Keystroke dynamics Based Technique to Enhance the Security in Smart Devices

Farman Pirzado[1] Shahzad Memon[2] Lachman Das Dhomeja[3] Awais Ahmed[4]

Abstract—Nowadays, smart devices have become a part of our lives, hold our data, and are used for sensitive transactions like internet banking, mobile banking, etc. Therefore, it is crucial to secure the data in these smart devices from theft or misplacement. The majority of the devices are secured with password/PINbased user authentication methods, which are already proved a less secure or easily guessable user authentication method. An alternative technique for securing smart devices is keystroke dynamics. Keystroke dynamics (KSD) is behavioral biometrics, which uses a natural typing pattern unique in every individual and difficult to fake or replicates that pattern. This paper proposes a user authentication model based on KSD as an additional security method for increasing the smart devices' security level. In order to analyze the proposed model, an android-based application has been implemented for collecting data from fake and genuine users. Six machine learning algorithms have been tested on the collected data set to study their suitability for use in the keystroke dynamics-based authentication model.
Index Terms—Keystroke dynamics; Smart Devices; user authentication.

I INTRODUCTION

Now a day's mobile phones are one of the most important things for people. Mobile phones are not only used for calling or sending text messages. They are also used in many confidential transactions such as (E-Banking, E-Commerce, and social networking). Therefore, it is becoming more important to secure mobile phones. As far as the security of smart devices is concerned nowadays, the majority of devices use integrally weak authentication mechanisms based on usually passwords and PINs. This method is based on the user name and password secrecy. The password usually consists of common words and phrases associated with a particular user. That is universally considered weak because it can be easily hacked by different password hacking methods like guessing, phishing, etc. [1] Another user authentication is a pattern lock, which is a graphical password. It includes 3 x 3 grids of small dots. On that small grid, the user is required to draw a graphical pattern with his finger for authentication, and it can also be easily broken by a computer attack named smudge attack. An increase in the deficits of password-based authentication still the majority of smart devices use weak authentication mechanisms based on PIN or password, which do not ensure an appropriate security level for access to the stored information and to the available services. It is important to implement a strong authentication system for smart devices. As compared to physiological biometric systems, behavioral biometrics systems are considered more secure and unique as it is not possible to copy the behavior of the user to the system. The keystroke dynamics is one of the types of behavioral biometric; it monitors the behavior of a user by analyzing the user behavior based

[1]*Mohammad Ali Jinnah University Karachi, Pakistan | farman.ali@jinnah.edu*
[2]*University of Sindh Jamshoro, Pakistan | Shahzad.memon@usindh.edu.pk*
[3]*University of Sindh Jamshoro, Pakistan | lachman@usindh.edu.pk*
[4]*Mohammad Ali Jinnah University Karachi, Pakistan | awais.ahmed@jinnah.edu*

on the typing patterns of the user. Keystroke dynamics can be used to improve the security level of pin-based user authentication in smart devices. Keystroke dynamics is the behavioral biometric, with dynamic keystroke user can be easily authenticated via his key typing unique feature such as key press time, the difference between two keys pressed and over all typing speed of the user. With the help of keystroke dynamics, we can strengthen the security of smart devices; even if a hacker hacks your password, he needs to know how to type that password [2]. However, KSD undergoes many implementation challenges such as low accuracy, low permanence, user situation, and emotion because of which, as per our literature survey, research on KSD is very challenging and is in its infancy.

2 BACKGROUND AND RELATED WORK

In the literature, Significant research efforts have been conducted over the years in an attempt to improve the quality of keystroke dynamics as an additional authentication method for smart devices.

Nowadays, smart devices are an important part of our daily life; they are not only used for general communication but also for private communication and important financial transactions. This makes security issue one of the important concerns in smart devices. There exist a number of user authentication methods, including Password/PIN, pattern lock,etc. All of these suffer from various drawbacks and limitations, as discussed before [3].

A biometric identification trait is unique for every single user, and so biometrics can provide secure means for user authentication in smart devices. The term biometric comes from two Greeks words 'bio' and 'metrics'; the former means' life,' and the latter means "measurement." Therefore, biometric authentication has to do with measuring and analyzing the biological characteristics of an individual. Biometrics is classified as being physiological biometrics and behavioral biometrics. The former involves physical characteristics of the user (Thumb, Iris, Retina, etc.) and the latter behavioral characteristic of the user (Keystroke dynamics, signature, etc.) [4], as shown in the figure 1 below

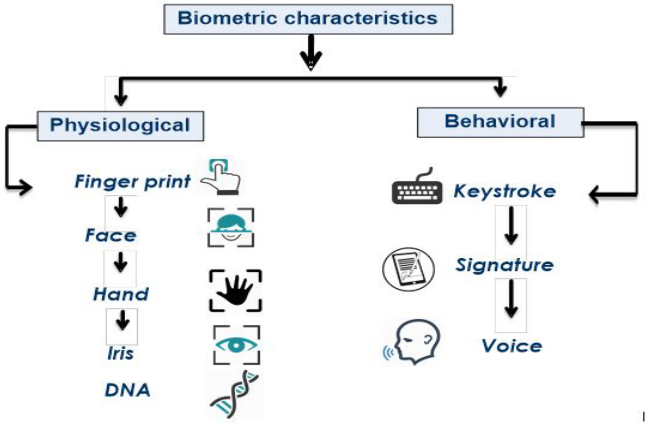

Figure 1: Categories of Biometrics

However, KSD undergoes many implementation challenges such as low accuracy, low permanence, user situation, and emotions, which are discussed below [5]

- ***Low accuracy***

One of the significant issues faced by keystroke applications in its implementation is low accuracy. This issue is fundamental stroke dynamics (KSD) authentication is caused by the large variation in typing style caused by many external factors such as injury, fatigue, or any other distraction. KSD undergoes many issues due to these reasons and challenges, but research has not been stopped; it is being carried out, and one can hope for improved and better results in the future.

- ***Lower permanence***

While researchers have put forward many methods, KSD suffers from lower permanency as compared to all other biometric systems. It is because of a human's typing pattern that may regularly change following the customization towards a password.

- ***User's situation & emotions***

Under such a different user situation like walking, driving, etc. can also cause apprehension in user "s major dynamicity in keystroke value. They were thus contributing to the implementation challenges of keystroke dynamics. Such situations and user "s happy or angry emotions will play a significant role in the typing behavior of a user. To solve the above problem and challenges in KSD, different models have been introduced in literature based on different methods such as machine learning classifiers, neural networks, and statistical. However, KSD is still in its early stages in mobile devices, and much research needs to be done to make it useful and accurate. In our proposed, model six different classification and regression algorithms of machine learning are applied to the data set of keystroke values, and six different results are generated and compared.

Dr.T.Pandikumar et al.] have proposed the model based on KSD; they have only used a random forest classifier algorithm to achieve the better performance of the proposed system. While in our proposed model, we have applied six different classifiers. To test the performance of these models on the data set we have collected from the users. To evaluate the performance of the system, the authors have used two parameters such as false acceptance rate (FAR) and false rejection rate (FRR), Whereas in our proposed work, four different performance parameters are used to evaluate the performance of the proposed model, including accuracy level and classification error.

In 2020 [6], to improve the authentication accuracy of Keystroke dynamics, researchers assess the feasibility of KSD based biometric verification from a model that is based on sequences of typed characteristics with a Gaussian mixture model (GMM).. At the same time [7] suggested KSD based authentications multi-factored with PIN-based authentication using the Novel feature-scoring method. Another article published in 2020 that focuses on user touch time and force features extracted from the piezoelectric force-touch panel of smart devices; in this works,

researchers used a Support vector machine, Artificial neural network, and Random forest in order to judge error rate in Keystroke dynamics using different classifiers. [8]. Apart from applying different techniques to judge the accuracy of the Keystroke dynamics, there is also a work that focuses on the variation of Keystroke value according to user position; considering it as a research problem, authors presented a three-step authentication model based on three user position, i.e., sitting, walking and relaxing in. [9]. Later [10], Researchers also focus on designing a data mining system by Appling the dimension reduction technique on KSD data and applied two data mining algorithms on data to find out the accuracy, i.e., K – nearest neighbor, Bay classifier, and Decision tree. Currently, a neural network is used as a good problem-solving approach in research; therefore, authors represent a novel analysis for the KSD authentication using timing and no timing feature using neuralnetworks. Researchers also focus on designing a data mining system by Appling the dimension reduction technique on KSD data and applied two data mining algorithms on data to find out the accuracy, i.e., K – nearest neighbor, Bay classifier, and Decision tree. [3]. In addition, Neural networks also help researchers to explore the effectiveness of employing KSD to differentiate between authentic and fake users of mobile using deep learning techniques based on conventional neural networks. [11]. Authors also focus on comprehensive analysis using KSD using Neural network; different neural network classifiers are used in this work in order to find different performance parameters like false acceptance rate, low error rate, and equal error rate [12].

Another research forwarded in [13] by [Dr T.Pandikumar, Abraheem fekde].In their proposed research, they used artificial neural networks for training and classification purposes. Working on the same, they developed an artificial keyboard for collecting the keystroke values from smart devices as well. Whereas it is our proposed research, we have developed an android application to collect the keystroke values from the users. In international journal of soft computing and engineering, M.Karnan and N.Krishnaraaj, 2012 [14], They brought another security model for smart devices. These both the researchers, for user authentication, introduced a hybrid model that is based on the fusion of six biometrics, including keystroke, fingerprint, and palm print. In this model, data from all six biometrics are captured in order to develop a template, which is later used to authenticate genuine users. In our proposed research, we introduced a hybrid model based on the fusion of two user authentication approaches, including password/PIN with keystroke dynamics. On the other hand, as in work, a shorter text has been used by the Authors. In this, only once a password, along with an extended length of text for continuous verification of the user using the vector machine (SVM), was used. In addition, the results of this showed that they could validate authentic users while rejecting imposters with a slight error rate [13]. The user's prediction was calculated by using "SVM." Based on this prediction, a number of sham and genuine users were selected. All 34 users who were selected in the experiment were identified using two-keystroke dynamics features flight and dwell time of 12 most common digraphs. The above researchers applied only one classification algorithm support vector machine, whereas in the proposed research, six different classification algorithms are applied, and the performance of the system is evaluated. Moreover, in our proposed work, for experimental setup we have collected 50 positive samples and 50 negative passwords in order to train classifiers. Apart from the difference in the number of users, they have developed a desktop application to collect keystroke values, while in our proposed work, we developed an android application for data collection. As far as the performance of the proposed model is concerned, a threshold value was set to detect the genuine and fake user, while we extracted

four different performance parameters to evaluate the performance of the proposed model.

3 PROPOSED MODEL

The proposed model show in figure 2 indicates that the user will enter his/her user name and password in a smart device, the sensors are monitored, and keystroke features are collected in the data acquisition phase. After collection of keystroke features such as key holders, time features are normalized, and a dataset is designed for the classification model.

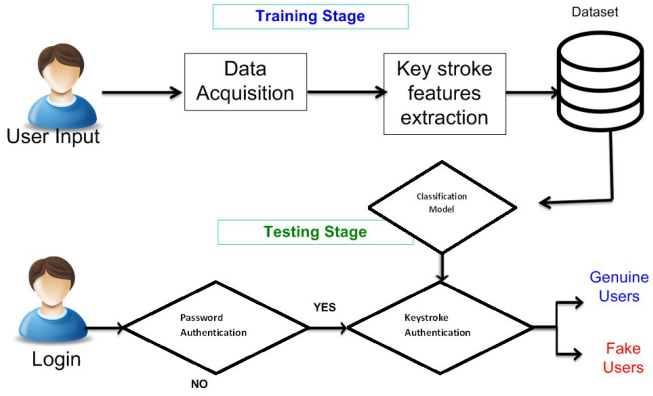

Figure 2: Proposed System

The classification model will accept the dataset as input for authentication, and it will apply a classification algorithm to the given data set, which will provide the two results.

According to the classification algorithm applied in the classification model on a given dataset, if keystroke values don't match with the features stored in the classification model, the user is rejected if the keystroke values match with the data set in the classification model, the user is authenticated.

• ***User Registration:*** The registration phase asks a user to type his/her user name and password. Simply, in this, a user requires only to log in to the system, then the android application will automatically extract the keystroke features such as (Key hold time (KHD), Key up downtime (KUD), and Total typing speed (TTS)of the user. Dealing with hardware keyboards is very different from dealing with touch screens. There are many features of keystroke dynamic in our proposed system following features have been extracted:

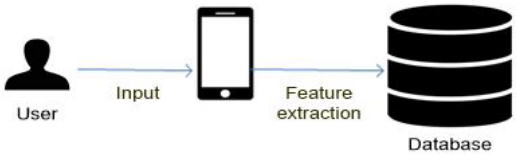

Figure 3: User Registration

- Key hold time (KHT).Time between key down and key up for single character.
- Key Up down Time (KUD).Time between key up of a character and key-down of the next character.
- Total typing speed (TTS).Total time to press a single character.

• ***Data acquisition:*** For data acquisition, an android application is developed using android development toll Eclipse, which runs as a stand-alone application on an android smart device, asking users to type their password. Here touch sensor of the smart device is used to collect keystroke features of users.

4 PERFORMANCE PARAMETER

The performance of any biometric system, including KSD, is measured in a false acceptance rate (FAR) and false rejection rate. The false acceptance rate (FAR) is defined as "the percentage of invalid inputs which are incorrectly accepted," while false rejection rate (FAR) is defined "as a percentage of valid inputs which are incorrectly rejected." [15]. The above algorithms generate different FAR and FRR, which determine the accuracy and performance of the proposed model. Since different values in the operating threshold may result in variation in values of FRR and FAR, the receiver operating system characteristic ROC curve will show the trades off between the FAR and FRR. In addition to FAR and FRR, two other performance parameters, i.e., accuracy level and classification error, are generated by the algorithms being applied.

5 DATA ANALYSIS

Over the last six decades, many classification methods have been applied in keystroke dynamics study. Keystroke dynamics can be perceived as a pattern recognition problem, and the most commonly deployed methods can be broadly categorized as statistical and machine learning approaches [12].

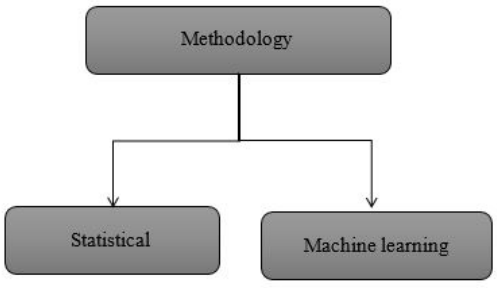

Figure 4: Data Analysis Approaches

Statistical methods are commonly used for data analysis based on different statistical parameters. This approach elaborates in carrying out a study including planning, designing, and analyzing, which results in expressive understanding and reporting of research conclusions. There are many generic statistical measures available such as mean, median, and standard deviation. Machine learning is the branch of artificial intelligence, which is based on the idea that a system can learn from data, can identify patterns, and can make decisions and predictions. MATLAB

also provides immediate access to prebuilt functions, toolboxes, and special applications for classification, regression, and clustering. It is an acronym for matrix laboratory used to solve different mathematical and machine learning problems. It is used as a data analysis tool in research worldwide. For a large amount of data, machine learning is used to find patterns from data and build models. That model can predict future outcomes based on data sets; MATLAB also provides immediate access to prebuilt functions, toolboxes, and special applications for classification, regression, and clustering. Data is acquired using a self-designed android application, which will ask users to type their password. Data has been collected from users 50 positive samples and 50 negative passwords to train classifiers using a self-designed android application. The data set of genuine and fake users has been developed in Ms. Excel according to the format that classification algorithms support for further data analysis in MATLAB.

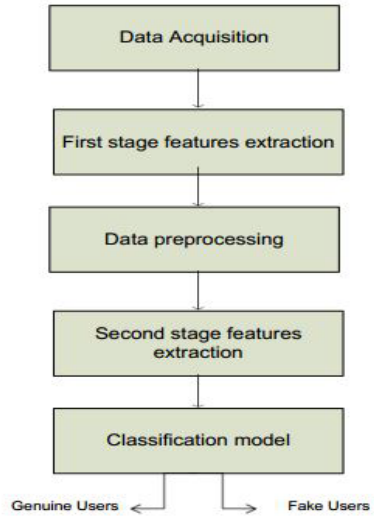

Figure 5: Data Analysis Process

As the range of values from the first stage of feature extraction is different, therefore; normalization will be applied to transform all the values in the range of o – 1.

- The first stage features include the numeric value of keystroke features such as key hold time, key up-down time, and full typing speed.
- MATLAB converts these first stage features into second stage features, such as statistical, frequency domain, and non-linear features, to accurately predict results.
- A statistical feature includes the first difference of mean values, Standard deviation, the first difference of standard deviation values, mode, median, root mean square, etc.
- Frequency domain features include spectral density, Fourier coefficients, magnitude, and peak value.
- Non-linear features include point-care geometry features, which will be extracted after transforming the first stage features into given vectors.
- Based on the second stage features, the feature space will be constructed, and data set will be labeled with the ground truth such as the genuine user (1) or fake user (0).
- The system will then be trained using different classification algorithms, such as decision tree, linear discriminant functions, logistic regression, support vector machine, Knearest neighbor, random forest, and ensemble methods.

There are many machine-learning algorithms used to identify and classify patterns and make correct decisions based on data provided. The classification algorithms, which are used in this research, are described below in Table I.

Table 1: Machine-learning algorithms [16] & [17]

S.No.	Classification Algorithm	Description
1	Support Vector Machine (SVM)	A discriminative classifier formally defined by a separating hyper lane.
2	Random Forests	This algorithm creates forests with number of trees. The more the trees in forests the high the accuracy.
3	Decision tree	It creates a training model which is used to predict the values from a given data set.
4	K-nearest	It classifies a new class based on similarity measures. Used for statistical and pattern recognition.
5	Logistic Regression	Used to predict a binary outcome 0/1 from a given set of independent values.
6	Adaptive Boosting	Also known as AdaBoost used for classification or regression. Often sensitive to noisy data outliers.

6 RESULTS

In this section all the results are discussed and presented in graph. Total number of passwords in data set is 10 and for each password we have collected 50 positive samples and 50 negative passwords in order to train classifiers. As a result the data set contains 1000 of 500 of positive and 500 of negative.

A Support Vector Machine

The first classification algorithm that is applied to the data set is the support vector machine. As shown in the graph, SVM generated 89% accuracy level when applied on the collected data set, whereas the true positive rate is 86% and the false positive rate is 7%.

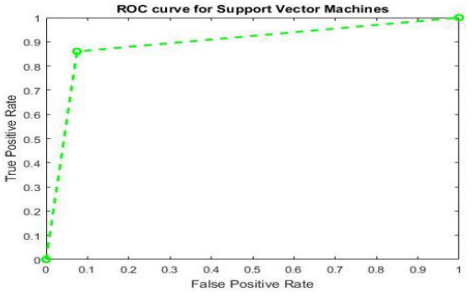

Figure 6: Support Vector Machine

B Random Forest

As shown in graph in figure 4.3, Random Forests algorithm generated 97% accuracy level when applied on the collected dataset, whereas the true positive rate is 98% and false positive rate is 3%.

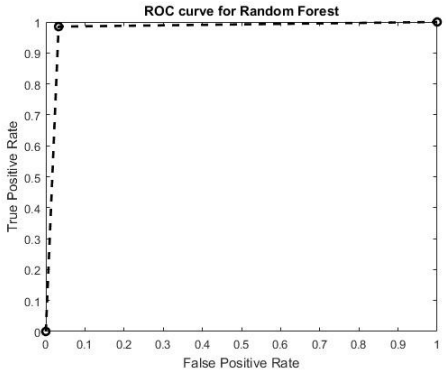

Figure 7: Random Forest

C Decision tree

As shown in figure 4.4, decision tree generated 97% accuracy level when applied on collected data set, where as true positive rate is 97% and false positive rate is 3% only.

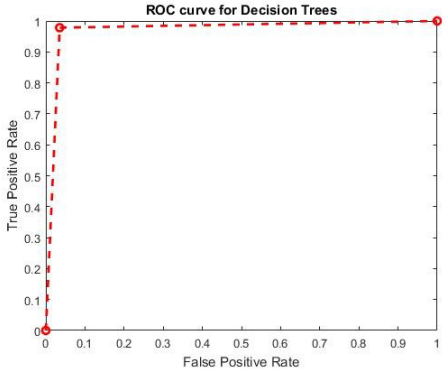

Figure 8: Decision tree

D Adaptive boosting (Ad boost)

Adaptive boosting is also known as adaboost: used for both classification and regression. After applying this algorithm on dataset 77% of accuracy level is accomplished with 2% of classification error and 16% of false positive rate and 71% of true positive rate as show in the graph below.

E k-Nearest neighbor

This algorithm classifies a new class based on similarity measures is used for statistical and pattern recognition. By applying K- nearest algorithm the accuracy level that is achieved is 93% with classification error of 23%. This algorithm generates the true positive rate of 88% and 1% of false positive rate.

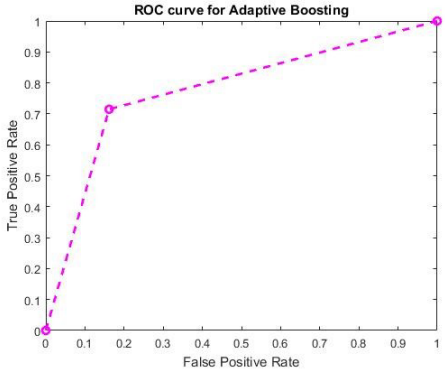

Figure 9: Adaptive boosting (Ad boost)

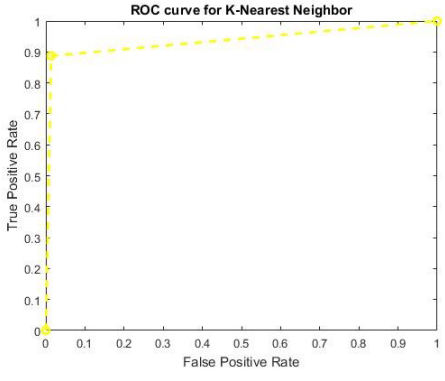

Figure 10: k-Nearest neighbor

F Logistic regression

Results of this algorithm are shown in the graph given below. It generates the accuracy level of 66% with classification error rate of 13%. It gives the 35% of false positive rate and 69% of true positive rate.

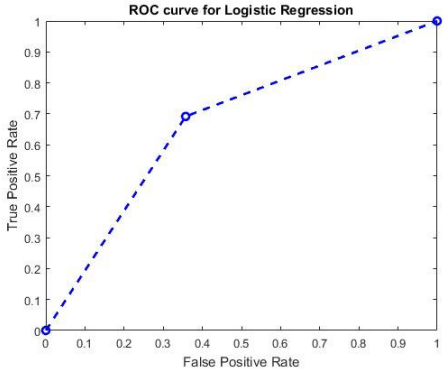

Figure 11: Logistic regression

Six different classification algorithms, when applied on collected data set, produced the six different results as shown in the table II.

The proposed work is explored to design and develop a model based on Keystroke dynamics for user authentication in smart devices. As shown below in table II, the two algorithms Decision Tree and Decision tree provide maximum accuracy level of 97% and SVM provides 80% of accuracy level. Classification error is the proportion of classification a classifier gets wrong therefore if classification rate decrease the false acceptance rate also decreases and performance of a system increases.

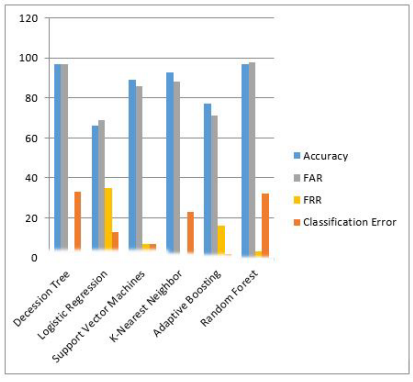

Figure 12: Comparative Analysis S.No Algorithm Accuracy False

Table 2: Results

S.No	Algorithm	Accuracy	False Positive Rate	True Positive Rate	Classification error
1	DT	97%	3%	97%	33%
2	LR	66%	35%	69%	13%
3	SVM	89%	7%	86%	7%
4	K-NN	93%	1%	88%	23%
5	AB	77%	16%	71%	2%
6	RF	97%	3%	98%	32%

The results suggest that using machine-learning algorithms can play a vital role in order to implement keystroke dynamics with traditional PIN based authentication in smart devices.

7 FINDINGS

In order to find out the performance of the proposed model, different classification machine learning algorithms were applied to the collected data set. The findings are mentioned below:

- Results suggest that the use of machine learning algorithms in keystroke dynamics based authentication models is a promising technique as they provide a good accuracy level as shown in results that two algorithms, Decision Tree and Random Forests, provide a maximum accuracy level of 97% and has a minimum false positive rate of 3%.
- Another advantage of implementing KSD in the smart device is that it is easily implementable since it can be implemented at the software level without requiring any additional hardware and hence a cost-effective method for user authentication.
- KSD can prove itself a good additional user authentication method along with password in smart devices.

8 CONCLUSION

The main objective of this research is to propose keystroke dynamics- based technique to enhance the security in smart de-vices. For achieving our goals, an android application has been developed, which collects keystroke features of authorized and unauthorized users. After collecting data (from genuine and fake users), the data set is developed for further analysis and results. Later on, different classification algorithms of machine learning are applied to data set using the tool of MATLAB. Results show that implementing machine learning algorithms provide acceptable levels in performance measures, as a good factor of authentication. Six different classification algorithms were applied to the same data set using MATLAB, and six different results were generated. The results show that implementing KSD using machine learning algorithms as an additional security method, along with a password, can enhance the security of smart devices. Since the proposed model uses keystrokes dynamics along with the password and that keystroke values are unique for individual users, it provides a strong security model. In case of the password being hacked, the hacker still has to know the typing behavior. Moreover, unlike other authentication methods (fingerprints, face recognition, etc.) that require dedicated hardware for implementation, keystroke dynamics are easily implementable since they can be implemented at the software level. Therefore, it has proven itself capable of providing additional security with other authentication methods such as user name and password. In this research, we have proposed and designed a user authentication model for smart devices based on keystroke dynamics and username and password.

9 FUTURE WORK

As the research in the field of keystroke dynamics suggests that it is a promising technique to be used as an additional security method with other authentication mechanisms such as user name, etc., it is still in its infancy and research phase. The proposed keystroke-based user authentication model can further be investigated with respect to the following directions:

- Currently, four keystroke features (i.e., key hold time, key up downtime, key down-down time, and total typing speed) have been used for user authentication in the proposed model. However, it may be interesting to investigate the accuracy of the proposed model by exploiting more keystroke features such as finger pressure, latency, etc.

- Neural network-based techniques and other data analysis techniques may be applied for further investigation of the proposed model.
- The proposed model may be tested under other operating systems such as AppleIOS, Blackberry, and Symbian, to see how the proposed model performs on these platforms.

REFERENCES

[1] M Karnan and N Krishnaraj. A model to secure mobile devices using keystroke dynamics through soft computing techniques. International Journal of Soft Computing and Engineering (IJSCE) ISSN, pages 2231–2307, 2012.

[2] Mohd Anwar and Ashiq Imran. A comparative study of graphical and alphanumeric passwords for mobile device authentication. In MAICS, pages 13–18, 2015.

[3] Asma Salem, Ahmad Sharieh, Azzam Sleit, and Riad Jabri. Enhanced authentication system performance based on keystroke dynamics using classification algorithms. KSII Transactions on Internet & Information Systems, 13(8), 2019.

[4] Anil K Jain, Karthik Nandakumar, and Abhishek Nagar. Biometric template security. EURASIP Journal on advances in signal processing, 2008:1–17, 2008.

[5] Yu Zhong and Yunbin Deng. A survey on keystroke dynamics biometrics: approaches, advances, and evaluations. Recent Advances in User Authentication Using Keystroke Dynamics Biometrics, pages 1–22, 2015.

[6] Himanka Kalita, Emanuele Maiorana, and Patrizio Campisi. Keystroke dynamics for biometric recognition in handheld devices. In 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), pages 410–416. IEEE, 2020.

[7] Dong In Kim, Shincheol Lee, and Ji Sun Shin. A new feature scoring method in keystroke dynamics-based user authentications. IEEE Access, 8:27901–27914, 2020.

[8] Anbiao Huang, Shuo Gao, Junliang Chen, Lijun Xu, and Arokia Nathan. High security user authentication enabled by piezoelectric keystroke dynamics and machine learning. IEEE Sensors Journal, 2020.

[9] Baljit Singh Saini, Parminder Singh, Anand Nayyar, Navdeep Kaur, Kamaljit Singh Bhatia, Shaker El-Sappagh, and Jong-Wan Hu. A three-step authentication model for mobile phone user using keystroke dynamics. IEEE Access, 8:125909–125922, 2020.

[10] Shri Kant, Alok Katiyar, and Shubhii Shuklla. Smart mobile device authentication using keystroke dynamics based behavior classification. [11] Emanuele Maiorana, Himanka Kalita, and Patrizio Campisi. Deepkey: Keystroke dynamics and cnn for biometric recognition on mobile devices. In 2019 8th European Workshop on Visual Information Processing (EUVIP), pages 181–186. IEEE, 2019.

[12] Asma Salem and Mohammad S Obaidat. A novel security scheme for behavioral authentication systems based on keystroke dynamics. Security and Privacy, 2(2):e64, 2019.

[13] Hayreddin C¸ eker and Shambhu Upadhyaya. User authentication with keystroke dynamics in long-text data. In 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), pages 1–6. IEEE, 2016.

[14] Ricardo N Rodrigues, Glauco FG Yared, Carlos R do N Costa, Joao BT ˜ Yabu-Uti, Fabio Violaro, and Lee Luan Ling. Biometric access control´ through numerical keyboards based on keystroke dynamics. In International Conference on Biometrics, pages 640–646. Springer, 2006.

[15] Naveed Riaz, Ayesha Riaz, and Sajid Ali Khan. Biometric template security: an overview. Sensor Review, 2018.

[16] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. New advances in machine learning, 3:19–48, 2010.

[17] Ayon Dey. Machine learning algorithms: a review. International Journal of Computer Science and Information Technologies, 7(3):1174–1179, 2016.

[18] T Pandikumar, Abraham Fekede, and Capt Zinabu Haile. Enhancing performance and usability of keystroke dynamics authentication on mobile touchscreen devices using features extraction scheme. International Journal of Engineering Science, 13415, 2017.

[19] Amund Tveit, Magnus Lie Hetland, and Haavard Engum. Incremental ˚ and decremental proximal support vector classification using decay coefficients. In International Conference on Data Warehousing and Knowledge Discovery, pages 422–429. Springer, 2003