**Proposed Architecture for Authenticating Internet Banking Users based on Keystroke Dynamics**

**1Siddhivinayak Kulkarni, 2Tanaya Joshi, 3Khushboo Biyani, 4Shivani Bhide**

**Department of Computer Engineering**

**M.I.T College of Engineering**

**India, Maharashtra, Pune-411016**

**e-mail: siddhivinayak.kulkarni@mitcoe.edu.in**

**Abstract:**

An individual can be uniquely identified from a group based on certain characteristics by using biometrics. Several authentication and validation systems rely on biometrics because of its capability to efficiently authenticate a user. Physical and behavioral biometrics is the two main types of biometrics. This paper focuses on the deployment of behavioral biometrics in the field of Online Banking in order to add an extra layer of security to the existing authentication systems. Such a system would rule out the possibility of passwords being forcefully acquired and the account being accessed by impostors. This paper proposes the use of keystroke dynamics as behavioral biometrics, which makes use of the rhythm and manner in which an individual types characters on the keyboard or keypad. A unique biometric template would be created to store the measured typing dynamics of a user. For the data classification purpose, machine learning algorithms like K-Nearest Neighbors (KNN) with multiple distance measures, Neural Networks, and Support Vector Machine (SVM) can be used. In this proposed system, access to the user’s account details would be provided based on his recorded keystroke dynamics. In case of any discrepancy, the user would be alerted. This system has a lot of scope for further research. For instance, a second layer of authentication using physical biometrics like fingerprint recognition or retina scan could be added for further access. The potential of such an authentication system in order to add an extra layer of security to one’s account and in reducing and detecting frauds is immense.

***Keywords****: Authentication; Biometrics; Keystroke dynamics; KNN; Machine learning.*

**1. Introduction:**

The online banking industry is growing at a phenomenal rate and will continue to do so because of easy access to banking transactions. It is imperative to increase the level of security involved in the internet banking industry since a large number of customers perform transactions online on a daily basis [17].

Internet Banking Security refers to the use of various techniques and algorithms to ensure integrity, confidentiality and availability of data. Ensuring Internet Banking security has become an onerous task due to the advanced attacking tools available with the ever-increasing number of perpetrators. It is essential to adopt new, efficient techniques to reduce the number of frauds prevalent in this domain. This paper aims at focusing on a framework which uses behavioral biometrics for authentication in online banking thereby increasing security to a great extent.

Biometrics refers to certain distinguishable characteristics which can be used to uniquely identify an individual from a group. Biometric identifiers are measurable metrics. Physical biometric identifiers could be fingerprint, DNA, face recognition [1]. Behavioral biometric identifiers could be typing dynamics, voice recognition etc.

Keystroke dynamics, keystroke biometrics or typing dynamics, is the detailed timing information which describes exactly when each key was pressed and when it was released as a person is typing at a computer keyboard. The behavioral biometric of Keystroke Dynamics uses the manner and rhythm in which an individual types characters on a keyboard or keypad [2].

The keystroke rhythms of a user can be measured to develop a unique biometric template of the user's typing pattern for future authentication. Raw measurements available from almost every keyboard can be recorded to determine Dwell time and Flight time. The recorded keystroke timing data is then processed through a unique neural algorithm, which determines a primary pattern for future comparison. Similarly, vibration information may be used to create a pattern for future use in both identification and authentication tasks.

Data needed to analyze keystroke dynamics is obtained by keystroke logging. Normally, all that is retained when logging a typing session is the sequence of characters corresponding to the order in which keys were pressed and timing information is discarded.

* 1. Keystroke dynamics features:

1. Flight time

It is the time calculated as the difference between the time of key-press and key-release. There are four types of latencies based on the order of a press and release.

1. Press-press (Time taken to press the first key – time taken to press the other key)

2. Release-press (Time taken to release the first key – time taken to press the other key)

3. Press-release (Time taken to press the first key – time taken to release the other key)

4. Release-release (Time taken to release the first key – time taken to release the other key)

2. Dwell time:

It is the time taken by the user to press and release a single key.

**2. Literature Review:**

Biometric recognition is the task of identifying an individual on the basis of his/her physiological or behavioral traits. Necessity of developing fool proof security systems has provided biometric research much needed impetus. Some of these biometric measures have shown tremendous potential in forensic applications. In this paper a review of these emerging biometric modalities is presented [1]. Kofigagbla [2] demonstrated the design and implementation of a biometric authentication scheme by means of keystroke dynamics to secure web applications. The scheme illustrates the effect of the template size and the number of biometric features used to improve performance of FAR, FRR and EER. The enrollment process is devoid of spurious outliers that may introduce significant variations to template data. Due to the dependence of the system on the user’s psychological characteristics, the adaptation module introduces dynamism into the template features per changes in the legitimate user’s login template variants.

The keystroke dynamics samples are collected in a web-based uncontrolled environment (OS, keyboards, browser, etc.). Such kind of dataset is important since it provides us more realistic results of keystroke dynamics’ performance in comparison to the literature (controlled environment, etc.). Second, we present a statistical analysis of well-known assertions such as the relationship between performance and password size, impact of fusion schemes on system overall performance. In this paper [4], the keystroke bio metrics is used with the application in news reporting system. It will detect the person who send the news is the reporter or some other person who hacked the user name and password of the system. First the pattern of the reporter is stored with the server system. Server after receiving the text then it matches with the text pattern information in it. The main purpose of the system is to develop a secure, cheap and effective security system for securing the computer applications and data based on typing biometrics called typing patterns. An overview of keystroke dynamics and how it can work with different approaches has been detailed in the paper based on biometric authentication [5].

The paper’s [6] objective is to collect a keystroke-dynamics data set, to develop a repeatable evaluation procedure, and to measure the performance of a range of detectors so that the results can be compared soundly. Data was collected from 51 subjects typing 400 passwords each, and the system was implemented and evaluated 14 detectors from the keystroke- dynamics and pattern-recognition literature. The three top-performing detectors achieve equal-error rates between 9.6 and evaluation methodology—constitute a benchmark for comparing detectors and measuring progress. This paper [7] proposes a novel kNN type method for classification that is aimed at overcoming its shortcomings. The method constructs a kNN model for the data, which replaces the data to serve as the basis of classification. The value of k is automatically determined, is varied for different data, and is optimal in terms of classification accuracy. The construction of the model reduces the dependency on k and makes classification faster. Experiments were carried out on some public datasets collected from the UCI machine learning repository in order to test the method. Another technique of applying keystroke dynamics as a continuous process and not just for one stage verification was detailed by Bours et al. The authentication function is called for every set of keys the user presses thus making it a continuous process. Security is one of the major concerns in e-commerce, this issue has been discussed in the paper [10] related to on-line payments and how to overcome it. Statistical analysis based on both behavioral and physical biometrics has been performed [13].

In this paper [14], a new distance metric that is effective in dealing with the challenges intrinsic to keystroke dynamics data i.e. scale variations, feature interactions and redundancies, and outliers is proposed. Their keystroke biometrics algorithms based on this new distance metric are evaluated on the CMU keystroke dynamics benchmark dataset.

**3. Proposed Methodology:**

In this paper, we will focus on an authentication system for Internet Banking security. We propose to use behavioral biometrics for authentication [3]. The behavioral biometrics i.e. typing dynamics would be used to generate a password unique to each user. This number would determine the access of the user to his account and online transactions. Keystroke dynamics are recorded and stored corresponding to the user. These are then re-checked and used during further transactions. In case a variation in dynamics is encountered, alerts would be sent to the account holder as well as another person nominated by the person. Also, the transaction would be recorded and stored. Efficient implementation of this system could help in reducing internet banking frauds to a great extent. Figure 1 discusses the overall functional block diagram for proposed research.

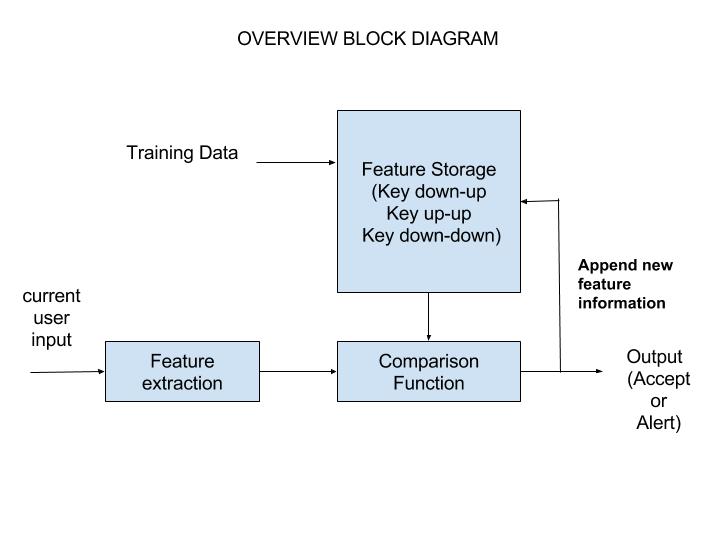


Fig 1.Overall functional block diagram

**3.1 Data set:**

In supervised machine learning, the machine is given example inputs and their desired outputs by a teacher/supervisor i.e. it is provided with the training data set. The main aim of this technique is to map the output to the input. In our process, the system would be given data obtained from the user typing method and by extracting the values of factors like Dwell time (the time a key pressed) and Flight time (the time between "key up" and the next "key down").

**3.2 Feature storage:**

The training dataset given to the machine is stored. This storage data is then later accessed for comparison with the new features extracted from test data.

**3.3 Feature extraction:**

Features like the Dwell time(Time when a key is pressed), Flight time (Time amid key-up and next key-down), time between key-up and next key-up, time between key-down and next key-down and time between key-down and next key-up are extracted from the test input. These features, along with the stored training dataset, are then given as an input to the comparator. The comparison function evaluates the exactness between the current input and the training data set. The output from the comparator would be used to determine if the input features match the features of the account holder or not.

**3.4 Comparison function:**

The comparison function is given two inputs viz. the training dataset and the testing data. Different algorithms like KNN, SVM etc. can be used for classification purpose which would eventually determine if the user is to be given access to the account i.e. accepted or if the user needs to be alerted i.e. rejected. The comparison function has various components and is shown in Figure 2.

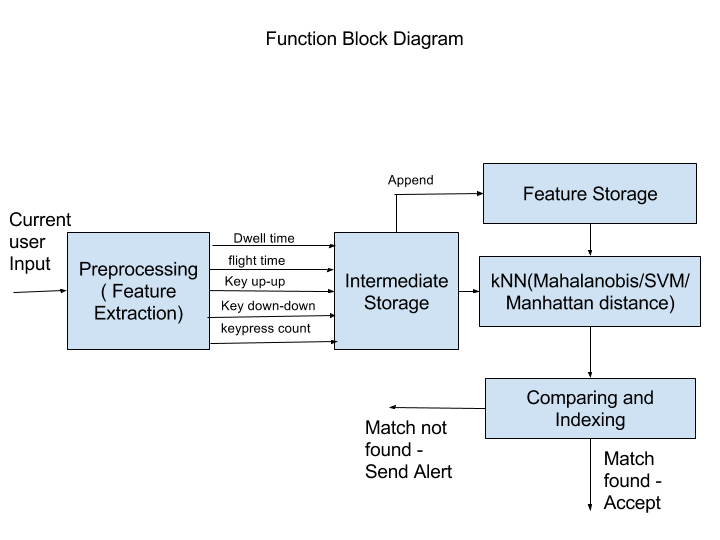


Fig.2.Functional block diagram of comparison function

**3.4.1 Pre-processing:**

In the pre-processing phase, the features are extracted and then stored. These features include dwell time, flight time, key up-up, key down-down and key-press count.

**3.4.2 Intermediate storage:**

This phase consists of the temporary storage which stores the extracted features. Any new features obtained are appended to the permanent storage. Data from the intermediate storage is fed as an input to the classification algorithm used.

**3.4.3 Classification algorithms:**

Classification is an important data mining technique which is used to segregate data into groups based on certain attributes. Classification is a technique which comes under supervised machine learning. Machine learning, which is a sub-domain of Artificial Intelligence, refers to the construction and designing of algorithms through which a machine can learn from the data. Thus, it makes decisions or predictions based on the input data and past incidents rather than on static program instruction.

Various classification algorithms like K-Nearest Neighbors (KNN), Support Vector Machine (SVM) can be used.

**3.4.3.1. K-Nearest Neighbors:**

K-Nearest Neighbors (KNN) is a classification algorithm which uses supervised machine learning techniques [9]. There are two phases involved:

1. Training phase:

The training examples consist of vectors in a multidimensional feature space, each having a class label. In the training phase of KNN, the feature vectors and class labels of the training samples are stored.

2. Classification phase:

‘k’ is a user-defined constant which indicates the number of neighbors. In the classification phase, an unlabeled vector i.e. a test point is classified by being assigned the most frequently occurring label among the k training samples which are nearest to the test point. [7]

Euclidean distance is a commonly used metric for continuous variables. Euclidean distance is the straight-line distance between two points in a Euclidean space.

Formula for calculating Euclidean Distance:

1. d(p,q) = d(q,p) =

Equation 1.1

For text classification which uses discrete variables, another metric can be used, such as the Hamming distance.

Formula for calculating Hamming Distance:

2. DH =

x = y => D = 0

x y => D = 1

Equation 1.2

Algorithm:

1. For each training example <x, f(x)>, add the example to the list of training examples.

2. Given a query instance xq to be classified, Let x1, x2, x3,...,xk denote the k instances from training examples that are nearest to xq.

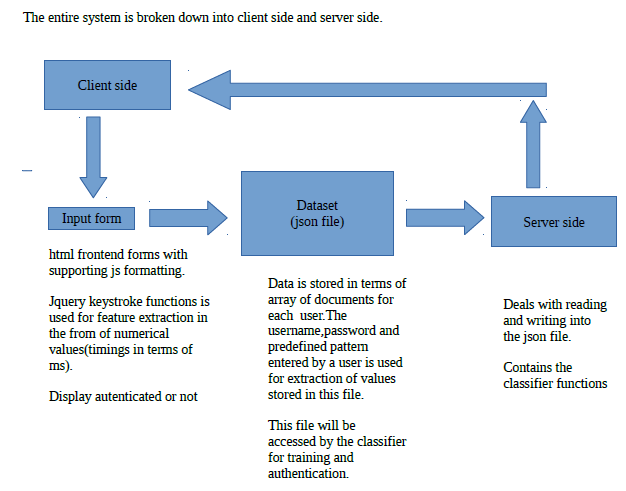
3. Return the class that represents the maximum of the k instances.

**3.5 Comparing and Indexing:**

The value obtained by using the classification algorithm would give us the similarity between the test and the training data. By performing indexing on these values and choosing the value nearest/closest to the test data i.e. that value which is the most similar, we would be able to either accept or reject the test data. If the test data is rejected, the user would be alerted since it would indicate a mismatch. In case the test data is accepted, the user would be granted access to the user account.

**4. Implementation of the proposed system:**

The proposed system is a distributed client-server system. The online banking users have to enter their details on the client side form. The data entered on the form is sent to the server where the classification takes place. Node.js can be used on server-side, while on the client side HTML is used to create a user-friendly GUI.

****

**5. Data Templates :**

Dataset has been taken from 6-7 users and each user has 8-10 templates each in order to train the classifier to authenticate the user appropriately.

Sample Data Template:

Username: XYZ,

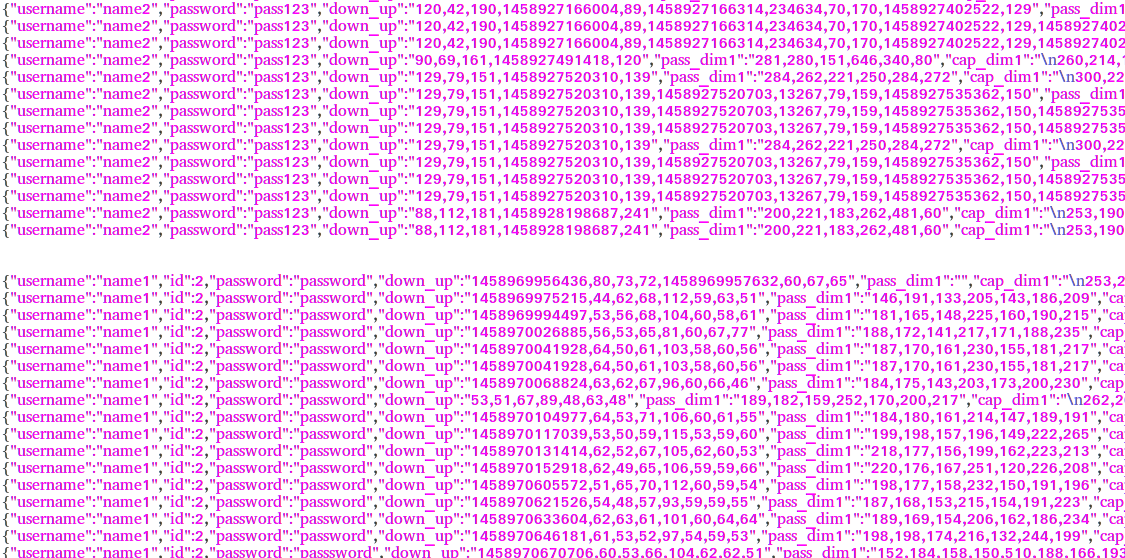
Password: PQR,

Key down\_up: 156,223,167,190;

Key up : 226, 321,200,152;

Key down: 180,154,210,320;

Screenshot of the templates:



**5. Advantages:**

* Keystroke dynamics can measure time in terms of milliseconds. At this level of accuracy it is very difficult to reproduce one person's pattern by another.
* Physical biometric systems work on a combination of hardware and software platforms. However, for keystroke dynamics only software is required for implementation. Hence, this system is inexpensive.
* The user need not be aware about the extra layer of security and can function without any interference.
* As keystroke dynamics are unique to every person, they cannot be replicated. Hence security is enhanced.

**6. Disadvantages:**

* Typing rhythm of a person can change based on their state of mind. Hence, this system is not completely reliable.
* A lot of data needs to be stored to match the patterns of the user and define a range of different parameters.

**7. Conclusion and Future research:**

In this paper, we have studied the use of behavioral biometrics, namely keystroke dynamics, in authentication systems related to financial services. We have studied and analyzed the classification algorithms which could be used in this system. Also, we have discussed the potential of keystroke dynamics in increasing security and thereby reducing Internet Banking frauds to a considerable extent.

The potential of the authentication system using biometrics in Internet Banking security is ever-increasing. Further research could be done to identify emotional states using keystroke dynamics. Artificial bots could be trained using the acquired data. Also, this process could be extended to other verification systems (for non-financial purposes) to enhance security and reduce frauds.

**References:**

[1]. A survey of emerging biometric modalities (SushilChauhan, A.S. Arora, AmitKaul) Elsevier Procedia Computer Science Volume 2, Pages 213–218, Proceedings of the International Conference and Exhibition on Biometrics Technology, 2010)

[2]. Applying Keystroke Dynamics for Personal Authentication (George Kofi Gagbla, thesis, 2005)

[3]. The BeiHang Keystroke Dynamics Authentication System (Juan Liu , Baochang Zhang , LinlinShen ,Jianzhuang Liu , Jason Zhao ,University of Hong Kong, 2009)

[4]. Open password system with press time and inner time (Gaddam S SaiVeena , TadepalliSivani , SravaniBandi , NishaGajjala ,VemulaSharanya,IJAIEM Volume 2, issue 12, 2013)

[5]. Keystroke dynamics as a biometric for authentication (Fabian Monrose a, Aviel D. Rubin, elsevier, 2009)

[6]. Comparing Anomaly-Detection Algorithms for Keystroke Dynamics (Kevin S. Killourhy Roy A. Maxion, Carnegie Mellon University, 2009)

[7]. KNN Model-Based Approach in Classification (GongdeGuo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer, 2003)

[8]. Continuous Authentication using Biometric Keystroke Dynamics (Patrick Bours and Hafez Barghouthi, NISK, 2009)

[9]. Nearest Neighbor Classification (Kilian Q. Weinberger, Journal of Machine Learning Research, 2009)

[10]. On-line Payment and Security of E-commerce (Yang Jing,Proceedings of the 2009 International Symposium on Web Information Systems and Applications (WISA’09)Nanchang,P.R. China,May22-24,2009,pp.046-05)

[11]. Keystroke Dynamics Benchmark Dataset (Kevin Killourhy and Roy Maxion, 2009)

[12]. Web-Based Benchmark for Keystroke Dynamics Biometric System (RomainGiot and Mohamad El-Abed and Christophe Rosenberger, 2012)

[13]. Biometric Systems: A Statistical Analysis (RomainGiot and Mohamad El-Abed and Christopher Rosenberger, Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012 Eighth International Conference, pages 11-15, 2012)

[14]. Keystroke Dynamics for User Authentication (Anil K. Jain,Prof.,Michigan State university, Computer Vision and Pattern Recognition Workshops(CVPRW), 2012 IEEE Computer Society Conference, pages 117-123,2012)

[15]. http://docs.opencv.org/doc/tutorials/ml/introduction\_to\_svm/introduction\_to\_svm.html

[16].Support vector machine applications in bioinformatics. (Johann Wolfgang Goethe-Universität, InstitutfürOrganischeChemie und ChemischeBiologie, Frankfurt, Germany Applied Bioinformatics [2003, 2(2):67-77])

[17] Electronic Banking Fraud; The need to enhance security and customer trust in online banking (RashadYazdanifard, Wan FadzilahWanYusoff , Alawa Clement Behora,̧ Abu Bakar Sade, International Journal in Advances in Information Sciences and Service Sciences 2011; 3(10.61):505-509)