Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd

ISSN,e-ISSN: 2548-1304
Volume: IDAP-2023, pp:143-150, 2023
https://doi.org/10.53070/bbd.1345519
Research Paper

Keystroke Biometric Data for Identity Verification: Performance Analysis of Machine Learning Algorithms

Erhan YILMAZ*[1], Ozgu CAN[2]

[1,2]Department of Computer Engineering, Ege University, Izmir, Türkiye

(erhan.yilmaz@itu.edu.tr, ozgu.can@ege.edu.tr)

Received:Aug.18,2023 Accepted:Aug.21,2023 Published:Oct.18,2023

Abstract— Cyber-attacks are on the rise in today's environment, where traditional security measures are ineffective. As a result, the adoption of cutting-edge tools such as artificial intelligence technology is critical in the fight against cyber threats. User behaviors, such as keyboard dynamics, provide potential data that can be used for protection against cyber-attacks. Keystroke dynamics is one of the fastest and most cost-effective methods that can be used to detect user behaviors, as it can be captured using standard user keyboards. The analysis of this data and protection against cyber-attacks is made possible through machine learning algorithms. Based on keyboard dynamics, this study analyzes the performance of k-Nearest Neighbor (k-NN), Support Vector Machines (SVM), Random Forest (RF), and Neural Network (NN) methods for user behavior analysis and anomaly detection. The findings shed light on the significance of artificial intelligence in cyber security by examining the accomplishments of several machine learning algorithms. The study's findings may serve as a foundation for future research and novel solutions in the realm of cyber security.

Keywords : *Keystroke analysis, user behavior analytics, anomaly detection, machine learning*

1. Introduction

With the rapid advancement of technology, computers, and internet usage have gained significant importance in our daily lives. The increasing use of computers has also led to a rise in cybersecurity concerns. Cyber-attacks can cause both financial and intangible damage, posing a threat to both individual users and organizations (Anderson & Moore, 2006). The theft of sensitive data and service disruptions resulting from cyber-attacks can damage the reputation and customer trust of organizations and further lead to problems in terms of legal obligations (Gordon, Loeb & Sohail, 2003). However, the ever evolving and complex nature of cyber-attacks, the data inadequacy and lack of standardization in the field of cybersecurity, and the effect of human factor, cause various challenges in cybersecurity (Li & Liu, 2021). The difficulties encountered in preventing cyber-attacks have led to the adoption of new approaches in the field of cybersecurity. Machine learning techniques have proven to be one of the most effective tools for analyzing and learning from large amounts of data to detect and mitigate cyber threats (Buczak & Guven, 2015).

Machine learning is a branch of artificial intelligence that aims to improve the performance of a system by learning from data without explicit programming. Machine learning techniques can be used to analyze large-scale and complex data generated by user behavior, extracting valuable patterns and insights. By combining machine learning techniques with user behavior analysis, various cybersecurity tasks such as user authentication and anomaly detection can be achieved (Sarker, Kayes, Badsha, Alqahtani, Watters & Ng, 2020). The analysis of keystroke dynamics has become an important research topic in the fields of user authentication and behavior analysis (Revett, 2009; Banerjee & Woodard, 2012). Keystroke dynamics analysis can be employed to identify the unique characteristics and habits of users in keyboard usage, thus providing valuable information for security and user experience (Joyce & Gupta, 1990; Gunetti & Picardi, 2005). In this method, the user's keystroke dynamics are analyzed to perform the authentication process. Keystroke dynamics involve examining factors such as pressure applied while pressing keys, keystroke speed, and inter-keystroke intervals. The data obtained from keystroke dynamics can be utilized to enhance the security of passwords and other authentication methods (Bergadano, Gunetti & Picardi, 2002).

Anomaly detection algorithms are utilized to identify both insider and external attackers through the analysis of keystroke dynamics. Anomaly detection refers to the technique used to identify behaviors or events that deviate from normal processes (Chandola, Banerjee & Kumar, 2009). By utilizing data obtained from keystroke dynamics, anomaly detection can be performed to enhance user security without disrupting the user experience, enabling the detection of potential harmful or suspicious actions (Ahmed & Traore, 2005).

This study aims to identify users based on their keyboard usage habits. For this purpose, the keystroke dynamics comparison data set was used to measure the extent to which machine learning algorithms successfully identify users. The study intends to compare the utilization and performance of machine learning algorithms in user behavior analysis and anomaly detection using keystroke dynamics data.

The organization of the study is as follows: Section 2 explores the keystroke dynamics and provides the recent literature, Section 3 presents the machine learning algorithms used for user behavior analysis and anomaly detection, Section 4 discusses the experimental results, and finally, Section 5 and 6 conclude the study and outlines the future directions.

2. Keystroke Dynamics

The trace left by an individual while typing a text using a keyboard can be utilized to analyze the user's typing habits. Among the user's writing habits are factors such as different typing speeds, keystroke durations, and consecutive keystroke orders when different users press the same letter or number. Dynamic keyboard keystroke pattern analysis determines unique keystroke patterns that emerge when a person types each letter or number by analyzing every aspect of their keyboard usage. The analysis of dynamic keystroke patterns on a keyboard enables the identification of distinctive patterns that manifest when an individual inputs each letter or numeral. This examination involves a comprehensive assessment of all aspects pertaining to the usage of their keyboard (Porwik, Doroz & Wesolowski, 2021).

Numerous studies focus on authentication through keystroke dynamics. The study conducted by Ivannikova, David, and Hämäläinen (2017) centered around the construction of an authentication system, which relied upon the analysis of keystroke dynamics originating from the innate writing patterns exhibited by individuals. The study utilized data obtained from real users and employed the machine learning algorithm for identity verification.

The feasibility of password authentication through keystroke dynamics, as an alternative to inconvenient methods such as fingerprint scanning, retinal scanning, and voice recognition, is explored by Muliono, Ham, and Darmawan (2018) in their research. Consequently, the deep learning methods resulted in an accuracy rate of 92.6%.

The potential of utilizing keystroke dynamics for identity verification is investigated by Aversano, Bernardi, Cimitile, and Pecori (2021). Also, a cost-effective approach is proposed. The proposed approach eliminates the need for additional sensors or specialized equipment, unlike other biometric-based authentication methods. The study achieved an accuracy rate of 99.7%.

In the field of gathering keyboard keystroke dynamics from users, an agent has been developed by Akşit, Aydın, and Zaim (2022). Besides, the utilization of artificial intelligence techniques and deep learning algorithms for authenticating user identities is explored. The results of the studies demonstrate the potential of utilizing keyboard keystroke dynamics as a secondary authentication method alongside usernames and passwords, enabling the detection of malicious attackers compromising computer systems.

The identification of users through the utilization of keystroke dynamics, thereby eliminating the need for an external authentication device, has been investigated in a recent study by Kar, Bamotra, Duvvuri, and Mohanan (2023). For this purpose, artificial neural networks and convolutional neural networks algorithms are used and an accuracy rate of 95.05% is achieved. Also, the study proposes that keystroke dynamics could serve as a viable alternative to two-factor authentication devices.

3. Material and Methods

The study utilized the "CMU Keystroke Dynamics-Benchmark Data Set" created by Killourhy and Maxion (2009). The "CMU Keystroke Dynamics-Benchmark Data Set" is a valuable resource widely employed for investigating and analyzing keyboard dynamics. This dataset is used in studies related to the timing and characteristics of user actions performed using a keyboard.

The content of the dataset consists of over 2000 keyboard samples collected from 51 different users. Each user was asked to type a 10-character password (.tie5Roanl), which included letters, numbers, and special characters, 400 times. The keystroke dynamics were recorded while users typed the password. Each data instance in the dataset includes information such as the duration of key press for each key pressed by a user while typing the password, the time interval between pressing two keys, and the duration between releasing one key and pressing the next key. The information that can be derived from keyboard keystrokes is illustrated in Figure 1.

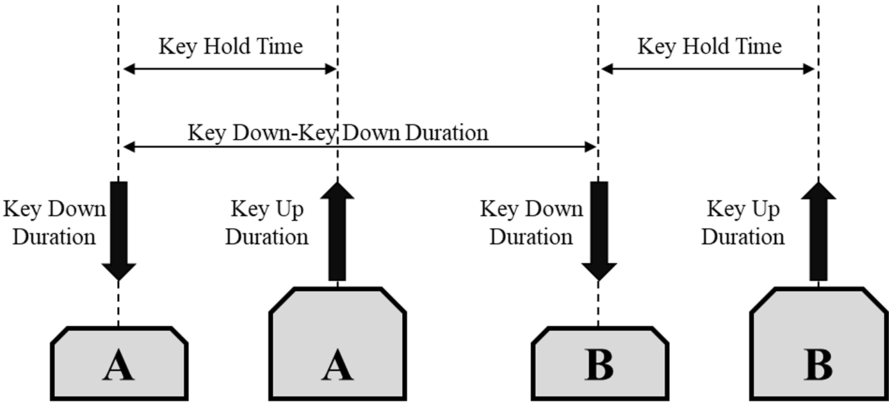

Figure 1. Information obtained through the recording and analysis of individual keyboard inputs

The obtained durations illustrate the user's typing speed and timing of keyboard interactions. The dataset reflects users' unique typing habits and behaviors. Figure 2 presents the graph depicting the key presses and inter-key transition durations for the s002 user during a single password entry within the dataset.

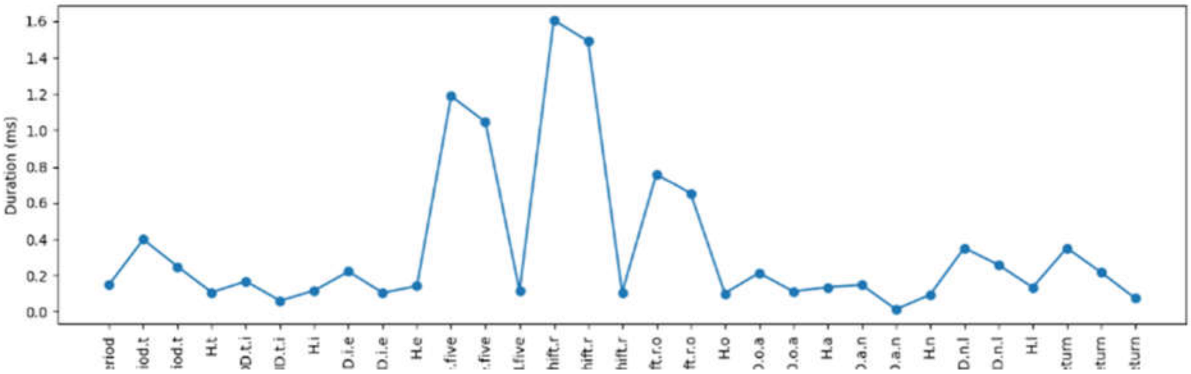

Figure 2. Keystrokes and inter-key transition times of user s002 during a single session

The comparison of the durations of pressing the letter "a" in the password entries of users s002 and s003 in the dataset is illustrated in Figure 3.

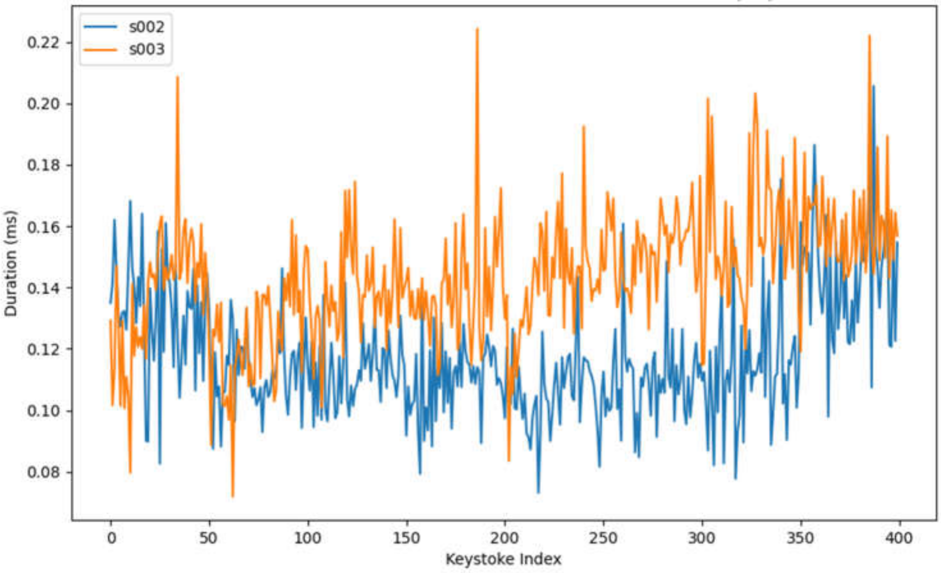

Figure 3. Comparison of the key down duration of the 'a' key for users s002 and s003

In this study, a series of classifiers are trained using keystroke biometrics to enhance user authentication performance. In the data preprocessing stage, the features in the dataset were scaled using a standard scaling method, and then the labels are encoded using label encoding. The aim of this process is to improve data processing and classifier performance.

After these processes, the dataset is split into training and testing data. The training set is used for the classifiers to learn user writing behavior. The testing data, on the other hand, is used to evaluate the performance of the trained classifiers. The classifiers used in the study are k-Nearest Neighbor (k-NN), Support Vector Machines (SVM), Random Forest (RF), and Neural Network (NN).

The k-NN algorithm determines the nearest "k" neighbors surrounding an instance and predicts the majority class. The fundamental principle of the k-NN algorithm is based on the assumption that similar examples tend to belong to the same class (Cover & Hart, 1967). SVM is a classification algorithm used to find the best separating hyperplane between different classes. The SVM algorithm enables to find the optimal separation hyperplane and uses the support vectors on this hyperplane to separate examples into different classes (Cortes & Vapnik, 1995). RF is a learning algorithm formed by combining multiple decision trees. Each tree is trained independently, and the classification results are obtained through voting or averaging (Breiman, 2001). NN is an artificial intelligence model inspired by biological neural systems. This algorithm utilizes a network structure where artificial neurons come together to perform complex computations. NNs generate outputs by multiplying input data with weights and using activation functions (LeCun, Bengio & Hinton, 2015).

The performance of the trained classifiers is evaluated using accuracy rate on the test set data. This evaluation is conducted to determine how successful the classifiers were in the process of user authentication. High accuracy rate indicates that the classifiers can successfully verify users, while low accuracy rate may indicate the need for improvement or the use of a different method for the classifiers.

This study is an important step in increasing the usability of keystroke biometrics in the field of user authentication. The obtained results demonstrate the potential of keystroke biometrics for conducting user authentication processes securely and effectively.

4. Results

In this study, four different classification algorithms are applied to the keystroke dataset: k- Nearest Neighbors (k-NN), Support Vector Machines (SVMs), Random Forest (RF), and Neural Networks (NN). The main goal of the study is performing user authentication by using keystroke data. The experiments are conducted using Python programming language and Visual Studio Code software.

The dataset consists of keystroke data from various users. The features of the data are determined as all columns except for user identity, session index, and repetition count. The developed Python code first underwent preprocessing by applying the standard scaling method to the data. Subsequently, the labels were encoded using Label Encoder. Further, the data is divided into training and test sets, with the test set size set to 20% and a randomness degree of 42. Definitions are made for each of the k-NN, SVM, RF, and NN algorithms to be used in classification, and the necessary parameters are adjusted. Specifically, for the neural network (NN) model, the parameters were tuned using the Adam optimization algorithm. The Adam optimizer was chosen due to its effectiveness in adapting learning rates for each parameter during training, thus facilitating faster convergence and potentially leading to improved classification performance. The optimization process aimed to find the optimal combination of parameters that would yield the highest accuracy and predictive capability for each classifier. Each classifier algorithm is trained using the training dataset. The trained classifiers are then used to make predictions using the test data. The accuracy and ROC curve are calculated to determine how accurate and inaccurate the trained models' predictions were for each classifier algorithm. The accuracy scores of each classifier are compared, and the results are presented. Figure 4 shows the workflow diagram for the performed operations.

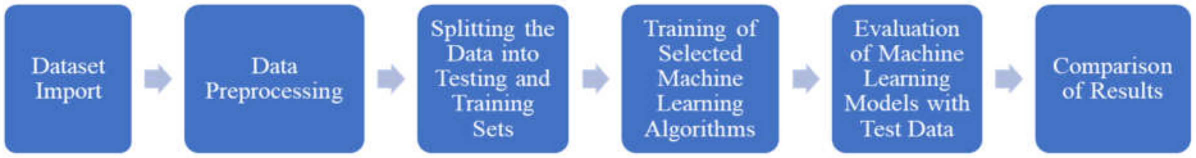
Dataset Import
Data Preprocessing
Splitting the Data into Testing and Training Sets
Training of Selected Machine Learning Algorithms
Evaluation of Machine Learning Models with Test Data
Comparison of Results

Figure 4. Data processing flowchart

According to the experimental results, the highest accuracy rate of 93.31% was achieved by the RF algorithm. The k-NN algorithm has an accuracy rate of 84.51%, the SVM algorithm has an accuracy rate of 88.14%, and the

NN algorithm has an accuracy rate of 88.28%. The results demonstrate that the RF algorithm outperforms the other algorithms in terms of accuracy. A comparison of the accuracy data obtained from the experiments is presented in Table-1.

Table 1. Accuracy Results

Machine Learning Algorithm	Accuracy
RF	93.31%
k-NN	84.51%
SVM	88.14%
NN	88.28%

The experimental results are evaluated using the ROC curve, which is a powerful tool for assessing the performance of the classification models employed. Figure 5 illustrates the ROC curve, showcasing the relationship between the model's sensitivity and false alarm rate. The logarithmic scale applied to the x-axis in Figure 6 enhances the visualization, particularly when the False Positive Rate (FPR) spans a wide range of values. This logarithmic representation allows for a clearer examination of the classifiers' performance, especially at lower FPR levels, which are crucial in real-world applications. The ROC curve provides valuable insights into the models' success and aids in evaluating the reliability of the presented results, highlighting the trade-off between sensitivity and specificity for each classifier. The ROC curve analysis aligns with the accuracy rates obtained, supporting the finding that the random forest classifier outperforms the other algorithms. By examining the area under the ROC curve (AUC), the overall discriminatory power of the models can be measured. A higher AUC indicates better performance, with a value of 1 representing a perfect classifier. The results suggest that the random forest classifier provides an effective balance between sensitivity and specificity, making it a robust choice for user authentication based on keystroke dynamics.

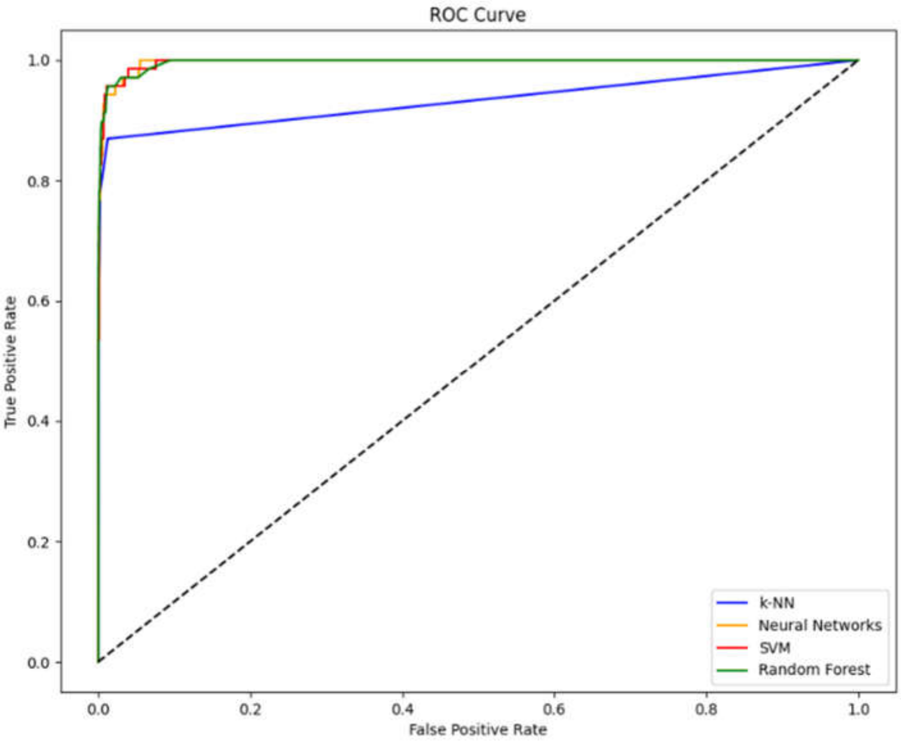

Figure 5. ROC Curve Graph of Machine Learning Methods Used in the Study

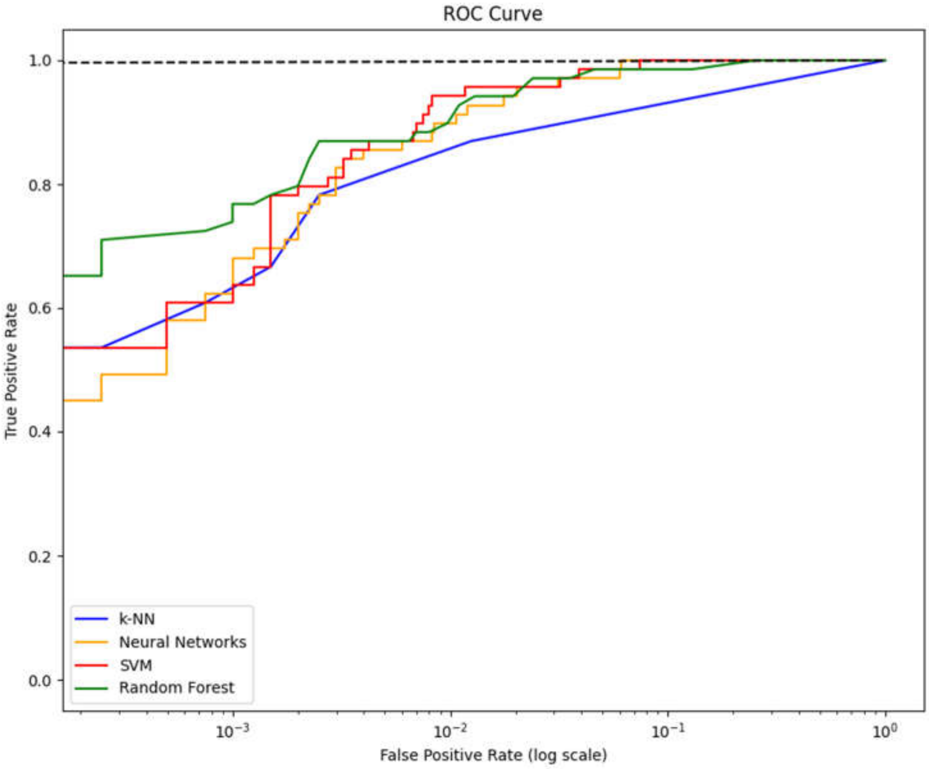

Figure 6. ROC Curve Graph of Machine Learning Methods Used in the Study (Log Scale)

5. Discussion

The use of keystroke biometrics for multifactor authentication purposes has been previously suggested in various research. For this purpose, keystroke dynamics are used by Monrose and Rubin (2000) to authenticate users based on their typing patterns. The study presents promising results with low error rates and indicates the potential of keystroke biometrics for user identification.

In this study, the effectiveness of Random Forest classifier's performance in various biometric applications including keystroke dynamics is demonstrated. In the academic paper by Saini, Kaur, and Bhatia (2017), the authors demonstrated that the highest level of precision can be achieved by exclusively employing Random Forest techniques while utilizing a dataset derived solely from numpad keystroke data. The high accuracy rates of the remaining classifiers, such as k-NN, SVM, and NN, also support the potential of keystroke dynamics as a reliable biometric feature for identity verification (Killourhy & Maxion, 2009; Darabseh & Namin, 2015).

The ROC curve analysis further validates the performance of the classification models employed in this study. The curve illustrates the relationship between the model's sensitivity and false alarm rate, providing a crucial tool to assess the model's success and evaluate the reliability of the presented results (Fawcett, 2006). The ROC curve analysis is consistent with the accuracy rates obtained, indicating that the random forest classifier is the most effective model for user authentication based on keystroke dynamics.

However, this study has some limitations. The research is conducted on a single dataset, which may not be representative of the broader population. Therefore, the performance of different machine learning classifier algorithms on larger datasets will be explored to further validate the findings of this study in the future.

In conclusion, this study contributes to the growing body of research on keystroke anomaly detection and biometrics. For this purpose, the potential of keystroke dynamics is demonstrated as an effective biometric feature for user authentication. As a result, the high accuracy rates achieved by the classifiers, particularly the random forest classifier, highlight the potential of keystroke biometrics as an additional security layer in remote access scenarios. Nevertheless, further research is warranted to validate the findings on larger datasets and explore additional behavioral cues to ensure generalizability and improved performance. By addressing these aspects, keystroke dynamics can emerge as a valuable component of multifactor authentication systems, bolstering cybersecurity measures and safeguarding sensitive data from potential threats.

6. Conclusion

The goal of this study is to perform identity verification using keystroke biometric data. For this purpose, data preprocessing steps that involve standard scaling and label encoding methods are used to prepare the data. Thereafter, the data is randomly divided into training and test sets. Finally, identity verification is carried out by using four different classifiers which are k-NN, support vector machine, random forest, and neural networks. Furthermore, the accuracy rates of the classifiers are measured. In this context, the accuracy rates of the other classifiers are measured as follows: 84.51% for k-NN, 88.14% for support vector machine, 88.28% for neural networks, and 93.31% for the random forest. As a result, the random forest classifier yielded the highest accuracy rate. The experimental results demonstrate that keystroke data is an effective biometric feature for identity verification.

Consequently, this study demonstrated that keystroke data can be used for multifactor authentication purposes. In addition, it has been shown that the random forest classifier has the highest accuracy rate, although the accuracy rates of the k-NN, support vector machine and neural network classifiers are also quite high. The results of this study highlight the potential of keystroke biometrics as an additional security layer against cyber attackers in remote access scenarios.

The experimental studies of this research are conducted on a single dataset. In future work, experiments will be conducted to determine the performance of different machine learning classifier algorithms on larger datasets. Additionally, a more comprehensive experimental design will be employed to understand the user behavior during the authentication process across different devices and use case scenarios. Also, besides the keystroke dynamics features used in this study, the use of other data obtained from user computer interactions (such as mouse interactions, user interface clicks, etc.) in subsequent studies will be explored to achieve more accurate results in user identity recognition.

Acknowledgment

We extend our sincere appreciation to Kron Technologies for their generous support in this study. Their resources, flexibility, and belief in the importance of academic pursuits have been instrumental in this research. We are also grateful to our colleagues at Kron Technologies for their valuable insights and the inspiring environment they have created.

References

Anderson, R., & Moore, T. (2006). The economics of information security. science, 314(5799), 610-613.

Gordon, L. A., Loeb, M. P., & Sohail, T. (2003). A framework for using insurance for cyber-risk management. Communications of the ACM, 46(3), 81-85.

Li, Y., & Liu, Q. (2021). A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments. Energy Reports, 7, 8176-8186.

Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 18(2), 1153-1176.

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big data, 7, 1-29.

Revett, K. (2009). A bioinformatics based approach to user authentication via keystroke dynamics. International Journal of Control, Automation and Systems, 7, 7-15.

Banerjee, S. P., & Woodard, D. L. (2012). Biometric authentication and identification using keystroke dynamics: A survey. Journal of Pattern recognition research, 7(1), 116-139.

Joyce, R., & Gupta, G. (1990). Identity authentication based on keystroke latencies. Communications of the ACM, 33(2), 168-176.

Gunetti, D., & Picardi, C. (2005). Keystroke analysis of free text. ACM Transactions on Information and System Security (TISSEC), 8(3), 312-347.

Bergadano, F., Gunetti, D., & Picardi, C. (2002). User authentication through keystroke dynamics. ACM Transactions on Information and System Security (TISSEC), 5(4), 367-397.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

Ahmed, A. A. E., & Traore, I. (2005, June). Anomaly intrusion detection based on biometrics. In Proceedings from the sixth annual IEEE SMC information assurance workshop (pp. 452-453). IEEE.

Porwik, P., Doroz, R., & Wesolowski, T. E. (2021). Dynamic keystroke pattern analysis and classifiers with competence for user recognition. Applied Soft Computing, 99, 106902.

Ivannikova, E., David, G., & Hämäläinen, T. (2017, July). Anomaly detection approach to keystroke dynamics based user authentication. In 2017 IEEE Symposium on Computers and Communications (ISCC) (pp. 885-889). IEEE.

Muliono, Y., Ham, H., & Darmawan, D. (2018). Keystroke dynamic classification using machine learning for password authorization. Procedia Computer Science, 135, 564-569.

Aversano, L., Bernardi, M. L., Cimitile, M., & Pecori, R. (2021). Continuous authentication using deep neural networks ensemble on keystroke dynamics. PeerJ Computer Science, 7, e525.

Akşit, N., Aydın, M. A., & Zaim, A. H. (2022). Siber Güvenlikte Klavye Davranış Analizi. İstanbul Ticaret Üniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi, 5(1), 109-122.

Kar, S., Bamotra, A., Duvvuri, B., & Mohanan, R. (2023). KeyDetect--Detection of anomalies and user based on Keystroke Dynamics. arXiv preprint arXiv:2304.03958.

Killourhy, K. S., & Maxion, R. A. (2009, June). Comparing anomaly-detection algorithms for keystroke dynamics. In 2009 IEEE/IFIP International Conference on Dependable Systems & Networks (pp. 125-134). IEEE.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Monrose, F., & Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. Future Generation computer systems, 16(4), 351-359.

Saini, B. S., Kaur, N., & Bhatia, K. S. (2017, January). Keystroke dynamics based user authentication using numeric keypad. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence (pp. 25-29). IEEE.

Darabseh, A., & Namin, A. S. (2015, October). On accuracy of classification-based keystroke dynamics for continuous user authentication. In 2015 International Conference on Cyberworlds (CW) (pp. 321-324). IEEE.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.