IDENTIFYING MALICIOUS USERS USING KEYSTROKE PATTERNS

D.Lakshmi Padmaja, Harshitha Palvai, Sravya Patharlapalli, Bala suresh Markanti

ABSTRACT:
The principle point of this paper is to classify the user as genuine or malicious dependent on keystroke designs. Random Forest algorithm is utilized for characterization. Numerous cyberattacks are going on around the globe, roughly there are 1.5 million digital attacks yearly which imply more than 4,000 attacks every day, 170 attacks each hour, or almost three attacks each moment. This strategy is to recognize computer users utilizing keystroke and mouse designs. Machine Learning algorithm helps in the arrangement of users. A model of each user's typical composing style was made and contrasted and later composing tests. Utilizing the samples the machine characterizes the user as real or malicious. The accuracy of this model is determined.

KEYWORDS: Machine Learning, Random Forest, Keystrokes.

1.INTRODUCTION:

Existing protection mechanisms defend computer systems and networks from unauthorized with the assist of passwords. However, those get admission to controls may be compromised through distinct cybersecurity strategies including phishing wherein is used to steal login credentials user statistics, malware, etc. which might also additionally motive extremely good harm to the organization

In data protection, intrusion detection is that the technique of tracking activities throughout a laptop or community and examine them to hit upon indicators of viable incidents, which might be violations or threats of violations of protection policies, suitable use, or protection practices. An intrusion detection gadget (IDS) automatizes this technique.

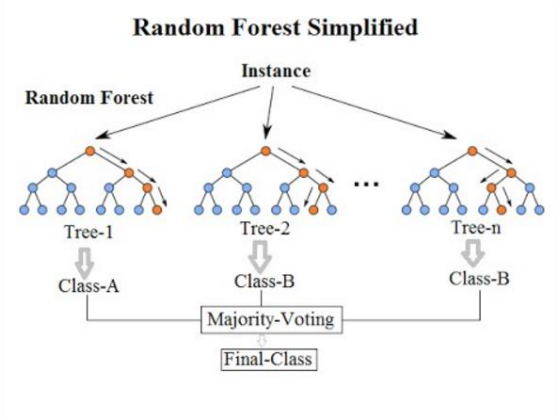
This technique allows maximum agencies to research the conduct of their personnel and test whether or not the proper individual is logging into the gadget or not. Using the sample that he/she will classify the individual as a real user or malicious user by use of the Machine gaining knowledge of classifying. Hackers can assault machines and login as real users this version allows them to hit upon the individual. This is one of the authentication strategies that allow within side the first step of user evaluation.

Keystroke dynamics are frequently implemented in ways: static text or dynamic text. Static text most effectively plays an evaluation of constant expressions as, for instance, a password. While, in dynamic text, the evaluation takes place for any text it truly is typed through the user. Keystroke dynamics within side the static textual content calls for much less attempt to be carried out and it additionally reached decrease mistakes quotes within side the literature

In this form, there might be special timing data that depicts exactly while each password is changed into squeezed and keeping in mind that it changed into dispatched as somebody is composing on a PC console that is estimated to extend a remarkable biometric format of the user's composing test for user's verification. Anything composed on a console while signing into a couple of records, the user validation is made on whether the secret word changed into composed gradually or quickly, the letters had been completely composed on the equivalent spot, or if there might be any put off even as composing any character.

Random Forest which is an AI model is utilized to classify. Random Forest assembles various decision trees and unions them to get a more precise and stable forecast. We can know from its name, which is used to make a

forest by how and make it arbitrary. There is a nearby association between the number of trees inside the woodland and thusly the results it can give: the greater the number of trees, the more accurate the result. It is a serious learning adaptation of the decision tree it gatherings decision trees, normally prepared with the "bagging" strategy. The bagging strategy is called the Bootstrap aggregating technique which is an incredible measurable strategy for assessing an amount from an information test. An ensemble strategy is a procedure that consolidates the expectations from different AI calculations together to make more precise forecasts than any individual model. One major favourable point of random forest is, it can be used for solving both classification as well as regression problems in Machine learning.



2. LITERATURE SURVEY:

Research has been wiped out a search for applications of keystrokes within the Information Security field. a research which was done by S.M Furnell, J.P. Morrissey, P. W.Sanders, and C. T.Stockel [8] used neural networks as an algorithm for implementing keystroke dynamics. they need to be implemented in two practical systems that are supported by static and dynamic techniques. During a static technique, they used neural networks whereas in dynamic they used the statistical analysis method. The dynamic technique was supported arbitrary text input which features a great scope on real-time supervision. The general result for the research had some errors which may be modified. The methods which they used are often implemented in real-life for statistical and dynamic approaches. Further study was supported using genetic algorithms alongside subsequent modification of a dynamic approach.

Gaussian Mixture Models [6] also can be wont to classify whether the given test suit belongs to the user or not. Since many studies have proved that digraph patterns present in keystroke data are generated by the Gaussian method. The advantage of random forest over other methods is it takes as an input the uncertainties within the features and therefore the labels and treats these as probability distribution functions instead of deterministic quantities. This treatment allows it to naturally represent missing values within the data set, which other algorithms fail to try to do.

In the research which was done by Mindaugas Ulinskas, Marcin Woźniak, Robertas Damaševičius [1] in July 2017 they proposed a model using statistical features and k-Nearest Neighbour (KNN) classifier to discriminate between different consecutive key typing sessions. The accuracy of detecting Fatigue was nearly 91%. The methodology was also utilized in the medical field i.e. EEG-Based Estimation of Mental Fatigue. Electroencephalogram (EEG) is a test that identifies electrical movement in your cerebrum utilizing little, metal circles (cathodes) joined to your scalp.

In the examination which was finished by Manuel Rodrigues, Sérgio Gonçalves, Davide Carneiro, Paulo Novais, and Florentino Fdez-Riverola they suggested that utilizing keystroke examples and mouse

developments we will identify mood of students in E-learning [7]. Nowadays E-learning is updating rapidly, in normal classrooms Teachers can detect the mood of students whereas during a virtual environment it's difficult to do. The model is to build up a unique pressure assessment model that, while utilizing this setting data, will permit instructors to adjust systems inside the search for expanded accomplishment in learning.

In the examination was finished by Fabian Monrose, Michael K., Reiter Susanne Wetzel's methodology they joined keystroke elements with the secret phrase of users this gives greater security and it turns out to be hard for online of disconnected programmers to hack the secret key since it naturally adjusts the adjustments in users composing patterns[9]. Utilizing experimental information and a model execution of the plan, they gave proof that the methodology is reasonable m practice, m terms of straightforward use, improved security, and execution [4].

KH Lam et al proposed a paper for "Clinical measures in various sclerosis (MS) and to discover the unwavering quality and legitimacy of keystroke elements for clinical purposes [10].

Tom Olzak proposed "Keystroke Dynamics: Low Impact Biometric Verification" to distinguish the numerous boundaries like how it functions, history, and cost worries invalidation for biometric confirmation [11].

Mayur Mahadev Sawant et al distributed a survey paper to discover what is research has done on keystroke elements [12].

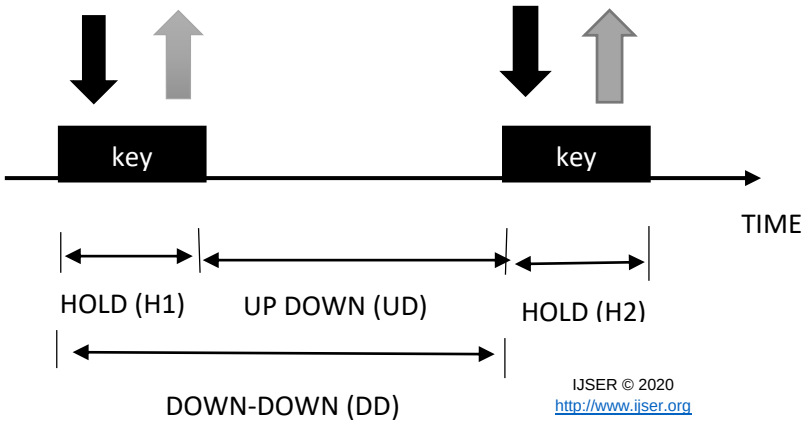
3. PROPOSED METHODOLOGY:

The primary advantage of using Random forest is in classification problems it will avoid the overfitting problem and it helps in identifying features i.e. it helps in feature engineering. Considering wrong features may lead to a decrease in accuracy which affects prediction.

There are two distinctive processes are involved in this methodology one is feature extraction and second is classification of the extracted features. In the first part, many features are extracted for the acknowledgment of a user. These highlights ought to speak to how the user carries on as far as keystroke elements.

To prepare and test a model we have to have a collection that can be set up by utilizing python libraries. In Python, there are packages like keyboard and mouse which help in listening to the patterns of keyboard and mouse. Utilizing mouse.hook technique and keyboard.record strategy from packages we can store the patterns, timing, and the secret key that we type. Aside from the content itself, the keyboard gives the moments in which each key is squeezed and delivered. From this fundamental information, features are extracted and used as input for the classification algorithm.

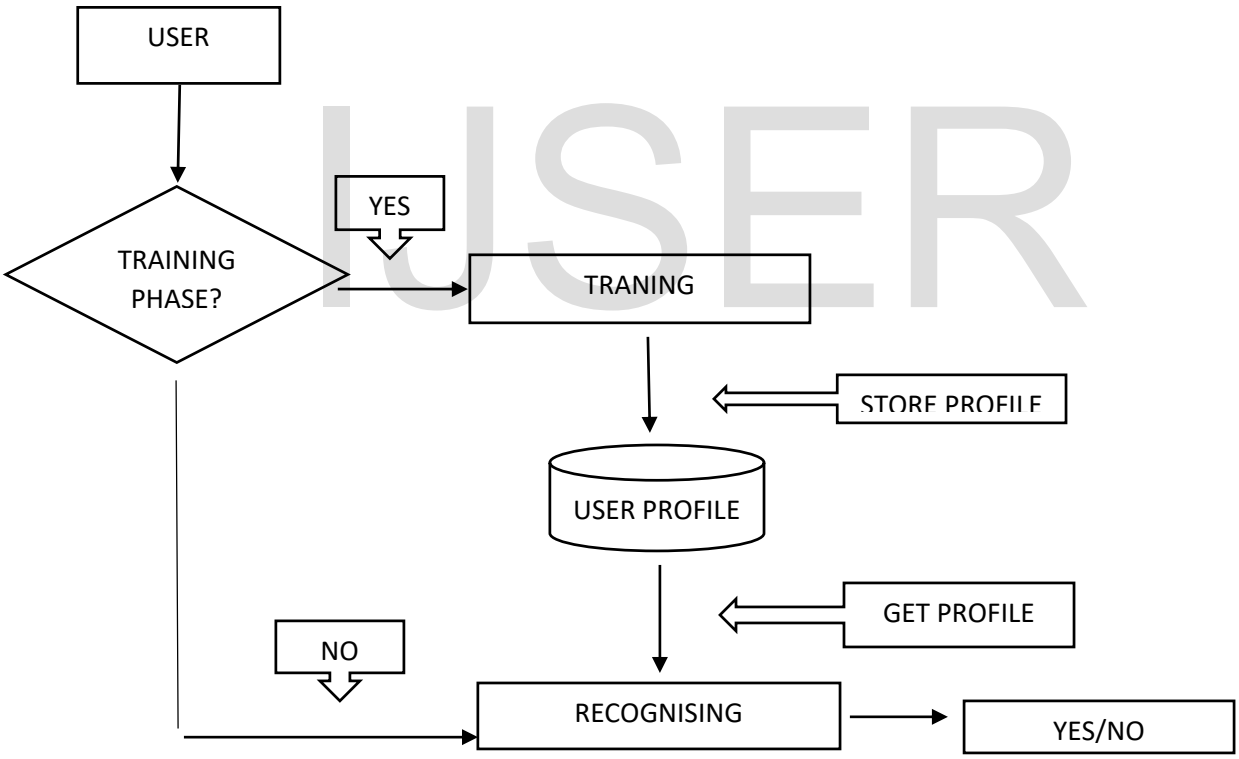
Working Model:



Making an informational collection helps in distinguishing the client in light of the fact that various people have various methods of composing speed and passwords. To make an informational index we figure a few highlights, for example, Hold time, Up-down time, and Down-down time. Hold time is the time between the press and the arrival of a key. Up-down time is the time between key delivered and other key squeezed. Down-Down is the time between squeezing two back to back keys.

Subsequent to making the informational collection it standardizes all the estimations of the informational collection which helps in preprocessing the information and mark the informational collection as 0 and 1(0 for some unacceptable user and 1 for the correct user). Subsequent to preparing the information place the information in a random forest model. In this model, they have utilized 20 decision trees to settle on the choice. The yield will be given dependent on the yield of choice trees. The most extreme votes will be thought of. Features such as hold time, up-down time, down-down time for the user sent to the model.

Working Flow Chart for the Model:



In the initial step after typing then it is checked whether it is in the preparing stage or testing stage. When the model is in the preparing stages it stores the user's profile, for example, the keystroke examples of the user. At the point when the model is in the testing stage then it gathers the examples of a user, for example, a hold time, key up-down time, key down-down time at that point, then the model compares the testing data i.e. the data which was given with already stored data by getting users keystroke patters. At that point, the classifier predicts the user as genuine or malicious by giving a yield i.e. output as yes or no.

4. RESULTS:

Once we have generated model we need to analyse results using confusion matrix

Confusion matrix:

Actual/ Predicted class	Class=true	Class=false
Class=true	True Positive	False Negative
Class=false	False Positive	True Negative

True Positive (TP) is correctly predicted values which mean that the user is genuine and the output from the model is also the same i.e. genuine user. False Negative (FN) when the user is genuine but the model predicted the user as malicious. False Positive (FP) when the user is malicious but the model predicted him as genuine. True Negative (TN) these are correctly predicted i.e. the user is malicious and the model predicted him as malicious.

Accuracy is the ratio of correctly predicted observation to total observations i.e.

Accuracy = (TP+TN) / (TP+TN+FP+FN)

Precision is the ratio of correctly predicted true values to total true values

Precision = (TP) / (TP+FP)

The recall is the ratio of correctly predicted true values to all values in the actual class

Recall = (TP) / (TP+FN)

F1-score is that the weighted average of precision and recall

F1-Score = 2*(Recall * Precision) / (Recall + Precision)

The percentage of training data and testing data is 80 and 20 respectively. The accuracy of the given method is 83%.

Table 1 : Classifier-Precision-Recall

SN0	CLASSIFIER	PRECISION	RECALL
1	KNN(K nearest neighbour)	1	1
2	Random Forest	1	0.83
3	SVM(Support Vector Machine)	1	1

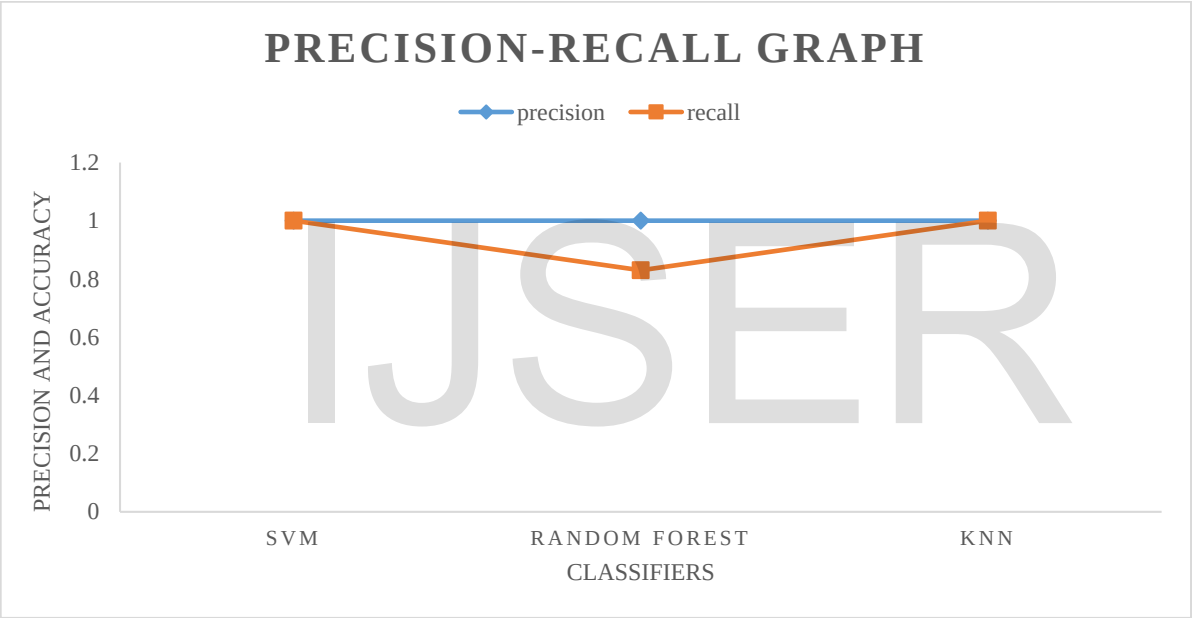
The table compares the values of precision (Positive prediction values) and accuracy with different classifiers such as KNN(K-nearest neighbour), Random Forest, and SVM(Support Vector Machine). The Precision for three classifiers is the same but the recall value of the random forest is less compare to other classifiers.

Table 2: Classifier-Accuracy

SNO	CLASSIFIER	ACCURACY (PERCENTAGE)
1	KNN(K nearest neighbour)	100
2	Random Forest	83
3	SVM(Support Vector Machine)	100

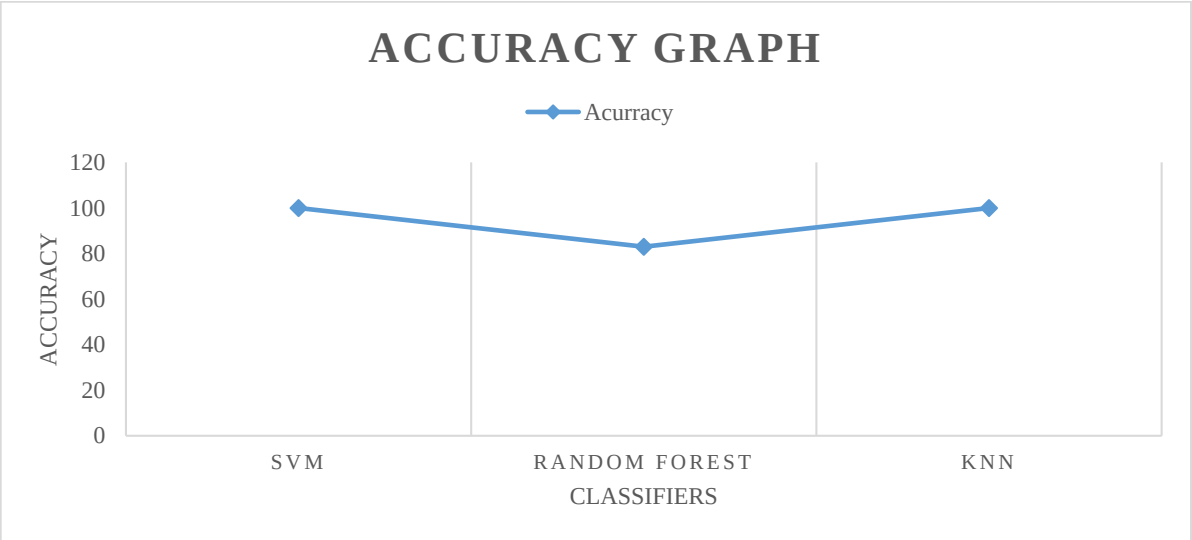
The above table compares accuracy by using different algorithms such as KNN(K nearest neighbour), Random Forest, SVM(Support Vector Machine). The accuracy of random forest is 83% whereas other classifiers are 100%.

Figure 1: Graph Representation for Table 1 Results



This graph is drawn by using Table 1 which represents about precision and recall using different classifiers

Figure 2: Graph Representation for Table 2 Results



This graph is drawn by using Table 2 which represents accuracy using different classifiers Though the accuracy rate of other algorithms is higher the main advantage of random forest is it handles all the uncertainties in features and label and treats them using probability distribution and have less computation-intensive. The maths which is behind Random Forest is simpler compared to KNN and SVM.

5. CONCLUSION:

Not just utilizing the Random forest calculation the model can be prepared utilizing ay order calculations, for example, KNN, SVM, Logistic regression, etc and deep learning as well i.e. neural networks. This calculation helps in the grouping of users and aims in giving users access consequently. It sends an alarm if the individual isn't real. We can likewise consider different boundaries like time, geolocation, login, gadget subtleties. Utilizing time we can get subtleties like login time and logout time, utilizing gadget subtleties we can get the IP address, mac address, and CPU subtleties which can likewise be considered as features.

In spite of the fact that it has its own points of interest, it has hindrances moreover. Acknowledgment exactness by keystroke elements could even be influenced inside the presence of consoles with various qualities inside a comparable climate. Taking everything into account, it is typical that such differentiations don't basically incapacitate the affirmation execution, in this manner, actually enable authentic user recognizing confirmation. This can be contrasted with the biometric identification during which, regardless of the pen utilized, the framework stays prepared to separate between genuine and malicious users.

Besides, cautioning rates (when a genuine user is evaluated as a malicious i.e. in case of False negative) in keystroke elements are normally high and don't satisfy guidelines in some access control systems.

6. REFERENCES:

[1] Mindaugas Ulinskas, Marcin Woźniak and Robertas Damaševičius,"Analysis of Keystroke Dynamics for

Fatigue Recognition",Computational Science and Its Applications – ICCSA 2017,pp 235-247

[2] Paulo Henrique Pisani and Ana Carolina Lorena, "A systematic review on keystroke dynamics", 10 July 2013,

Journal of the Brazilian Computer Society volume 19, pp 573–587

[3] Wasif Afzal and Richard Torkar, "On the application of genetic programming for software engineering

predictive modeling: A systematic review", September 2011, doi 10.1016/j.eswa.2011.03.041, volume 38, Issue

9, pp 11984-11997

[4] Pierre Thiffault and Jacques Bergeron,"Monotony of road environment and driver fatigue: a simulator study",

2003, doi 10.1016/S0001-4575(02)00014-3, Volume 35, Issue 3, pp 381-391

[5] Fabian Monrosea and Aviel D.Rubinb ,"Keystroke dynamics as a biometric for authentication", February 2000,

doi 10.1016/S0167-739X(99)00059-X, Volume 16, Issue 4, pp 351-359

[6] Danoush Hosseinzadeh and Sridhar Krishnan, "Gaussian Mixture Modeling of Keystroke Patterns for Biometric

Applications", IEEE Transactions on Systems, Man, and Cybernetics, Part C, volume 38, Issue 6

[7] [Manuel Rodrigues, Sérgio Gonçalves, Davide Carneiro, Paulo Novais and Florentino Fdez-Riverola,

"Keystrokes and Clicks: Measuring Stress on E-learning Students", Part of the Advances in Intelligent Systems

and Computing book series (AISC, volume 220), pp 119-126

[8] Steven M. Furnell, Joseph P. Morrissey,Peter W. Sanders and Colin T. Stockel,"Applications of keystroke

analysis for improved login security and continuous user authentication",IFIP International Conference on ICT

Systems Security and Privacy Protection-1996,Part of the IFIP Advances in Information and Communication

Technology book series (IFIPAICT),pp 283-294.

[9] Fabian Monrose, Michael K. Reiter, Susanne Wetzel,"Password Hardening Based on Keystroke Dynamics ",

ACM 1999, doi 10.1145/319709.319720, pp 73-82

[10] KH Lam , KA Meijer, FC Loonstra , EME Coerver , J Twose , E Redeman, B Moraal, F Barkhof , V de Groot,

BMJ Uitdehaag and J Killestein, “Real-world keystroke dynamics are a potentially valid biomarker for clinical

disability in multiple sclerosis”, https://doi.org/10.1177/SAGE-JOURNALS-UPDATE-POLICY, Multiple

Sclerosis Journal 1–11,2020.

[11] Tom Olzak, “Keystroke Dynamics: Low Impact Biometric Verification”, September 2006,

https://www.adventuresinsecurity.com/Papers/Keystroke_Dynamics.pdf

[12] Mayur Mahadev Sawant et al.,”Keystroke Dynamics: Review Paper”, International Journal of Advanced

Research in Computer and Communication Engineering, Vol. 2, Issue 10, October 2013, ISSN (Print) : 2319-

5940,ISSN (Online) : 22781021.