement of model architectures to support a large scale and real-time keystroke dynamics analysis, or on distributed computations or other new methods of data management.

3.Real-Time Authentication and Low Latency

One problem that still persists in keystroke dynamics is how to attain authentications in real time or with low latency. It has been found that depending on the nature of the models that employ CNNs for feature extraction and RNNs for sequence analysis, the computational complexity is high and causes delay which is undesirable in real-time. To reduce the time, it takes to process the text other than by using simple techniques such as removing stop words, working on mini models is an area that lacks research. Also, it can be rather interesting to investigate the optimizations for the data flow from the keystrokes input to the authentication decision and make sure that the system do not take too much time.

4.Robustness Against Adversarial Attacks

Like any other biometric system, keystroke dynamics is not resist to this kind of attacks including mimicry attacks and replay attacks where by an attacker tries to imitate the keystroke pattern of a authorized user. However, the existing literature concerning the defence against these attacks is scarce to a considerable extent, while numerous studies are aimed at the application of the conventional security techniques without taking into consideration the specific characteristics of keystroke dynamics. Further studies could focus on enhancing the current approaches towards building more complicated anomaly detection techniques, the used of clustering and pattern recognition, the combination of keystroke dynamics with other types of biometrics to form a stronger multi-modal system as well as the using of adversarial learning as a regular training approach. [1]

5.Data Privacy and Ethical Concerns

The collection and analysis of keystroke data present privacy and ethical issue. users are not conscious that their typing behaviour is being logged. There is hence the need to address some of the privacy aspects arising from keystroke dynamics solutions to enhance the likelihood of acceptance and utilization of the systems. The research in this domain could build on the application of data protection methods like differential privacy, or federated learning to collect, and process the keystroke data while preserving privacy without compromising on the performance of this system. However, there is a lack of clear ethical principles and rules regarding the usage of keystroke dynamics in different applications especially those who are sensitive as in surveillance or employee monitoring cases.

6.Limited Availability of Diverse Public Datasets

The models regarding the keystroke dynamics for user authentication are much rested with difficulties concerning the restricted sets of available large public datasets. All most all existing datasets are small-scale or collected in rather restricted settings, which could fail to include the nature of variability of typing behaviours in real worlds adequately. Improving the availability of high-quality keystroke datasets or creating a new generation of methods that can create keystroke data that simulates all types of typing behaviour could be very beneficial. Additionally, more research could target increasing the data variance originating from various devices, typing conditions, and groups of users. [4]

7.Integration with Existing Authentication Systems

Adding keystroke dynamics to existing solutions which may include passwords, a second factor, or multiple factors of authentication also have their difficulties in terms of integration. Keystroke dynamics can extend the security layer but to do so without making the interactions complicated for users is a difficult task. It may be beneficial for research to examine the creation of regular procedures and open APIs that can enable the integration of keystroke dynamics into existing authentication systems, together with interfaces that enable the use of keystroke dynamics to be introduced to conventional organization frameworks smoothly.

8.Long-Term Stability and Adaptation

keystroke dynamics, like any other biometric modality, keystroke dynamics may vary with time, perhaps by aging, change in typing style as well as physical ailment. One of the most significant difficulties is to maintain the effectiveness of the keystroke dynamics system in the long run, without carrying out re-enrolment or retraining very often. More research could be directed towards how to achieve incremental learning and update the user profiles and at the same time learning from the dynamics of typing ceaselessly. This might include methods like repeated learning in which the model is time to time trained on new data or transferring from an existing model to a new set of data.

Research	Individual Variability	Real-Time Processing	Scalability	Privacy	Integration	Adaptation	Generalization
Research A	Yes	No	No	Yes	No	Yes	No
Research B	Yes	Yes	No	Yes	No	Yes	Yes
Research C	No	No	Yes	No	Yes	No	Yes
Proposed Project	Yes	Yes	Yes	Yes	Yes	Yes	Yes

2.4 Research Problem

Keystroke dynamics authentication systems are significantly affected by user typing behaviour variability. This is specifically so because typing behaviour is not uniform and can change with the mood of the typist, level of tiredness, or even the type of typewriter being used. This variability gives rise to discrepancies within the usual flow of authentication and therefore is a security risk.

Variability in Typing Behaviour

As keystroke dynamics authentication systems rely on users' typing behaviour, they are highly sensitive to variations in the data. Emotional condition, tiredness, stress, or shift in the typing environment or gadgets can all alter the typing style of the users. This variability can lead to inconsistency and irregular rhythm of keystrokes and so creating significant problems with reliable user authentication at regular times. The system has to distinguish between normal variations in typing characteristics and security risks, as is still the case in many current systems.

Need for Robust Real-Time Processing

For keystroke dynamics authentication the ability to analyze the typing pattern in real time is important. The system needs to be able to recognize the user types on the keyboard and it has to do this without any noticeable lag. This is possible by using the most sophisticated algorithms and processing methods that can accommodate the large data in terms of keystrokes. The issue is to use the time efficiently while making the system capable enough to predict the accuracy of authentication in real time when load is applied to the network and the computational power.

Generalization Across Different Populations

There are probably many factors that will have an impact on keystroke dynamics such as age of the user, gender, ability level of typing, and physical state. Such differences may lead to different behaviours of the specified authentication system depending on the demographic variable in question. Taking time to adapt the system to be general across the users is also important in the population. The problem is to create models which distinguish these differences but at the same time do not reduce the general system performance and reliability.

Data Privacy Concerns:

Keystroke dynamics and hence the collection and analysis of the data collected under this aspect entails dealing with private biometric information that are very sensitive in nature. Since this data is generally sensitive, measures must be taken to safeguard the privacy of users, in regards to storing, transmitting as well as processing this data. However, following the regulations and standards, including GDPR is a must if one wants to avoid breaking the law. This concern is on how to provide a good architecture that effectively implements the key technologies of extra security that include encryption, anonymization, and access control to ensure that the user data is protected while at the same time providing an optimal system functionality and operational performance.

3. Objectives

The goal of this project is therefore to design a keystroke dynamic which is capable of authenticating a user to a level of 95 % accuracy and above. To this end, the project will employ a CNN for feature extraction and an RNN including LSTM and bidirectional RNN for analysing sequential typing behaviour. The purpose of this approach is to improve the system's performance to identify any user using the system based on their typing patterns.

Much attention will be paid to its ability to work in real time, it is expected that the system will provide authentication results as soon as possible, preferably within 2 seconds. It will also be flexible enough to remain highly efficient no matter how users' typing habits change in the future: whether they type faster or slower, adopt new styles or make other changes.

Data management is important, hence, it will propose standard ways on how to collect, preprocess, and extract features from online data for training and testing the models efficiently.

Security will be important and appropriate measures taken in ensuring that the collected biometric data is secured to the highest level and that privacy of users' biometric data will be accorded to the highest level of protection as enshrined under local and international policies and acts. This will involve creation of a system that can accommodate more users and data volumes in the future without a negative impact on the current or future performance; the system should be dependable with rare occurrences of failure; where a failure occurs, the system should be capable of handling the errors effectively.

User participation will also be enhanced through a user-friendly interface for the end-users and the administrators of the system that will require little training. Further, proposed system will have compatibility with the existing authentication and security systems in the organization making it to work in one platform for multi applications.

4. Methodology

4.1 System Overview and Diagram

A system overview of the Keystroke Dynamics model demonstrates how accuracy improves security and robustness for user authentication. It shows how to develop a hybrid model that combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNN is used the analyse the spatial data and RNN is used the analyse the temporal data. Furthermore, this includes phases such as data collection, data preprocessing, feature extraction, sequence analysis, and model integration for authentication decisions. The diagram represents the flow of these components.

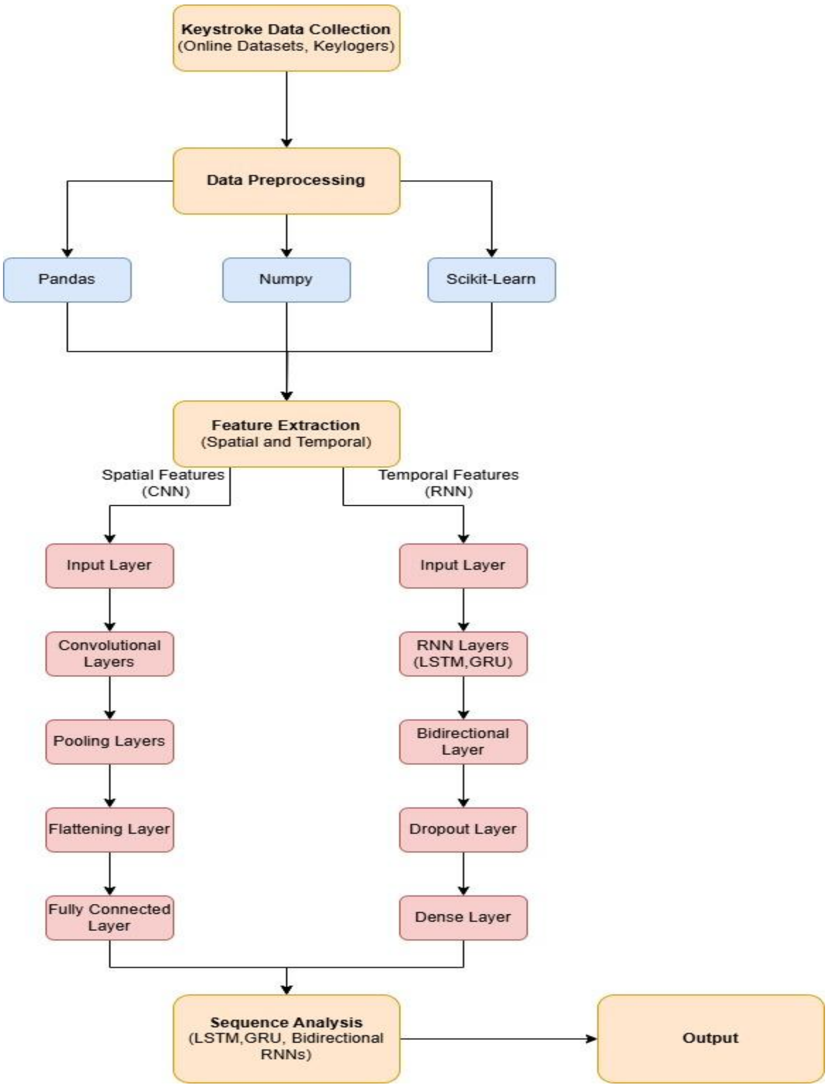


Figure 1- System Architecture

4.2 Datasets

To develop the keystroke dynamic authentication model will be using multiple datasets.

The following datasets will be used,

A Carnegie Mellon University Benchmark Dataset was found online which contains 51 subjects on and each subject includes 400 times typed passwords. There are 34 variables and 20,400 observations in the data set. [5]

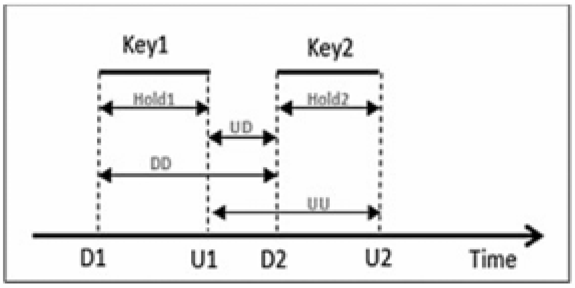


Figure 2 - CMU Benchmark Dataset Features

Link - CMU Dataset

Benchmark Dataset found on Kaggle.com its contains a table with 34 columns. [6]
Link - Keystroke Dataset From Kaggle.com

Following datasets are also using this project.
BKSD Dataset - Link
Pressure-sensitive keystroke dynamics data - Link
University of South-Eastern Norway Dataset - Link

Furthermore, all of these datasets include various features and parameters like flight time, dwell time, hold time, key press time, key release time, repetition count, error rate and more.

4.3 Methodology

Data Collection

The data collection phase collects the raw keystroke dynamics data using online datasets and keyloggers. Which contains various parameters like flight time, dwell time, error rate, keypress time, and keypress patterns and more. Additionally, using keyloggers captures keystroke dynamics data in real time.

Data Preprocessing

In this step doing the preprocessing process contains normalizing the keystroke data, cleaning the data to remove noise, manage missing values, converting the keystroke dynamics data to suitable format of feature extraction and handle timing information using pythons tools like NumPy, Pandas and scikit learn. This step ensures the accuracy of keystroke dynamics data, quality of the data and more.

Feature Extraction

In this phase extract the special features from pre-processed keystroke dynamics data using both CNN and RNN. Using the CNN extract the spatial features in the keystroke data and using RNN extract the temporal features in the keystroke data.

The CNN feature extraction process starts with the input layer, which obtains the pre-processed keystroke dynamic data. Then convolutional layer finds the spatial features within the keystroke data and applies various filters. Following this, the data moves into a pooling layer. This layer reduces the dimensionality of the feature maps and gains the most valuable features. After Pooling, the data moves on to the Flattening Layer. In this layer, the 2D feature matrix is converted to 1D vector and the data is prepared for the fully connected layer. Then data moves on the final layer called fully connected layer. In this layer keystroke dynamics flattened vector data transfer through more fully connected layers to integrate the spatial features. [7] [8]

The RNN feature extraction process starts with also input layer, which obtains the pre-processed keystroke dynamic data. The input layer moves the processed data to the RNN layers such as Long Short-Term Memory and Gated Recurrent Units. Using LSTM and GRU capture the temporal sequences and patterns in the pre-processed keystroke dynamics data. [9] After the RNN layers data moves to the Bidirectional RNN layer, using Bidirectional RNN layer extract features data from both backward and forward directions. its means using Bidirectional RNN layer capturing both past and future typing patterns in keystroke data. [10] Then data moves to the Dropout layer. This layer applied to prevent overfitting by dropping some units during the training process and to ensure that the model generalizes well to unseen data. After that data moves to the final layer. its called Dense Layer. This layer helps to the synthesizing the temporal features of keystroke dynamics data captured by the previous layers. [11]

Additionally, both RNN and CNN feature extraction process are parallel process. In this process extract various of features. CNN may extract from keystroke data such as keypress duration, flight time, dwell time, trigraph time, key transition pattern, typing speed, error rates and more. RNN may extract from keystroke data such as Temporal Sequences, Typing Rhythm, Long-term Dependencies, Typing Patterns, Repetitive Sequences, Anomalous Sequences, Adaptive Behaviour and more. This above process are supported by TensorFlow, Keras, or PyTorch. After the feature extraction process then data moves to the next process called Sequence Analysis.

Sequence Analysis

The sequence analysis process is analysing the temporal patterns in keystroke dynamic data using combination of CNN and advanced RNN technologies like LSTM, GRU and Bidirectional RNNs. First extract the spatial features from keystroke data and identify the critical pattern in the keystroke data using CNN. [12] This identified feature and temporal features are transfer through the RNN models. where TensorFlow and Keras or Pytorch are used to train and fine-tune LSTM, GRU, and Bidirectional RNN layers. This hybrid model enables the comprehensive, robustness analysis for spatial and temporal keystroke features. This process ensures the more accuracy and robustness of the authentication mechanism.

Model Integration and Authentication Decision

The final stage is the combination of the outcome from the CNN and RNN into one model for the authentication decisions. This is made by the use of decision-making tools and models like TensorFlow and Scikit learn, the models that can be used include the thresholding, Support Vector Machines (SVM) or the Random Forest kinds of models. Combining the spatial and temporal features offer the function more accurate and reliability as well as better authentication to increase its safeguard and integrity. [3]

4.4 Feature Selection

Features in keystroke dynamics are dynamic features that are significant characteristic features attained from typing behaviour that aid in identifying different users. These features characterize different aspects such as the behaviour of typing and they are used in construction of models used in analysing as well as authenticating users based on their typing behaviour. All of them offer some or the other perspective of typing and reveal the aspects of typing behaviour at various levels. Here is the detailed description of the particular features being applied to this project.

Flight Time: The time between key down and key up, that is from the time keyboard key is depressed to the time when it is released. [13]

Dwell Time: Describes how long for a particular key is kept held down till its action is performed. [13]

Digraph Timing: The time interval between two keys of sequence simultaneously pressed.

Trigraph Timing: The timing between any three keys being pressed in succession.

Key Transition Patterns: Manipulation of the speed, rhythm and duration in which the column and row keys are depressed and/or released.

Frequency of Specific Keystrokes: Frequencies of when specific keys all or combinations of keys are typed.

Typing Speed: The rate in which the keys of the computer keyboard are typed on an average by a person.

Error Rates: The number of errors or corrections, for instance backspaces or retyped keys.

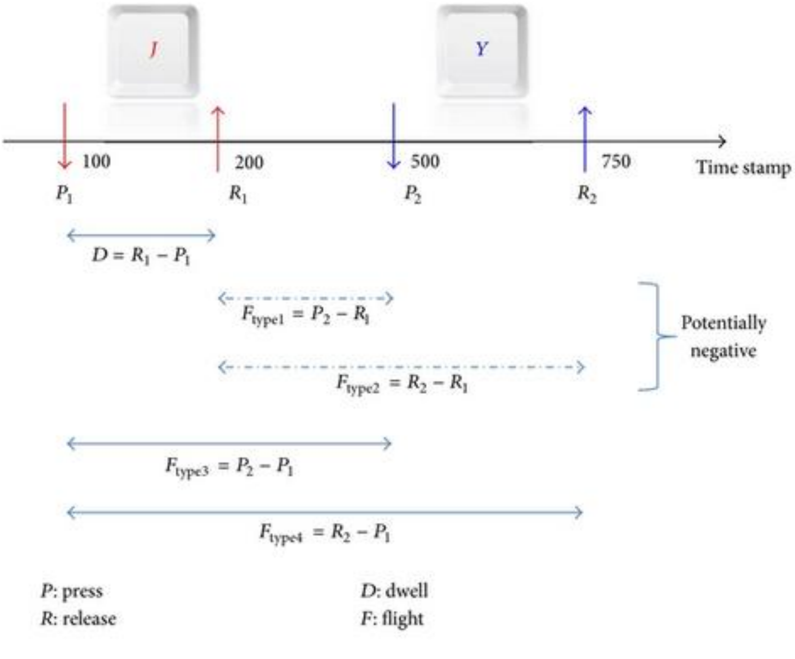


Figure 3- Different keystroke events

5. Technologies and Tools

5.1 Data pre-Processing tools

- **Pandas**

  Pandas is fast, powerful and easy python library that used for the data preprocess tasks such as normalization, manager missing values, and cleaning. [14]

- **NumPy**

  NumPy is an open-source Python library that provides support for matrices and large multi-dimensional arrays. which is used for handling and preprocessing keystroke data. [15]

5.2 Data Analysis Tools & Technologies

- **TensorFlow/Keras**

  TensorFlow/Keras is the primary framework that developing and implementing deep learning and machine learning models. These tools can be used to develop both CNN and RNN models. TensorFlow allows to develop and optimize the complex neural networks. It helps to handle the intricate patterns in the keystroke dynamic data using large scale data processing and GPU acceleration. [16]In the other hand Keras is TensorFlow's high-level API that helps to simply the model development process and provides rapid prototyping and experimentation. In this project, Keras utilized to set up both CNN and RNN layers. CNN layers for extracting spatial features from keystroke dynamic data. RNN layers for using LSTM, GRU and Bidirectional RNNs, to extract and analyse the temporal features in keystroke dynamic data. [17]The combined use of both TensorFlow and Keras ensures the model accuracy and performance.

- **Scikit-learn**

  This library is for data preprocessing which involves normalization, scaling and encoding of data to make the keystroke data ready for analysis. It is useful in identifying and selecting the appropriate features in the database originating from keystroke dynamics. Further, Scikit-learn has a set of machine learning and performance measures for models necessary when developing, training, and, specifically, evaluating the models used in keystroke dynamics. [18]

- **Long Short-Term Memory (LSTM)**

  LSTM is improved RNN version designed by Hochreiter & Schmidhuber. [9] LSTM is an improved RNN version designed by Hochreiter & Schmidhuber. LSTM is a special type of memory to store and output data. It can store and learn long-term sequences. A traditional RNN has a single hidden state that passes through time, while an LSTM is more complex. In the LSTM gain input from the three different states such as current input state, the short term memory from the previous memory cell and long term memory. The LSTM use 3 gates. They are input gate, forget gate and output gate. Input gate used for the choose what keystroke data will stored in the long term memory, Forget gate used for the decide what data keep or discard from the long-term memory and output gate is used for the take the outputs. [19]

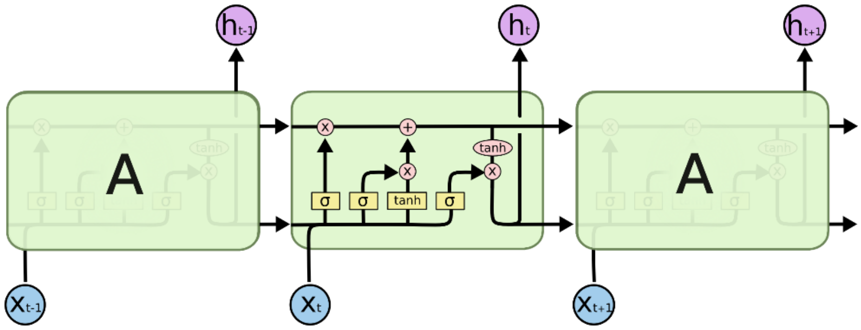


Figure 4- LSTM Architecture

➢ **Gated Recurrent Unit (GRU)**

GRU is simply version of LSTMs. Because the GRU control the data process in the memory using single update gate. GRU faster and easy to use than LSTM. But LSTM more effective. GRU has two gates such as update gate and reset gate. [19]

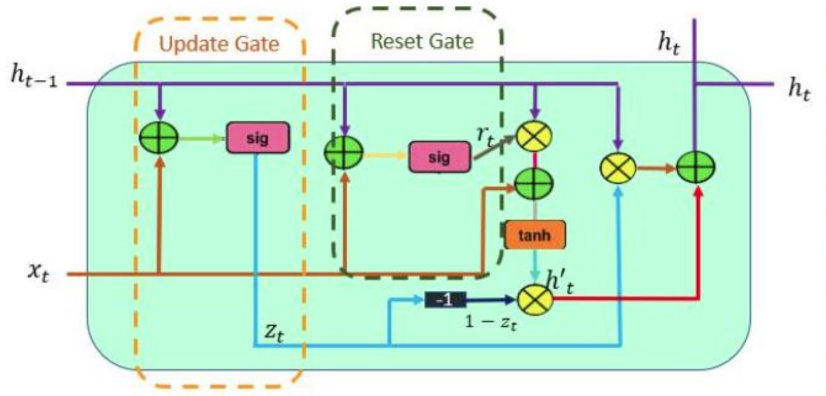


Figure 5- GRU Architecture

➢ **Bidirectional RNN**

Bidirectional RNNs is neural network that allows too process both forward and backword directions. The goal is the analyse both past and future data from input data.

Bidirectional RNN has two recurrent hidden layers such as input sequence forward and processes it backward. [20] Collect the output of these hidden layers and input them to prediction making final layer. Additionally, recurrent neural network cell also using the Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU). [12]

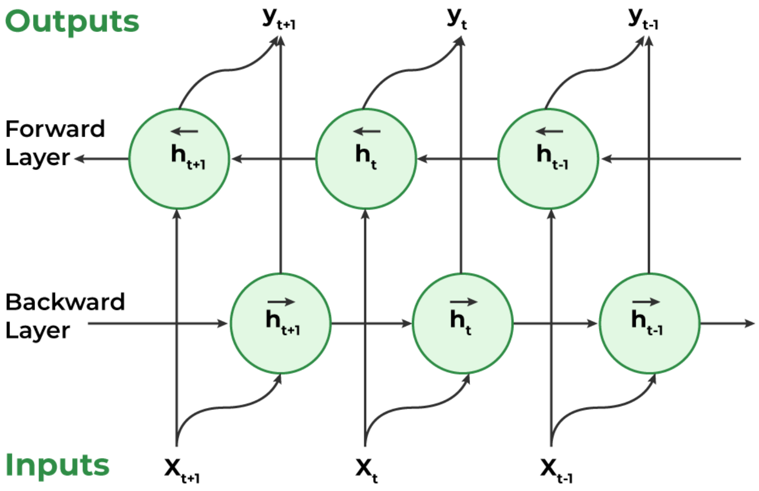


Figure 6- Bidirectional RNN Architecture

5.3 Model Evaluation and Visualization

- **Matplotlib**

  use for Data visualization and result presentation. It assists in graphing different facets of the keystroke data including the distribution of keypress durations, time intervals between keystrokes, and feature significance. Matplotlib provide better and visually appealing analysis of the data and the model and the results which can be obtained through histograms, scatter plots and line graphs. [21]

- **Seaborn**

  Seaborn is used to graphics and analysis of the datasets. Seaborn is a high-level interface developed on Matplotlib with which you get sophisticated and relatively more stylish graphics. It is useful in developing intricate structures like the heat maps, the pair plots, and the violin plots which would come in handy in the analysis of such features, distributions, and correlations in the keystroke data. The predefined themes along with colour palate enhances the appearance of visualizations and makes them interpretable and hence increases the power of communication of the analytical insights from the data. [22]

5.4 Machine Learning Development

- **Jupyter Notebook**

  Jupyter Notebook is tool that allows to combine the code, text in one documents and visualizations. Because of this, it is especially beneficial for data science projects such as the keystroke dynamics analysis where one is able to load, clean, visualise and build models directly from a single platform. This makes it possible to tinker, log the work, and also enlighten others as to some results. [23]

- **Google Colab**

  Google Colab is an online tool that comes with a Jupyter Notebook interface, all located on the Google Cloud. In doing keystroke dynamics analysis, users can write and execute Python code without install any setup locally. One feature that makes Google Colab for learning is that it comes with free GPU and TPU that helps to train machine learning models. It also enhances interaction, since notebooks are considerably easy to share as well as make changes for the purpose of collaboration in real-time. [24]

- **Visual Studio Code**

  Visual Studio Code better known as VS Code is an efficient and effective code editor used for writing and editing code, this platform is used in programming and development. These are Terminal in-built, Source code debugger and Git in-built, making it a robust environment for writing and manipulation of code. Also, its vast extensions marketplace to enable the editor to be tailored to specific needs such as the ability to support individual programming languages as well as frameworks. It is more effective for the purpose of data science and machine learning in the coding as well as experimentation. [25]

6. Requirements

6.1 User Requirements

Functionality

- **User-Friendly Interface**

The system should include a user attractive interface that does not require extensive user training or calibration and requires minimal user input beyond normal typing. The application also includes settings to set the user needs things, a panel for results showing settings, see history, see previous logins to systems, and more.

- **Multi-Platform Support**

  The system must be supports various platforms like desktop, mobile and tab to ensure reliability of user authentication. Additionally, the system must be support various operating systems such as Windows, IOS, Android, Linux.

- **Adaptability**

  Individual users have different typing patterns, and the system must learn and adapt to changes in the user's typing patterns over time.

- **Integration with Systems**

  The solution must effectively integrate with existing authentication systems and other authentication mechanisms like mouse dynamics, voice, and gait analysis to enhance the overall system security. [26]

- **Scalability**

  The system must manage a range of users and data volumes, enabling both small- and large-scale deployments.

Performance

- **Accuracy and Reliability**

  The system must ensure accurate the user authentication process with reducing false positives and negatives. It is important for performance to be constant across a range of situations and environments in order to maintain reliability.

- **Real-Time Processing**

  The system must provide fast authentication result, processing keystroke data without long delays.
- **Resource Consumption**

  The system must be lightweight and efficiently operating with limited CPU power, GPU power and memory.

Security

- **Security and Privacy**

  The system must securely manage, transfer and store the keystroke data with protecting sensitive data.
- **Compliance**

  The system must be ensuring the relevant regulations and standards for data collecting, data processing and data storing.

6.2 Functional Requirements

Keystroke Dynamics Accuracy

The system must achieve a high level of accuracy in recognizing individuals based on their keystroke dynamics, with a target accuracy rate of 95% or above. This ensures that the system can successfully identify various individuals based on their unique typing patterns. The system should maintain a low error rate, reducing both false positives and false negatives. This is crucial for ensuring the reliability and accuracy of the authentication mechanism.

Real-Time Analysis

The system must be able to evaluate keystroke dynamics data and output authentication results in real time with maintaining the latency minimum level ideally within 2 seconds to ensure fast user authentication.

Data Pre-Processing

The system will utilize keystroke dynamic data collected from online datasets and keylogging. This approach allows a diverse set of keystroke patterns, enhancing the model's robustness and generalizability. The system must properly preprocess the raw keystroke data obtained from these online sources. This includes tasks such as data normalization, noise reduction, managing missing values, cleaning and feature extraction to ensure data quality and consistency before it is used in model training and evaluation.

Model Training and Evaluation

The system must be able to train hybrid models such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models using the pre-processed keystroke data. This requires setting up and fine-tuning the model to increase its authentication capabilities and learn from the data. Furthermore, system must include comprehensive evaluation methods like precision, recall, and F1-score to evaluate the model performance and accuracy. It helps to recognize unique typing patterns efficiently.

User Management

The system must be able to enrol new users by capturing and storing their keystroke dynamics data. The system must maintain user profiles like manage user access and updating keystroke data. The system able to adapt to changes the user’s typing behaviour over time and maintain accurate.

Result Reporting

Depending on the keystroke dynamics system, the decision-making system should be able to offer detailed reporting on the authentication result, the identification results as well as the confidence level. This feature will inform the users with comprehensive understanding of the authentication process thus allowing them to make conclusions on the effectiveness of the system. Furthermore, you can export the results and data in any format like CSV so can use the contents in Excel, reporting tools, other systems, and databases.

Security and Privacy

The biometric data has to be protected and this entails that adequate measures though encryption are put in place to protect both the data in transit and the data at rest to ensure that user’s privacy is observed whereby unauthorized personnel does not access the biometric data. Also, there is the need to incorporate security measures such as access controls that restrict the entry point of the particular data and some functions to accredited persons to avoid abuse by other unauthorized persons and should conform to the current security policies.

6.3 Non- Functional Requirements

Scalability

The system must accommodate more users to the system and data proportional to it and at the same time the performance should not be compromised in anyway. It has also to be designed in such a way that one can easily add more resources or more servers based on the growth of the system especially when handling huge amounts of keystroke data.

Reliability

The system should be very reliable, virtually free of breakdowns, and has to remain up and running at all times despite contingencies. It must be able to incorporate error-checking mechanisms and adjust to compile results that are correct all the time and irrespective of the number of users and workload on the system. Data backup and the recovery procedures need to be employed constantly to avoid possible loss of data.

Security

Encryption policies should be incorporated into the system wherein data must be protected from other people while in transit and when placed on storage devices. It should also apply maximum level of access control which only enables some special personnel to access or alter the data. Concerning the protection of the user's privacy, the system must obscure the data wherever feasible and meet the Privacy regulations.

Usability

This is something that should not necessitate training to be run hence the need to have a friendly user interface. It should be able to give quick response to users' prompts to input data and for authentication processes.

Maintainability

Ensuring that the system can be managed and amended where require without much problems. It should therefore be designed in a way which can be adapted or updated in some areas without affecting the rest of the overall system. Documentation should be clearly documented form the installation of the system all through to the troubleshooting process. Furthermore, there ought to be the right support structure that should deal with matters arising from time to time and also make modifications regularly for the enhancement of the systems' security and operation.

7. Feasability Study

7.1 Technical Feasibility

Availability of Tools and Technologies

The project incorporates familiar tools and technologies. The primary language that is Being used is Python because of flexibility it offers plus it has support for the machine learning platforms as in TensorFlow, PyTorch, and Scikit-learn. When it comes to data handling the following will be used Pandas and NumPy. These are tools that are easily accessible and widely backed up and hence ease the deployment of the system. [17]

Expertise and Knowledge

These skills include machine learning, data science, and software development that the team has needed. They have studied CNNs and RNNs, which are going to be the key to this project. This is the case since they come from a cybersecurity background and hence, make them appreciate the intricacies of behavioural biometrics.

Data Availability and Suitability

Sources from both academic and online sources for keystroke dynamics datasets already available will be employed. Such databases are key press patterns or necessary for training the model's databases. This utilises pre-existing datasets so as to substantially avoid the gathering of new data for big-data based algorithms of distinctions and offers a viable starting point for developing and evaluating the models.

System Architecture and Design

Its modularity and scalability are some of the critical features of the system. It will be pre-processed to normalized and cleaned before the extraction of features using CNN as well as the RNN. The hybrid model incorporates all these features for an analysis of keystroke data. The system has been divided into modules. Therefore making it possible to adopt development, testing and maintenance of the system.

Integration and Deployment

The system is intended to be able to plug into existing authentication systems by following standard protocols and application programming interfaces. It can be implemented in the local area network or as remote solution as well as be utilized as a cloud-based solution or the use of a combination of both making it versatile in its use.

Risk Mitigation

That is why within the framework of this project, the corresponding measures to prevent risks are reflected in the plan. Pre-existing frameworks in Machine learning will help in cutting down the time to be taken and also the mistake to be made. Computer services will be delivered via the Cloud to meet impacts related to computing and guarantee system accessibility.

7.2 Economic Feasibility

Cost of Data Acquisition

The project will use previously collected keystroke datasets that can be sourced online and most of which can be gotten free. This approach has the huge advantage of reducing the costs for data collection. there is no need to build databases from scratch, no need to spend days or weeks in the field to collect the data.

Software Tools and Libraries

Major tools and libraries used in the project shall be Python, TensorFlow, and Scikit-learn but they are open source. These tools are open source and hence reduces the amount of cost that will be spent on software to some level. Also, such languages as Python use IDEs like Visual Studio Code that also include free versions, thus also reducing the costs.

Computing Resources

The training and testing of the machine learning models will require a lot of computations which will be addressed by cloud computing services. E.g., AWS, Google Cloud, or Microsoft Azure provide computing services that can be easily adjusted in size and used based in the actual costs without any subscription. The flexibility of this system means that the project will only spend what is necessary to meet the objectives and will avoid paying for more computing capability than is needed.

Hardware Requirements

No specialised or high-end hardware will be used at the project site. Organs Blade servers or adequate power machines for development, testing and small implementation scales are acceptable. The system has little need for costly investment on various pieces of hardware, thus fate advocating for the economic feasibility of the project.

Maintenance and Updates

Updates and subsequent maintenance are expected to cost very minimal amount since the project will rely on easily accessible open-source tools from the community. The great flexibility of the used system architecture also means that changes and improvements can be made without great effort, which keeps the costs of the system in the long run low.

7.3 Schedule feasibility

The project timeline, which runs from July 2024 to May 2025, is designed to ensure the Keystroke dynamic authentication system is developed efficiently and completed on time.

Data Collection (Weeks 1 - 8)

The first process is collecting the keystroke data using online datasets and using keyloggers.

Data Preprocessing (Weeks 9 - 17)

In this process collected keystroke dynamic data will be moved to preprocessing tasks such as normalization, cleaning, managing missing values and notice reduction. This process ensures the keystroke data quality and reliability.

Feature Extraction (Weeks 18 - 26)

This process will extract features from the pre-processed keystroke data. Convolutional Neural Network (CNN) will be used to extract spatial features and Recurrent Neural Network (RNN) will be used to extract temporal features of the keystroke dynamics data.

Sequence Analysis (Weeks 27 - 35)

The sequence analysis process is analysing the temporal patterns in keystroke dynamic data using combination of CNN and advanced RNN technologies like LSTM, GRU and Bidirectional RNNs. This process improves the accuracy and robustness of the keystroke dynamics authentication model.

Model Integration and Authentication Decision (Weeks 36 - 44)

The final stage is the combination of the outcome from the CNN and RNN into one model for the authentication decisions.

Testing and Validation (Weeks 45 - 49)

Testing and validation ensure the keystroke dynamics-based authentication system works effectively and accurately. The system will be tested with known datasets to check its accuracy in authenticating the users.

Final Report and Presentation Preparation (Weeks 50 - 54)

The final report and presentation will summarize the entire project, highlighting key findings, methodologies, results and live demostrate

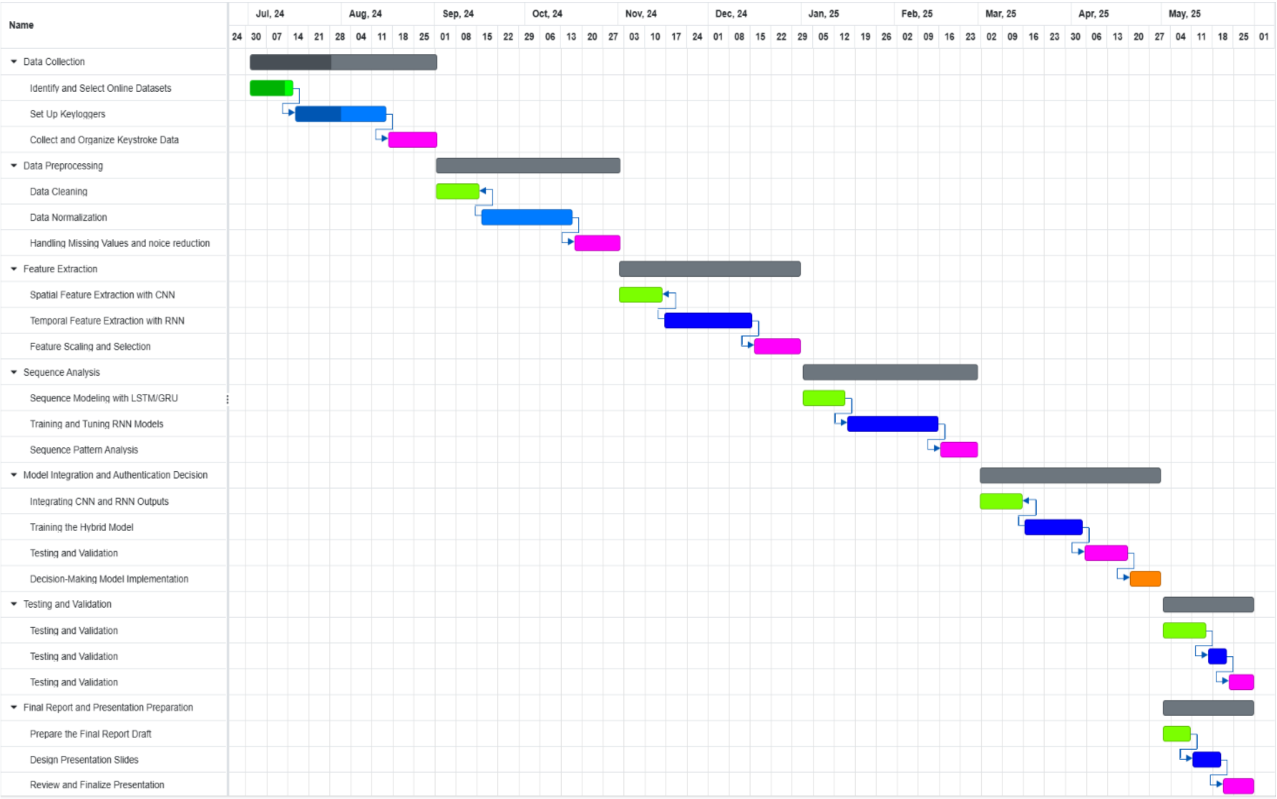


Figure 7 - Gantt Chart

8. Budget

8.1 Data Collection

- Cost of purchasing and accessing the keystroke dynamics dataset like CMU Benchmark Dataset, BKSD Dataset and more. Cost of using keylogers
- Estimated Cost: $50

8.2 Hardware

- Cloud computing service like Azure, AWS for model training, feature extraction, and other process.
- Estimated Cost: $100

8.3 Software

Development Tools and Libraries

- Costs of machine learning and deep learning libraries such as TensorFlow, Keras, and scikit-learn.
- Estimated Cost: $50

Feature Extraction & Sequential Analysis process

- Additional cost required for the extracting relevant features from keystroke data and analysing the keystroke data using advance RNN technologies.
- Estimated Cost: $10

Integrated Development Environment

- Cost for tools and plugins for development and coding process
- Estimated Cost: $0

Version Control

- Cost for version control and collaboration tools like GitHub or Bitbucket
- Estimated Cost: $0

Visualization Tools

- Cost for the Visualization Tools and libraries like Matplotlib, Seaborn
- Estimated Cost: $20

8.4 Other Expenses

Documentation and Reporting

- Cost for create the documents, printing, binding and presentation.
- Estimated Cost: $10

Contingency Fund

- Provision is made for unexpected expenses or additional requirements or others that may arise during the project.
- Estimated Cost: $50

Expense Category	Estimated Cost
Data Collection	$50
Hardware	$100
Development Tools and Libraries	$50
Feature Extraction & Sequential Analysis process	$10
Integrated Development Environment	$0
Version Control	$0
Visualization Tools	$20
Documentation and Reporting	$10
Contingency Fund	$50
TOTAL	**$290.00**

Figure 8 – Budget

9. Commercialization

9.1 Market Opportunities

Healthcare and Personal Security

As a variant, keystroke dynamics can be used in the healthcare field to control the state of health by identifying shifts in the typology which can signify diseases. This technology can be implanted in wearable gadgets and smart phones to ensure that not only is access effected safely but also to monitor users' activity as they seek medical, safety and health attention.

Smart Environments and IoT

In smart environments and IoT, the keystroke dynamics can play a positive role in the security and the user experience by providing the smart building system with the authentication. This enables changes of aspect such as brightness and warmth individually for the identified face. Also, the integration of keystroke dynamics into IoT devices can successfully enhance the access to devices and applications and add even more tangible features to smart homes and make them more secure.

Financial Sector

Keystroke dynamics can improve the security for the online financial institutions such as in banking and trading systems. Through constant monitoring of typing characteristics, financial institutions can be in a position to identify fraudulent dealings for example, those relating to; transactions and or account log-ins. For this reason, it increases security and protection of the financial data and transactions beyond proximal authentications.

Government and Défense

In government and defence area, the keystroke dynamics can be used for authentication to access the top-secret information and secured systems. It enhances security for high-risk facilities where there is need for protection of national security so that only approved individuals can access sensitive information and restricted areas.

9.2 Business Models

Licensing Model

License the technology involving the identification of keystroke dynamics to software developers in the cybersecurity industry cybersecurity industry, banking institutions as well as educational management platforms. [4]

Subscription Model

Introduce the keystroke dynamics system as one that will require users to pay for the service. This model can interest businesses that need solutions that keep on being worked on and developed right from the time they are bought. Develop different packages of subscriptions where the degree of services, analytics, and reports concerning keystroke data will differ. [4]

Partnerships and Joint Ventures

Partner up with companies that make authentication hardware or cybersecurity package with an interest in incorporating keystroke dynamics into their product line. Consult with researchers who are already involved in the advancement of the said technology and work with them to refine the technology and enter the market. [4]

Direct Sales

Market and sell the keystroke dynamics system for use with enterprises, cyber security companies, and technology companies directly. Address special requirements of various industries and provide implementation services and strategies. [4]

9.3 Growth Strategy

Market Entry Strategy

Initiate to conduct research with major organizations in the fields of cybersecurity, finance, education to establish the feasibility of the keystroke dynamics system and get first-hand feedback. These pilot projects will assist in fine-tuning the system based on the use by other centres and show the real application of the concept. Also ensure, that the technology will be presented in industry exhibitions and conferences as this will attract potential partners and customers hence a larger market for the technology. [27]

Marketing and Promotion

Use advertising that focuses on the favourable benefits of the keystroke dynamics system especially to the different businesses. This is in affording more secure and efficient means of user authentication and identity. Produce and circulate case and success stories as a way to increase people’s confidence on the system, and at the same time, show its function in practices. The materials can be utilised to establish this confidence and prove the application of the technology to the potential customers. [27]

10. Description of personnel and facilities

Student Number	Name	Feature
IT21340864	**E.M.N Edirisinghe**	❖ Collect Keystroke Data using online datasets and using keyloggers
		❖ Normalize the keystroke data to improve data quality
		❖ Apply Noise Reduction techniques to clean and manage missing values
		❖ Apply CNNs to extract spatial features from keystroke patterns.
		❖ Apply RNNs to capture temporal features and patterns in keystroke dynamics.
		❖ Integrate Features from CNN and RNN into a Hybrid Model
		❖ Build and evaluate the hybrid model's performance in accurately identifying and authenticating users.

11. References

[1] Y. W. a. Q. L. X. Chen, “Improving Keystroke Dynamics Authentication Using Deep Learning,” 2019.

[2] S. Bhatt, “Keystroke dynamics for biometric authentication,” 2013.

[3] A. S. a. S. K. A. Kumar, “Deep Learning for Keystroke Dynamics Authentication:,” 2022.

[4] S. G. a. S. Kumar, “A Survey on Keystroke Dynamics-Based Authentication Systems,” 2016.

[5] R. a. B. V. a. C. S. Chandok, “Behavioural Biometric Authentication using Keystroke Features with Machine Learning,” 2022.

[6] kaggle, “Keystroke Dynamics - Benchmark Data Set,” [Online]. Available: https://www.kaggle.com/datasets/carnegiecylab/keystroke-dynamics-benchmark-data-set.

[7] A. T. a. P. Verma, “Keystroke Dynamics based Recognition Systems using,” 2022.

[8] L. Craig, “convolutional neural network (CNN),” techtarget, 2024. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network.

[9] geeksforgeeks, “What is LSTM – Long Short Term Memory?,” 2024. [Online]. Available: https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/.

[10] geeksforgeeks, “Bidirectional Recurrent Neural Network,” 2023. [Online]. Available: https://www.geeksforgeeks.org/bidirectional-recurrent-neural-network/.

[11] T. P. M. M. FARHAD MORTEZAPOUR SHIRI, “A Comprehensive Overview and Comparative Analysis on Deep,” 2023.

[12] geeksforgeeks, “Bidirectional Recurrent Neural Network,” 2023. [Online]. Available: https://www.geeksforgeeks.org/bidirectional-recurrent-neural-network/.

[13] A. B. J. T. Pin Shen Teh, “A Survey of Keystroke Dynamics Biometrics,” 2013.

[14] pandas.pydata, “pandas,” pandas.pydata, [Online]. Available: https://pandas.pydata.org/.

[15] numpy, “numpy,” [Online]. Available: https://numpy.org/.

[16] tensorflow, “Keras: The high-level API for TensorFlow,” [Online]. Available: https://www.tensorflow.org/guide/keras.

[17] keras, “About Keras 3,” [Online]. Available: https://keras.io/about/.

[18] scikit-learn, " Preprocessing data," [Online]. Available: https://scikit-learn.org/stable/modules/preprocessing.html.

[19] analyticsindiamag, "Difference Between LSTM Vs GRU in Recurrent Neural Network," 2024. [Online]. Available: https://analyticsindiamag.com/ai-mysteries/lstm-vs-gru-in-recurrent-neural-network-a-comparative-study/.

[20] N. Malingan, "Understanding Bidirectional RNN," 10 Januray 2024. [Online]. Available: https://www.scaler.com/topics/deep-learning/bidirectional-rnn/.

[21] oreilly, "Visualization with Matplotlib," [Online]. Available: https://www.oreilly.com/library/view/python-data-science/9781491912126/ch04.html.

[22] seaborn, "An introduction to seaborn," [Online]. Available: https://seaborn.pydata.org/tutorial/introduction.html.

[23] jupyter, "jupyter," [Online]. Available: https://jupyter.org/.

[24] colab, "Welcome to Colab!," [Online]. Available: https://colab.research.google.com/.

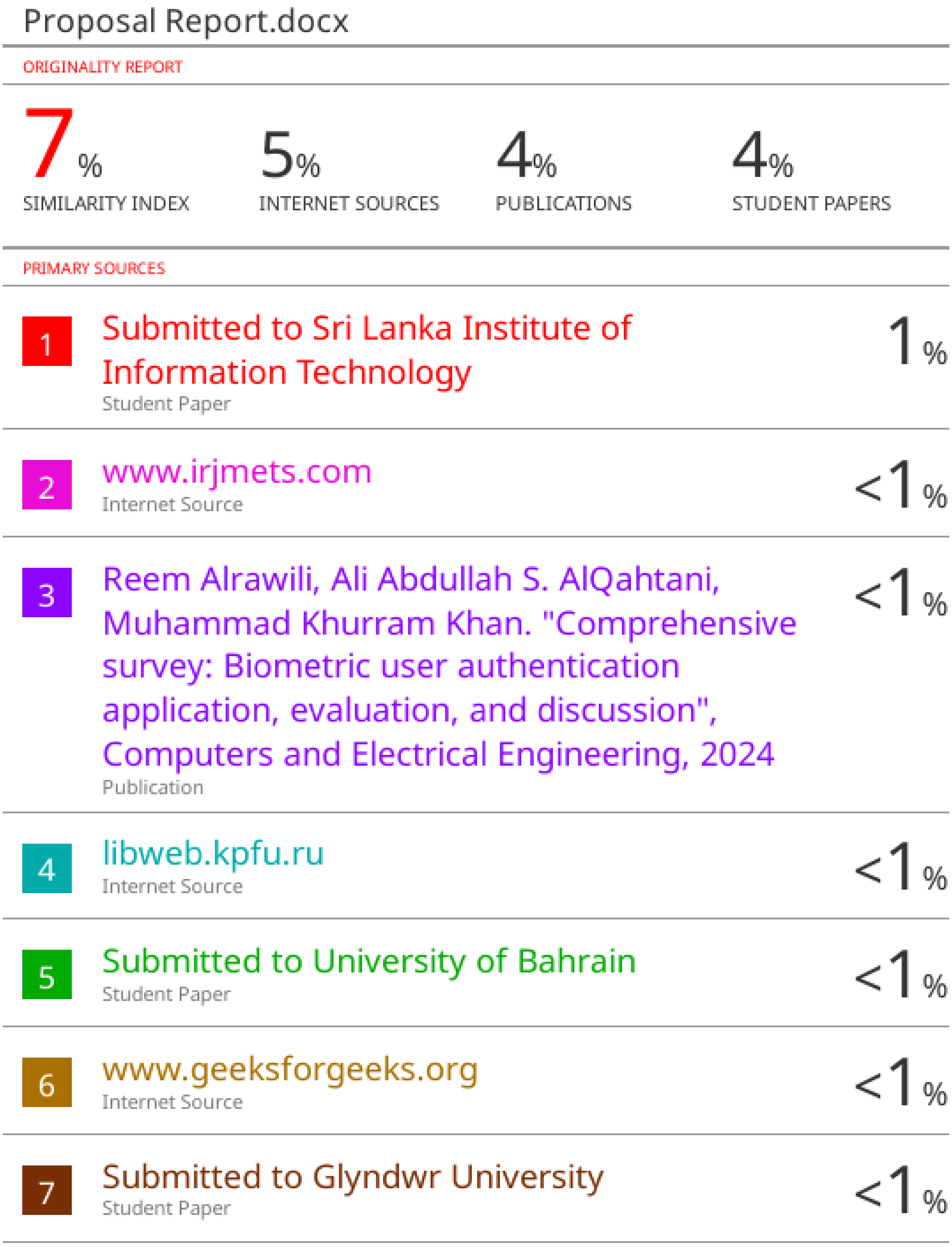
[25] visualstudio, "Why did we build Visual Studio Code?," [Online]. Available: https://code.visualstudio.com/docs/editor/whyvscode.

[26] R. R. a. S. Y. a. D. H. a. J. Dermoudy, "The design and evaluation of adaptive biometric authentication systems: Current status, challenges and future direction," no. 2023.

[27] D. H. Ahmed Wahab, "Utilizing Keystroke Dynamics as Additional Security Measure to Protect Account Recovery Mechanism," 2021.

[28] C. M. University, "Keystroke Dynamics - Benchmark Data Set," [Online]. Available: https://www.cs.cmu.edu/~keystroke/.

12. Plagiarism Report



Proposal Report.docx

ORIGINALITY REPORT

7%	5%	4%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Sri Lanka Institute of Information Technology Student Paper	1%
2	www.irjmets.com Internet Source	<1%
3	Reem Alrawili, Ali Abdullah S. AlQahtani, Muhammad Khurram Khan. "Comprehensive survey: Biometric user authentication application, evaluation, and discussion", Computers and Electrical Engineering, 2024 Publication	<1%
4	libweb.kpfu.ru Internet Source	<1%
5	Submitted to University of Bahrain Student Paper	<1%
6	www.geeksforgeeks.org Internet Source	<1%
7	Submitted to Glyndwr University Student Paper	<1%

Figure 9- Plagiarism Report