International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

# Non-Conventional Factors for Keystroke Dynamics as a Support Factor for Authenticating User

**Nataasha Raul, Radha Shankarmani, Padmaja Joshi**

***Abstract:*** ***Keystroke Dynamics Are Most Frequently UsedIn Situation Where Password Security Is The Major Concern. In The Event That The User's Secret Word Is Disclosed, AndThe Keystroke Rhythm Of The Genuine Client Is Known, The Application Will Not Accept The Fake User In Spite Of Having Valid Credentials. This Paper Focuses On Strengthening The Keystroke Dynamics By Introducing Non-Conventional Features For Authenticating A User Using Static Keystroke Dynamics. The Paper Also Discusses The Improvement In Far, Frr, Err, And Accuracy After The Application Of The Proposed Non-Conventional Feature Set Along With The Existing Timing Feature Set.***

***Index Terms: Features, Keystroke, Non-Conventional, Static, User.***

## I. INTRODUCTION

Keystroke dynamics is explored as a support factor for authentication for many reasons. Firstly, unlike other biometric techniques, keystroke dynamics do not require additional hard- ware as users will be using a computer keyboard for typing which is analyzed for their identification. Secondly, it is not costly to implement, as the hardware required is the keyboard. Finally, typing rhythms will be available even after the user is authenticated, because keystroke information exists as a mere consequence of users using computers through keyboard [1]. A lot of research is available on keystroke dynamics and is based on either free texts, password or passphrases. In general, the typing rhythm of the user are often taken either statically (at login time) or continuously (while using a computer). The static keystroke model uses a single defined expression, such as a password and/or a username. Static Keystroke model helps in strengthening the authentication process with username and password [2]. As the length of the password increases, the count of keystroke timing features also increases, which further increases the training time of an algorithm in a static keystroke authentication mechanism. In a continuous keystroke authentication mechanism, the user typing is continuously monitored. There is no constraint on the length of the keys typed as compared to fixed phrase in static keystroke mechanism. In continuous keystroke mechanism, the non-conventional features such as usage of Caps Lock key, Shift Key and word typed per minute can also be extracted along with the timing features.

**Revised Manuscript Received on February 28, 2020.**
***** Correspondence Author
**Nataasha Raul*,** Sardar Patel Institute of Technology, Mumbai, India. **Radha Shankarmani,** Sardar Patel Institute of Technology, Mumbai, India. Padmaja Joshi, CDAC, Mumbai, India.



The work in this paper focuses on applying non-conventional features along with timing features for the Static Keystroke Authentication mechanism, to reduce the training time which is required to train the user profile and to improve the performance of the system. The paper is arranged as follows. Section II deals the basics of keystroke dynamics. Section III covers the study and research done by different authors on Keystroke Dynamics. Section IV includes the proposed approach for verifying users using static keystroke dynamics as a support factor. Section V describes experimentation carried out for the proposed approach. Section VI focuses on the results and observations from the experiment done. The paper is concluded with concluding remarks in Section VII.

## II. KEYSTROKE DYNAMICS B ASICS

The process in keystroke dynamics consists of collecting raw keystroke timing data from the user. The captured keystroke information, also known as features, is stored and used to create user profile database. In classification, the captured keystroke features are compared with the stored profile database of the user and the decision is generated. The classification is in binary form for keystroke dynamics based authentication, i.e. either authenticated or not authenticated. The keystroke dynamics try to capture the characteristics of an individual through his or her typing pattern. Features based on key press and key release timings are mainly used to identify the pattern of an individual's typing. The features are extracted from the KeyPress and KeyRelease time as shown in Fig 1. These features are explained ahead.

Dwell time is the difference between the KeyPress and KeyRelease of a single key as shown in Fig 1.

Flight time is the difference between the current keyre- lease and the next keypress as shown in Fig 1.

Di-graph is the time taken by the user to type two consecutive keys. The same is shown by Di-graph in Fig 1.

Tri-graph is the time taken by the user to type three consecutive keys. The same is shown by Tri-graph in Fig.1.

## III. LITERATURE SURVEY

In keystroke dynamics, the features are first extracted to represent the user typing characteristics, then are classified for authentication and identification purposes. The user profile is created on the basis of the set of features selected for the individual. Different researchers used different set of features to authenticate users using keystroke dynamics. Typically, the diagram is a series of two characters, including punctuation, numerals, letters, and space.

Digraph latencies within the virtual letter pairs were then used to produce a score [3]. Features like digraphs and trigraphs rely
only on the word context [4]. To solve the

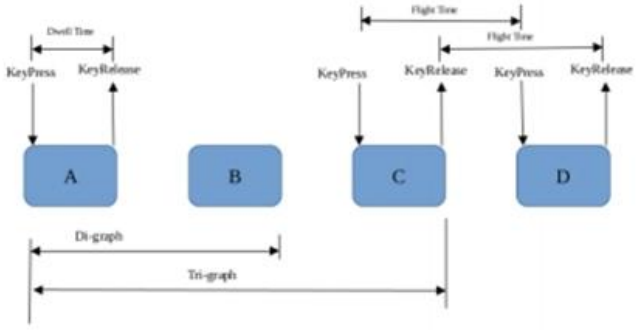



**Fig. 1: Keystroke Timing based features**

dependencies Dowland and Furnell [5] used digraphs, trigraphs and n-graphs features along with the keyword latencies i.e. AutoID, Left character, Right character, Latency, and Timestamp. The most effective results were achieved by using digraph.Feature set identification which helps to form a unique profile for each individual is the most significant side of machine learning-based keystroke dynamics [3]. The user identification and classification used to authenticate the user is a necessary step in the keystroke dynamics. The classification algorithm learns the typinmg pattern of each user only when the user profile data undergoes a training. Linear Classifier, Support Vector Machines (SVM), Decision Trees, Boosted Trees, Random Forest, Nearest Neighbor (NN) and Neural Network are the types of classifiers in machine learning [3]. The features selection is an important factor and also an area of research.

It is observed from the studies that different feature sets were used by different authors for authenticating users using keystroke dynamics. D. R. Gentner et. al. [6] and Roth, et. al. [7] used digraph and trigraph timing features for authenticating users using static keystroke dynamics. Alsultan, et. al.[8] used timing features as well as non-conventional features for authenticating a user using continuous keystroke dynamics.

Data after classification must be evaluated for its correctness. It is essential to provide standardized testing methods and performance metrics for analysis. The datasets used plays a very vital role during this testing as it is a known fact that the training dataset is used for testing, cannot comment on the performance of the classification. [3] The data set chosen has a significant effect on the performance of the system. Various studied have reported a major increase in output for comparable and identical algorithms, taking into account the fact that most researchers have used their own dataset. To address this issue, Killourhy and Maxion compiled and published the keystroke dynamics benchmark dataset. [9].

Recently, a number of datasets have been implemented in this area to allow researchers to compare the performance of their algorithms or models with other existing algorithms / models. GREYC keystroke benchmark dataset[10], GREYC12 static keystroke dynamics benchmark Dataset [11], Si6 Labs keystroke rhythm Dataset [12], Beihang University static keystroke dynamics benchmark Dataset [13], GREYC- NISLAB keystroke dynamics soft biometrics Dataset [14], Sapientia University dataset [15], Clarkson University's dataset [16], RHU keystroke dataset [17] and CMU dataset [9]. Although these datasets help to have a common benchmark to compare with other algorithms, when any new features are added, the data for the newly introduced features must be collected. Studies have shown that not much work has been done using non-conventional authentication features, particularly static ones, and that different researchers have also used different features set to identify the user using keystroke dynamics. There was therefore a need for us to create our own dataset and then find the appropriate dataset features that could be effective in uniquely identifying the user.

## IV. PROPOSED METHODOLOGY

As mentioned earlier, the majority of the research for authenticating a user using a static analysis of keystroke dynamics focuses on time-based features. However, it is our observation that people have a particular style of using a keyboard. Some people always use Caps Lock for typing capital letters, while some use Shift Key. Also, there are people who shuffle among left and right Shift Keys while typing. In addition to the personal choice or convenience of using these special keys, the user may also depend on the password that is to be typed. For this reason, we drafted a password policy which is as follows:

The length of the password should be of at least "10" characters. It must contain at least one uppercase character and a special character (! @,#,$,%,&, etc).

Sample combination of password following the password policies is shown in Table I.

**Table I: Password Table**

| Description | Password(s) |
|---|---|
| Keeping the uppercase character and the special case character together. (Any order) | I@ndia2019, indiA@2019, indi@A2019. |
| A number of uppercase characters together | INDia@2019, inDIA@2019 |
| Alternate upper and lower case characters | iNdIa@2018, InDiA@2018 |
| A number of special characters | oM@nySpec!@ls |
| Start with numbers | 2018India@, 18India@19 |
| Continous switching between right and left part of the Qwerty keyboard | PaleMapOW19@ |
| All characters on one side of the keyboard | DeafRex@12, PinKill!89 |
| Half upper-case Half lower case | HALFhalf@27 |

For example:
If a password contains uppercase character and the special case character together,
then the usage of Shift and CapsLock keys may be interesting and vary from user to user as for special character Shift key has to be used. For typing A, it may be interesting to observe which key the user uses in which scenario CapsLock, Left Shift or Right Shift Key.
If the password contains consecutive uppercase characters such as INDia@2018, inDIA@2018, then in such cases the user may use Caps Lock or Shift key or combination of the two keys to type consecutive capital characters.
If the password contains alternate upper and lower case characters such as iNdIa@2018, InDiA@2018, number of special characters such as SoM@nySpec!@ls, the possibilities of the users to use Shift Key increases in such scenario is an assumption. It will be interesting to see what the typing pattern of an individual is in such scenarios.
If the password contains characters where there is continuous switching between right and left part of the QWERTY keyboard such as PaleMapOW19@, in such case, the user typing behavior may be different as compared to usual behavior. As the user may use Right Shift Key when using the right section of the keyboard and Left Shift Key when using the left section of the keyboard.
On similar lines, it is observed that a set of people regularly use the number keys from top whereas some people always use the number keys associated with Num Lock. These observations lead that different keys like Caps Lock, Shift or a different set of number keys can be used as a feature set for keystroke dynamics. These are considered as non- conventional features as these are not capturing time, related to key press and release. In our proposed method, the key usage data of every individual will be captured. Timing features and Non-conventional features will be calculated with the help of the captured data for every user. Following are the features considered:
A. Timing Features:
Hold Time is the difference between the keypress and key release [18].
Up-Up Time is the difference between two successive key releases [19].
Up-Down time is the difference between a key release and the next keypress [19].
Down- Down time is the difference between two successive key presses [18].
Down-Up time is the difference between a keypress and the next key release [18].
The collected user samples will consist of a variety of features that increases the computational complexity. All of these features may not be useful and some of them will be irrelevant while forming a training set.
In order to keep the computational complexity under control, a module can be implemented that identifies user-specific features that are more crucial than the other features.
The identified relevant features will form a training set that can be used as a reference during the authentication phase. During the authentication/testing phase, the typing rhythm of the given input sample from the user will be compared with the reference template of the training set to verify the user.
B. Non-conventional Features:
As described earlier other than timing features, other keys usage of the user is also captured.
Shift Key Usage (Left/Right): As can be seen, every individual has different habits of using a keyboard especially while typing capital characters or special characters. The variation includes usage of CapsLock, Left Shift Key or Right Shift key. This usage in a particular scenario is captured for every individual repetitively to understand the user's keystroke behavior. This feature will capture user habits of using the Shift key [20].
Caps Lock Key Usage: This feature captures user's habits related to the use of CapsLock [8].
First key release occurs after the second key release: This is applicable only if the user uses the Shift key. For example, this situation occurs when the user uses Shift key and then presses any other key, and releases the Shift Key before releasing the other key [8].
First key release occurs before the second key release: This is applicable only if the user uses the Shift key. For example, this situation occurs when the user uses Shift key and then presses any other key, and releases the Shift Key after releasing the other key [8].
Number Key (Top/Right) Usage: There are two different aspects of a user's habits related to Numbers Key usage. Some users always use only the top panel whereas some always use the right panel of the keyboards and some alternate between the two.
This feature captures this aspect.
The proposed workflow model is depicted in Fig. 2. Like any machine learning application, it consists of two phases: (1) Training Phase and (2) Testing Phase.
In the training phase, the keystroke data of every individual is captured, through which the timing features, as well as proposed non-conventional features, are extracted. The user profile of every individual is then created using both the timing and non-conventional features.
This data set is then used to train the system for a particular user, for a particular key phrase. This profile is used to register the user behavior.
In the testing phase, user keystroke data is taken at run time, through which the timing features and non-conventional features are extracted.
These features set is then tested for Error to marginal value with the identified pattern for that user. If the marginal value is in the accepted range then the system predicts authorized/unauthorized users with accuracy after comparing it with the registered user behavior data set. But if the marginal value is not in the accepted range the features extracted are sent to the training phase.

## V. EXPERIMENTATION

The user keystroke timing (in millisecond) is captured when the user enters the given password. During this timing

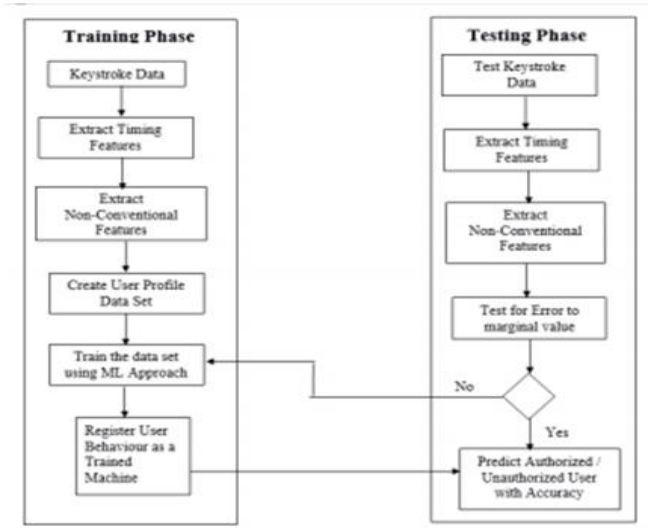



**Fig. 2: Workflow for Keystroke Behavioral Analytics as a Support Factor for authenticating user**

period, consistency in their typing rhythm is checked. During the collection of individual password entries, it is ensured that the entry is not an outlier. The outlier is the value that deviates largely from the rest of the values taken from the user. The first three entries are considered to be genuine. Thus, if the first three entries are not close to perfect, the entire dataset will be affected. Entries after third are stored only if they are not outliers. The outliers are detected using the Nearest Neighbor algorithm. The nearest distance of each previously recorded training sample is calculated. Three nearest points are considered for experimental purposes. Thus, we get n such distances. The average value of n distances recorded is chosen as a threshold value for incoming training samples. Only if the Kth nearest distance of incoming training sample is less than or equal to the threshold value, it is considered a genuine training sample.

In order to build the training set, data were collected from 30 users where each user was asked to type 'India@2018'as a password for 15 times. During this period their consistency was checked as mentioned above. The user needs to first register. The register page consists of only a username and password. Once registered the CSV file of the user is created.

The CSV file contains the timing features of the user for the first time. The user after registration needs to login for at least 10 times so as to create a training set that can be used for authentication. Every user must make sure the same username and password is entered each time during the training period.

**Fig. 3: K best Features selection with different value of K's**

Feature selection is a technique used to choose those features within the dataset that contribute most to the target variable. In other words, we choose the best predictors for the target variable. K-nearest algorithm was chosen to select K best features from sample sets such that the features will help to improve accuracy scores and can also boost the performance on very high-dimensional datasets. The only limitation with the K-nearest algorithm is, it asks for K to be defined before running the algorithm on the set. To decide the value of K, experimentation was done by assigning different values to K, for observing which value of K will yield good results. The result is summarized in Fig. 3. As shown in Fig. 3, for every user with a different value of K, different digraphs features set is more appropriate as compared to all the digraphs of the key pressed and released by the user for typing the password 'India@2018'. Table II shows the contributing features set when k=15 for users 'darshan' and 'Deepshree' while typing password as 'India@2018'.

**TABLE II: Two users with different Contributing Features Set for the same typed password India@2018.**

| Users | Contributing Features when k=15 |
|---|---|
| Darshan | I.n.UpUp, i.a.UpUp, a.shiftLeft.UpUp, @.2.UpUp, I.n.DownDown, a.shiftLeft.DownDown, @.2.DownDown, I.n.DownUp, a.shiftLeft.DownUp, @.2.DownUp, I.n.UpDown, i.a.UpDown, a.shiftLeft.UpDown, @.2.UpDown, and 0.1.UpDown |
| Deepshree | I.n.UpUp, a.shiftLeft.UpUp, @.2.UpUp, I.n.DownDown, a.shiftLeft.DownDown, @.2.DownDown, I.n.DownUp, a.shiftLeft.DownUp, @.2.DownUp, I.n.UpDown, n.d.UpDown, i.a.UpDown, a.shiftLeft.UpDown, shiftLeft.@.UpDown and @.2.UpDown |

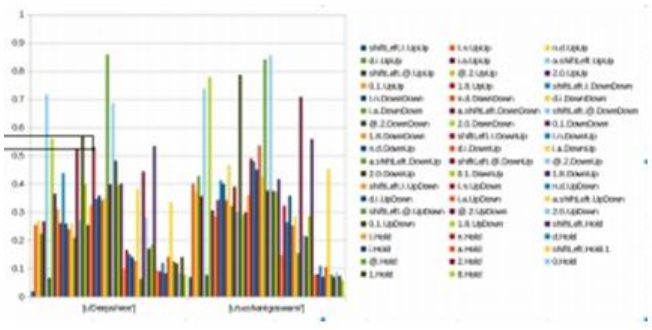


**Fig. 4: Timing Diagram of conventional features of two Users typing the same password at different times.**

The number of features and also the set of features extracted for user 'darshan' and 'Deepshree' are different and thus these features set can be used for training instead of using all the digraphs of the password length.

It has been observed that two different users typing the same password can be differentiated through their type patterns as represented in
Fig. 4. It can be seen from Fig. 4 that two users typing the password India@2018"have some variations in their pattern with the press and release of the Shift key and Numbers key press. Finding usage patterns of non-conventional features can become additional support along with timing features to identify the user. As can be seen from Fig. 5, two users have completely different usage patterns of non-conventional features. Thus, if the typing pattern of the non-conventional features is recorded, it is expected to be more supportive to differentiate the users.

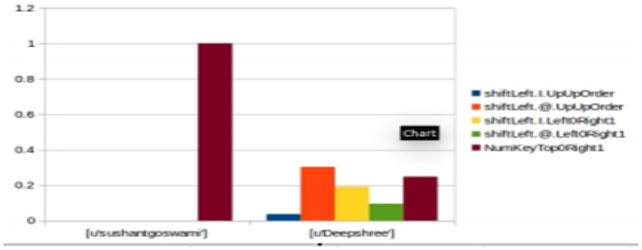



**Fig. 5: Timing Diagram of non-conventional features of two User typing the same password at different time.**

For checking the performance of the K best features algorithm linear regression, Gaussian Mixture model, Support Vector Machine and Random Forest algorithm were used. The features identified in this process are timing features. The results for K as 15, 20 and 30 are summarized in Table III. It can be seen that for K = 15 Support Vector Machine works best. The non-conventional features classification was carried out using different algorithms such as Naive Bayes, Random Forest, Decision trees and Logistic Regression. It was observed that Logistic Regression gives less error count with highest accuracy of 90.50% as compared to other algorithms for classifying non-conventional features as shown in Table IV. Lesser the error count more is the accuracy.

**TABLE III: Accuracy of K best selected features using different classification algorithms**

| Algorithms | Accuracy(K=15) | Accuracy(K=20) | K=30 |
|---|---|---|---|
| Linear Regression | 70.83% | 69.16% | 65.22% |
| Gaussian | 75.75% | 69.24% | 74.39% |
| Random Forests | 77.50% | 76.96% | 80.68% |
| Support Vector Machine | 81.14% | 79.22% | 78.97% |

**TABLE IV: Accuracy of different classification algorithms for Non-Conventional features**

| Algorithm | Error Count | Accuracy (%) |
|---|---|---|
| Logistic Regression w/ CV | 41 | 82% |
| Logistic Regression | 22 | 90.50% |
| Naive Bayes | 64 | 72.41% |
| Decision Tree | 50 | 78.44% |
| Random Forest | 45 | 80.60% |

**TABLE V: Results of Proposed Approach**

| Type | FAR | FRR | EER | Accuracy |
|---|---|---|---|---|
| Timing features | 0.5 | 0.03 | 0.26 | 64.89% |
| Non-conventional Features | 0.4 | 0 | 0.23 | 67.36% |
| Both Timing (60%) and Non-conventional features (40%) | 0.07 | 0.1 | 0.08 | 91.48% |
| Both Timing (80%) and Non-conventional features (20%) | 0.12 | 0.13 | 0.22 | 87.23% |

Thus, the user profile was created and trained by using the extracted relevant timing (conventional) features and non-conventional features, as discussed above. The user profile was created by using K best algorithm with the value of k=15. The timing features were trained using SVM algorithm and non- conventional features were trained using Logistic Regression.

## VI. RESULTS

The results of our proposed approach were evaluated on the data collected from 30 users. To evaluate the performance of our proposed approach, we evaluated four different kinds of results. First, only the timing features of the users were trained using an SVM classifier. Second, only the non-conventional features were trained using Logistic Regression. Third, 60% of the timing features and 40% of non-conventional features were trained using an SVM classifier. Fourth, 80% of the timing features and 20% of non-conventional features were trained using SVM classifier. The results are shown in Table V.It can be seen from Table V FAR, FRR and EER all improve. Equal Error Rate (EER) is lesser when Timing and Non-conventional features are used together. It is thus evident when both the timing features and proposed non- conventional features are used to authenticate a user using keystroke dynamics, the accuracy can be improved drastically.

## VII. CONCLUSION AND F UTURE W ORK

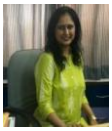
Keystroke dynamics is used as a support factor for authentication. This paper mainly focuses on the timing and non-conventional features for static keystroke dynamics where the pattern is identified in a given phrase. Currently, static keystroke dynamics use only timing features and continuous keystroke dynamics use both timing and non-conventional features, to model the user typing behavior. In this paper, we proposed to use non-conventional features along with the conventional time-based features for user identification for static keystroke dynamics. The experiment was carried and it was observed that the combined use of timing and non- conventional features improves FAR, FRR as well as ERR.

## REFERENCES


1. A. Jain, A. Ross, and U. Uludag, "Biometric template security: Challenges and solutions,"in Signal Processing Conference, pp.1-4, 2005.
2. S. Mondal and P. Bours, "Combining keystroke and mouse dynamics for continuous user authentication and identification, "in Identity, Security and Behavior Analysis (ISBA), IEEE International Conference, pp. 1-8, 2016.
3. N. Raul, R. Shankarmani, and P. Joshi, "A comprehensive review of keystroke dynamicsbased authentication mechanism, "in International ConferenceonInnovative Computing andCommunication ICICC, 2019.
4. Y. Zhong and Y. Deng, "A survey on keystroke dynamics biometrics: approaches, advances, and evaluations," in Recent Advances in User Authentication Using Keystroke Dynamics Biometrics. Science Gate Publishing, pp. 1-22, 2015.
5. P. S. Dowland and S. M. Fumell, "A long-term trial of keystroke profiling using digraph, trigraph and keyword latencies,"in Security and Protection in Information Processing Systems, ed: Springer, pp. 275-289, 2004.
6. D. R. Gentner et al., "A glossary of terms including a classi fication of typing errors," in in Cognitive aspects of skilled typewriting, ed: Springer, pp. 39-43, 1983.
7. J. Roth, X. Liu, A. Ross, and D. Metaxas, "Investigating the discriminative power of keystroke sound, "in IEEE Transactions on Information Forensics and Security, vol. 10, pp. 333-345, 2015.
8. A. Alsultan, K. Warwick, and H. Wei, "Improving the performance of free-text keystroke dynamics authentication by fusion,"in Applied Soft Computing, 2017.
9. K. S. Killourhy and R. A. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics,"in Proc. IEEE/IFIP International Conference on Dependable Systems Networks (DSN), 125–134 , 2009.
10. R. Giot, M. El-Abed, and C. Rosenberger, "Greyc keystroke: a benchmark for keystroke dynamics biometric systems,"in Biometrics: Theory, Applications, and Systems, IEEE 3rd International Conference, pp. 1-69, 2009.
11. C. R. R. Giot and M. El-Abed, "Web-based benchmark for keystroke dynamics biometric systems: A statistical analysis,"in Intelligent Information Hiding and Multimedia Signal Processing, Eighth International Conference, pp.11-15 , 2012.
12. L. Bello, M. Bertacchini, C. Benitez, J. C. Pizzoni, and M. Cipriano, "Collection and publication of a fixed text keystroke dynamics dataset," in XVI Congreso Argentino de Ciencias de la Computaci´ n, 2010.
13. Y. Li et al., "Study on the beihang keystroke dynamics database,"in Biometrics, International Joint Conference on, pp. 1-5, 2011.
14. S. Z. S. Idrus, E. Cherrier, C. Rosenberger, and P. Bours, "Soft biometrics database: a benchmark for keystroke dynamics biometric systems," in Biometrics Special Interest Group (BIOSIG), IEEE Conference, pp. 1-8, 2013.
15. M. Antal et al., "Keystroke dynamics on android platform,"in Procedia Technology, Vol. 19, pp.820-826, 2015.
16. E. Vural, J. Huang, D. Hou, and S. Schuckers, "Shared research dataset to support development of keystroke authentication," in Biometrics, IEEE International Joint Conference on, pp. 1-8, 2014.


## AUTHOR'S PROFILE

**Nataasha Raul** is pursuing PhD in Technology from Mumbai University. She has done B.Tech in I.T. and Master of Engg. in Information Technology. Working as an Assistant Professor in Sardar Patel Institute of Technology, Mumbai from last 13 years. She has published 20 papers in National / International Journals and Conferences. Her area of research and interest is Security, Machine Learning and Algorithms.



**Dr. Radha Shankarmani** has pursued her Ph.D. (CSE) at JNTU, Hyderabad Master of Engineering in Computer Science, NIT (Formerly REC),Tricky. Bachelors of Engineering in Electronics and communications, P.S.G. College of Technology, Coimbatore, Tamil Nadu. She has published any papers in National / International Journals and Conferences. Her area of interest is Database Management Systems, MIS, Software Engineering, Simulation modeling, Software Testing, Software Architecture and OOAD, Software Project Management, Data Analytics, Business Intelligence and e-Business.

**Dr. Padmaja Joshi** is working as Senior Director at C-DAC, Mumbai. She is an active member of various national committees such as Enterprise Architecture for e-Governance (IndEA), Strategy and planning committee for Blockchain Technology, State Committee on Enterprise Architecture, etc. She also has contributed a chapter on security in IndEA. She is the project lead of the national authentication solution with single sign on facility in India. She is a known speaker on technical subject like e-Authentication, blockchain, cyber security etc and has published papers in international conferences and journals. She has started a workshop on "International Workshop on Reverse Engineering (IWRE)" and a track on "Mobile Software Engineering "in International Software Engineering Conference. Her areas of interest are blockchain technology, e-authentication, Fraud management, Keystroke dynamics, e-governance, object oriented technology, Mobile Cloud Computing, Mobile Software Engineering, and cyber security