المؤتمر الوطني الخامس لكليات الحاسب الآلي بالجامعات السعودية
The Fifth National Conference of Saudi Computers Colleges 2022
21-22 December 2022

Detect keyloggers by using Machine Learning

Sarah M. Alghamdi ∗ , Elham S. Othathi †, and Bassma S. Alsulami‡
Computer Science Department, King Abdulaziz University Jeddah, Saudi Arabia

Abstract

In today's world, the field of information technology is rapidly evolving. Maintaining security and privacy is a major problem for cyber professionals. According to studies, the quantity of new malware is rapidly increasing. A keylogger is a highly sophisticated malware that records every keystroke made on the machine, allowing the attacker the potential to steal enormous amounts of critically sensitive information invisibly without the authorization of the message's owner. Identifying keylogger is important to avoid data loss and sensitive information leaking. Anti-viruses can detect keylogger via heuristic and behavior analysis, but if the keylogger is not a Known threat, antivirus or anti-malware software cannot detect it as a virus.

Objectives

The research objective is to find a model capable of detecting and preventing keylogger–spyware harm on the system after training the model on the keylogger's main features and behavior to ensure the CIA triad on the system and enhance the security in our information and digital systems.

The following objectives support this aim:

1) Study the keylogger works, types, characteristics, and methodology Keylogger used.

2) Classify the dataset with different types of machine learning.

3) Increase devices security by preventing Keylogger software.

Classification

Classification will be done to segment the data tested with the data obtained from several classifiers, SVM, kNN, Decision Tree, Random Forest will be implemented in this paper.

Support Vector Machine

This type can be used in the two categories of regression classification. It includes a separating hyperplane used to differentiate between the plots or classes. The selection of the hyperplane is made according to the best separating area; this algorithm is effective to be used in high dimensional spaces. Assume input data is x j = (x j1 ...x jn), while · is a map that maps function space to the space of a mark y with several vectors, mathematically labeling (x 1 , y 1), ... (x m , y m).

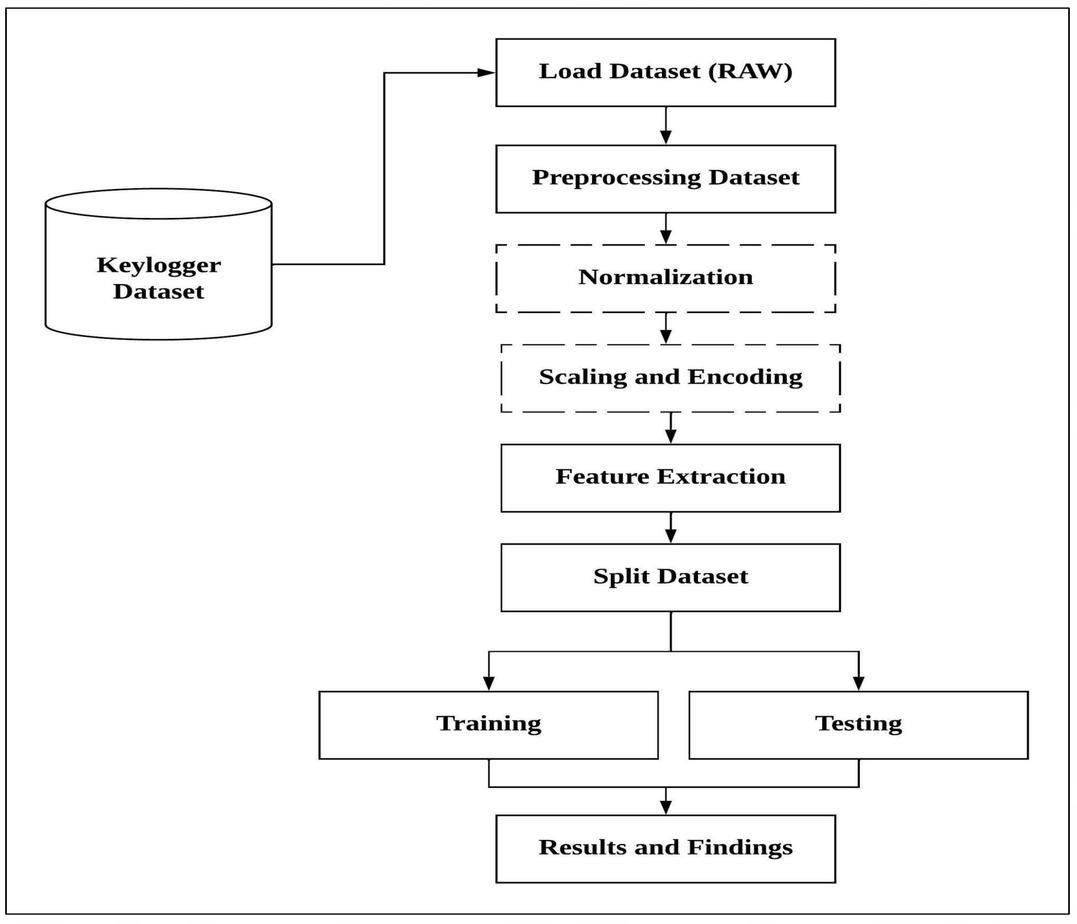
Decision Tree

A decision tree is defined as a conceivable results tree shaped from planning the related decisions to externally and unequivocally speaking to make a decision. It employs a model of decisions usually applied in machine learning and information mining. It has obstructed various machine learning zone, including; regression and classification. The decision tree begins from an unsociable node that branches into possible results and incomes like that. The X-variables are the decision nodes. Each node has its attribute; variables a and b are boundaries of attributes that divide decisions into three tree paths by names or numbers. The variables in a class are the leaves of a tree, so the analyzed object should be categorized.

Random Forest

Based on the bootstrap data sample, RF For tree construction, an ensemble of classification trees is used to achieve bagging and random variable selection. Each split is picked at random from all of the variable possibilities. Randomness is injected into various random subsamples by growing each tree and partially randomly determining dividers. Every tree is cultivated to produce a low lip. The low correlation for individual trees is maintained by bagging and random variable collection. The algorithm gives rise to a forest of the ensemble by averaging over a wide collection of low-life, high variance, yet low correlation trees.

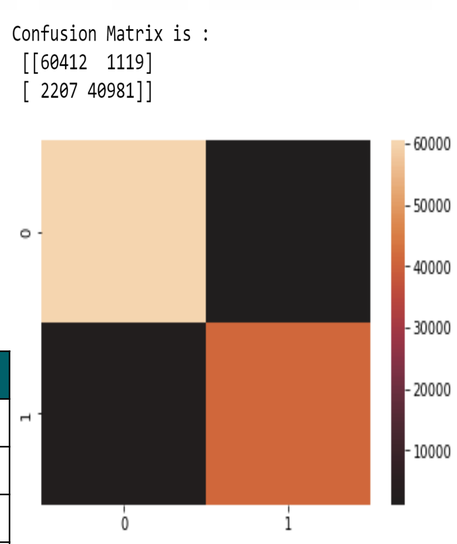
Methodology



Result

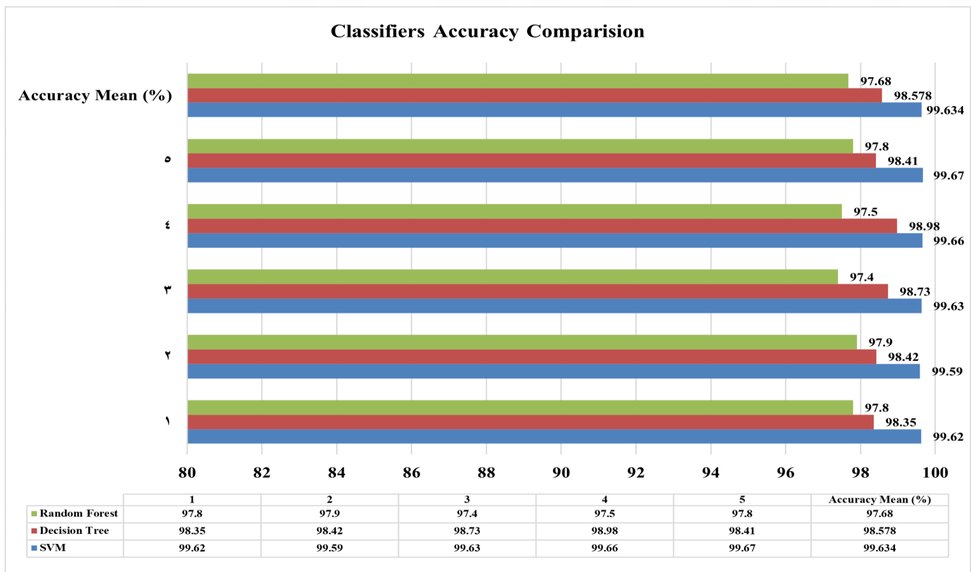
In the proposed method, we used several evaluation methods to evaluate the results and obtain the findings, such as confusion matrix, accuracy, precision, and recall. The findings have been studied and evaluated using the mean of the main five runs, the standard deviation, and the variance between the findings in the main five runs.

Random Forest				
Run	Accuracy	Precision	Recall	F-score
1	0.98	0.96	0.96	0.98
2	0.98	0.96	0.96	0.98
3	0.98	0.96	0.96	0.98
4	0.98	0.96	0.96	0.98
5	0.98	0.96	0.96	0.98



Decision Tree				
Run	Accuracy	Precision	Recall	F-score
1	0.97	0.97	0.96	0.97
2	0.97	0.97	0.96	0.97
3	0.97	0.97	0.96	0.97
4	0.97	0.97	0.96	0.97
5	0.97	0.97	0.96	0.97

Support Vector Machine				
Run	Accuracy	Precision	Recall	F-score
1	0.95	0.95	0.95	0.95
2	0.94	0.95	0.95	0.95
3	0.95	0.95	0.95	0.95
4	0.95	0.95	0.95	0.95
5	0.94	0.95	0.95	0.95



Conclusion

We can detect the various hazardous activities performed by keyloggers in the system using this machine-learning approach. The keylogger generates a variety of keystrokes and patterns. We built our framework based on these patterns and the keys being pressed. In this study, a keylogger detection model based on machine learning was suggested to detect keyloggers and spyware, and the model was trained using keylogger and spyware datasets to identify host behavior when keyloggers were operating on the system. The findings will be based on numerous metrics and provided based on the categorization report and confusion matrix to determine system performance in identifying keylogger spyware. In the proposed method, we used several machine learning algorithms to classify the keylogger dataset, and the classifiers are SVM, Random forest, and decision tree. The RF achieved the best accuracy of 99.6% and the highest run time, while the Decision tree scored the lowest accuracy of 94% and highest run time. Future work will continue to study the problem by using more advanced technology in classification using deep learning and monitoring the network to understand the problem better and enhance the security of the host and network.