Enhancing Social Engineering Detection Using Behavioural Biometrics and Machine Learning

MSc Research Project
MSc IN CYBERSECURITY

PRANAV UDAY
Student ID: x23173980

School of Computing
National College of Ireland

Supervisor: ARGHIR NICOLAE MALDOVAN

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name**:**	PRANAV UDAY		
Student ID:	x23173980		
Programme:	MSc IN CYBERSECURITY	Year:	2024 – 2025
Module:	MSc RESEARCH PROJECT		
Supervisor:	ARGHIR NICOLAE MALDOVAN		
Submission Due Date**:**	12/12/2024		
Project Title:	Enhancing Social Engineering Detection Using Behavioural Biometrics and Machine Learning		
Word Count:	8127	**Page Count**:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: PRANAV UDAY

Date: 12/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Social Engineering Detection Using Behavioural Biometrics and Machine Learning

PRANAV UDAY
X23173980

Abstract

In order to enhance the detection of social engineering through the use of behavioural biometrics, it is necessary to make use of various human behavioural features in order to identify and combat potential threats. The most recent study has an emphasis on a number of methods, including keystroke dynamics and the merging of social behavioural information, both of which have the potential to significantly improve detection capability and robustness. In this paper, we suggest the use of artificial intelligence to develop behaviour-based detection algorithms for cybersecurity. This approach tackles both the key concerns of protecting users against complicated SE attacks and ensuring that authentication goes smoothly. In conclusion, this research attempts to improve cybersecurity detection by combining behavioural principles with artificial intelligence technology. This is done in response to the growing number of security concerns. Keystroke dynamics and touch analytics are used to train a variety of machine learning models, which are then constructed from scratch. After evaluating the performance of the models, it was determined that the random forest model produced the highest level of accuracy.

Keywords: Cyber Security, Application Design, Machine Learning, Biometrics

1 Introduction

Social engineering has turned into one of the fundamental risks in cybersecurity attacking the psychological and behavioural loopholes to gain access. While cyclic nature relies on exploiting loopholes in the target system, social engineering attacks use human characteristic and thus are difficult to notice and avoid (Mahanta & Maringanti, 2023). The necessity for efficient detection methods has only increased over the years as digitalization encroaches on industries such as finance, health sector, and critical infrastructure (Ramaraj et al., 2024; Budžys et al., 2023). Present day countermeasures comprise staff training, security awareness programmes, and other technical controls like the use of spam filter and multi-factor authentication as necessary baseline defence that, however, may lack efficiency when faced with the advanced social engineering strategies (Azhar et al., 2023; Edwards et al., 2024).

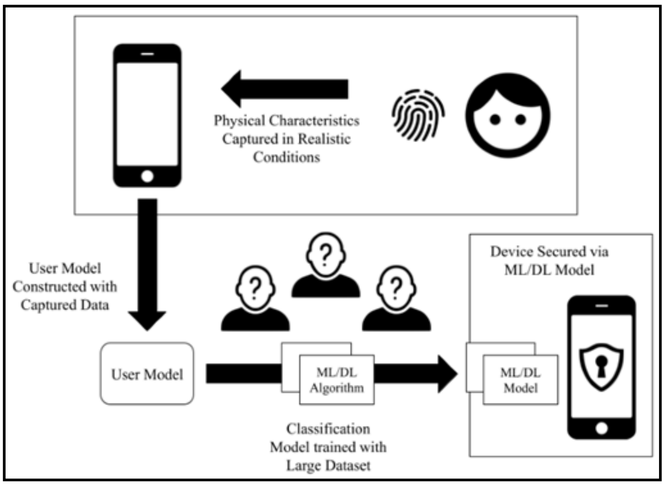
Physical Characteristics Captured in Realistic Conditions
User Model Constructed with Captured Data
Device Secured via ML/DL Model
User Model
ML/DL Algorithm
ML/DL Model
Classification Model trained with Large Dataset

Figure 1: A machine learning and deep learning framework for biometrics mobile authentication

1.1 Motivation

The combination of AI and ML with behavioural biometrics is seen as a potential for improving social engineering protection. Continuous user verification approach includes biometrics of the behaviour like the Keystroke dynamics, Mouse movements, and Touch-stroke recognition which are subclasses of behavioural biometrics as proposed by Madavarapu et al., 2024; Fakhouri et al., 2024. Thus abstracted, these behavioural parameters record specific aspects of user behaviour in real-time, obviating personal intrusion for passively monitoring for unusual or unusually suspicious responses to social engineering threats (Rebeca, 2023). Interestingly, there is reasonable prediction of such patterns by machine learning, with deep leaning algorithms such as Long Short-Term Memory (LSTM) networks showing proficiency that approaches 95.6% accuracy in user verification (Madavarapu et al., 2024). Bansal & Ouda, (2024**)** presented an integrated model that incorporates behavioural biometrics for improved social engineering detection. This model is developed by drawing from various views. The purpose of their study is to transform raw biometric data into visual data in a format that would feed into the Convolutional Neural Network (CNN) is to improve the extraction of features and the detection of social engineering techniques that are inherent to the data.

Moreover, powerful AI solutions, including TimeGAN, are used to produce real-like behavioral data for detecting and fine-tuning the mentioned threats' techniques and enhancing the model's robustness against further enhancement of social engineering attacks (Gupta et al., 2023). This framework improves the process of authentication and significantly considers privacy and usability hence flexibility of this framework in the face of social engineering threats.

1.2 Research Objective

In my study, I attempt to suggest the concept of behaviour-based detection systems as applicable to the field of cybersecurity. The approach outlined in this work effectively responds to both important issues concerning strengthening user's protection against elaborate SE attacks and preserving their smooth and uninterrupted authentication. At the end, this research aims at enhancing cybersecurity techniques by incorporating behavioural essentials with artificial intelligence technology in posing refreshing detection with the increasing security threats. This thesis therefore fits into social engineering detection through the fusion of multiple behavioural biometric modalities including Analyzed Keystroke Dynamics, Analyzed Touch Patterns, and Social Behavioural data for a much-enhanced detection mechanism. It adds flexibility in the model update to the latest social engineering strategies to guarantee constant enhancement in a short period. Using different ensemble models like Random Forest and XGBoost the result is more accurate and generalized compared to previous works that use basic models. Also, the project focuses on a real-time basis for attack identification so that users can be informed immediately to prevent losses. These results establish this method as superior to previous work and to provide the field of behavioural biometrics for cybersecurity with the advancement needed. In detail:

- **Integration of Multiple Behavioural Biometric Modalities**: In contrast to many similar studies in which only a limited type of behavioural biometrics is examined (for example, keystroke dynamics or touch analytics), this work examines multiple forms of behavioural biometrics. While keystroke dynamics and touch patterns are significant in identifying an attacker’s fingerprint, the integration of the collected social behavioural data increase the robustness of recognizing social engineering attacks.
- **Dynamic Model Adaptation**: Unlike most research studies, this project focuses on the creation of models that are capable of changing according to new social engineering techniques. Most of the previous work has employed models of a fixed architecture with predetermined features, while this work proposes the active updating of the detection models considering updated patterns and behaviours in order to maximize their performance.
- **Advanced Ensemble and Non-linear Models**: Unlike previous studies where only more basic techniques such as Decision Trees or Logistic Regression ire employed, this research adopts the more contemporary and accurate ensembles like Random Forest and XGBoost as the models because of their ability to work through complicated and non-linear patterns within data. The application of these models leads to even better generalization and accuracy, leaving alone the case of using smaller and imbalanced datasets.
- **Real-time Detection Focus**: The other novelty is the aim to identify and prevent currently unfolding real-life attacks that use social engineering techniques, which can be supported by means of behavioural biometric data for real-time monitoring and gaining feedback at once. This is important if one wants to address the problem before it reaches a point where it can inflict severe harm, something that probably makes this study more prospective than some other equivalent studies.
- **Evaluation and Comparison with Prior Research**: In addition to the use of generic evaluation measures such as accuracy, precision, recall, and F1 score, this project examines how the proposed models perform relative to previous work done by Momoh et al. in 2023. This is to highlight advances made over previous concepts.

1.3 Research Questions

The research question I like to describe in this thesis is How do machine learning methods improve behavioural biometrics-based social engineering attack detection accuracy and speed? In the following, I will discuss in detail about the different previous papers, followed by the methodology discussion. In the results and analysis, I will discuss in detail about the results of machine learning models.

2 Related Work

2.1 Social Engineering Threats and Prevention in Cyberspace

The research in Momoh et al. (2023) examines human facets in cyberspace whereby the subject implementing testbed contains financial institutions to evaluate the management of social engineering attacks. It discusses scholarly papers to measure social engineering occurrences, justify their relevance, and assess the current protective strategies. Several forms of social engineering attacks and their effects on human factor in financial contexts are discussed in the paper. One of the highlights of the authors is that organizations and companies should have wide-ranging awareness programs, mock phishing tests, and regular sessions for new information about attacks. The main objective is the improvement of the cybersecurity in financial companies and security of transactions and information in the context of their growing digitalization. This research also hopes to add its value in enriching information security research area by emphasizing the role of behaviour towards the protection of certain financial institutions (Momoh, Adelaja, & Ejiwumi, 2023).

The behaviour biometrics, such as keystroke dynamics, and human activities discussed in this thesis classified into classical and deep learning techniques. They propose a new way of working with raw biometric data in the form of time series which has been visualized as 2D colour images to enable the use of Convolutional Neural Network for angiogram feature extraction. This work seeks to evaluate this approach on the EER performance metric of this approach and demonstrate how it can enrich the security aspect without losing user familiarity in regular databases. However, the thesis also describes limitations of biometric systems mainly in terms of presentation attacks and uses temporal adversarial generators, (TimeGAN) to create synthetic behavioural data mimicking real users. Despite the outcomes obtained, research depicts biometric authentication as successful in addressing numerous issues in diverse fields, such as banking and cybersecurity, the increase in cyber risks demands credible protection of personal and important information (Momoh et al., 2023; Khan et al., 2023).

Also, it reveals how social engineering uses social media with great prejudice, in which people's information is used falsely for destructive ends (Rebeca, 2023). The study presents strategies from three perspectives: by technical measures provided by the SM, users' responsibility, and organizational rewards and training. Based on a compilation of existing literature and surveys with cybersecurity experts, suggestions for imprinting methods alongside AI and targeted measures against social engineering attacks were made (Bergmann & Solheim, 2024).

Furthermore, a discussion of social engineering threat also favours on the educational approach to prevent threats focusing on the multi-factor authentication, constant updates, and security-mindedness in organizations to eliminate the effects of these kinds of attacks (Azhar et al., 2023). This paper focused on social engineering which, unlike other forms of attacks, uses deception to take advantage of psychological vulnerabilities, and represents a major threat to organizations, as it can lead to data leakage and loss of reputation. Identified herein is the historical perspective of social engineering strategies, accompanied by a section on multi-levelled security that could address these threats. This framework focuses on strengthening security control, training the staff, and using behavioural analytical systems or AI systems for immediate threat identification (Edwards et al., 2024).

2.2 Behavioural Biometrics and Machine Learning for Cybersecurity

In the study of Madavarapu et al., the author stresses the improvement of cybersecurity through the implementation of behavioural biometrics supplemented by machine learning with the aid of efficient verification systems. Instead of focusing on the content of passwords typed by a user for example, the researchers introduce more efficient verification processes which consider real-time changes in the typing style and mannerisms including but not limited to keystroke dynamics and mouse movements. From the experiments which they performed on real life datasets, the authors clearly show that higher accuracy of 95.6% can be achieved by LSTM models when compared with other machine learning types like SVM (88.7%), KNN (84.5%), and Random Forest (92.3%). However, LSTM training is longer as compared to the other networks, yet LSTM presents enhanced flexibility and efficiency in comprehending patterns in sequential data and progress behavioural biometric recognition systems.

This paper focuses on the increasing phenomenon of social engineering attacks in cybersecurity, where a computer hacker takes advantage of an individual's weakness instead of using a loophole; specifically, it targets the internal human factor within an organization's network (Mahanta & Maringanti, 2023). It splits out social engineering techniques like phishing, pretexting, baiting, Quid pro quo, shows how social engineers work, working on urgency, trust, and fear. The chapter also touches on measures to reduce the mentioned risks including increased staff training and security awareness, and technical controls including spam filtering, and two factor authentication. For instance, the use of electronic health records and patient data in the care delivery processes of various specialties has been widely discussed as there are many parties enrolled in it and data security and data integrity may be put at risk (Ramaraj et al., 2024). As a remedy for this, a high-level behavioural security approach is presented which incorporates a one-class Support Vector Machine (SVM) model to different types of a normal user that accesses the healthcare systems. Due to the ability to identify anomalies and adjust the authorizations based on the user interactions this framework successfully contains the unauthorized access and thus increases the security of the vitally important patient data. Altogether, these works help to progress the insights of more secure and trustworthy technologies in both cybersecurity and healthcare (Budžys et al., 2023). The paper also emphasis on the issue of insider threats in critical infrastructure protection by presenting a theoretical method of utilizing deep learning networks to perform user authentication. The methodology converts behavioural biometric data into images and employs keystroke patterns together with Siamese neural networks to improve the intrusiveness of the system and make user authentication quicker (Budžys et al., 2023).

Furthermore, the thesis analyses the weaknesses of biometric systems and more specifically, presentation attacks. This work uses temporal adversarial generators (TimeGAN) to generate the synthetic behavioural data and evaluates the proposed approach by qualitative and quantitative perspectives to showcase that the proposed approach can be used to evaluate the authentication systems (M. K. Gupta et al., 2023; Bansal & Ouda, 2024). In the context of IoT, this article involves HCI based and natural habit based behavioural biometrics for identification of the user. It raises the degree of user identification as a key competence within IoT environments, thus defines a number of touch-stroke, swipe, and voice recognition modalities. The survey is expected to improve both security and utility of IoT applications in response to current recognition schemes while recognizing the future research area (Khan et al., 2024).

2.3 AI-Driven Techniques and Future Directions in Cybersecurity

Similarly, in the same context, Fakhouri et al. (2024) carry out a similar study on the complexity of social engineering tactics that bypass the advanced securities systems by exploiting human vulnerabilities. The currently used attack characteristics, which use social networks, mobile applications, and artificial intelligence are discussed, as well as the role of artificial intelligence in the fight against such attacks. They note that machine learning has a significant challenging task of picking out dependencies, revealing how NLP is helpful in recognizing the scam, and explaining how analytical prediction models can help to foresee additional attacks. They demonstrate that utilization of AI improves efficiency of the distinguished mechanisms with regard to detection and prevention of social engineering threats, which behavioural analytics can identify tendencies of potential manipulative activities. They also argue that such AI tools and capabilities must be trained actively and must update constantly according to trends in the cyber threats. Behavioural biometrics remains a young method of user identification or verification where the interactions between user and IT system are analysed and while this increases security and provides better customer experience, it is highly invasive concerning user privacy (Momoh et al., 2023; Khan et al., 2023).

Furthermore, it recaps more advanced social engineering tricks caused by technology progress and demonstrates how AI and cloud services can bolster security systems. Applying artificial intelligence and processing big data allow to recognize the phishing attacks, investigate the users' actions, and provide the automatic reactions to the incidents. It also looks at the ethical perspective of AI in cybersecurity and also puts into consideration the necessity of more research partnerships between academies, industry and government to enhance invention in tackling the social engineering threats (Momoh et al., 2023; Piugie, 2023). This thesis also provides an extensive overview and classification of behavioural biometrics, clearly differentiating the field of IT user identification and authentication. It presents an archetypical approach which covers types like keystroke dynamics and human actions basing on classical machine learning and deep learning. A new method is described for processing raw biometric data in the form of time series converting it to 2D colour images but retaining the properties of the behavioural signals. This method enables the application of 2D convolutional networks to produce highly optimised deep feature vectors for user authentication that resulted in suitable performance indicators such as Equal Error Rate (EER).

Finally, the examination of the kind of attacks relative to the authentication systems leads to the identification of various techniques of social engineering attack for example phishing, and vishing. Therefore, the paper stresses the need for increasing the public awareness and carrying continuous training among them as main methods of risk management in this sphere highlighting the need for effective methods of the behavioural biometrics' authentication based on the such frameworks as B2auth: for real-world use-cases and more efficient data collection (Borowiec et al., 2023; Mahfouz et al., 2024).

3 Research Methodology

3.1 Dataset Description

3.1.1 Keystroke Dynamics Challenge Dataset

Keystroke dynamics pertains to the examination of a user's keystroke patterns. This can enable non-intrusive real-time user authentication. Analysing the intervals between two key presses the duration of key press, and the time elapsed from key release to the subsequent key press might yield significant insights regarding the user. Upon the user's subsequent login, the authenticity of the user can be verified by contrasting their present typing pattern with prior patterns to determine if the user is legitimate or fake. Insights on user typing patterns are gained by examining features like press-press duration (PPD), hold duration (HD), release-press duration (RPD) as shown in Fig. 3. The graphic illustrates the duration assessments for pressing two keys: A and B. The smaller keys denote the key press event, whereas the larger keys signify the key release event.

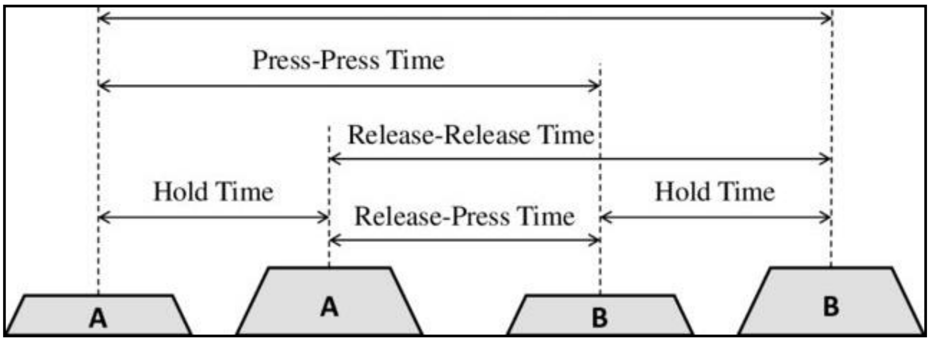

Figure 2: Features showing various metrics of Key press dynamics

3.1.2 Touch Analytics Dataset

This dataset comprises the unprocessed touch data from 41 individuals engaging with Android smartphones, together with a collection of 30 derived attributes for each touch stroke. The possibility of continuous user authentication through the interaction with a smartphone's touchscreen is being investigated. A set of 30 behavioural touch features are available, which can be extracted from raw touchscreen logs, such that distinct subspaces of this feature space are populated by different users. The dataset consists of results of systematic experiments that were conducted to test whether consistency over time is exhibited by this behavioural pattern. In the dataset, Touch data was collected from users interacting with a smart phone through basic navigation manoeuvres, such as up-down and left-right scrolling. Detailed descriptions of each of these columns are given below.

A. Phone ID: indicates the phone and the experimenter that recorded the data reaches from 1-5
   1: Nexus 1, Experimenter E
   2: Nexus S, Experimenter M

3: Nexus 1, Experimenter R
4: Samsung Galaxy S, Experimenter I
5: Droid Incredible, Experimenter E

B. user ID: anonymous users
C. doc id: This number denotes the document that the user viewed on screen during data collection. Each document signifies a distinct session, indicating that the user has set aside the device while transitioning between several pages. The intervals between document IDs 1-5 and 6-7 are many minutes, respectively. Data for document IDs 6 and 7 was obtained 7 to 14 days subsequent to the collection of document IDs 1-5.
   1: Wikipedia article
   2: Wikipedia article
   3: Wikipedia article
   4: Image comparison game
   5: Image comparison game
   6: Wikipedia article
   7: Image comparison game
D. time[ms]: absolute time of recorded action (ms since 1970).
E. action: can assume three values 0: touch down, 1: touch up, 2: move finger on screen, A stroke is characterized as all activities occurring between 0 and 1, provided there is a xy-displacement between these actions.Clicks are actions that occur between 0 and 1 without any displacement.
F. 'phone orientation', 'x-coordinate', 'y-coordinate', 'pressure', 'area covered', 'finger orientation' are the values returned from the Android API at the current action.

3.2 Data Pre-Processing

3.2.1 Keystroke Dynamics Challenge Dataset

The preprocessing phase focused on transforming raw keystroke dynamics data into meaningful features for user classification based on behavioural biometrics. This step was crucial in ensuring that the data was clean, structured, and prepared for machine learning models. The process began with**,** where the training and testing datasets were examined to understand their structure and user distribution. This initial inspection provided valuable insights into the dataset’s scope and prepared the groundwork for subsequent feature engineering.

3.2.2 Touch Analytics Dataset

The preprocessing phase was integral to preparing the raw touch interaction data for machine learning tasks. This process involved generating meaningful features, normalizing the data, and selecting relevant attributes to enhance model accuracy and interpretability. The dataset was loaded using libraries such as pandas, numpy, matplotlib, and seaborn for data manipulation and visualization, while relevant modules from scikit-learn were imported for preprocessing, scaling, and model evaluation.

3.2.3 Feature Engineering

It plays a vital role in this process, as temporal features were derived to capture unique typing dynamics. Key metrics included Press-to-Press Duration (PPD**)**, which measured the interval between consecutive key presses; Release-to-Press Duration (RPD**)**, representing the gap between a key release and the next press; and Hold Duration (HD**)**, which captured the time a key was held down. These features encapsulated distinct behavioural patterns, making them critical for distinguishing between users. While for the second dataset, to capture the behavioural patterns of touch interactions, several features were engineered. The **T**ouch Duration was calculated as the time difference between consecutive touch actions where action = 0 (touch start) and action = 1 (touch end). This value was stored as a new feature, touch_duration, in seconds. Additionally, Stroke Length was computed as the Euclidean distance between consecutive x and y coordinates, while Speed was derived by dividing stroke length by touch_duration. Intermediate variables such as next_action and next_time, used in feature computation, were removed after the final features were derived to maintain data cleanliness.

3.2.4 Exploratory Data Analysis (EDA)

Potential of these features are further validated through visualizations such as histograms and swarm plots, the feature distributions and inter-user variations were analyzed, confirming their relevance for classification tasks. Following EDA, the dataset underwent a transformation to prepare it for modelling. It was reshaped into a long format to allow for focused analysis, and subsets of PPD, RPD, and HD were consolidated into unified datasets. This restructuring ensured that the data was efficiently organized for machine learning applications.

Finally, the data was refined by removing redundant columns and aligning features with user labels to facilitate efficient model training. This pre-processing pipeline effectively transformed raw keystroke data into a structured format that highlighted user-specific typing behaviours. By creating robust and interpretable features, it laid a strong foundation for accurate user classification and enabled further analysis of behavioural biometrics.

3.2.5 Data Loading and Setup

The dataset was structured and inspected to ensure compatibility with subsequent preprocessing steps. Visualizations were created to understand the distribution of features, and initial exploration provided insights into potential data issues.

3.2.6 Feature Selection

The final set of features included spatial attributes like the x-coordinate and y-coordinate, physical attributes such as pressure, area covered, and finger orientation, and engineered attributes like touch_duration, stroke_length, and speed. These features were selected for their relevance to distinguishing user behaviour. The target variable for classification was identified as the user ID.

3.3 Data Splitting and Scanning

To prepare the data for machine learning, it was split into training and testing sets using an 80:20 ratio. Standardization was applied to all feature values using StandardScaler, ensuring that the input variables were normalized for consistency across machine learning models.

3.4 SMOTE (Synthetic Minority Oversampling Technique)

It is a popular method for handling class imbalance in datasets. Instead of duplicating minority class samples, SMOTE generates synthetic samples by interpolating between existing samples of the minority class. It selects a random sample from the minority class and creates synthetic data points along the line joining the selected sample and one of its nearest neighbours. This approach helps in improving model performance by providing balanced training data without introducing redundancy.

4 Design Specification

Various machine learning models, namely, XG boost, Random Forest, LightGBM, Logistic regression, Decision tree, AdaBoost and ANN are applied to the problem. The machine learning models are trained and then tested, and the respective performances are compared.

4.1 XGBoost

Extreme Gradient Boosting is a scalable, distributed library for gradient-boosted decision trees (GBDT) in machine learning. It offers parallel tree boosting and is the preeminent machine learning package for regression, classification, and ranking tasks.

4.2 Random Forest

The random forest approach consists of an ensemble of decision trees, with each tree constructed using a bootstrap sample, which is a data sample selected with replacement from the training set. One-third of the training sample is allocated as test data, referred to as the out-of-bag (oob) sample, which will be addressed subsequently. Feature bagging introduces additional randomization, enhancing dataset variety and diminishing the correlation among decision trees. The prediction will vary based on the nature of the situation. In a regression task, the individual decision trees will be averaged, whereas in a classification job, the predicted class will be determined by a majority vote, namely the most prevalent categorical variable. Ultimately, the out-of-bag sample is utilized for cross-validation, thereby finalizing the prediction.

4.3 LightGBM

LightGBM is a quick, distributed, high-performance gradient boosting framework utilizing decision tree techniques, employed for ranking, classification, and several other machine learning problems. LightGBM employs a leaf-wise tree splitting method, in contrast to other boosting algorithms that utilize a level-wise approach. It selects the leaf to bifurcate that it anticipates would produce the most significant reduction in the loss function. Leaf-wise selection of splits is determined by their impact on global loss rather than solely on the loss of a specific branch, which frequently results in the formation of lower-error trees more rapidly than level-wise methods.

4.4 Logistic Regression

Logistic regression is employed for binary classification, utilizing the sigmoid activation function, which accepts independent variables as input and generates a probability value ranging from 0 to 1. It quantifies the likelihood that a specific input is associated with a particular class by utilizing the logistic function (sigmoid) on a linear amalgamation of input attributes.

4.5 Decision Tree

A decision tree is a diagrammatic framework utilized for making decisions or predictions. The structure has nodes that signify conclusions or evaluations of attributes, branches that denote the results of these decisions, and leaf nodes that indicate ultimate outcomes or forecasts. Every internal node represents a test on an attribute, each branch signifies the outcome of the test, and each leaf node denotes a class label or a continuous value. Root Node denotes the complete dataset and the primary decision to be undertaken. Internal Nodes indicate decisions or evaluations regarding attributes. Every internal node possesses one or more branches. Branches indicate the result of a decision or evaluation, resulting in another node. Leaf Nodes indicate the ultimate judgment or prediction. No additional divisions transpire at these nodes. Attributes that can be chosen for the root are Entropy, Information gain, Gini index, Gain ratio, Reduction in variance.

4.6 AdaBoost

AdaBoost's involves the iterative training of a weak classifier on the training dataset, with each subsequent classifier assigning greater weight to misclassified data points. The final AdaBoost model is determined by aggregating all the weak classifiers utilized during training, with weights assigned based on their respective accuracies. The model with the highest accuracy receives the greatest weight, while the model with the lowest accuracy is assigned a lesser weight.

A boosted classifier has the form,

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

Each (f_t) is a weak learner that accepts an object (x) as input and outputs a value denoting the item's class.

4.7 Artificial Neural Networks (ANN)

An Artificial Neural Network comprises an input layer, an output layer, and one or more hidden layers. The input layer acquires data from the external environment that the neural network must interpret or learn from. The data subsequently traverses one or more hidden layers that convert the input into information pertinent for the output layer. The output layer delivers a response from the Artificial Neural Networks based on the **input data** received. In most **neural**

networks, units are interconnected across layers. Each link possesses weights that dictate the impact of one unit on another unit. As data is transmitted between units, the neural network progressively acquires knowledge about the data, ultimately producing an output from the output layer.

5 Implementation

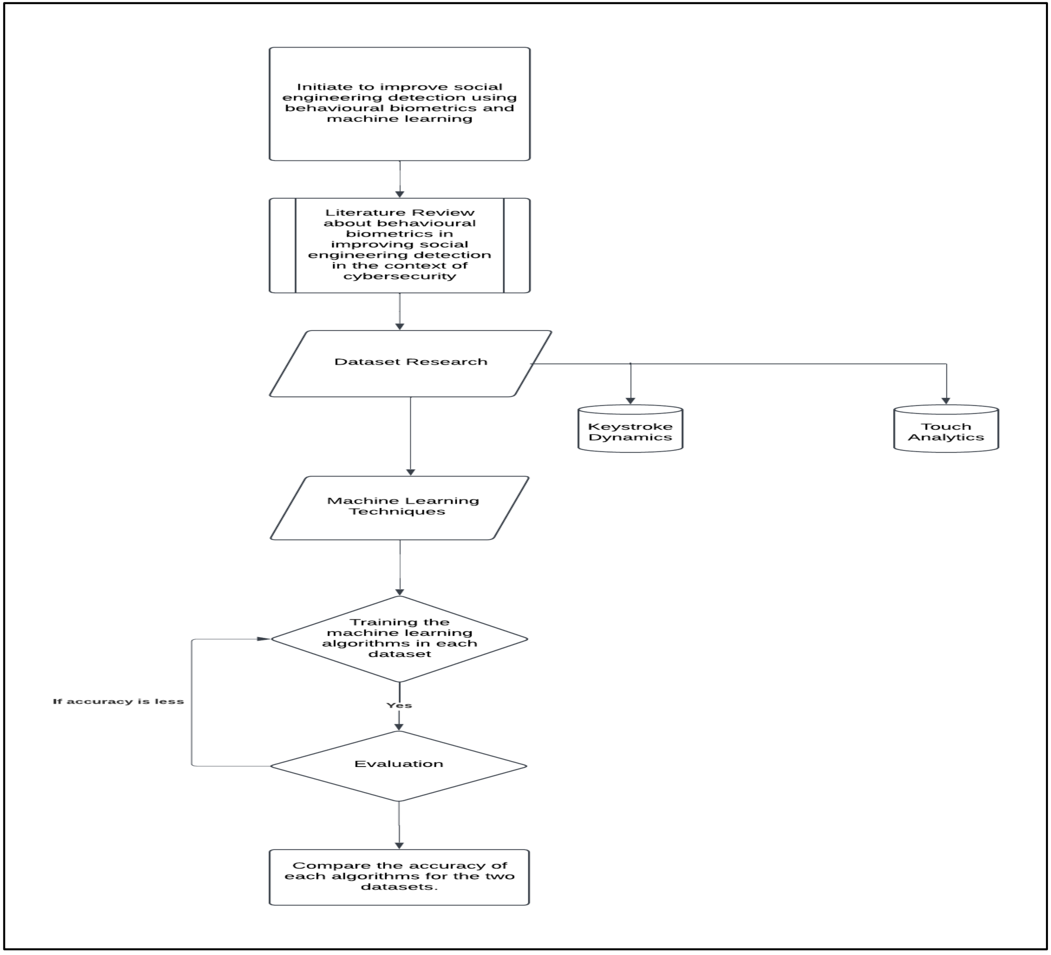
Initiate to improve social engineering detection using behavioural biometrics and machine learning
Literature Review about behavioural biometrics in improving social engineering detection in the context of cybersecurity
Dataset Research
Keystroke Dynamics
Touch Analytics
Machine Learning Techniques
Training the machine learning algorithms in each dataset
If accuracy is less
Yes
Evaluation
Compare the accuracy of each algorithms for the two datasets.

Figure 3: Steps and Methods used for the research implementation.

5.1 Algorithm Flow

***Step 1 – Load the dataset into memory*:** The objective is to do exploratory data analysis (EDA) to get insights into data distribution, identify abnormalities, and detect missing values. Employ bar plots to illustrate the distribution of the target variable 'label' for the purpose of visualization.

Step 2 – Data pre-processing process: This step was crucial in ensuring that the data was clean, structured, and prepared for machine learning models. The process began with data loading and inspection, where the training and testing datasets were examined to understand their structure and user distribution

Step 3 – Feature Engineering: Raw data is transformed into usable information for machine learning models by feature engineering. For the Key Stroke dataset, the features **Press-to-**Press Duration (PPD), Release-to-Press Duration (RPD), and Hold Duration (HD) and Touch analytics dataset the features are Touch Duration, Stroke Length and Speed.

Step 4 – Feature Selection: Machine learning utilizes feature selection to determine a subset of relevant features (variables, predictors) for model creation. It reduces feature space dimensionality, speeds up the learning method, and improves predictive accuracy and comprehensibility to improve model performance. The features are selected based on their relevance in determining user behaviour.

Step 5 – Data Splitting**:** This process prepares data for training or testing. The train_test_split instruction divides the data set into features, target, and test (80%) and training (20%) data.

Step 6 – Model Training**:** In this step the various ML models - XG boost, Random Forest, LightGBM, Logistic regression, Decision tree, AdaBoost and ANN, are developed and trained.

Step 7 – Model Evaluation: Each of the developed machine learning model is evaluated and accuracy of the models are compared to obtain the best performing model.

6 Evaluation

6.1 Dataset Analysis

6.1.1 Keystrokes Dynamics Challenge Dataset

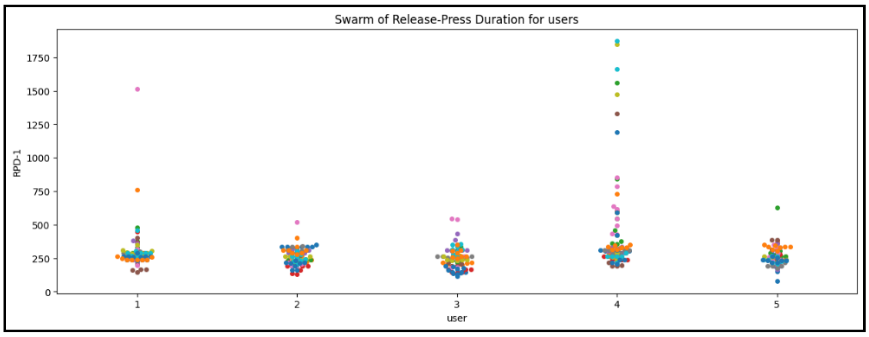

Figure 4: Scatterplot for PPD Vs RPD

The scatterplot displays the relationship between RPD-1 (x-axis) and PPD-1 (y-axis). Each point represents a data observation, and the colour intensity (from light to dark) corresponds to categories or a grouping variable, likely labelled from 1 to 5. The points exhibit a strong positive linear relationship, suggesting that as RPD-1 increases, PPD-1 also increases proportionally. Variations in colour intensity might indicate different groups, classes, or levels within the data. Points are more densely clustered in the lower ranges of RPD-1 and PPD-1, implying a higher concentration of values in these areas.

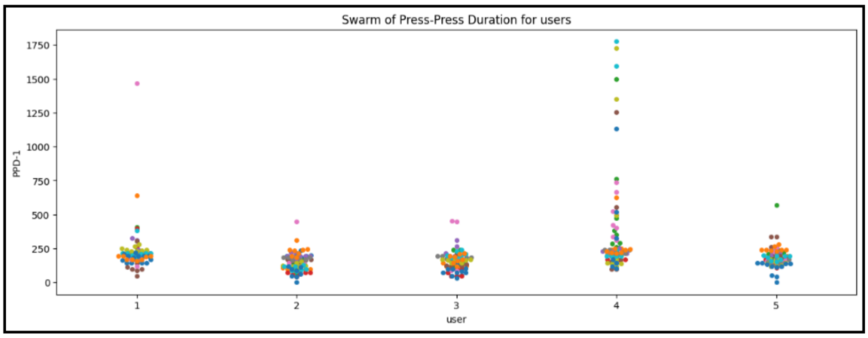

Figure 5: Using the swarm plots identifying the importance of the features

Most users have RPD values tightly clustered between 50 ms and 250 ms, showing consistent typing behaviour. User 4: Displays frequent outliers, with values exceeding 1000 ms and reaching up to 1750 ms, indicating delayed transitions from key release to the next press, possibly due to pauses or hesitations. Users 1, 2, 3, and 5: Exhibit more consistent patterns with fewer outliers, reflecting rhythmic and faster typing. Classification Insight: Variability in RPD values, especially outliers, highlights distinct typing behaviours. User 4's outliers can aid in identifying slower or inconsistent typists.

Most Users have PPD values clustered between 50 ms and 300 ms, indicating consistent key-to-key typing intervals. User 4: Again, shows significant outliers, with durations exceeding 1000 ms and up to 1750 ms, suggesting slower and inconsistent typing behaviour caused by pauses or distractions. Users 1, 3, and 5: Maintain tightly packed clusters, indicating steady and rhythmic typing. Classification Insight: PPD effectively captures typing rhythm. The distinction between User 4's variable patterns and the consistent behaviour of others supports its use in identifying typing styles.

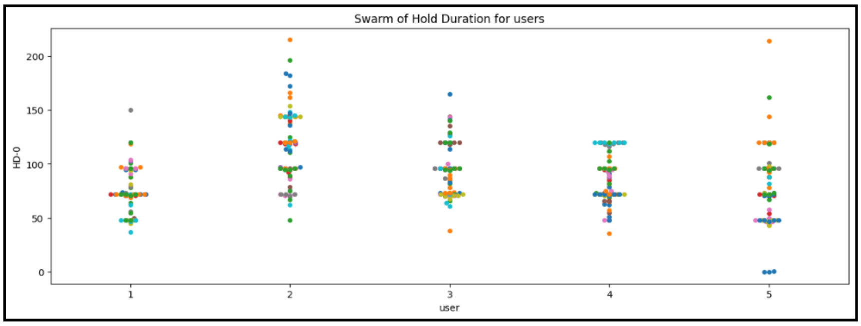

Figure 6: Swarm of Hold Duration for different users

Most Users have HD values distributed between 50 ms and 150 ms, reflecting uniform key hold times. Users 2 and 4: Show broader distributions with peaks near 200 ms, indicating variability in typing force or deliberation while holding keys. Users 1, 3, and 5: Exhibit tighter distributions with fewer outliers, suggesting consistent and uniform typing styles. Classification Insight: Variability in HD values, especially in Users 2 and 4, can help differentiate between users with dynamic typing and those with steady, rhythmic patterns.

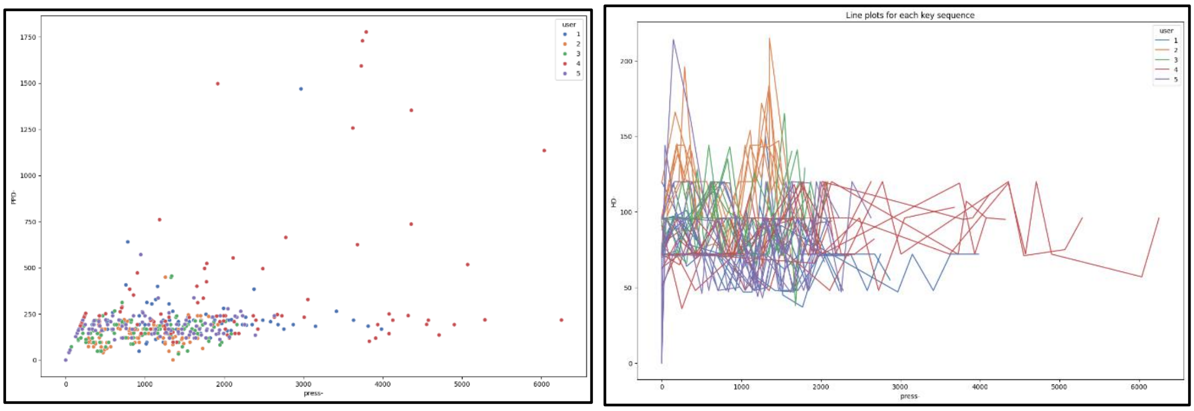

Figure 7: Scatterplot of press- vs PPD- by User and Line plots for each key sequence

This scatterplot visualizes the relationship between the feature press- (x-axis) and PPD- (y-axis), coloured by user groups (1 to 5). Spread and Clustering: The data points show some clustering in the lower ranges of both axes, indicating a concentration of users with shorter press and PPD durations. User Variability: Different users (colours) display varying distributions, particularly for higher values of PPD-, suggesting user-specific keystroke dynamics. Line plot behaviour pattern In hold duration of each user User 2: High variability with frequent peaks exceeding 150 ms and occasional spikes above 200 ms, indicating a dynamic typing style. User 4: Mostly consistent with occasional spikes, showing hesitations or deliberate pauses. Users 1, 3, and 5: Stable patterns clustered between 50 ms and 100 ms, reflecting rhythmic and steady typing. Dynamic Users (Users 2 and 4): High variability provides unique behavioural markers, making them distinct. Consistent Users (Users 1, 3, and 5): Uniform patterns are easier to classify due to their steady typing behaviour.

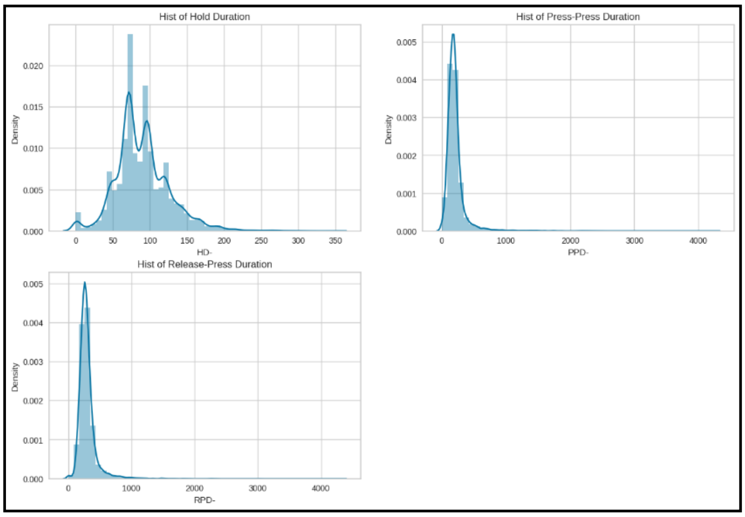

Figure 8: Histogram for HD, PPD, RPD identifying typing behaviour

Histogram of Hold Duration (HD*):* The histogram of Hold Duration (HD) reveals that most hold durations fall between 50 ms and 150 ms, with a noticeable peak around 100 ms. A small number of outliers extend beyond 200 ms, with some reaching as high as 350 ms. This clustering near 100 ms indicates a consistent typing behaviour among most users, reflecting their natural rhythm when pressing keys. On the other hand, the outliers represent longer key presses, which could signify moments of hesitation or deliberate, intentional typing. The variability in HD is crucial, as it can help distinguish users with dynamic typing behaviours from those with more consistent patterns, adding a layer of personalization to behavioural biometrics.

Histogram of Press-to-Press Duration (PPD*):* The histogram of Press-to-Press Duration (PPD) shows a tight clustering of values below 500 ms, with a sharp peak near 100 ms. A few outliers extend up to 4000 ms, representing instances of longer pauses between consecutive key presses. The tightly clustered majority indicates rhythmic typing with minimal gaps, characteristic of steady and fluent typing patterns. In contrast, the outliers highlight interruptions or slower typing speeds, which may occur due to pauses or distractions. PPD plays a critical role in capturing a user's typing rhythm, making it an essential feature for distinguishing between individuals with consistent versus variable typing styles.

Histogram of Release-to-Press Duration (RPD)*:* The histogram of Release-to-Press Duration (RPD) exhibits a distribution similar to PPD, with most values clustering below 500 ms and a prominent peak around 100 ms. Outliers extend to approximately 4000 ms, indicating slower transitions or pauses between releasing one key and pressing the next. The majority of users demonstrate quick transitions, reflecting natural and fluid typing behaviour. However, the presence of outliers highlights moments of slower transitions, which could stem from hesitation or deliberate key transitions. RPD variability provides a unique user signature and complements PPD by offering additional insights into a user's typing dynamics, enhancing the overall effectiveness of user identification based on behavioural biometrics.

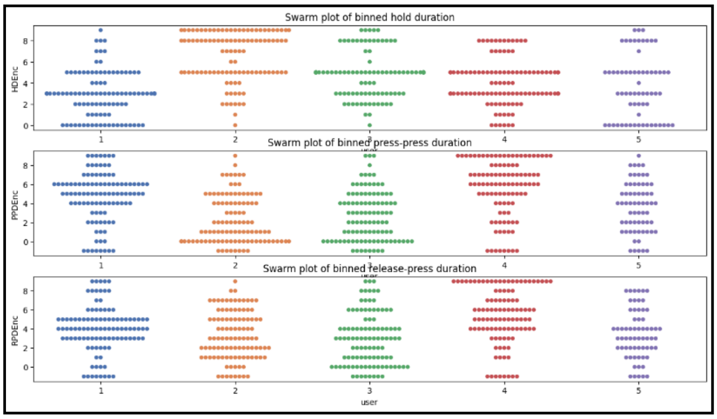

Figure 9: Swarm Plot of Binned Features by User

This plot contains three swarm plots, showing the distribution of binned durations for different features across users (1 to 5). Hold Duration (HDEnc): Represents the time a key is held before release. Each user exhibits a unique spread, with User 1 having a more compact distribution compared to others. Press-Press Duration (PPDEnc): Captures the time between successive key presses. The distributions vary significantly across users, reflecting individual typing styles. Release-Press Duration (RPDEnc): Measures the time from releasing one key to pressing the next. The distributions are distinct for each user, with varying spreads and central tendencies.

6.1.2 Touch Analytics Dataset

The bar chart shows the distribution of touch actions across three categories (`0`, `1`, and `2`), highlighting a significant imbalance. Action `2`, likely representing touch movement, dominates the dataset, while actions `0` (touch start) and `1` (touch end) occur much less frequently. This imbalance could impact model performance if action type is used as a feature, as the model may become biased toward the majority class (action `2`). To address this, techniques such as oversampling or down sampling for actions `0` and `1` may be necessary to ensure balanced representation and fair training. Analyzing the importance of these actions in the classification task can guide whether balancing is required.

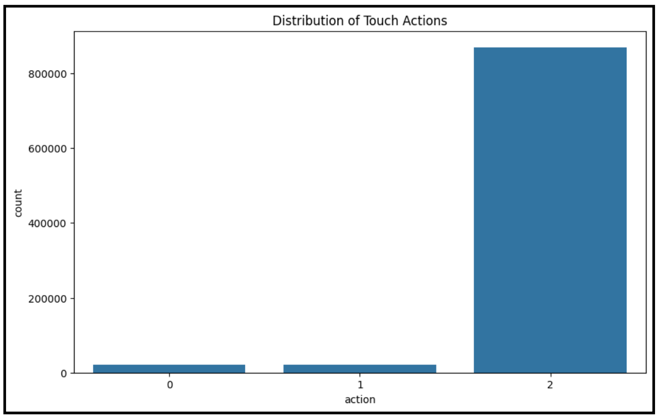

Figure 10: Imbalance Data distribution

This below plot Highlights distinct patterns in touch behavior. The orange curve (Area Covered) peaks sharply near 0.0, indicating that most touches cover minimal screen area, possibly reflecting light or precise tapping actions. In contrast, the blue curve (Pressure) shows a more uniform distribution, with values extending up to 0.8. This suggests varying levels of touch pressure, capturing differences in user interaction intensity. These variations in Pressure and Area Covered provide valuable insights for distinguishing user behaviors, making them important features for classification tasks

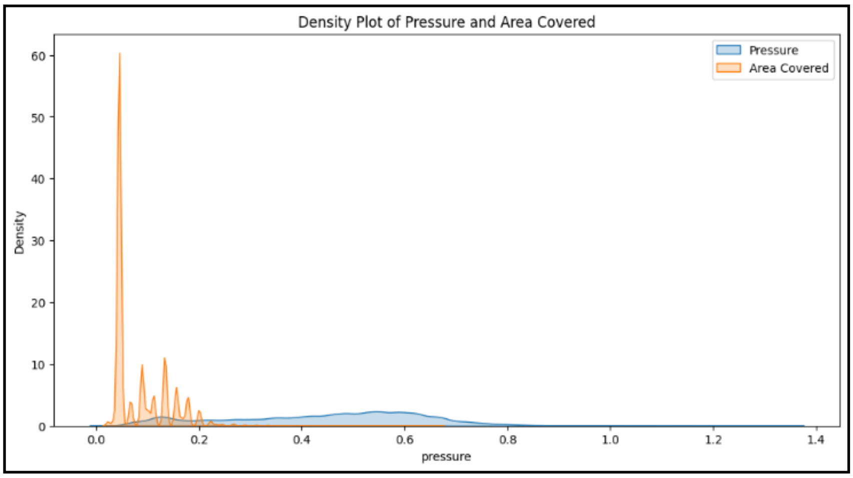

Figure 11: Density Plot of Pressure and Area Covered

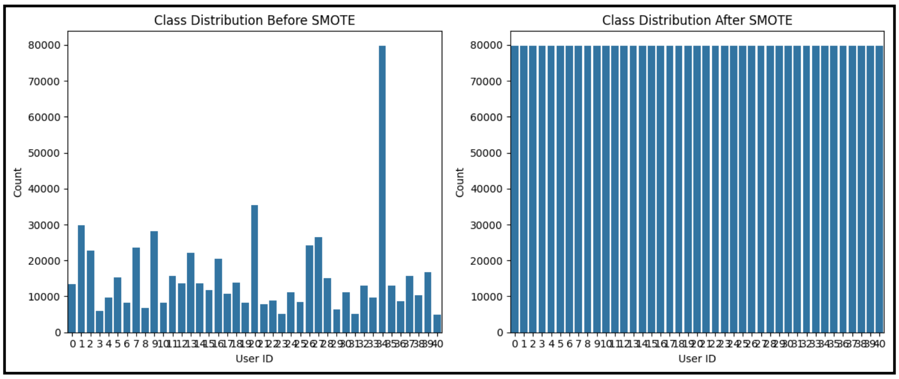

Figure 12: Application of SMOTE

Class Distribution Before SMOTE (Left Panel): In the graph, the dataset exhibits significant class imbalance, where certain user IDs (e.g., User ID 23) have a disproportionately large number of samples compared to others. This imbalance can lead to biased machine learning models that favour the majority classes, thereby reducing the accuracy and reliability of predictions for the minority classes.

Class Distribution After SMOTE (Right Panel): The graph on the right shows a uniform distribution of samples across all user IDs after applying SMOTE. This uniformity indicates that SMOTE successfully synthesized new data points for minority classes, balancing the dataset. A balanced dataset ensures that machine learning models are exposed to equal representation from all classes, improving their ability to classify both majority and minority classes effectively.

The feature importance bar chart identifies X-Coordinate as the most influential feature, followed by Pressure and Y-Coordinate, emphasizing the dominance of spatial attributes in distinguishing user behaviour. Moderate contributions from features like Area Covered and Stroke Length further support the significance of physical interaction attributes. However, features like Finger Orientation, Touch Duration, and Speed have minimal importance, suggesting limited relevance in this dataset or requiring further exploration. These findings highlight the critical role of spatial and intensity-based features, particularly X-Coordinate and Pressure, in user classification.

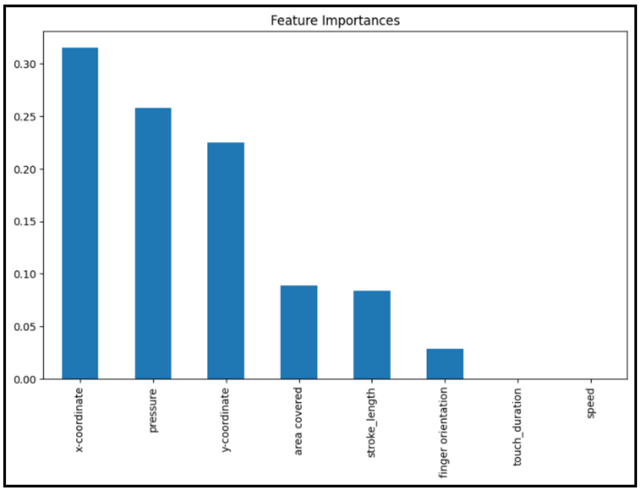

Figure 13: Identifying important feature in dataset

6.2 Results and Analysis

The Random Forest model, an ensemble-based learning approach, was designed to combine multiple decision trees to enhance prediction accuracy and reduce overfitting. Similarly, XGBoost, a gradient boosting framework, excels in identifying complex, non-linear patterns through iterative improvement of weak learners. Another ensemble-based model, AdaBoost, aimed to boost the performance of weak learners but with simpler implementations. In contrast, simpler models like Logistic Regression were tested to evaluate the dataset's linear separability, providing a baseline for comparison. The Decision Tree, while powerful as a single learner, served as a benchmark to understand how standalone models compare to ensemble methods. Gradient boosting frameworks such as LightGBM added a layer of efficiency and scalability to the experiments. Lastly, the Artificial Neural Network (ANN) leveraged its ability to capture non-linear relationships in the data through its hidden layers and activation functions.

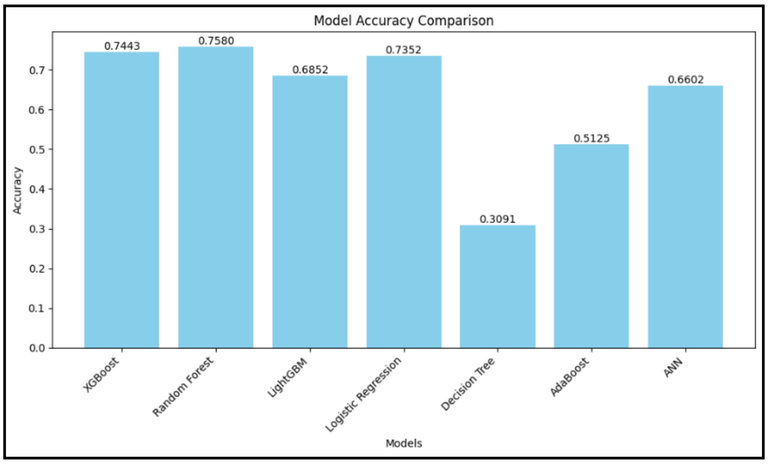

Figure 14: Performance Comparison of different machine learning models for the dataset 1

The evaluation revealed distinct performances across the models. The Random Forest emerged as the top-performing model, achieving an accuracy of 0.75, demonstrating its robust ensemble approach. The XGBoost model followed closely with an accuracy of 0.74, highlighting its ability to capture intricate patterns effectively. A surprising contender, Logistic Regression, achieved an accuracy of 0.73, suggesting the dataset may have substantial linear characteristics. The LightGBM model recorded an accuracy of 0.68, showcasing its gradient boosting capabilities but falling short of its competitors due to possible tuning limitations. The Artificial Neural Network (ANN) achieved a moderate accuracy of 0.66, indicating potential improvements through deeper architectures or better parameter optimization. On the other hand, AdaBoost performed moderately, achieving an accuracy of 0.51, while the Decision Tree struggled with an accuracy of 0.31, likely due to its inherent overfitting tendencies as a single learner. The Accuracy table of the dataset is illustrated below:

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.79	0.78	0.75	0.75
XGBoost	0.76	0.75	0.73	0.74
Logistic Regression	0.73	0.71	0.68	0.73
LightGBM	0.65	0.67	0.62	0.68
Artificial Neural Network	0.61	0.65	0.61	0.66

AdaBoost	0.99	0.99	0.99	0.51
Decision Tree	1.00	1.00	1.00	0.31

Table 1: Model comparison table of dataset 1

6.2.1 Touch Analytics Dataset

The dataset was analyzed using a trained machine learning model. A bar plot visualized the most significant contributors to user classification, revealing that features like x-coordinate, pressure, and y-coordinate played the most influential roles in distinguishing user behavior. The evaluation of machine learning models on the Touch Analytics dynamics dataset reveals distinct performance trends. Among the models tested, Random Forest achieved the highest accuracy of 0.79, demonstrating its capability to handle complex, non-linear patterns effectively. Its ensemble nature allows it to capture diverse feature interactions, making it a strong candidate for this classification task. Similarly, the Decision Tree, another tree-based model, performed well with an accuracy of 0.77. However, its standalone nature makes it more prone to over-fitting compared to Random Forest, which leverages multiple trees to enhance generalization.

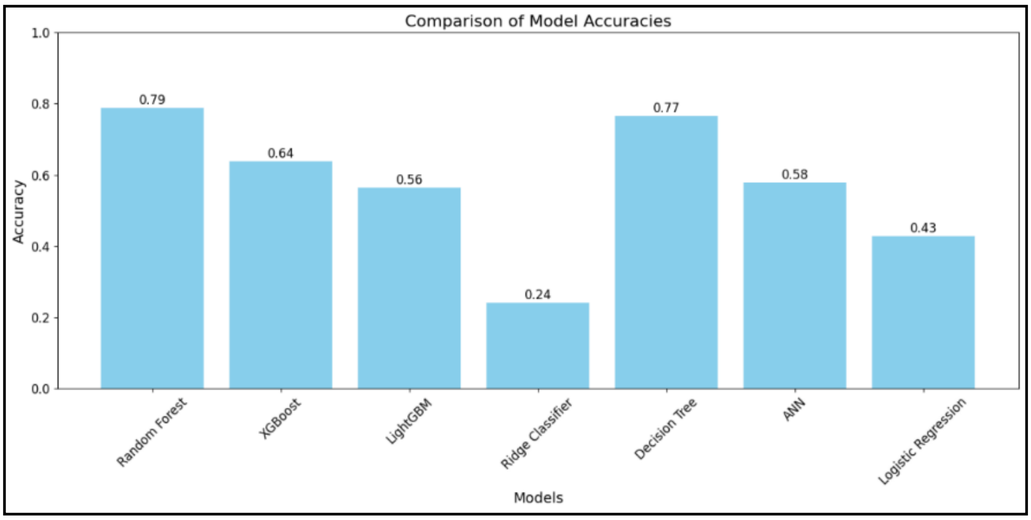

Figure 15: Performance Comparison of different machine learning models for the dataset 2

The XGBoost model recorded an accuracy of 0.64, reflecting its ability to model intricate data patterns. However, it underperformed compared to Random Forest, indicating that further optimization might be required. LightGBM, another gradient boosting framework, achieved an accuracy of 0.56, showcasing its efficiency but leaving room for improvement in feature interaction modeling. The Artificial Neural Network (ANN) demonstrated moderate performance with an accuracy of 0.58, suggesting potential for better results with deeper architectures or hyperparameter tuning. In contrast, simpler models like Logistic Regression and Ridge Classifier achieved accuracies of 0.43 and 0.24, respectively. These results highlight their limited ability to capture the non-linear nature of the dataset, emphasizing the importance of advanced models for this classification task. The results clearly underscore the strength of ensemble methods and non-linear models in addressing the complexities of behavioural biometric data. The Accuracy table of the dataset 2 is illustrated below:

Model	Precision	Recall	F1-Score	Accuracy
Decision Tree	0.766	0.76	0.766	0.77
Artificial Neural Network	0.561	0.579	0.559	0.58
LightGBM	0.563	0.564	0.557	0.56
XGBoost	0.626	0.638	0.623	0.64
Random Forest	0.793	0.788	0.790	0.79
Logistic Regression	0.372	0.429	0.362	0.43
Ridge Classification	0.13	0.24	0.13	0.24

Table 2: Model comparison table of dataset 2.

6.3 Discussion

The models in Table 2 demonstrate mixed performance compared to Table 1. While Random Forest and Decision Tree show improvements, LightGBM, Logistic Regression, XGBoost, and ANN exhibit declines. Ridge Classifier underperforms significantly in Table 2. Although, results from the past research papers remain the benchmark for high accuracy, but the accuracy tables from both the datasets highlights areas where further tuning and optimization can bring models closer to these benchmarks.

7 Conclusion and Future Work

In this research, I proposed using AI to create cybersecurity behaviour-based detection algorithms. My study shows that keystroke dynamics and social behavioural information merging can indeed increase detection capabilities and robustness. Touch analytics and keystroke dynamics train machine learning models, which are subsequently built from scratch. Necessary feature engineering and selection were carried out. The models were then tested, and the accuracy was compared. The random forest models yielded the highest accuracy. These results highlight the importance of advanced models for this classification task. The results clearly underscore the strength of ensemble methods and non-linear models in addressing the complexities of behavioural biometric data. To encapsulate, the answer to the research question is that from the dataset 1, the analysis underscores the dominance of ensemble-based models such as Random Forest and XGBoost in handling complex datasets. While simpler models like Logistic Regression offered competitive accuracy, standalone learners like Decision Tree highlighted their limitations, emphasizing the need for advanced frameworks in achieving optimal results. While from the dataset 2, the modelling is done taking into the account the comparison from the paper Momoh et al. (2023). Although the performance which they reported was 88.5% but when used for our particular dataset, the performance was 72.1% with the same hyper parameter settings.

However, the future work would consider the implementation issues of these applications in various domains, such as the limited-power IoT networks and essential infrastructures. New strategy called federated learning should be employed to protect user data while caring for the user’s trust. Real-time systems need to also incorporate machine learning models that improve with time as the more complicated cyber threats emerge while adopting NLP to interpret the attempts at phishing or scams. Another important contribution would be to extend the real-time behavioural analytics with educational tools for end-users, that would form a part of a new generation of hybrid frameworks. It was pointed out that these systems could be able to give prompt response such as during instance of phishing, or when a training module from an organization is incorporated. With such applications and assembling technical and human-oriented methodologies, cybersecurity strength grows noticeably.

References

Amjad, M.; Ahmad, I.; Ahmad, M.; Wróblewski, P.; Kami ´nski, P.; Amjad, U. Prediction of Pile Bearing Capacity Using XGBoost Algorithm: Modeling and Performance Evaluation. Appl. Sci. 2022, 12, 2126.

Azhar, M.B.M., Azlan, W.N.A.W.A., Mazri, W.N.A.W. and Radzi, S.M., 2023. Social Engineering and Cyber Threats: Exploring Techniques, Impacts and Strategies. International Journal of Accounting, Finance and Business, 8(50).

Bansal, P. and Ouda, A., 2024. Continuous Authentication in the Digital Age: An Analysis of Reinforcement Learning and Behavioral Biometrics. Computers, 13(4), p.103.

Bergmann, D. and Solheim, C., 2024. The Role of Social Media in Social Engineering Attacks: A Qualitative Study on Technical-, Individual-, and Organizational Measures to Mitigate Social Engineering Attacks in Social Media (Master's thesis, University of Agder).

Borowiec, Ł., Demidowski, K., Pecka, M. and Jonarska, A., 2023. The analysis of social engineering methods in attacks on authentication systems. Advances in Web Development Journal, 1, pp.83-106.

Budžys, A., Kurasova, O. and Medvedev, V., 2023, July. Behavioral Biometrics Authentication in Critical Infrastructure Using Siamese Neural Networks. In International Conference on Human-Computer Interaction (pp. 309-322). Cham: Springer Nature Switzerland.

Edwards, L., Zahid Iqbal, M. and Hassan, M., 2024. A multi-layered security model to counter social engineering attacks: a learning-based approach. International Cybersecurity Law Review, pp.1-24.

Fakhouri, H.N., Alhadidi, B., Omar, K., Makhadmeh, S.N., Hamad, F. and Halalsheh, N.Z., 2024, February. AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response. In 2024 2nd International Conference on Cyber Resilience (ICCR) (pp. 1-8). IEEE.

Guo, K.; Wan, X.; Liu, L.; Gao, Z.; Yang, M., 2021, Fault Diagnosis of Intelligent Production Line Based on Digital Twin and Improved Random Forest. Appl. Sci. 2021, 11, 7733.

Gupta, S., Maple, C., Crispo, B., Raja, K., Yautsiukhin, A. and Martinelli, F., 2023. A survey of human-computer interaction (HCI) & natural habits-based behavioural biometric modalities for user recognition schemes. Pattern Recognition, 139, p.109453.

Harilal, Athul & Toffalini, Flavio & Homoliak, Ivan & Castellanos, John & Guarnizo, Juan & Mondal, Soumik & Ochoa, Martín. (2018). The Wolf of SUTD (TWOS): A dataset of malicious insider threat behavior based on a gamified competition. Journal of Wireless Mobile Networks. 9. 10.22667/JOWUA.2018.03.31.054.

Khan, H.U., Malik, M.Z., Nazir, S. and Khan, F., 2023. Utilizing bio metric system for enhancing cyber security in banking sector: A systematic analysis. IEEE Access.

Khan, S., Devlen, C., Manno, M. and Hou, D., 2024. Mouse dynamics behavioral biometrics: A survey. ACM Computing Surveys, 56(6), pp.1-33.

Madavarapu, J.B., Mittal, M., Salagrama, S., Adnan, M.M., Rana, A. and Yadav, K., 2024, May. Behavioral Biometrics Authentication Systems: Leveraging Machine Learning for Enhanced Cybersecurity. In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) (pp. 1478-1483). IEEE.

Mahanta, K. and Maringanti, H.B., 2023. Social engineering attacks and countermeasures. In Perspectives on Ethical Hacking and Penetration Testing (pp. 307-337). IGI Global.
Mahfouz, A., Hamdy, A., Eldin, M.A. and Mahmoud, T.M., 2024. B2auth: A contextual fine-grained behavioral biometric authentication framework for real-world deployment. Pervasive and Mobile Computing, 99, p.101888.

Momoh, I., Adelaja, G. and Ejiwumi, G., 2023. Analysis of the Human Factor in Cybersecurity: Identifying and Preventing Social Engineering Attacks in Financial Institution.
Piugie, Y.B.W., 2023. Performance and security evaluation of behavioral biometric systems (Doctoral dissertation, Université de Caen Normandie).
Ramaraj, N., Murugan, G. and Regunathan, R., Securing Healthcare Data: A Behavioural Biometrics Approach using One-Class SVM.