INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078

(Volume 10, Issue 2 - V10I2-1164)
Available online at: https://www.ijariit.com

Keystroke Dynamics: A Machine Learning Approach to Behavioural Biometric Authentication

Swarangi Anant Sawant
swarangisawant19@gmail.com
Shree L. R. Tiwari College of Engineering, Mira-Bhayandar, Maharashtra

Sakshi Vasant Kalambe
sakshikalambe1601@gmail.com
Shree L. R. Tiwari College of Engineering, Mira-Bhayandar, Maharashtra

Rupali Pashte
rupali.pashte@slrtce.in
Shree L. R. Tiwari College of Engineering, Mira-Bhayandar, Maharashtra

ABSTRACT

With the ever-increasing dependence on digital services, ensuring the security of user accounts has become a paramount concern. Traditional authentication methods, such as passwords and PINs, have demonstrated vulnerabilities to various attacks. Keystroke dynamics, a behavioral biometric, offers a promising solution for adaptive authentication by analyzing typing patterns unique to everyone. This project explores the implementation of keystroke dynamics in adaptive authentication systems using machine learning algorithms. The primary objective is to create a robust, secure, and user-friendly authentication mechanism that continuously adapts to the changing typing behavior of users while maintaining a high level of accuracy. The proposed system employs a diverse dataset collected from users performing various typing tasks to train machine learning models. Features such as keystroke latency, flight time, and typing rhythm are extracted and used as inputs to the algorithms. Several popular machines learning techniques, including support vector machines, neural networks, and random forests, are employed to build classification models capable of distinguishing between legitimate users and unauthorized intruders. This project advocates for the adoption of keystroke dynamics in adaptive authentication systems, utilizing machine learning algorithms to create a secure and user-friendly experience. By combining behavioral biometrics with cutting-edge technology, the proposed approach offers a robust defense against unauthorized access, paving the way for more secure and convenient authentication methods in the digital era.

Keywords–*Keystroke dynamics, Behavioral biometrics, Adaptive authentication, Security, User accounts, Digital services, Authentication mechanisms, Passwords, Vulnerabilities, Typing patterns, Machine learning algorithms*

I. INTRODUCTION

A. Description

Keystroke dynamics is a behavioral biometrics modality that employs the characteristic typing patterns of users to verify their identity, generally as a part of a multifactor authentication scheme, but has also found other uses. For example, inferring physiological characteristics such as the dominant hand, identifying emotional states, revealing troll accounts on social media, and detecting early signs of neurodegenerative diseases. The methods and techniques of keystroke dynamics analysis have been applied both to fixed texts like passwords when authenticating users at login time, as well as to free text for continuous authentication. Timing information of key press and release events are the most widely logged source of keystroke dynamics data in conventional keyboards, although mobile devices can also provide pressure and acceleration values. Like those using any other biometric modality, keystroke dynamics authentication systems are not exempt from vulnerabilities and many attacks have been described in the literature that target them. Naïve implementations of keystroke dynamics verification are susceptible to presentation attacks with synthesized samples, which leverage partial or complete knowledge of the legitimate users' typing patterns to forge accurate imitations of their behavior. To make things worse, these kinds of attacks can be performed remotely and in an automated way, and can be enhanced with keystroke dynamics data leaked by legitimate applications through unintended side channels. The companion article [1] has presented several novel methods for the synthesis of forged keystroke dynamics samples given one or more authentic samples of free text by the legitimate user. The same article also presented a liveness detection method that employs the latter as adversaries to train a classification model, which can distinguish human-written samples from synthetic forgeries with high accuracy. With the objective of encouraging research about liveness detection in keystroke dynamics and making available our methods so other verification systems can be evaluated against them, we provide software and its source code that includes the following functionality:

- Forges synthesized samples of free-text keystroke dynamics, given their target texts and one or more authentic samples by the legitimate user to be impersonated.
- Trains classification models to distinguish human-written samples from synthetic forgeries, given a collection of samples from the legitimate user.
- Performs liveness detection with the above models, given a collection of samples of unknown origin to be verified and flagged as human-written or synthesized.

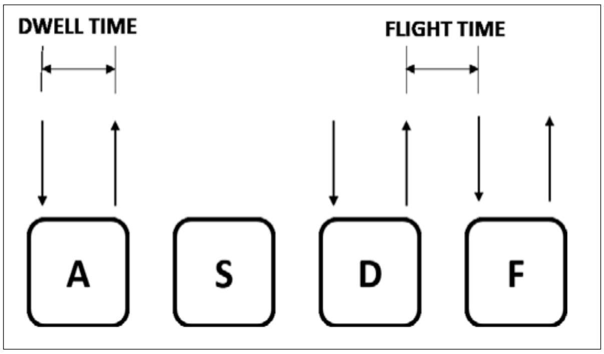

Fig. 1 Time Parameters of Keystrokes

B. *Problem Formulation*

In today's digitally interconnected world, ensuring the security of personal and sensitive information is paramount. Traditional authentication methods such as passwords and PINs are vulnerable to various forms of attacks, including brute-force attacks, phishing, and social engineering. As a result, there is a growing need for more robust and user-friendly authentication mechanisms that can reliably identify individuals. Keystroke Dynamics, a behavioral biometric authentication technique, offers a promising solution to this challenge. Keystroke Dynamics involves analyzing the unique typing patterns of individuals, including the timing and rhythm of keypresses and releases, to establish their identity. This method leverages the inherent differences in how individuals interact with keyboards, making it difficult for unauthorized users to mimic typing behavior accurately. The aim of this project is to develop and implement a Keystroke Dynamics authentication system using Machine Learning techniques. This system will analyze the keystroke patterns of users during the login process and determine whether the input matches the user's established profile.

C. *Motivation*

A project on keystroke dynamics using machine learning holds substantial promise by addressing critical challenges in contemporary digital landscapes. It is motivated by the need for enhanced security and user authentication, offering a biometric solution that can significantly mitigate identity theft and unauthorized access. This approach also prioritizes user convenience, eliminating the need for cumbersome password management and streamlining access to various systems. Furthermore, keystroke dynamics opens the door to personalization and accessibility, allowing for tailored user experiences and support for individuals with physical disabilities. The concept of continuous authentication is another compelling factor, as it provides an ongoing layer of security beyond the initial login. Machine learning advancements play a crucial role in making keystroke dynamics more accurate and efficient, leveraging the power of algorithms and data analysis. In the realm of data privacy and protection, it offers a means of authentication that reduces the reliance on storing sensitive user information, ultimately safeguarding user privacy, and reducing the risk of data breaches. By embracing this project, researchers can delve into an evolving field, fostering innovation and advancing the application of machine learning to keystroke dynamics. In doing so, they have the opportunity to make a tangible, real-world impact by fortifying security measures, improving user experiences, and promoting privacy, culminating in a project that serves as a significant and valuable contribution to the digital age's ever-evolving landscape.

D. Objectives

1) The primary objective is to create a robust, secure, and user-friendly authentication mechanism that continuously adapts to the changing typing behavior of users while maintaining a high level of accuracy.
2) Gather a diverse and representative dataset of keystroke patterns from users, ensuring it encompasses a wide range of typing scenarios and environmental conditions.
3) Design to extract relevant features from the keystroke data, including key press and release timings, inter-key time intervals, and typing rhythm, and train a machine learning model capable of accurately identifying users based on their keystroke dynamics.
4) Implement a secure system for creating, updating, and maintaining user profiles, allowing the model to adapt to changes in users' typing behavior over time.
5) Define and apply appropriate performance metrics to assess the model's accuracy, false positive rate, and false negative rate, ensuring it meets security and usability requirements.
6) Integrate the Keystroke Dynamics authentication system into the target applications or systems, ensuring a seamless and user-friendly login experience.

E. Proposed Solution

The general process for Keystroke based authentication methods requires the following stages:

1) Enrolment

In this phase, the keystroke data are processed and stored. This is done using the following steps:

- **Data acquisition:** Information is generated from the keyboard at the authentication device. Then, the existing system records the data and stores it as a course of events. These raw data are stored as enrolment samples required for later evaluations.
- **Pre-processing:** Pre-processing is essential as whenever the biometric features are extracted, they
- cannot be obtained in the same quality every time.
- **Feature extraction:** Its main function is to select the right features.
- **Storage:** The extracted data are stored in the database to compare the data during future evaluations. It proves very useful as changes can be easily recognized with the storage of these data.

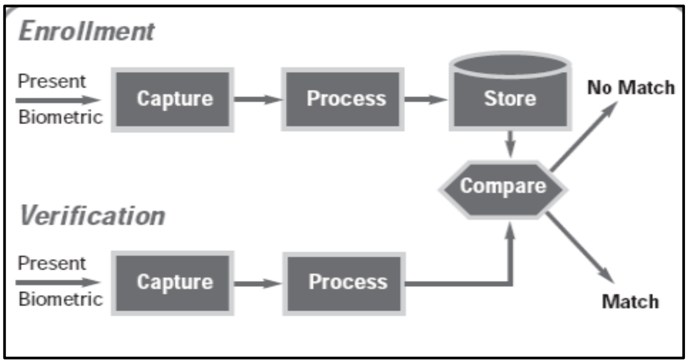

Fig. 2 Biometric Authentication System

2) Authentication

The user can access the system if the data are correctly stored in a database. For this, authentication needs to be done against the system. The authentication phase resembles the enrolment process. The data are stored as samples in the system after the data acquisition phase. In the final steps after the feature extraction, the actual data is compared and classified with the stored data. This is compared based on various biometric modalities like neural networks, distance measures, and probabilistic classifier. Biometric authentication has a disadvantage in that some users are falsely accepted known as FAR (false acceptance rate) and some people are falsely rejected known as FRR (false rejection rate). This means that some intruder trying to access the system can be accepted while on the other hand, the valid user can be rejected. The error rates should be as low as possible for better results and also need to be balanced for special cases as both cannot be zero at the same time.

Sometimes, the EER (equal error rate) is evaluated instead of FRR and FAR. This happens when FRR and FAR are equivalent. The main task is to find the right threshold value for comparison and filtering purposes. Some systems require low FAR (high threshold) such that no intruder can get into the system. This is in fact the best solution to make highly secured systems. The systems for which this protection mechanism is used may have a high usability thus it is important that the user does not need to authenticate himself several times. To achieve this, the threshold needs to be smaller as compared to the previous situation and the FRR should be lower as well. Particularly for mobile devices, the second approach is better. Due to these different arising situations FAR and FRR are required.

F. Scope

The scope of a project on "keystroke dynamics" typically involves studying and analyzing the unique typing patterns of individuals to achieve specific objectives. Keystroke dynamics, also known as typing biometrics or keystroke biometrics, focuses on the behavioural characteristics of how people type on a keyboard, which can vary from person to person. Here are some common aspects:

1. **Authentication Systems:** Develop robust user access control systems that utilize keystroke dynamics as a biometric authentication method to enhance security. This will involve creating an interface for users to enroll and authenticate using their typing patterns.
2. **Biometric Data Collection:** Design and implement methods to collect keystroke data for analysis. This includes data acquisition through keyboard or input device interfaces and ensuring data is stored securely.
3. **Machine Learning Models:** Develop machine learning algorithms for user identification and authentication based on typing patterns. Train and fine-tune these models to accurately distinguish between users and detect anomalies.
4. **User Adaptation:** Implement techniques that allow the system to adapt to changes in users' typing patterns over time. This may involve continuous model retraining and adaptation to accommodate factors like fatigue, injury, or natural changes in typing behavior.
5. **Security and Privacy:** Address security and privacy concerns in keystroke dynamics systems. This includes encryption of collected data, secure storage, and compliance with privacy regulations and standards to protect users' biometric information.
6. **Cross-Platform Compatibility:** Ensure that the keystroke dynamics system is compatible with various devices and input methods, including different keyboard layouts, languages, and input devices. The project should consider both desktop and mobile platforms.
7. **Intrusion Detection:** Build mechanisms to detect unauthorized access based on keystroke dynamics. This includes setting thresholds for deviation detection and implementing alarm systems or notification protocols in case of suspicious or unauthorized access attempts[2]

II. LITERATURE REVIEW

A. Literature Survey

Nahuel González," KSDSLD-A tool for keystroke dynamics synthesis & liveness detection",2022. [1]

The software presented in this article aims to detect synthetic forgeries in keystroke dynamics. It includes features such as sample forging, training classification models, and performing liveness detection. The software is based on finite context modeling and is provided as open-source with the objective of encouraging research in liveness detection in keystroke dynamics. The article references the companion article that describes the design and implementation details of the sample forging and liveness detection methods. The software has been used in academic publications and there are plans to extend it to support additional biometric information.

Alejandro Acien, Aythami Morales, John V. Monaco, Ruben Vera-Rodriguez, Julian Fierrez," TypeNet: Deep Learning Keystroke Biometrics,"2022. [2]

The TypeNet model is a deep learning approach for keystroke biometric authentication. It achieves high performance in recognizing subjects based on their typing patterns on both physical and touchscreen keyboards. The model was trained using large datasets of keystroke data, with participants typing English sentences as quickly and accurately as possible. The model outperforms previous methods based on traditional statistical techniques and deep learning architectures. It demonstrates the potential for operating at an Internet-scale and shows promising results in free-text keystroke dynamics. The model's performance is evaluated using metrics such as Equal Error Rate (EER) and True Acceptance Rate (TAR).

Dong In Kim, Shincheol Lee, and Ji Sun Shin," A New Feature Scoring Method in Keystroke Dynamics-Based User Authentications,"2021. [3]

The paper presents a new feature selection method for keystroke dynamics-based user authentication in the context of data mining for the Internet of Things (IoT). The proposed method is a filter-based feature selection method that aims to improve the classification performance of keystroke dynamics authentication by selecting the most relevant features. Experimental results show that the proposed method outperforms existing feature selection methods by up to 21.8%. The method also ensures user privacy by utilizing only mean values from imposter data. The authors suggest that the method can be applied to other data-mining datasets, such as IoT sensor data sets.

Arafat Rahman, Muhammad E. H. Chowd, Amith Khandakar Serkan Kiranyaz Kh Shahriya Zaman Muhammad Abdul Kadi," Multimodal EEG and Keystroke Dynamics Based Biometric System Using Machine Learning Algorithms,"2020. [4]

This paper tested the dataset and algorithm in different scenarios like closed set generalized and personalized classification for all the modalities (EEG, keystroke, and the combination of EEG and keystroke) with different feature sets and augmentation techniques. We also reported the Cumulative Matching Characteristic (CMC) curve for template matching in identification scenarios and the Equal Error Rate (EER) with Receiver Operating Characteristic (ROC) curve for authentication scenarios where our method achieved better performance for the fusion of EEG and keystroke than individual modality

III. SYSTEM ANALYSIS

A. Requirements

Certainly, here are the hardware and software requirements for a "Keystroke Dynamics Using Machine Learning" project presented as single points:

1) Hardware Requirements

- **Computer:** A modern laptop or desktop with a multi-core CPU and at least 8GB of RAM.
- **Keyboard:** for data collection.
- **Data collection devices** (e.g., webcam and microphone for voice authentication).

2) Software Requirement

- Python programming language.
- Integrated Development Environment (IDE) for Python development (e.g., Jupyter Notebook, Visual Studio Code, or PyCharm).
- Machine learning libraries, including NumPy, Pandas, Scikit-Learn, TensorFlow, PyTorch, and Matplotlib or Seaborn for data manipulation, machine learning algorithms, deep learning, and data visualization.
- Data collection and preprocessing software for keystroke dynamics.
- Database Management System (e.g., MySQL or PostgreSQL) for user profile management.
- Version control using Git and platforms like GitHub or GitLab.
- Text editor for documentation and README files.

3) Additional Considerations

- Choice of Operating System (Windows, macOS, Linux).
- GPU for faster model training (optional).
- Consider cloud services (AWS, Google Cloud, Azure) for additional resources.
- Collaboration tools for team projects (e.g., Slack, Trello).
- Implement data privacy and security measures as needed, including encryption and access controls.

B. Use case Diagram for Proposed System

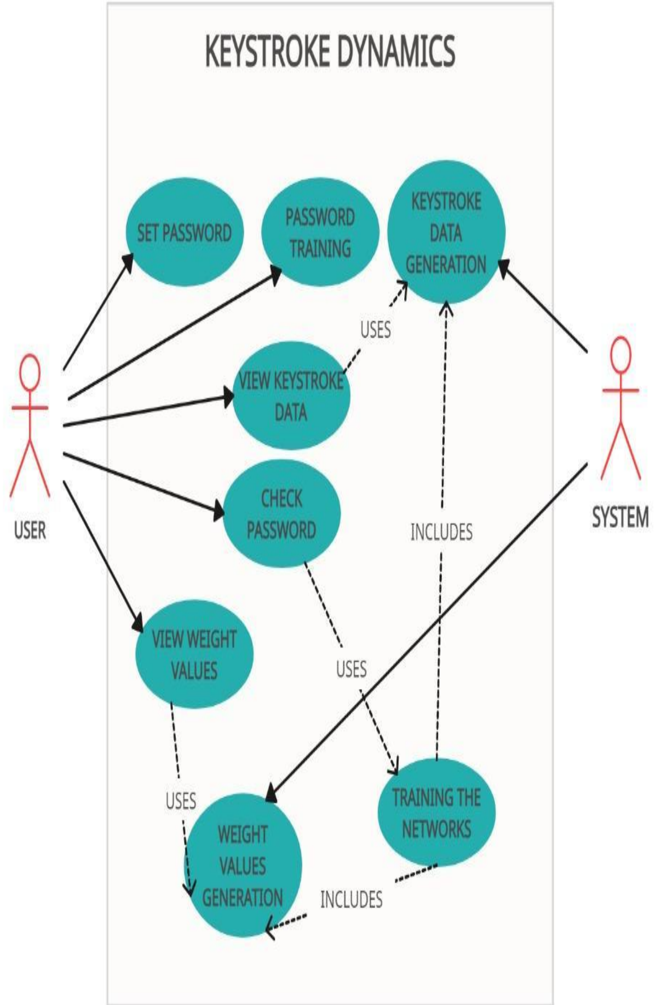

Fig.3 Use case Diagram

C. Flow Chart

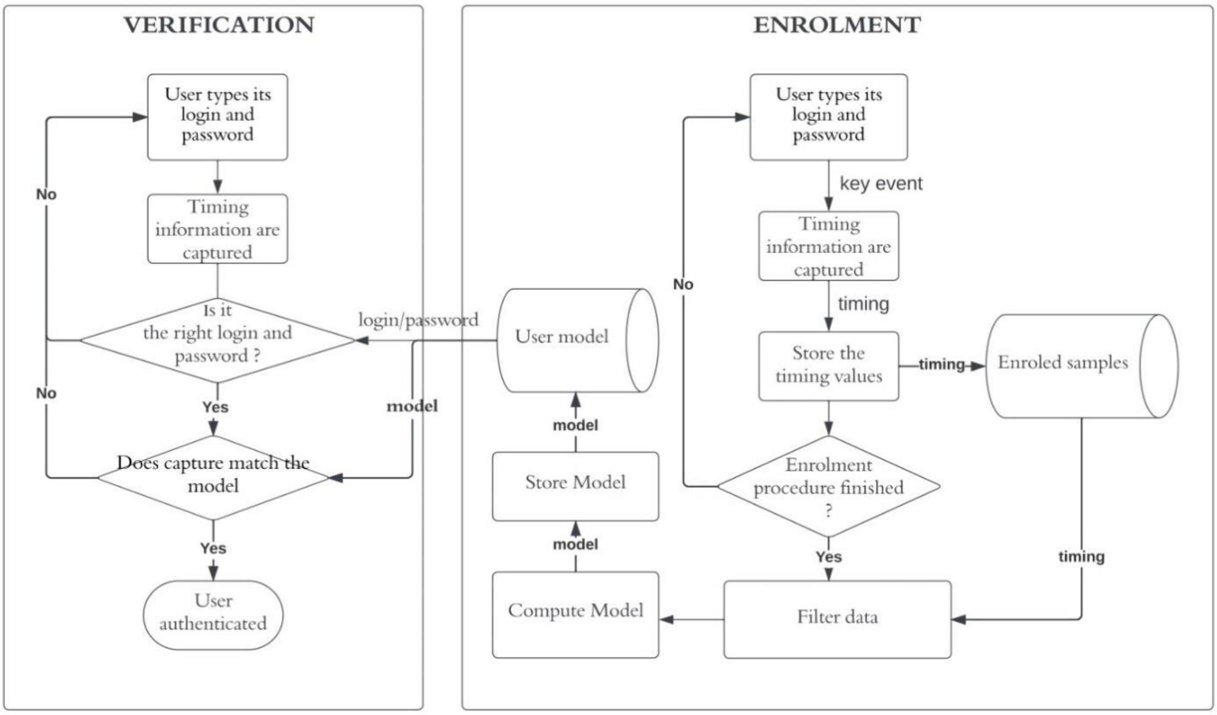

Fig. 4 Flow Chart of the Proposed System

IV. IMPLEMENTATION

A. Algorithms Used

1) Logistic Regression

Logistic regression serves as a statistical technique employed to forecast binary outcomes, like affirmative or negative responses, by scrutinizing past observations within a dataset.

In essence, a logistic regression model forecasts a reliant data variable by assessing the correlation amid one or more prevailing independent variables. For instance, it could foresee whether a political aspirant will triumph or fail in an election or if a high school pupil will secure admission to a specific college. These binary results facilitate straightforward decision-making between two alternatives.

Moreover, logistic regression models can accommodate multiple input criteria. In the context of college admissions, the logistic function might factor in aspects such as a student's GPA, SAT score, and extracurricular involvement. By leveraging historical data on previous outcomes with analogous input criteria, it assigns scores to new instances based on their likelihood of belonging to one of two outcome categories.

This statistical method has emerged as a pivotal tool in the realm of machine learning, empowering algorithms to classify incoming data grounded on past data. With each iteration and the influx of pertinent data, these algorithms refine their predictive capabilities within datasets.

Furthermore, logistic regression can contribute to data preparation endeavors by enabling datasets to be organized into predefined categories during the extract, transform, load (ETL) process, thereby priming the information for subsequent analysis.

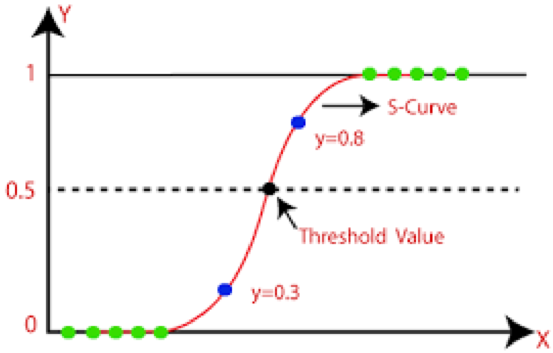

Fig. 5 Logistic Regression

Implementing a logistic regression model in keystroke dynamics involves several steps. Keystroke dynamics refer to the pattern and timing of keystrokes as a form of biometric authentication. Logistic regression can be used to classify keystroke patterns into different categories, such as genuine users or impostors. Here's how you might implement it:

Data Collection: Gather a dataset of keystroke dynamics. This dataset should include features such as key press timings, key release timings, and possibly other metrics like typing speed, typing rhythm, etc. Additionally, you'll need labels indicating whether each instance is from a genuine user or an impostor.

Data Preprocessing: Preprocess the data to make it suitable for training. This may involve steps such as normalization (scaling features to a similar range), handling missing values, and possibly feature engineering (extracting additional relevant features).

Feature Selection: Select the relevant features that are most informative for distinguishing between genuine users and impostors. This step may involve techniques such as statistical tests, feature importance ranking, or domain knowledge.

Splitting the Data: Divide the dataset into training and testing sets. The training set will be used to train the logistic regression model, while the testing set will be used to evaluate its performance.

Model Training: Train the logistic regression model using the training data. The model will learn the relationship between the input features (keystroke dynamics) and the target variable (genuine user or impostor). This step involves optimizing the model parameters to minimize a loss function, typically using techniques like gradient descent.

Model Evaluation: Evaluate the trained model using the testing data. Common evaluation metrics for binary classification tasks like keystroke dynamics authentication include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

Hyperparameter Tuning (Optional): Fine-tune the hyperparameters of the logistic regression model to improve its performance. Hyperparameters include parameters like the regularization strength, learning rate, and convergence criteria.

Model Deployment: Once you're satisfied with the performance of the logistic regression model, you can deploy it for real-world authentication tasks. This involves integrating the model into your authentication system and handling new keystroke dynamics data in real time.[3]

V. RESULTS

	id [PK] integer	email character varying (100)	username character varying (50)	password character varying (50)	keystroke_timing jsonb
1	3	sakshikalambe1671@gmail.com	Diksha	diksha123	[{"key": "d", "up_time": "12162.60", "down_time": "12094.50"}, {"key": "i", "up_time": "12855.60", "down_time": "12795.2
2	5	nilimakalambe1976@gmail.com	Nilima	Nilima123	[{"key": "N", "up_time": "67856.70", "down_time": "67771.20"}, {"key": "i", "up_time": "68319.10", "down_time": "68254.0
3	6	dikshakalambe31@gmail.com	Diksha1	Diksha1234	[{"key": "D", "up_time": "30393.50", "down_time": "30337.30"}, {"key": "i", "up_time": "30900.30", "down_time": "30826.5
4	7	sejalmore123@gmail.com	Sejal	Sejasl13	[{"key": "S", "up_time": "16124.00", "down_time": "16068.80"}, {"key": "E", "up_time": "16644.10", "down_time": "16592.
5	8	sahalmanasia@gmail.com	Sahal	Sahal789	[{"key": "S", "up_time": "15617.10", "down_time": "15551.40"}, {"key": "a", "up_time": "16036.50", "down_time": "15933.
6	9	sabamansuri@gmail.com	Saba	Saba@123	[{"key": "S", "up_time": "32038.30", "down_time": "31968.70"}, {"key": "a", "up_time": "32494.10", "down_time": "32385.
7	10	anagha123@gmail.com	Anagha	Anagha1123	[{"key": "a", "up_time": "25182.20", "down_time": "25094.80"}, {"key": "n", "up_time": "25827.40", "down_time": "25744.
8	13	yogipatil@gmail.com	yogi	yogi123	[{"key": "y", "up_time": "14033.00", "down_time": "13966.60"}, {"key": "o", "up_time": "14358.30", "down_time": "14301.4
9	14	Sanu@gmail.com	Sanu	Sanu123	[{"key": "s", "up_time": "17967.60", "down_time": "17907.20"}, {"key": "a", "up_time": "18358.40", "down_time": "18287.
10	15	patu@gmail.com	patu	patu5678	[{"key": "p", "up_time": "18969.20", "down_time": "18898.30"}, {"key": "a", "up_time": "19200.60", "down_time": "19108.
11	16	SAKSHIKALAMBE1761@GMAIL...	Sakshi	Sakshi123	[{"key": "s", "up_time": "11423.80", "down_time": "11357.20"}, {"key": "a", "up_time": "11868.50", "down_time": "11750.
12	17	sakshi.v.kalambe@slrtce.in	Rohan	Rohan@123	[{"key": "n", "up_time": "38599.20", "down_time": "38461.60"}, {"key": "@", "up_time": "40161.60", "down_time": "40028.
13	19	swara4634@gmail.com	BCD	bcd	[{"key": "b", "up_time": "7594.40", "down_time": "7527.00"}, {"key": "c", "up_time": "7811.20", "down_time": "7741.70"}, {
14	20	sakshikalambe1602@gmail.com	ABC	abc123	[{"key": "a", "up_time": "8154.50", "down_time": "8028.50"}, {"key": "b", "up_time": "8439.50", "down_time": "8374.30"}, {
15	21	abc@gmail.com	ABCD	ABCD	[{"key": "a", "up_time": "270169.30", "down_time": "270025.50"}, {"key": "b", "up_time": "270709.10", "down_time": "270

Fig. 6 Database

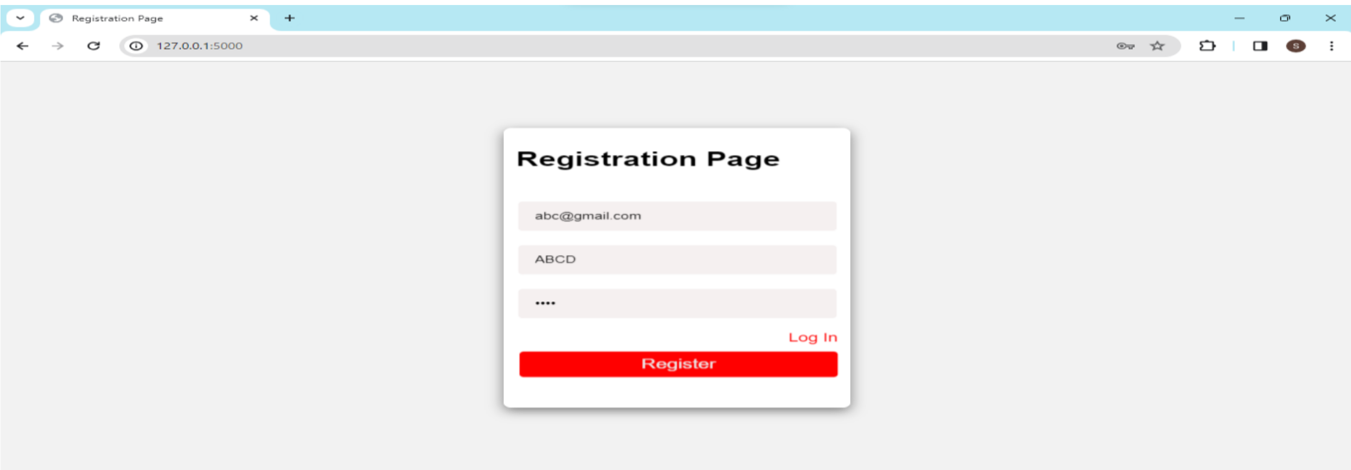

Fig.7 Registration Page

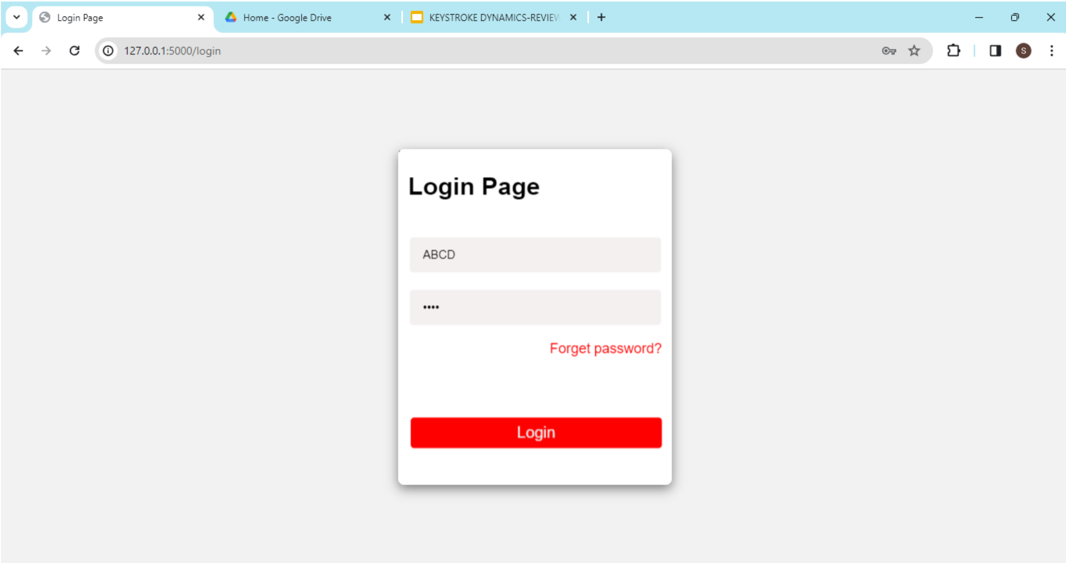

Fig.8 Login Page

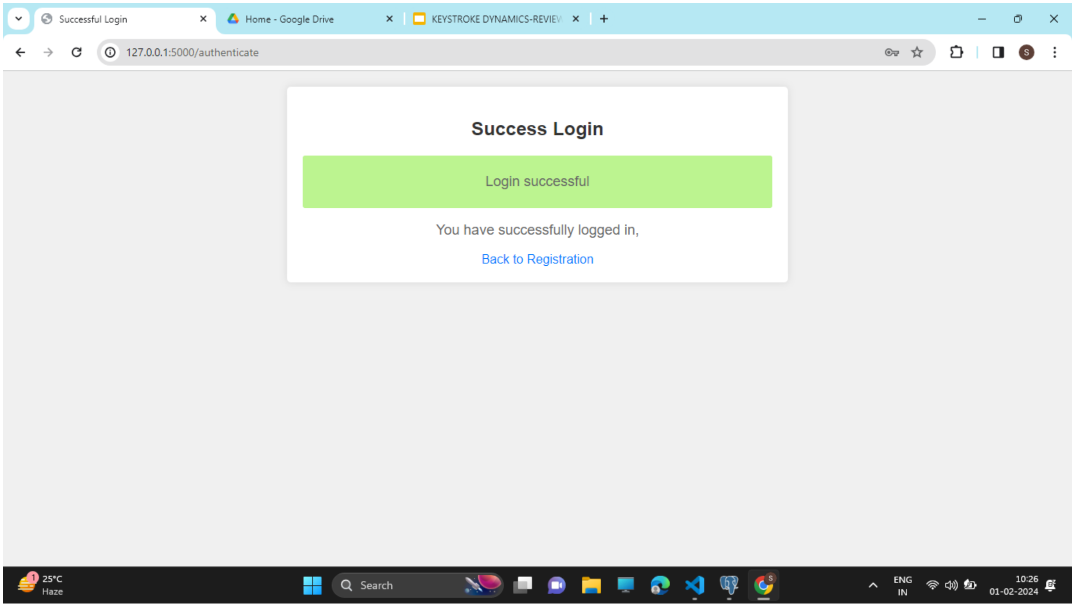

Fig.9 Successful Login Page

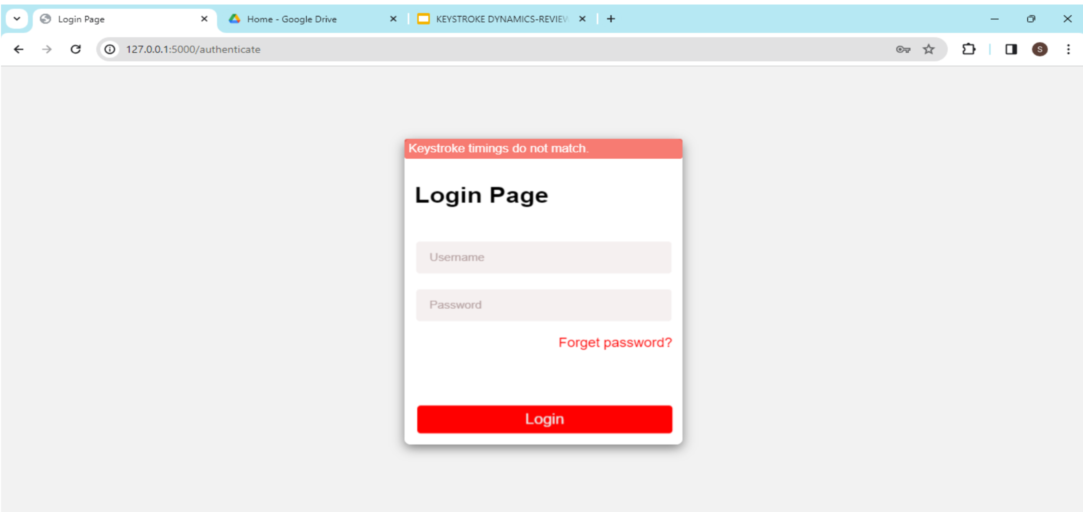

Fig.10 Unsuccessful Login

VI. CONCLUSION

A. Conclusion

Keystroke dynamics, as a biometric authentication method, holds significant promise in enhancing security, privacy, and user convenience. Research and development in this field have made considerable strides in recent years, leading to innovative applications and improvements in accuracy. However, challenges related to security, privacy, and usability persist and require ongoing attention.

In conclusion, the "Keystroke Dynamics Using Machine Learning" project has showcased the promise of keystroke dynamics as a biometric authentication method. It has the potential to enhance security while being user-friendly, as it doesn't require additional hardware. However, the project also underlines the importance of stringent data privacy measures and ethical considerations when handling user data. The project's success is closely tied to the quality and quantity of training data and the choice of machine learning algorithms. Keystroke dynamics can find applications beyond security, such as user behavior analysis and assistive technologies. Future research should focus on adapting the models to changing typing styles, exploring multi-modal biometric authentication, and addressing privacy concerns. This project has illuminated a path towards more secure and user-friendly authentication methods with keystroke dynamics and machine learning at their core.

B. Future Scope

Future developments will also focus on privacy preservation, standardization, machine learning enhancements, cross-platform compatibility, ethical and legal frameworks, and collaborative research efforts, ensuring keystroke dynamics continue to play a significant role in enhancing authentication and security across various domains and industries.

Based on the results of the present study, our future study will focus on developing the different data features of keypads, such as button size, character addition, efficient arrangement, and user classification. This will help generate a unique keypad for every user, which will contribute to improving the user classification performance because it will collect features that can only be collected from the normal user. Even if a smartphone user's PIN, as the ultimate password, has been hacked, there will be no need to worry because this technology will help protect their private information in an even safer way.

REFERENCES

[1] Nahuel González," KSDSLD — A tool for keystroke dynamics synthesis & liveness detection",2022.

[2] A. Acien, A. Morales, J. V. Monaco, R. Vera-Rodriguez and J. Fierrez, "TypeNet: Deep Learning Keystroke Biometrics," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 57-70, Jan. 2022, doi: 10.1109/TBIOM.2021.3112540.

[3] D. I. Kim, S. Lee and J. S. Shin, "A New Feature Scoring Method in Keystroke Dynamics-Based User Authentications," in *IEEE Access*, vol. 8, pp. 27901-27914, 2020, doi: 10.1109/ACCESS.2020.2968918.

[4] A. Rahman *et al*., "Multimodal EEG and Keystroke Dynamics Based Biometric System Using Machine Learning Algorithms," in *IEEE Access*, vol. 9, pp. 94625-94643, 2021, doi: 10.1109/ACCESS.2021.3092840.