LOVELY
PROFESSIONAL
UNIVERSITY

User Authentication By Using Advanced Keystroke Biometrics

A Dissertation Proposal submitted

By
Deepika Sharma
To

Department of Computer Science

In partial fulfilment of the Requirement for the
Award of the Degree
of
**Master of Technology in
Computer Science and
Engineering**

**Under the guidance of
Ms. Urvashi Garg**

Ass. Professor,
Computer Science Engineering domain
School of Civil Engineering

(April 2015)

ABSTRACT

Security is important issue these days. Every system or transaction requires security. As they are more prone to unauthorized access. One way to protect these sensitive objects is Username-Password mechanism. But with development in technology the Username-Password used to give security is also not so securable. Because one can easily find the password using shoulder surfing and brute force method. So to make it strong one technique is introduced i.e. keystroke biometrics. This technique depends on user's typing style. In this paper, the new technique is introduce for keystroke biometrics which is t-test and typing features considered are key hold time, key inter time, up and up time, total time etc. We work on the mean and standard deviation of the features that give us the better result. The main consideration of this paper is to reduce TYPE 1 error to 0 i.e. the decision is wrong and person is genuine and in our research area it is in the way that the genuine user never reject by keystroke system to access his account which is main problem in keystroke research area till now. With this t-test we are able to reduce TYPE 1 error to 0%.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during thesis work.

Firstly, I would like to thank my supervisor Ms. Urvashi Garg for being great mentor best adviser I could ever have. Her advice, encouragement and critics and innovative ideas, inspiration are cause behind the successful completion of this dissertation. I am highly obliged to all faculty members of computer science and engineering department for their support and encouragement. I would like to express my sincere appreciation and gratitude towards my friends for their encouragement, consistent support and invaluable suggestions at the time I needed the most.

I am grateful to my family for their love, support and prayers

Deepika Sharma

Reg.no 11311436

DECLARATION

I hereby declare that the dissertation proposal entitled, User Authentication By Using Advanced Keystroke Biometrics submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: 1-5-2015 **Deepika Sharma**

Reg no: 11311436

CERTIFICATE

This is to certify that ____Deepika Sharma_____ has completed M.Tech dissertation proposal titled ____User Authentication By using Advanced Keystroke Biometrics_ under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfilment of the conditions for the award of M.Tech Computer Science & Engg.

Date:

Signature of Advisor

Name: Urvashi Garg

UID:

TABLE OF CONTENTS

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1
INTRODUCTION

With increase in technology the use of computer in every field is also increased. As the demand of computer increases day by day, the security issue is also arises and also it becomes important issue to provide security to individual account access. Everyone wants that their systems, personal accounts are secured from hackers or imposter like in online banking. To make the system secure various username password facility is provided to each user. But now a days the password protected system also face problem that they are also hacked by imposters easily either by making a strong guess or by reentering the different passwords. So all this required a technique that can protect the system from unauthorized user.

The access of global information as well as resources becoming vital part of our lives. and this access increases day by day. Also the present generation uses technology for every aspect of their life. They totally dependent on technologies because it makes life simple and more convenient but this increase also give chances to malicious attacks and intruders to perform some hazards. Our private information is also suffers from risk. So it is mandatory to secure these resources and information from intruders. For this purpose we require some low cost technique which verifies user’s identity. To protect from these attacks password is used. Unfortunately, now a day’s only password is failed to secure the applications. so there is need of some special techniques.

Authentication is the way of ensuring the identity of someone. It is the way of ensuring the truth of data. It is method of conforming that someone who try to access one’s account, is that someone or something really meant for it. Authentication can be provided to someone by various methods. The actual motive of providing authentication to someone is for security purpose that who so ever authenticate only that person can access his personal data. Mostly the authentication is provided by password that is self generated by person. But now with the advancements of technology password is also stolen by hackers and these results in hazards. So there is requirement to protect password with some other techniques that describes actual physical identity of person like his thumb recognition, eye recognition, typing speed and many more.

Authentication includes three factors:

- **Knowledge:** The thing that user knows is knowledge like password, personal identification number, security questions.

- **Ownership**: the thing that owned by someone or which belongs directly to user like identity card, token etc.

- **Biometrics**: something done by user or that user has like voice, typing pattern, facial recognition, retina recognition, signature etc.

All of the above methods have some drawbacks like *password* is however a cheap method but it can be stolen easily where *tokens* are expensive and also are difficult to handle so for user authentication every enterprise prefer *biometrics authentication* because it gives the actual physical identity of user.

Biometric authentication measures the physical characteristics of user. This is used in computer field for authentication purpose. Biometrics are used to portray the individuals. It is used for identification purpose and also it restricts the access control to a particular person by conforming human characteristics. Today all enterprises uses this type of authentication for security purpose like banking sector, IT industries, medical field, research organizations. However some of the biometrics techniques are expensive because of their hardware but as they provide high level of security organizations are ready to apply these techniques.

Biometric characteristics are of two types:

First characteristics is physiological that comprises of palm and retina analysis, facial, DNA, iris recognition. These characteristics are physically related to users. The things that directly belongs to users.

Second characteristic is *behavioral characteristics* that include Keystroke dynamics also called typing pattern, voice and gait. These characteristics results because of user's personal behavior or routine.

In this research we are interested in behavioral characteristics

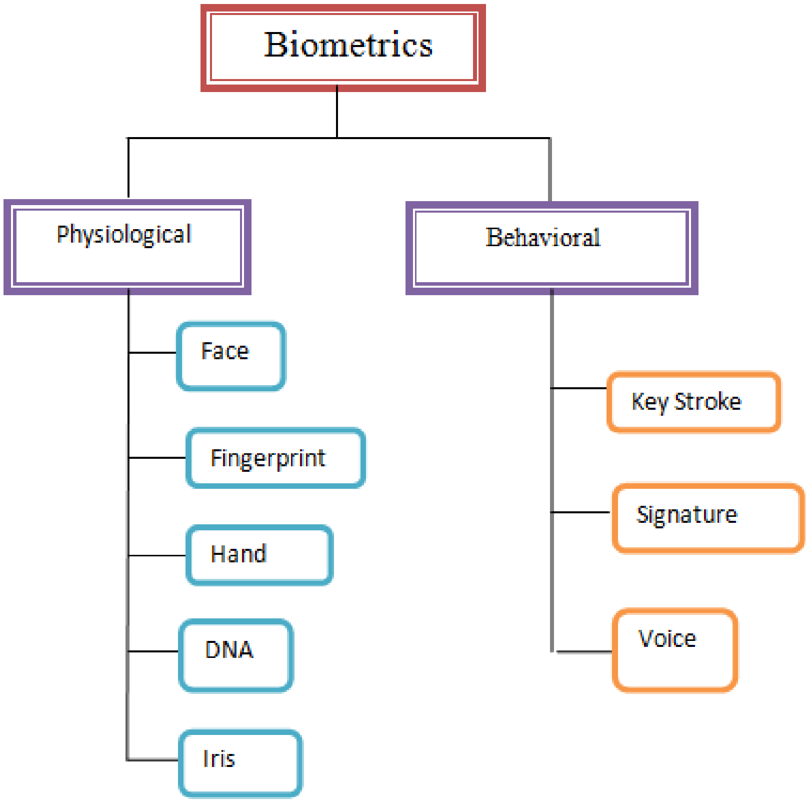


Fig.1. types of biometrics

Biometric system has two modes:

- Verification
- Identification

Verification: In Verification mode the system do one to one judgment of captured biometric data with existing data template in database for verifying user identity.

Three steps involved in verification process:

In first step the reference template for each user are calculated and stored in database.

In second step some samples are matched with reference template in order to find the genuine and imposter user and store in database.

Third step is called comparison or testing phase. In which user personal identity is used to check which template is used for authentication.

Identification: in this mode system perform one to many judgment against the template in order to find the individual's identity. The system is successful in identify user if value of comparison fall under the limit of threshold.

Identification mode can be used for both negative recognition and positive recognition.

Where verification mode can only use for positive recognition. In positive recognition there is no need to provide information by user. But in negative system has to established the identity of user.

Keystroke Dynamics:

As I have chosen the keystroke dynamics as my research area from behavioral biometrics. The reason for choosing it is this technique is used to provide user authentication. It uses typing style on keyboard to identify the authenticate user. Key stroke recognition is the technique used to protect the system by permitting imposters to enter into system. The system based on keystroke dynamics captured features regarding pressing of key and also releasing of key. Keystroke biometrics works with one basic hypothesis that every person has its unique typing style. No one can copy that style as it is natural one. So that's why it is use to protect the system. These systems are totally based on the typing speed of users. It is data processing technique used to analyses the way of user typing. This technique works by integrating the keystrokes features in existing password and record all the features for original user and store into database as a unique template and if any imposter unfortunately get login information, he/she try to login into the system with this login information then our keystroke recognition system denied their access by comparing all features with existing features.

The advantage of using keystroke recognition system is that it is cheapest method as it does not require any additional hardware. In this system user has to fill password, when he presses the key all keystroke features are recorded and compared with previously recorded features and then user is allowed to enter into system. All these features help us to distinguish between the different users. Another benefit is the error rates can be adjusted at level that imposter cannot get access control.

If we compare Keystroke with physiological characteristic then it may happen that someone can forced to person to violate the security by taking his fingerprint access. But this never happen in keystroke recognition as it checks typing rhythm of user which varies from person to person. As a example suppose there is a employee in bank uses computer and make any transaction and then left the computer as logged in. at the same time someone else start using his access but with the help of Keystroke recognition we avoid these type of threat as its basic concern is on matching typing pattern of genuine user with imposter.

Advantages of Keystroke Biometrics:

- Individuality: Each and every user has its own distinct typing style. Because of this uniqueness no one can copy another person's typing style for which it can be used to measure the identity of user.

- Cheap method: As this is only one biometrics technique which is lowest in cost. Because it does not require no extra costly hardware. It can implement easily by using our keyboard and software that run as a backend for it.

- Lucidity and easiness: This is very simple and easy to use and beneficial for that user who does not have any practical knowledge as this method does not require any practical knowledge by user side.

- Copying avoidance: As typing style is unique for each user. So it is difficult to copy typing style of another person. In this way it avoid replication of typing patterns.

Process of keystroke recognition:

When user starts using the biometrics system the first phase the user phase is enrollment. During enrollment phase the biometric information is captured, processed and then stored in database. This is one of the interfaces between the real world data and system. The system interacts with the user and captured data from user and stored into database. The care should be taken while capturing data.

Enrollment phase consists of:

- Capture data
- Process data
- Stored in database

Next phase is verification phase. In this phase a new user come and start using system. The system again capture data from user, processed that data by using some technique. Then comparison function performs between these new data template and existing data template. Here a technique is implemented that can compare data.

In last decision should be made whether data matched or not matched.

Similarly verification phase consists of :

- Capture data
- Process data
- Comparison
- Decision

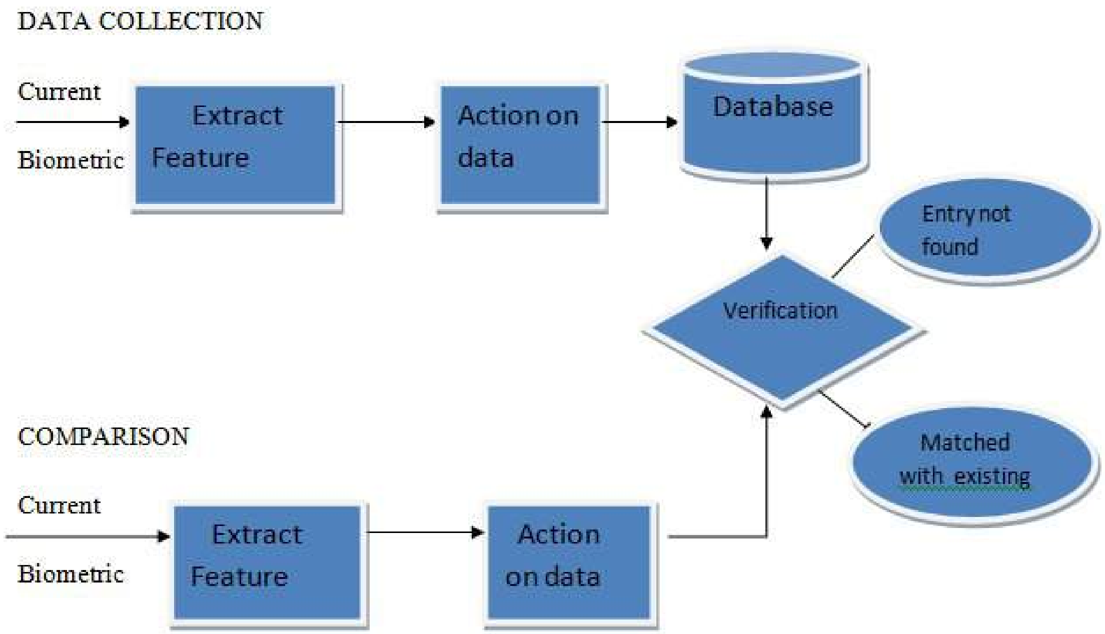


Fig.2. Process of keystroke

Chapter 2
Review of literature

First problem is stated in broad general way. Once the broad area selected, start survey in that field to decide the problem definition .review all the researches done in this field and single out the problem we want to study. After deciding the problem area surveys each research papers, journals, magazines available on the problem which we want to study. This survey helps us in narrow down the problem. After reviewing the literature and seeking their future work we decide our problem.

As for my dissertation, I review all the research paper in Data Mining field first and then I decide my problem area is “Key Stroke Recognition”. After Deciding Problem area I start all the literature survey regarding my Problem and decide the future work.

The Research Paper which I read is listed below:

In [13] this paper they used representation, extraction and classification for keystroke data. For data extraction they developed C++ toolkit for data analysation This toolkit serves as front end for main recognition engine. It is helpful in predicting system behavior and for producing graphical results for both Matlab and Gnuplot system. This toolkit divided the all users into different group by using clustering approach like intuition. Then factor set is used to determine the feature set on the basis of similarity in features.

And for classifier they use

a) Euclidean distance measure. The distance is calculated between the reference pattern and claimant pattern.

b) Then non weighted probability is used to find the score.

c) then for data that has high frequency of occurrences are measured by using weighted probability. The score is again calculated between known user and new one.

Thus in this paper they show that result rate is higher with weighted probability s compare to other two

In [21] they express time when key is down and time when key is up and the code of key.When user typing the string all these features are collected. They used totally four features and performed seven experiments. The conclusion of the experiment was measured by three kinds of users imposters, legitimate user, and observer imposter user.

Features used:

- Code of key pressed
- Latency of keys
- Duration

An implementation of this paper is as follows :

Whenever user try to login into system he has to type a string. When user start typing all features are calculated side by side, sample is created and then stored into database. If it is found that there is new account found then fresh template is made. But if account is found as existing one then system check the sample and with classification the decision should made that whether the user is imposter or legitimate. And if the user is exposed as imposter then another trial given to user.

In second turn if again system decided that sample is not matched with exising one then user is treated as unauthenticated.

So finally there are three basics trials in this paper:

- Successful trial
- First Failed attempt
- Second successful attempt
- Or two failed attempt

And in last one adaptation method is used for constructing new template and update it in database.

Experiments of this paper can done on three type of user:

- Genuine user
- Imposter user
- Observed imposter user

Result of this paper shown that type 2 error is 1.89% and type 1 error is 1.45%.

In [18] “BeiHang Keystroke Dynamics Authentication System” they represent a embedded system that provide password protection. It gather two public databases and provide results for three classification algorithm i.e. one class support vector machine, Gaussian classifier, nearest neighbour classifier. In this paper they present two keystroke dynamics system. One is for internet which is already used in China. Other one is embedded system in which they use simple electronic device for purpose of keystroke dynamics.

Keystroke dynamics for internet: in this first they discussed about kernel based method that how they used to capture the keystroke data. Then they discussed about HOOK method. And in last they discussed about web browsing method. And after this they give information about their proposed method that include three parts : kernel based keyboard driver, second is pre-processor and third part is authenticator.

Keystroke dynamics for embedded system: this is only the part of entire device. For feature extraction this paper introduce three transformation method like Fast Fourier Transform , Discrete Cosine Transform and Gabor Wavelets so that performance of system get increased

Beihang Database description:

There are two databases that called Database 1 and Database 2.

Database 1 is collected online where database 2 is collected from embedded system.

There are total 10 subjects that collect registration data from genuine user and use to train the model, sign in data from imposters, sign in data for legitimate user.

Then for final result they used fusion of features and classifier. In this paper result is shown in the form of ROC curve.

In [20] YogeshMeena, UrvashiGarg authors of User Authentication using Keystroke Recognition continued the previous work[6].they proposed a technique by using keystroke for providing user authentication. In this paper key hold time inter key time and other constraints used to verify the user. When user start type the password all features are calculated and matched with existing value in database by finding Euclidean distance and at last token system is used. In their results they reach type 2 error 0% but type 1 error 30%. As type 1 error tells that genuine user is not verified and he is not able to access his account and type 2 error states thet unauthorized user is allowed to use resource.

Features used in this paper:

- Inter time
- Key hold time
- Key type change time
- Number of backspaces
- Number of shifts
- Number of caps

Result of this paper is they get type 2 error 0% but type 1 error only 30%.

However this can be further improved by minimizing type 1 error and makes its accuracy levels more high.

In [8]his paper he uses fusion approach to provide authentication. Firstly he collects four types of Keystroke Features .Then convert them into score by using technique that mentioned in it

Features are:

- Time duration
- Inter key time
- Keystroke Pressure
- Typing sound
- Typing pattern difficulty
- Rate of typing
- Linguistic style
- Frequency errors

Methodology used:

In their methodology part they have two parts:

A. Matching

B. Fusion approach

In this paper first they calculate four type of features that discussed above from dataset then they find the similarity score by using GPD and DSM then in last for unite the score value they use fusion rules and make a decision.

Finally result of this paper shows that after combining three fusion techniques and keystroke features they get result of 1% EER.

Also they show the comparison of two technique and the results that come out are the feature D1 give better result with GSM. When D1 is combined with other features then this combination give better result as compare to any other combination without D1.

They also show that pairing of four features does not provide any advantage as given by group of two features. So overall they provide all combination for better result.

But there are also some drawbacks like some of the feature they discussed are not valid.As formulating pressure feature there is need of separate hardware which is pressure sensitive r that negates the main advantage of keystroke. Errors frequencies, typing difficulty are rarely practical as well as there is concern of noise in recording sound of typing.

In [10] they discuss problem of identity theft and provide the solution for that:

Features included:

- Duration of pressing key
- Order of proportional key.
- Speed of typing
- Shift key

Methodology used:

A. Feature acquisition

B. Feature extraction

C. Classifier

D. Signature database

In this paper they show the survey of all the techniques that are used in computers to secure it from theft .

First they describe the deployment method. for this they give the two types of verification that is login and continuous that can be provided at three levels

- Host
- Web- browser
- Client server

Then they show the challenges for keystroke biometrics in computers. These are:

- Collection of data
- Hardware that are used
- Changes in users mood
- Privacy
- Scalability

Then they show the future work that use of behavioral biometrics in mobile phone.

So in brief in this paper they show problem of theft in identity. And they provide many verification solutions like mouse and keyboard dynamics. Also they present different deployment configurations.

By proposing four new metrics and using statistical approach on these they were able to analyze keystrokes of 14 characters long for each user. After analyses the similarity metrics for generating decision table. With all these things they reached at 97% accuracy in results.

In [14]2012 this paper they introduce a new distance metrics which is succeed in finding outliers in keystroke data. also they present the comparison of this new distance metrics with the previous available distance metrics. This distance metrics made by combining the benefits of Mahalanobis distance and Manhattan distance to avoid the shotcomings of these two distances.

Firstly they use Mahalanobis distance to find identity matrix by normalizing keystroke features. This can fulfilled by using linear transform to keystroke data. Once the data is uncorrelated then Manhattan distance is computed. For this they used CMU keystroke benchmark data.

In this paper they show the limitations of using the Euclidian distance i.e.

- It is very sensitive to small variations.

- It does not deal with correlations between features.

Then they show briefly the advantages of Mahalanobis and Manhattan distance by using mathematical formula and graphs.

Proposed distance metrics:

After all this discussion they show the new distance metrics which they made by combining features of these two distances.

The reason for combining these two distances are also illustrate into this paper. The first distance i.e. Mahalanobis is good enough to deal with coorelations and variatios between features. But it is not good to deal with outliers.

Where second distance metrics manhattan is good to deal with outliers.

So for these two requirements they combine these two distance metrics so that only one distance deal with both problems lke outliers and correlations.

Classifier:

Then they use classifier for learning purpose. It can learn model for user ,reject outliers as imposter and accept inliers as legitimate user. The classifier which is used in this paper is nearest neighbor classifier with new distance metrics.this classifier take user as legitimate if distance of training data is below the threshold value. But if it is high then they user is not valid.

The result of this paper they get EER as 8.7% and error rate decreases to 8%

In [23] this paper they generate a new type of dataset which has both login and password i.e. imposed and according to user's choice. Also the dataset is collected in web-based uncontrolled environment.

The experiment of this work includes:

i) Acquisition protocol for acquiring features. In first session user has to enter the login of his or her choice. This process has three steps. each step requires that user has type login credential number of times. While typing user take care of one thing that he is not able to correct mistakes and if user press key backspace then field get blanked and he has to fill the string again.

ii) Next they have presentation of obtained dataset. The user they take to give them dataset belongs to 20-24 age.all the user give them data they require . the data should be real one.

iii) Further they include Experimental protocol which tells how the whole process works. EER is calculated for each person.they uses two common way to presenting results.

They used 20 samples for training and rest all are used for testing.First step in this protocol is computing distance which is based on Gaussian distribution assumption.
Each user has its own data and the system generate template for each user i.e. the mean and standard deviation values. Then distance between template and existing data is calculated.

Then they have Statistical validation for which they used Kruskal- Wallis test. This is one of the non parametric test, that can used for checking whether the sample belongs to same population or not. Or in general way this test is used to check hypothesis one is null which means sample belongs to population and another is alternative hypothesis.

iv) Last they used Score fusion to generate final results.

v) Then they present performance dependency on size of password. Here they discussed about whether the size effect the performance or not. For this they again use KW test. For this they calculate entropy of password .

The result of this paper is they increase the EER level from 10.01% to 16.09%.
In result section first they discussed about experiment for feature verification. For this they use KW test. This test give them result that there is no difference between data choosed by user or data imposed to user.

Next they show result regarding score fusion. In this they discussed about which feature give them better result. And from KW test it conclude that the worst result generated by pr feature.

And in last they conclude that use of all feature give better result. Because if they use only one set to perform experiment then they got worst result. So to get Best result they have to use all feature together i.e. pr,rr,rp etc.

In [11] their paper they classified Keystroke in Two classes : (a) based on Fixed text (b) based on Free text. The whole keys are divided into two parts one is left & other is right and then timing vector give differentiation of the legitimate user from imposters.
In this paper they show that the need of long string as they discussed that for free text based system taking small string is not enough for matching purpose. As the pattern that get matched are very small.

In their experiment they show the example of Intel password and they illustrate that flight time of two successive characters is more good result generator then dwell time. By this illustration they use flight time as main consideration in their work and also they use free text system where user has to enter the given string.

The experimental setup of this paper is as follows:
When user start typing text at login time the system generate time vector and parallel their flight time. Then Euclidian distance is calculated between two vectors to find the identity of user. When user is found as genuine user then his new timing vector is stored in the database for achieving better results

In [6] D. Shanmugapriya , Dr. G. Padmavathi are Author of research paper . This paper uses the feature of virtual Key Force along with commonly using Z-score method.

Features are:

- Dwell time
- Flight time
- Latencies
- Di Graph
- Virtual key force

Methodology used:

The entire methodology divided into three phases namely extraction phase use for capturing features , phase for normalizing, phase for selecting features. For selection phase they proposed machine learning technique. The name of techniques are ACO-ELM-ANP and GA-ELM-ANP.

An extraction phase use to extract the features that discussed in this paper from raw dataset. They use the password as a example that typed by 51 user about 400 times. And when user type all features is calculated.

After this there is turn of normalization phase in which z- score method is used to get data in predefined range.

Third phase is feature subset selections which identify more selective features. It reduces the differences in features which results in high accuracy and decreases the working out time. This phase is divided into two parts first is filter approach and second is wrapper approach. In first approach this phase does not use any learning algorithm. The second approach uses classification method.

Two techniques that are used in this paper are:

- Ant colony optimization : this technique is based on ant’s real behaviour that how they find shortest way from their nest to food and also they give marks to other ant to find the food source. In this way this technique works for many searching problems.
- Genetic algorithm: this is another searching technique. It calculated the fitness of all persons in one inhabitant. This technique generates new fitness and enters new results in database by using selection.

From the result of this paper it is observed that ACO-ELM-ANP selects less number of features for further processing. In this paper type 1 error increases if number of selected features increases.

The future work they introduce is to use Swarn intelligent technique in future work.

In [7] 2014Nandani Chourasia ,Author of “Authentication of the User by Keystroke Dynamics for Banking Transaction System” .

Features used:

- Flight time
- Press-press
- Press-release
- Release-press
- Release-release

In this paper they use Flight time and dwell time as a keystroke features dataset. The mathematical model in this paper is depend on the Euclidian distance find between features.

The design has two phases :

In first phase user has to registered himself for doing any banking transaction. For this user has to fill username and password ten times. At this time the threshold value is calculated and store in database.

In second phase user has to authenticate by system by finding Euclidian distance. According to variance of threshold value user has been provided the access of the account.

Then they show state transition diagram for design phase. And after this they describe hardware and software requirements for the research work.

They also described three types of user access:

i) Full access: this access is given to user if the variance is very less even nil.

ii) Partial access: this access is given to user if variance is not high then the standard value. And if this access is given to user then user only see the account detail and was not able to make any transaction.

iii) No access: this access is provided if variance is higher than standard value.

Result section

By using manhatton distance they show that if total number of attempts of false rejection rate are 860 then error rate is 22.32%

And if total number of attempts of false acceptance rate are 650 then error rate is 9.5%.
In case of Euclidian distance they show that if total number of attempts of false rejection rate are 860 then error rate is 6.62%
And if total number of attempts of false acceptance rate are 650 then error rate is 2.46%.

Finally in the end they show that Euclidian distance can give better result as compare to manhattan distance.

Chapter 3

Present Work

3.1 Problem Formulation:

This part show the problem on which we actually work. The process of defining a research problem area or particular problem is a basic building block of every research process. In my research work the problem formulation consists of minimize the type 2 error in previous work in all research paper by adding some advanced features in their work so that type 2 error minimize and original user always have a permission to access his account. Type 2 error is when user is right but system take wrong decision. In keystroke recoginition type 2 error occur because sometime user is not in good mood while typing the password and he type in slow spped and system calculate some wrong values for features. And treat valid user as invalid also not allow the user to access his own account .

This one of the big limitation of keystroke which I want to remove totally or reach it to 0%. So that whatever condition user type the data he always able to acess his own account.

The advanced features that I add to that paper are total time in typing, up and up feature, detection of shift whether it is left or not and Backspace and some previous used features

3.2 Scope of the Study

The study consists of survey in the area of Key Stroke Dynamics. This is the method to provide authentication to user. The scope of study is to find the problem in this area which is not solved by any researcher till now. The study covers all research papers, magazines and journals that give best knowledge about my problem area. This study helps to narrow down the problem. The scope of study related to the finding more and more information about keystroke recognition and explore something new in this area to improve the previous work. The study includes surveying research paper, their future work, concentrating on their conclusion, and generates some accurate conclusion from whole survey that helps in further research work. Till now so many researches are done in this field and researchers were trying to provide best solution in there research. The main focus of research is on removing type 1 and type 2 errors which are two main issue of security. But till now no one is able to reduce type 1 error completely. However researcher had got success in reducing type 1 error completely. So scope of this study is to read all researches that had done also study all the technique that used and then decide how to solve this problem in better way with some new features and new techniques.

3.3 Objective of the Study

The Objective of present study is to collect all information regarding the problem which I want to solve. To present the literature survey. Study the detail of area to narrow down the problem. To mention the all research paper those are surveyed in order to gather the relevant information and also mention the different journals and conferences in which the papers are published. To provide a brief comparison between different researches that has done in my problem area. To define a research work that has been done later. To elaborate the methodology used for implementing that research work. To estimates the possible outcomes from research. The main objective is to achieve a new finding or research in this area.

The main objective is to find new features in keystroke dynamics that are captured during typing of user. The type 1 error and type 2 error are computed on the basis of these features .Then the methodology used that used to compare new user's typing feature with existing one. So objective of this study is to find such a methodology and also present that methodology.

3.4 Research Methodology:

3.4.1 **Keystroke Feature Extraction**:

When a person types something by using keyboard then he has some unique feature associated with his typing style. Some person takes more time in holding key where some takes more time to release the key. All these features vary from user to user. So the basic building block of keystroke is getting these features:

First step is to decide some keystroke features that I have to include in my research:

- *Key hold time*: When user types something he holds the key for some instant of time and then release. We consider that time as a feature so for this we calculate the hold time of each user on the basis of our hypothesis that each user has different holding time.

  In other words the time taken between the pressing of key and then releasing that key is called key hold time or dwell time.

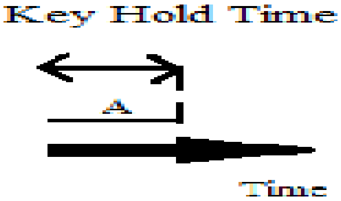


Fig. 3. Key hold time

- *Key Inter time*: We consider this fact that each user has different typing style. So inter time also vary from user to user. Some user release the key immediately and press next one. Where other users take time to release the key. All these difference helps us to calculate different inter times for user. Thus the time taken between one key is released and other is pressed.

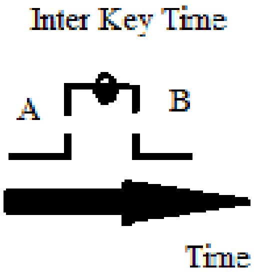


Fig 4. Inter key time

- *Up and up time*: This is the new feature introduce in the keystroke dynamics. It depends on two releasing factors. It calculates the variation between when user release first key and release second one.

  Thus it is the time taken between first key is released and second key is released.

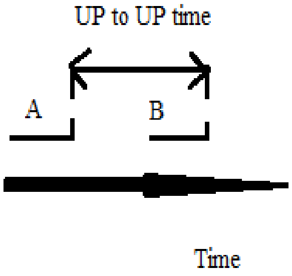


Fig 5. Up to Up tme

- *Total time*: the one new feature I decided to use is total time. The total amount of time which user takes to types a whole string. The speed of typing is genuinely different from one user to other. No user has same typing speed. One can type 14 characters in a minute and other only 5. So it depends on person who type. This feature is strongest among all features. It may leads to better accuracy.

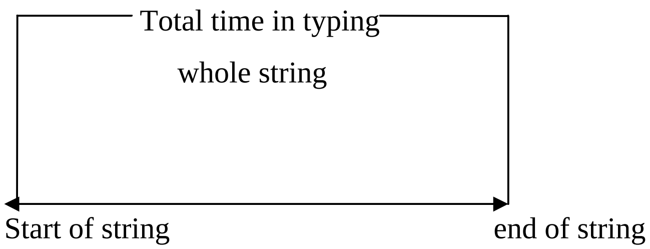


Fig.6. total time

- *Use of ASCII code*: to detect whether it is alphabet or numeric key we take help of ASCII code and also detect the use of caps lock.

- *Detection of shift key:* When there is need of special character user will pressed the shift key. But some user also press shift key when they need to write a capital letter word. So This feature is also used to detect which shift is pressed either left or right.

- *Use of backspace key*: this feature is used to detect when shift is pressed or how many time it is pressed.

- *Key pressed and Key released time*

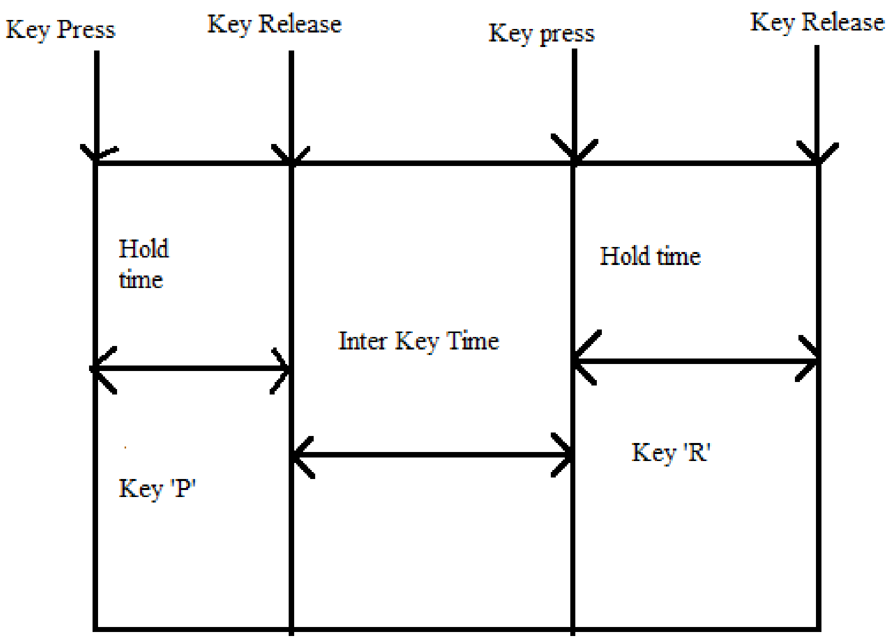


Fig 7 Explaining all key values

3.4.2 Flowchart for keystroke

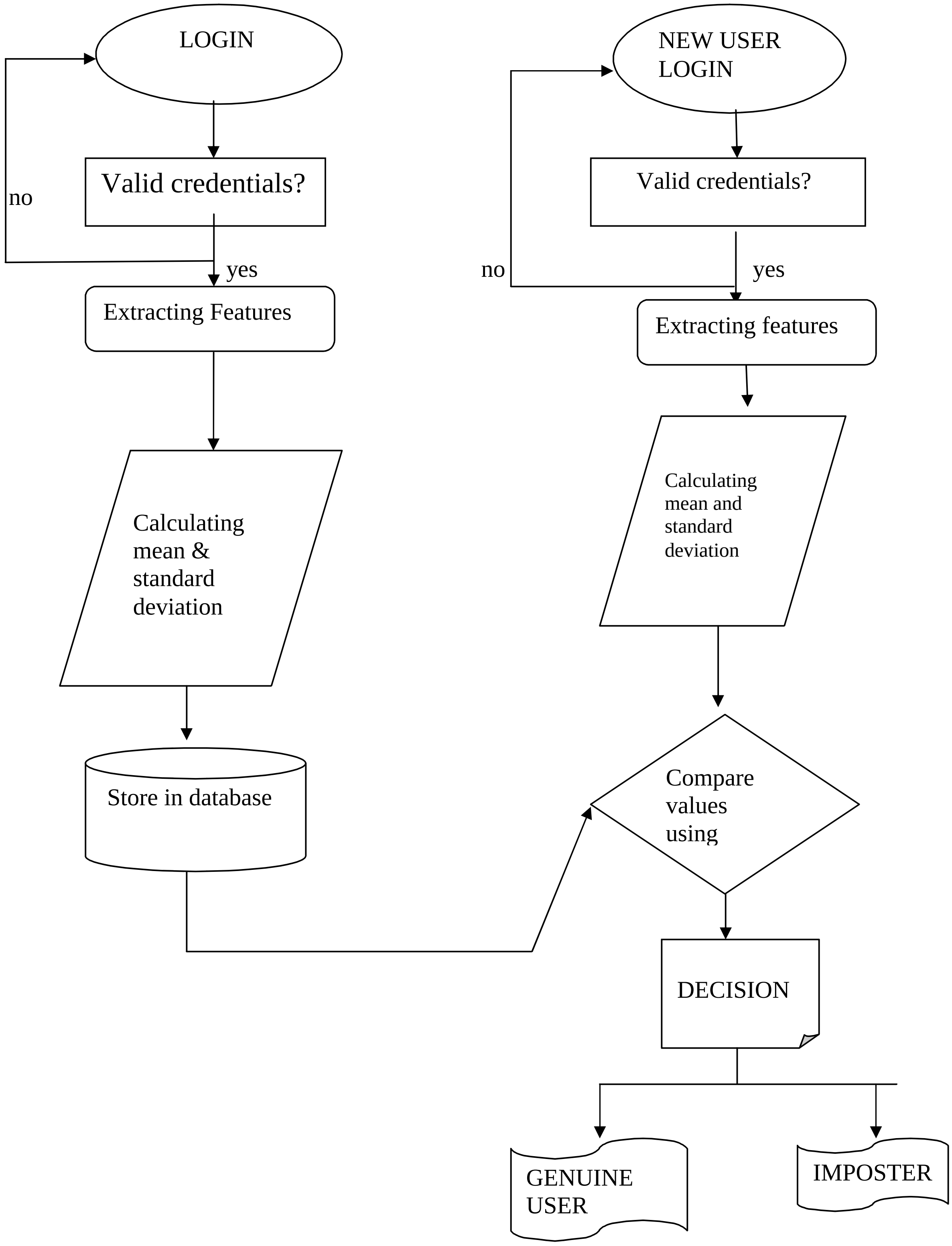


Fig. 8. Flowchart of keystroke

In the above flowchart first user has login to enter into system then his login credentials are checked. If the credentials are valid then user goes to next phase but if credentials are not valid user again go to login phase.

After filling valid credential next phase is of extracting features i.e. system extract key hold time, total time, up and up time, inter key time, detection of left and right shift, detection of backspace etc. After successfully calculating all features next phase is to calculate the mean and standard deviation of each feature and store it into the database.

Next part of the flowchart is for new user. When any user tries to access the account he enters the login credentials. After this all those feature that are calculated for valid user again calculated for this new user. After calculating features the mean value and their standard deviation is measured .Then by using t-test we compare these new values with the existing one store in database.

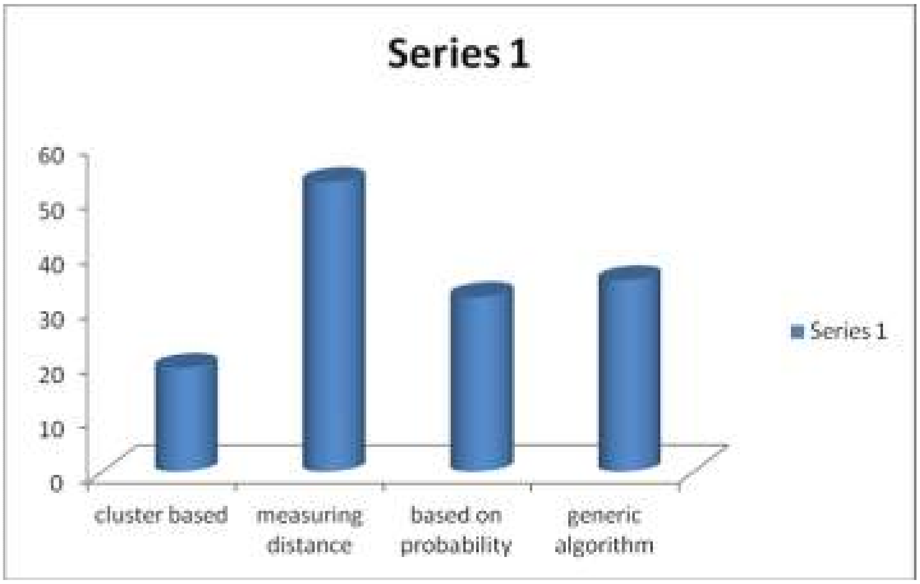
At last the decision is made that user is genuine or not. The decision can made if value is less than zero and greater than 1.28. the level of significance taken in it is 1%.

3.4.3 Survey of techniques used in keystroke biometrics:

As there are many techniques available for keystroke dynamics. So we do some little survey for al these methodologies. Commonly the methodologies that are used are based on statists, approach of machine learning and many more.

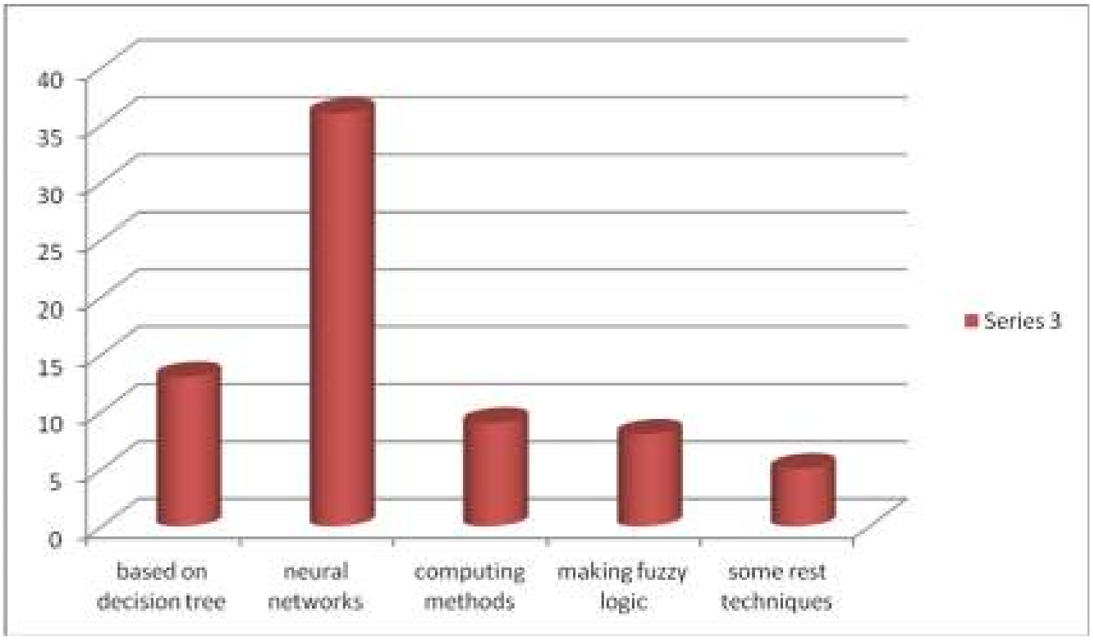
- First there is approach used statistical that include methods like finding mean, calculating standard deviation, t-test, k mean etc. The best thing about these methods are they are simple.

- Then machine training or learning is used. This holds methods like decision tree, neural networks, fuzzy logic, computing technique . all these techniques require to train a machine first. So these techniques having some troublesome in implementing them.



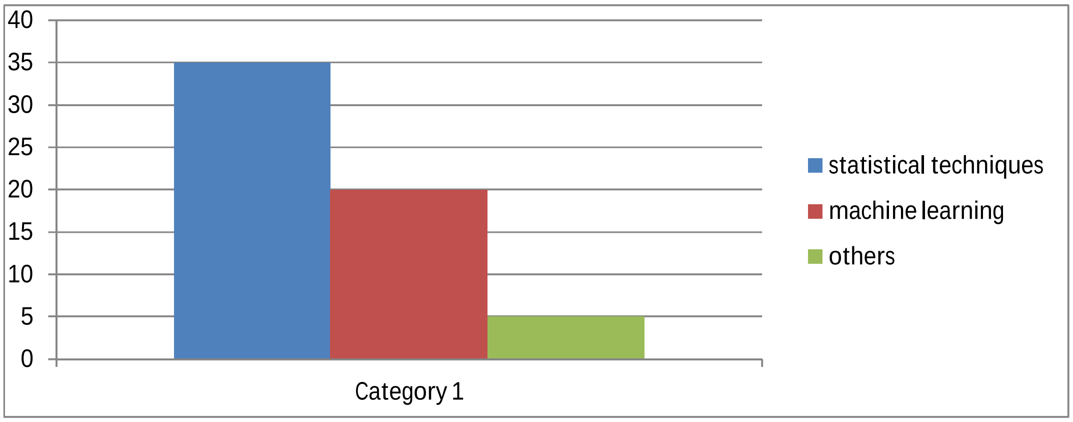
a

fig. 9(a) statistcal methodology graph



B

Fig 9(B) machine learning and other technique graph



C

Fig 9(c) comparison of three categories of technique

3.4.4 Proposed Technique

When a new user try to access an account this system can compare his typing features that is called claimant features with already existing features in database. If the features score are matched with existing one or it is nearer to the existing value than the user treated as genuine user otherwise the user is rejected and cannot access the account. For comparing these features and for accepting and rejecting the user, we required some methods.

The methodology used in this paper is t-test.

As we work on mean values of features and then their standard deviation of the feature so this is the reason why we adopt t-test as this is the method of comparing the mean of two sample that whether they belong to same population or not. And in our research this t-test helps to check whether the new coming user's calculated features are matched with existing one or not.

For all this process firstly we calculate mean of each feature and then the standard deviation of each features of genuine person and save in the database.

Then when new user try to access account first all features are calculated then mean and standard deviation is calculated and last t-test is used to compare these new mean and standard deviation values with existing one. And then decision is making whether the user is genuine or not.

3.4.3.1 t-test: t-test is also called student t distribution. T-test is used when we don’t know the exact value of standard deviation and sample is very small.It is hypothesis test which is followed if null hypothesis is supported. This test is used to determine whether two sample belongs to same population or not. The t-test works on the mean and standard deviation values.

There are different t-distribution available for all possible size of sample. As the size of sample increases the t-distribution become equal to the normal distribution.

Some assumption of t-test are taken while comparing mean of two samples that are independent. These are :

i) Both samples that have to be compared should follow normal distribution.

ii) If size of both samples is equal then t-test is highly robust

The formula used in this paper for t test is:

$$t=\frac{\overline{X1}-\overline{X2}}{\sqrt{S1^2/N1+S2^2/N2}}$$

where $\overline{X}_1$ is mean of new user

$\overline{X}_2$ is mean of existing or genuine user.

$S1^2$ is standard deviation of new user

$S2^2$ is standard deviation of genuine one

N1 is total length of string of new user

N2 is total length of string of genuine user.

3.4.5 Experimental Setup

This section consists of two parts:

- Extracting Features
- Comparison of user

3.4.4.1 Extracting Features

Now there is turn to collect data for research that holds all features discussed above for all trusted users.

To complete the data collection process we prepared a interface in which user has to fill correct username and password to continue further data collection process. This form has one username field and other password field and one ok button. Here is that form displayed below:

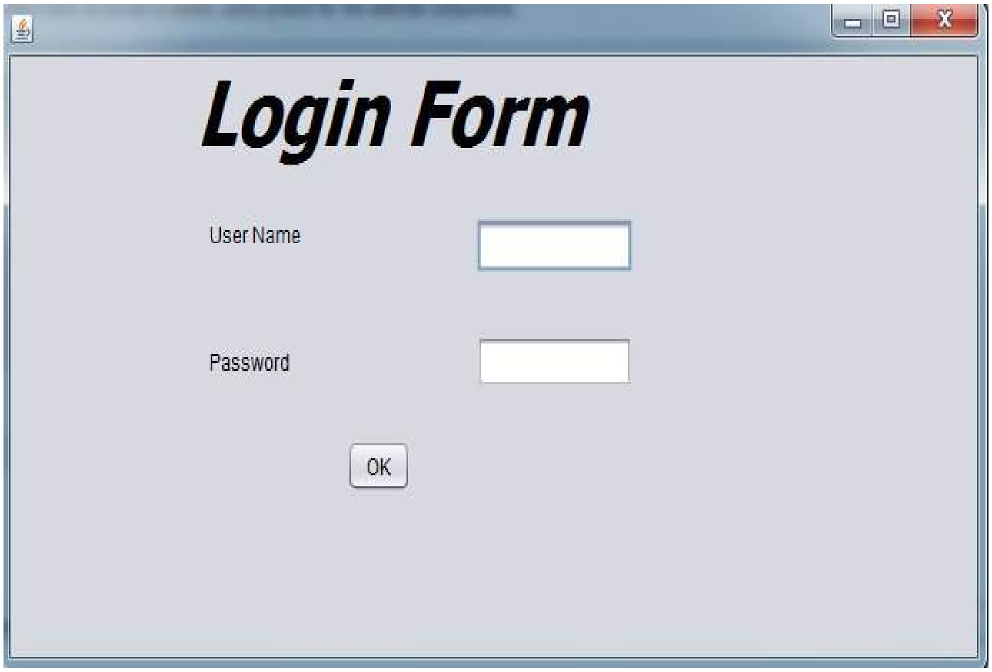
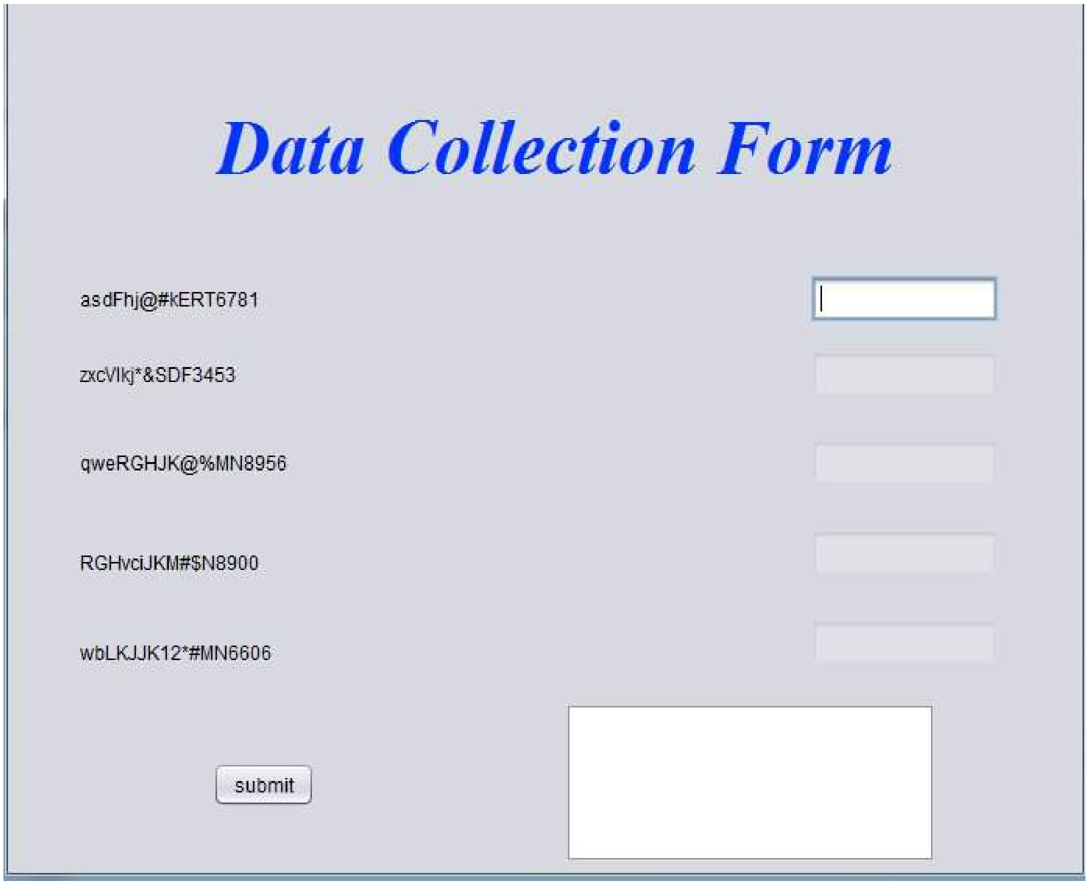


Fig.10. *login interface for data collection process*

In the above form user has to fill the correct username and password. If these credentials are correct then next form is displayed that include two columns first consists of label that show a string and second column show textbox in which user has to enter the same string which the label show. Only first textbox is enabled where all other four Textboxes are disabled by default. These textboxes are enabled if user has typed correct string in preceding textbox. When user enter the string and start typing, all key stroke features are calculated side by side like calculating the key hold time, key inter time, up &up time, total time and also calculate whether shift is used is left shift or right ,backspace is used or not, All these features are calculated and store into file.



Data Collection Form

asdFhj@#kERT6781

zxcVlkj*&SDF3453

qweRGHJK@%MN8956

RGHvciJKM#$N8900

wbLKJJK12*#MN6606

Fig.11. *Data collection form*

There are total five labels on form so user has to enter five times the different strings and we have to calculate all these features five times. The use of five label in our above form is for avoiding the limitation in typing style as user may be in different mood while typing his password. So to get accuracy we use these five labels. Then calculate the mean value of these features and standard deviation and store in file. After calculating all the features, the string that user entered is also matched with the label that show the string. If it matches then user is allowed to access. Otherwise the interface pop up the message the label not match.

Suppose there is first label show string “asdFhj@#kERT6781”. When user start typing this string the features are calculated

Table 1 Way of calculating features

Key inter time	Key hold time	Up & up time
release a and press s	a	release a & then up s
release s and press d	S	release s & release d
release d and press F	D	release d & release F
release F and press h	F	release F& release h
release h and press j	H	release h release j
release j and press @	J	release j release @
release @ and press #	@	release @ & release #
release # and press k	#	release # & release k
release k and press E	K	release k & release E
release E and press R	E	release E & release R

Release R and press T	R	release R & release T
release T and press 6	T	release T& release 6
release 6 and press 7	6	release 6 & release 7
release 7 and press 8	7	release 7 & release 8
release 8 and press 1	8	release 8 & release 1
	1	

When these three features are calculated next calculate the total time to type the whole string.

After that the use of SHIFT & BACKSPACE is detected.

When all features are calculated then the mean and standard deviation is calculated and stored into database.

3.4.4.2 Comparing User for user identification

Now there is turn to compare new user when he try to access the system. For this a new interface is prepared which again ask user for username and password. If user able to fill correct login credential then comparison form displayed otherwise user exit from system.

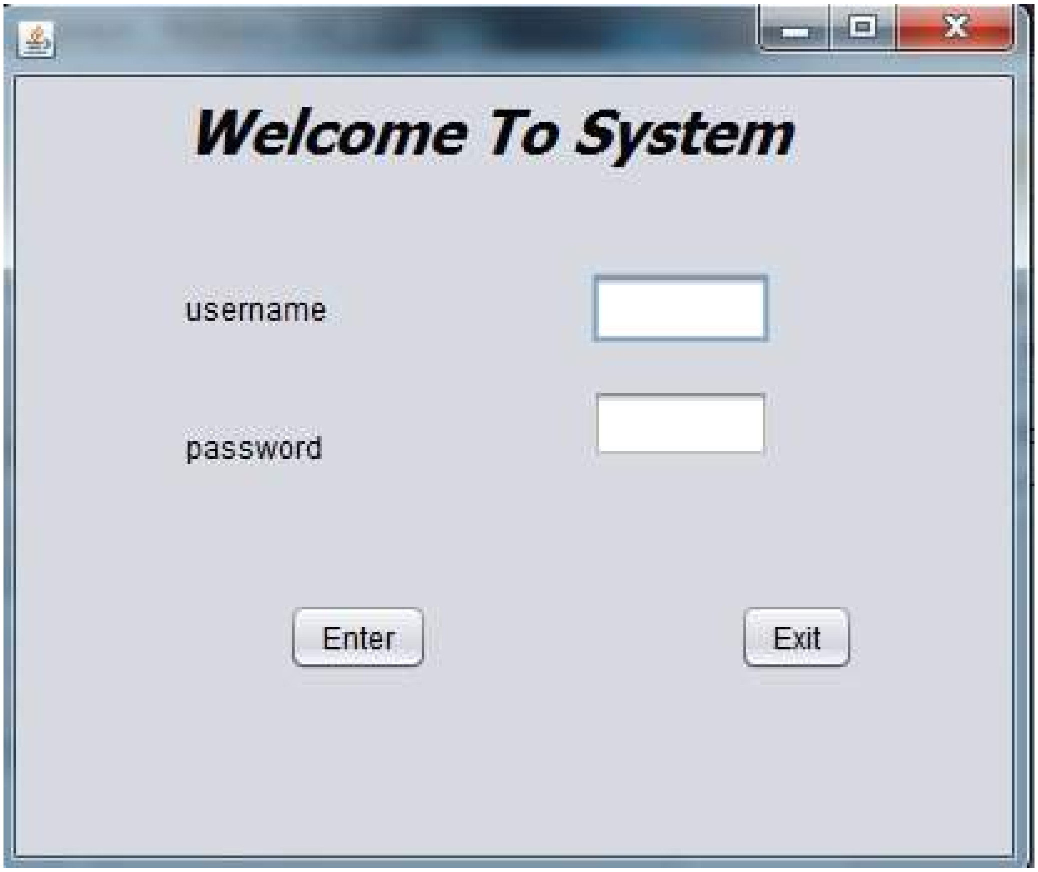
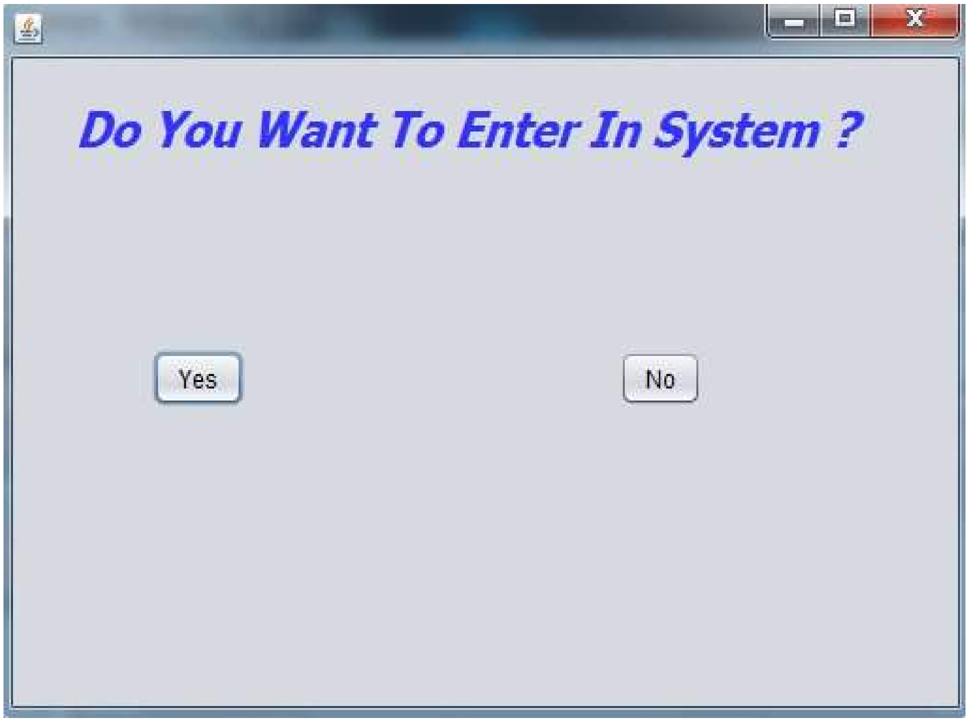


Fig.12. *Login interface for user comparison*

After filling exact credential a new form should open which again conforms from user that whether you want to enter into system or not.

This form includes two options. If user press YES button then system popup new forms that is form use for user identification or to compare new user with existing one.

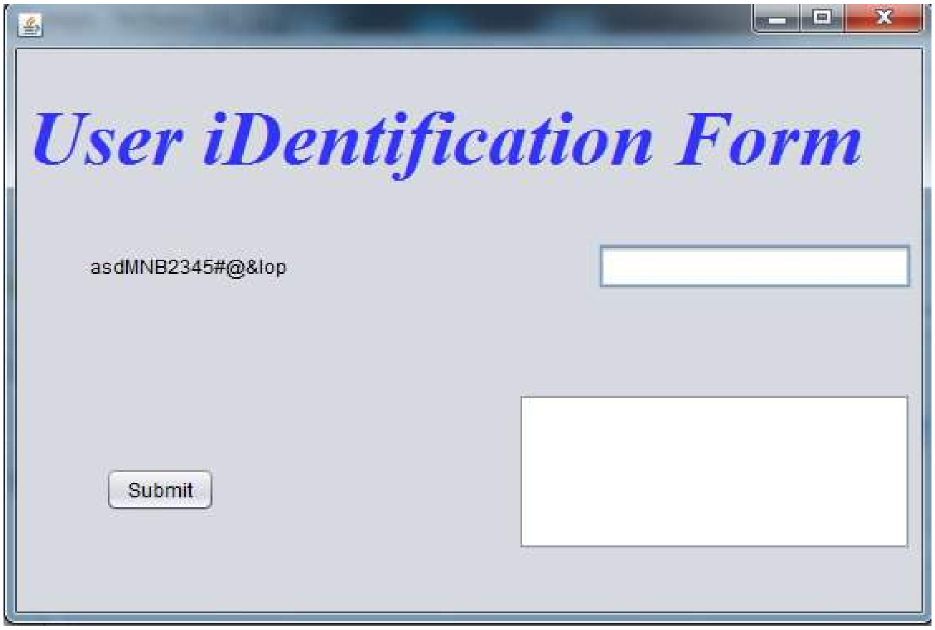
And if user press NO button then user get exit from system. And no form is displayed.



Do You Want To Enter In System ?

Fig.13. asking user to enter into system or not

After pressing YES button user faces a new form that is form which is used for comparison purpose.

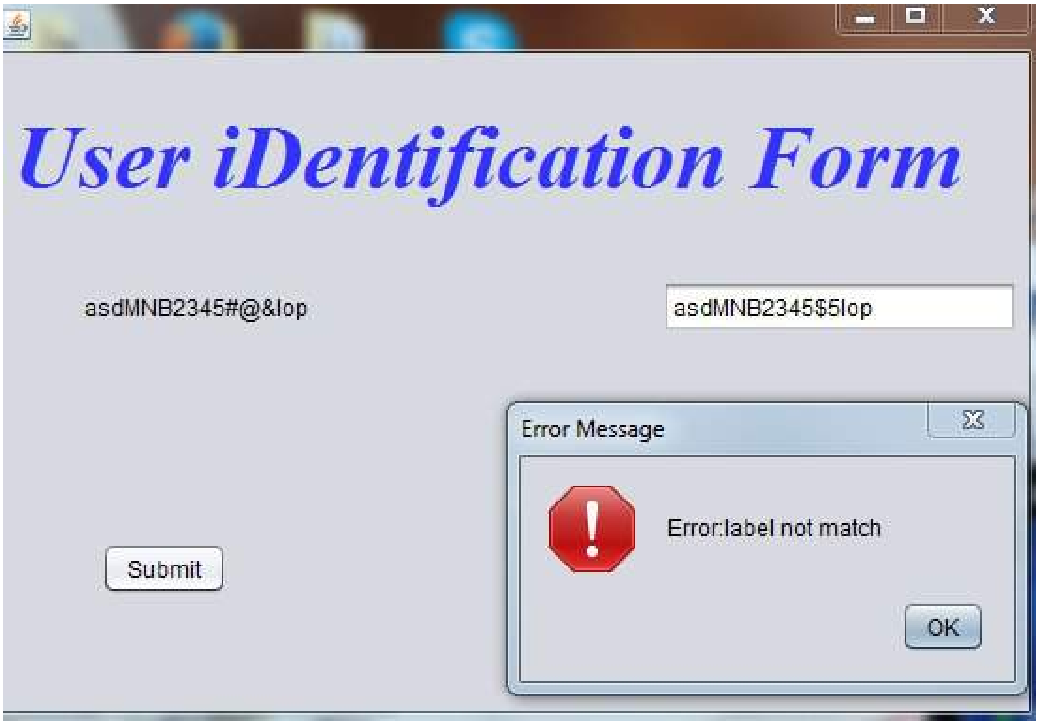


User iDentification Form

asdMNB2345#@&lop

Fig.14. *Comparison of new user with existing one*

This form has one label and one textbox.In this form user has to type label in textbox for which all features are calculated and then their mean and standard deviation are calculated and by using t-test these values are compared to existing values to make decision whether it is genuine user or imposter. User has to fill same string in textbox as it is shown in label. If string does not matched with label then it pop up the message string does not matched. Then user has to fill the string again.



User iDentification Form

asdMNB2345#@&lop

asdMNB2345$5lop

Error Message

Error:label not match

Fig.15.Warning for wrong string

From above figure it is clear that when user type wrong string in textbox. System pop-up the message that label not match. And user has to type all the string again.

After comparing features the decision is made that whether the user is genuine one or not. The decision is made by using t-test’s threshold value that is predefined through degree of freedom which is n1+n2-2 and for our research it is 30.

The error rate taken is 1%. So if the result after comparison comes less than 1.28 then user should allow accessing the account otherwise not.

Chapter 4

Results and Discussion

Now there is turn to show the result of our research.

Firstly before discussing result I explain the two types of errors which are reduced by keystroke.

4.1 Types of errors:

There are two types of errors for testing hypothesis.

- **Type 1 error**: Type 1 error is when person is true and we take decision to deny the request of valid user.. In other words we might be reject H_0 but in real H_0 is true. Also we can say like that we reject that hypothesis which should be accepted in any case.

- **Type 2 error**: it is the acceptance of that hypothesis that has to be rejected. In other words we might accept the hypothesis H_0 however it is actually not true i.e. the person is not valid but we take decision wrong and allow that imposter to access the account.

Table 2. Possible hypothesis test outcomes

	Exact situation	
Decision	H_0 true	H_0 False
Accept H_0	NO Error	TYPE 2 Error
	Probability = $1-\alpha$	Probability =β
Reject H_0	Type 1 error	No error
	Probability = α	Probability =$1-\beta$

Where H_0 is called the null hypothesis which is either accepted or rejected at the time of result. The probability that given in table for type 1 error is understood as level of significance for testing the hypothesis. If for understanding we take type 1 error is fixed at 5 percent level than it means that there are total 5 chances that we reject H_0 but actual H_0 is true. The type 1 error can be controlled by stable it at lower level i.e. at 1% level. But both type of errors may not condensed at the same time.

In our research work the type 1 error which we want to reduce as much possible is achieved by using t-test. The error rate that we allow is also very less that is 1% which is big achievement. For the genuine user we get result 0.05% which is less as compared to other researches. So we are able to reduce the main limitation of keystroke of type 1 error. Type 1 error i.e. the genuine user is denied by system is major limitation of keystroke till now and i am successfully achieving the better result in case of type 1 error. Type 1 error mostly reduced to 0% in my work.

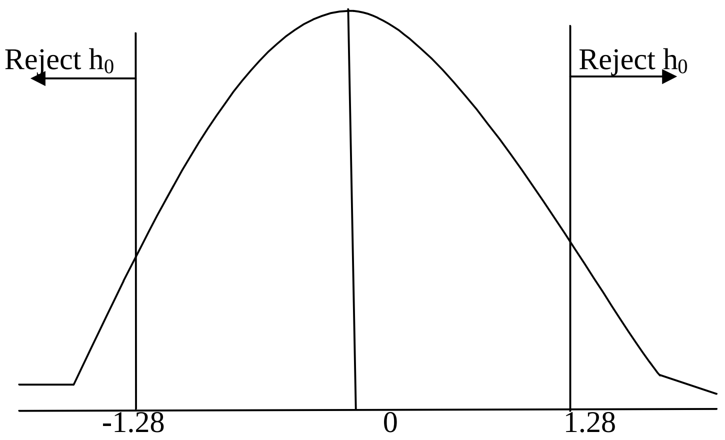


Fig.16. Rejection and acceptance of hypothesis

Above there is graph for either the hypothesis is accepted or the hypothesis is rejected. This graph shows that if the final result of t-test comes between 0 and 1.28 then that user treat as a genuine user. If the value is out of the range then the user must be imposter i.e. if value is greater than 1.28 and less than -1.28 then user must not be valid.

Basically we have to check two hypothesis on samples. These are:

- Null hypothesis :
  H_0: $u_x=u_y$ (sample belongs to population)

- Alternative hypothesis:
  H_1: $u_x \neq u_y$ (sample doe not belongs to population)

On the basis of these two hypothesis the sample is identified. And in our work we check these on new user's sample.

4.2 Critical values and decision rules:

Critical value divide the area of probability curve of test statics into two types of region one is critical region and other is acceptance region. Size of acceptance region is 1-α where size of critical region is α i.e. level of significance. First we show the graph for two tailed test.

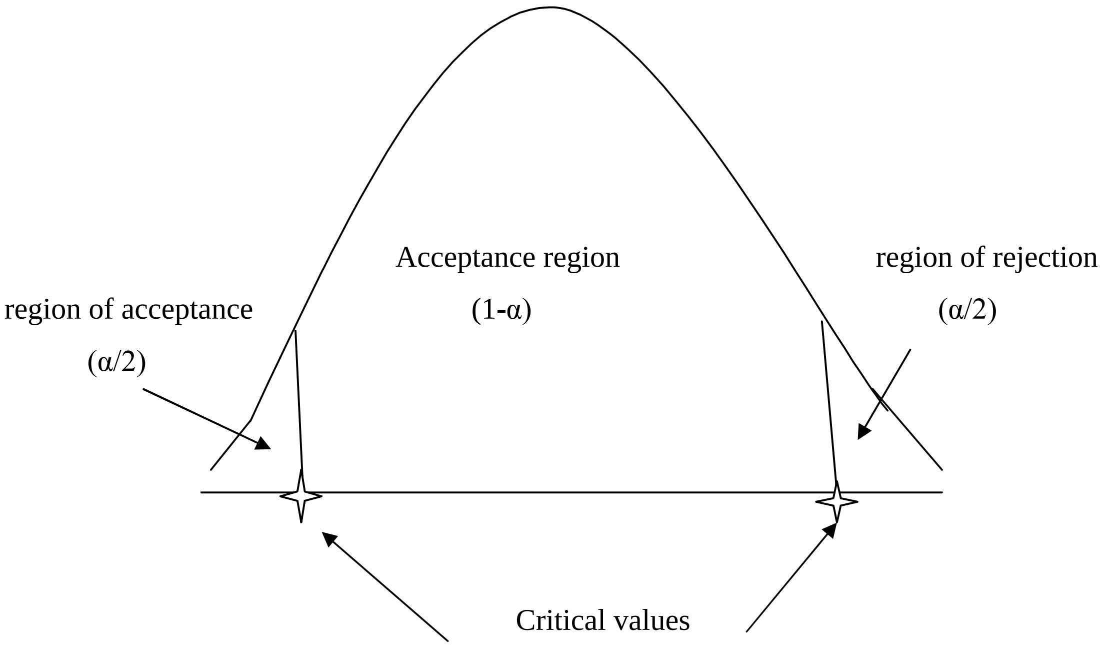


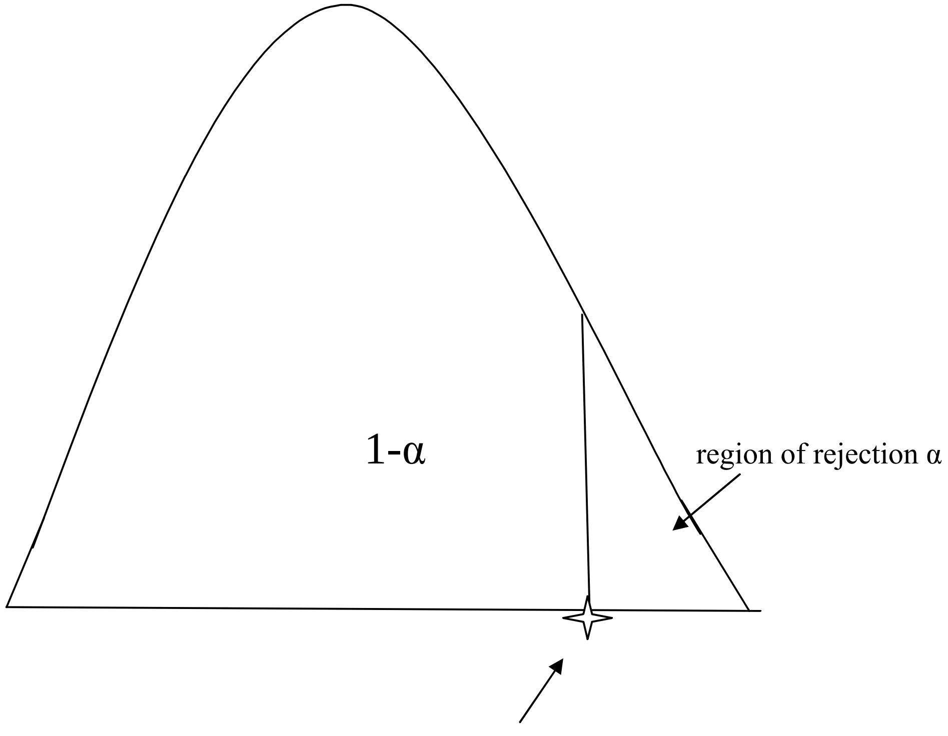
Fig 17. graph showing regions in case of two tailed test

The above figure show the critical region, acceptance region and also critical values.

For example if we take level of significance as 5% then its critical values are -1.96 and 1.96. and any values which is within these two critical values are accepted. Thus it interprets that we reject hypothesis if the value is below -1.96 and above 1.96.

For a right tailed test the critical values are on right side of the curve.

In this we test $H_{0:}$ u=u_0 against H_1 : u>u_0



Critical value

Fig 18. Graph for right tailed test

For example lets say level of significance is 5%. or $\alpha=0.05$ and the critical value is 1.96 then the rejection of hypothesis is occur if value is greater than the 1.96

For a left tailed test the critical values are on left side of the curve.

In this we test $H_{0:} u=u_0$ against $H_1 : u < u_0$

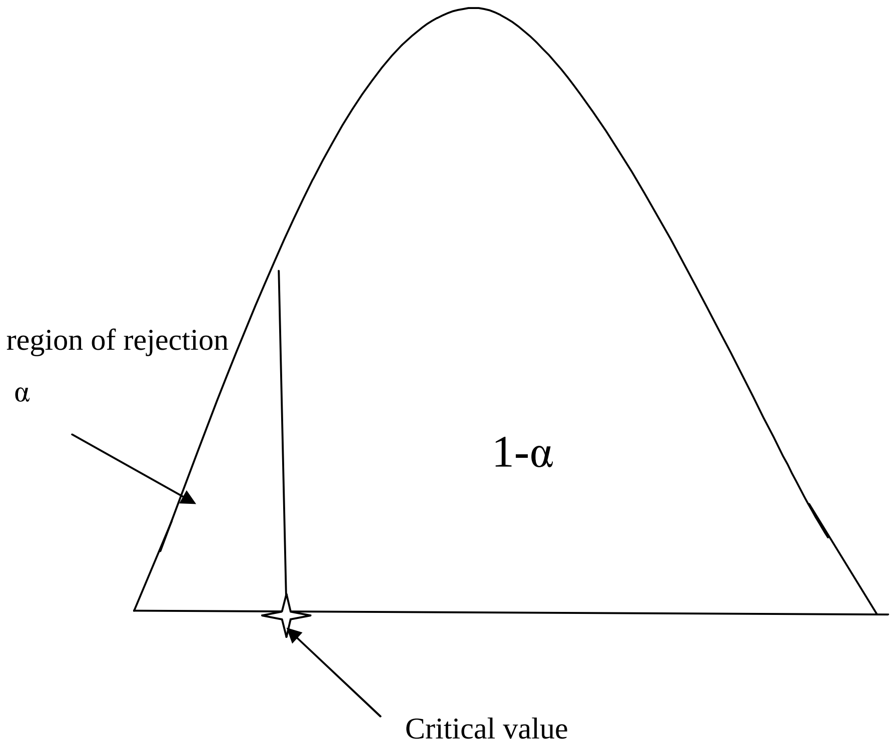
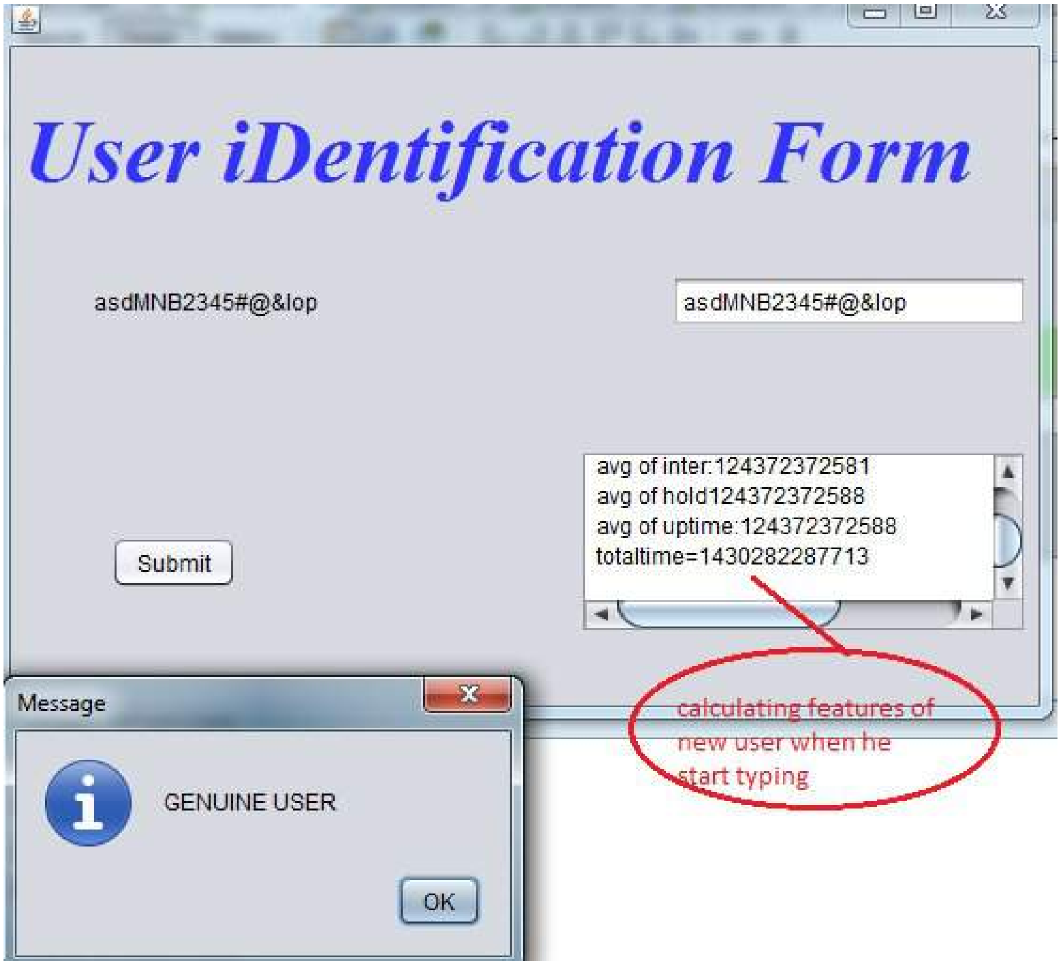


Fig 19. Graph for left tailed test

For example lets say level of significance is 5%. or $\alpha=0.05$ and the critical value is -1.96 then the rejection of hypothesis is occur if value is smaller than the -1.96

4.3 Result Figure:



User iDentification Form

asdMNB2345#@&lop

asdMNB2345#@&lop

```
avg of inter:124372372581
avg of hold124372372588
avg of uptime:124372372588
totaltime=1430282287713
```

Message

GENUINE USER

calculating features of new user when he start typing

Fig.20. result showing figure

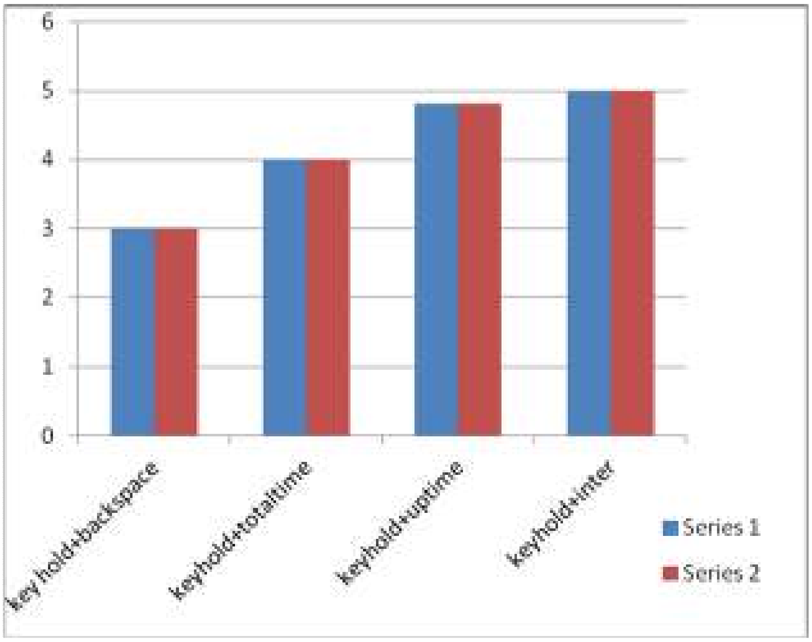
Above there is figure that show the result that the user is valid or genuine. In this first user fill the string in textbox same a the label shown. After that he press the submit button. While user typing in textbox his all features are calculated side by side and are shown in figure in one of the text area. when user press submit button then the t-test calculate its value and then in last give the decision that user is Genuine or imposter by pop up the message.

Table 3 Result table

Features	results of t-test at 1%error	Type2 error	Type1 error
Key hold time	0.0017368804	1%	0%
Inter key time	0.003545028	1%	0%
Up-Up time	0.0033047635	1%	0%
Total time	0.0023234678	1%	0%
Backspace	0.002123498	1%	0%
Left and right shift	0.0032134509	1%	0%

From above table it should be clear that the more accuracy in result is reached by key hold feature. Also our new features give a better result and with all these features we are able to achieve our type 2 error at 0%. That is genuine user never treated as imposter and should always allow to access the system.

4.4 Result Graph:

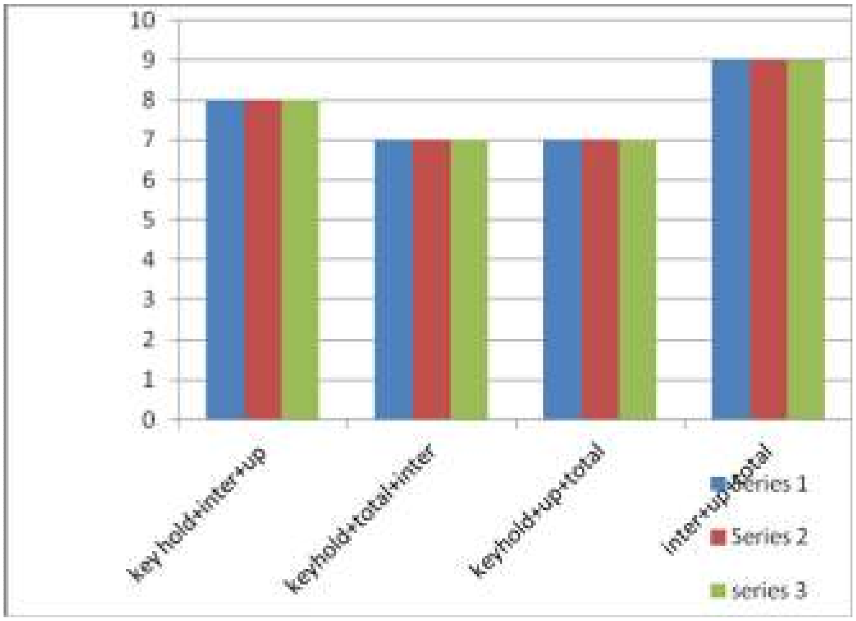


A

Fig.21. *Graph Representing result by combining two features*

In the above graph we combined two features and check their accuracy level i.e. we check that with which combination we get higher results. First we combine the key hold time with backspace and get accuracy of 3%. Next there is combination of key hold and total time and get accuracy of 4%. After this there is pair of key hold with uptime .this pair give us accuracy of 4.8%. In last the pair is key hold and inter time which results in better accuracy level i.e. 5%.

The reason for combining each features with Key-hold time is while calculating single features result that shown in above result table the best result we get from the key-hold time i.e. 0.0017. so thats why we combine each feature with it.



b

Fig.22. *Graph Representing result by combining three features*

In the above graph we combined three features and check their accuracy level i.e. we check that with which combination we get higher results. First we combine the key hold time with inter and up time and get accuracy of 8%. Next there is combination of key hold with total time and inter time and get accuracy of 7%. After this there is pair of key hold with uptime and total time. This pair gives us accuracy of 7%. In last the pair is inter time with up and total time which results in better accuracy level i.e. 9%.

The reason for combining each features with Key-hold time is while calculating single features result that shown in above result table the best result we get from the key-hold time i.e. 0.0017. so that’s why we combine each feature with it.

From above two graphs this is clear that by combining three features we get more accurate result. In case of three features more accuracy is achieved when we combine inter time with uptime and total time. In case of two features the more accurate result get when hold time is combined with inter time.

Chapter 5
Conclusion

As in all previous researches there must condition that sometime it may happening that actual user's access is also denied by system when features not matched with the stored values. This error is removed in this paper by adding some additional feature .so that no natural things like tiredness of user, some tensions etc. are become obstacles while entering the password. Also the type 1 and type 2 errors are minimized. As in today's time security is very important prospect of our life. Having a inexpensive technique that can provide a better security to us is one of the better solution. Keystroke do this work in no cost at all. As everyone has a keyboard the only need of this technique is software installed on the system and user get advantage from it.

The future work for this paper is calculating data on only one string and also uses the key pressure feature and tries to recognize the sound of typing of user.

Chapter 6
References

6.1 Websites

[1] www.google. Com

[2] www.wikipedia .com

6.2 Journal Articles

[3]P.Teh, A. Teoh, C.Tee, et al.: ‘A Multiple layer fusion approach on keystroke dynamics’, Pattern analysis & applications, 2011,vol. 14,no.1, pp 23-36

[4] K. Revett, F. Gorunescu, M. Gorunescu, et al.: ‘A machine learning approach to keystroke dynamics based user authentication’, International journal of Electronic Security and digital forensics, 2007, pp 55-70

[5] R.Giot, C. Rosenberger.: ‘A new soft biometric approach for Keystroke dynamics based on gender recoginition’, International journal of Information technology and Management, 2012,pp 35-49

[6] D. Shammugapriya and Dr.G.Padmayathi,: ‘A wrapper based feature subset selection using ACO-ELM-ANP and GA-ELM-ANP approaches for keystroke dynamic authentication’, International conference on signal processing image processing and pattern,2013

[7]Nandini Chourasia.: ‘Authentication for the user by keystroke dynamics for banking transaction system’, Proceedings of international Conference on Advances in Engineering & Technology, Goa, India, April 2014, pp 41-45

[8] Pin Shen Teh1, Shigang Yue2, Andrew B.J. Teoh3.: ‘Feature Fusion Approach on Keystroke Dynamics Efficiency Enhancement’, International journal of cyber – security and digital forensics, the society of digital information and wireless communication,2012, pp 20-31

[9]Rick Joyce and Gopal Gupta.: 'Identity authentication based on keystroke latencies', Communications of the ACM, 1990.

[10] Robert Moskovitch, Clint Feher, Arik Messerman, *et al*.: ' Identity Theft and computer ,behavioral biometric'

[11] Saurabh Singh, Dr. K.V.Arya,: 'Key Classification: A New Approach in Free Text Keystroke Authentication system'

[12] P.S.Teh, A.B.J.teoh, C.Tee,et al.: 'Keystroke Dynamics in password authentication enhancement', expert Systems with Applications, 2010 ,vol.37, no.12, pp 8618-8627

[13] Fabian Monrose,Aviel D. Rubin.: 'Keystroke Dynamics As a Biometric for Authentication', Future Generation Computer Systems,2000,pp 351-359

[14]Yu Zhong,Yunbin Deng, Anil k. Jain,: 'Keystroke Dynamics for user Authentication',Non Technical data-Releasable for foreign person,2012

[15] J. R. Young and R. W. Hammon,: 'Method and apparatus for verifying an individual's identity', United States Patent US 4,805,222, Feb 14, 1989.

[16] K.S. Balagani,V.V.Phoha, A.Ray,et al.: 'On the discriminability of keystroke feature vectors used in fixed text keystroke authentication', Pattern Recoginition Letters, 2011, vol.32,no. 7, pp 1070-1080.

[17]R.Zack , C. Tappert,S. Cha.: 'Performance of a long-text-input keystroke biometric authentication system using an improved k-nearest-neighbor classification method',IEEE Int'1 conference on Biometrics: Theory Applications and Systems(BTAS),2010, pp 1-6

[18] Juan Liu.Baochang Zhang, linlin Shen, et.al., "The beihang keystroke authentication system"

[19]Benjamin Ngugi, Beverly K.Kahn, Marilyn Tremaine.: 'Typing Biometrics: Impact of human Learning on Performance Quality', J.Data and Information Quality, 2011, Vol. 2 pp 11-17

[20]YogeshMeena, UrvashiGarg.: 'user authentication using keystroke recognition'

[21]Lívia C. F. Araújo, Luiz H. R. Sucupira Jr., Miguel G. Lizárraga,et al.: ‘ User authentication through typing biometrics features’, IEEE transaction on signal processing,vol.53,no. 2, Feb.2005

[22] Mariuaz Rybink and Piotr Panasiuk,: ‘User authentication with keystroke dynamics using fixed text’, conference on biometrics and kansei engineering, 2009

[23] Romain Giot,Mohamad EI-Abed, Christophe Rosenberger.: ‘Web Based Benchmark for Keystroke Dynamics Biometrics Systems: A Statistical analysis’, arXiv: 1207.0784vi [cs.LG], july 2012

Reference to book:

[24] C.R.Kothari, research methodology, new state international publishers.