A Comprehensive Review of Intrusion detection by leveraging the Machine Learning Techniques

Abstract

With advancement in digital devices and openness of critical network systems. The threat of cyber-attack and intrusion attack is a major concern to the critical resources of network security and systems. The cyber attacker takes the advantages of weakness and vulnerabilities exist in network system or users devices in order to exploit the various assets and steal valuable information. Over the past decades, Internet and computer systems have raised numerous security issues due to the explosive use of networks. Intrusion Detection System (IDS) provides the protection against any kind of network attack by detecting network intrusions from the suspicious traffic data. However, most of the intrusion detection data suffers from high dimensionality, due to this IDS leads to the degraded performance and lower prediction rate of any kind of new intrusion. Therefore, this work presents a comprehensive analysis of an intrusion detection system for network traffic data using machine learning techniques. The proposed system employed machine learning and data mining techniques in order to detect different kind of intrusion from network traffic on NSL-KDD dataset. The obtained results of proposed work show that the intrusion detection model for network system can efficiently and effectively identify intrusion behavior and malicious intension with higher accuracy.

Keywords: Intrusion detection, Machine learning, cyber-attacks, data mining, Network security, Artificial Neural Network

1. INTRODUCTION

In this section definitions of different key words that are used in this works are presented to have better understanding.

1.1 Data Mining

Usually, data mining also referred to as information discovery is the method of analyzing facts from distinctive views and abbreviation it into precious statistics that can be used to enhance revenue, cuts fees, or both. For analyzing information data mining software is a single wide variety of analytical gear. Data mining allow users to examine statistics from many different measurement or angles, classify it, and review the relationships recognized. Theoretically, data mining is the procedure for finding correlation or styles together with dozens of fields in massive relational databases. On the equal time as massive-scale information generation has been growing separate transaction and information systems, data mining provides the connection between the two. Data mining techniques are at the present time used in various applications such as finance, medical, telecommunication and other. Data mining tasks are generally categorized as classification and prediction; outlier analysis; cluster analysis. Among these the two most popular tasks, classification and prediction are widely used. There is various machine learning techniques which can used to perform following data mining functionalities:

- **Clustering**: Clustering can be invented as identification of comparable classes of objects. Via the usage of clustering techniques we are able to supplementary identify dense and sparse regions in item area and this can determine average distribution pattern and correlations among information attributes. There are three styles of clustering methods that is portioning method, Density based method, and Model based method.

- **Classification:** type is commonly carried out in statistics mining technique, which employs a set of pre-labelled examples to increase a model. Fraud detection and credit danger programs are in particular nicely suitable to this type of take a look at. This method frequently employs choice tree or neural community-primarily based classification algorithms.

- **Prediction:** The prediction is an character records mining strategies that discover courting among impartial variables and association among dependent and independent variables. For instance, the prediction analysis technique can be used in transaction to predict profit for the potential if we consider sale is an independent variable, profit could be a dependent variable.

- **Association:** affiliation and correlation is typically to discover frequent object set findings among huge facts units. This type of judgment helps businesses to create certain decisions; Association Rule algorithms need to be able to produce rules with confidence values less than one.

- **Decision Tree:** One of the most used data mining techniques is decision tree because its model is straightforward to understand for users. In selection tree technique, the root of the selection tree is a easy query or condition that has a couple of solutions.

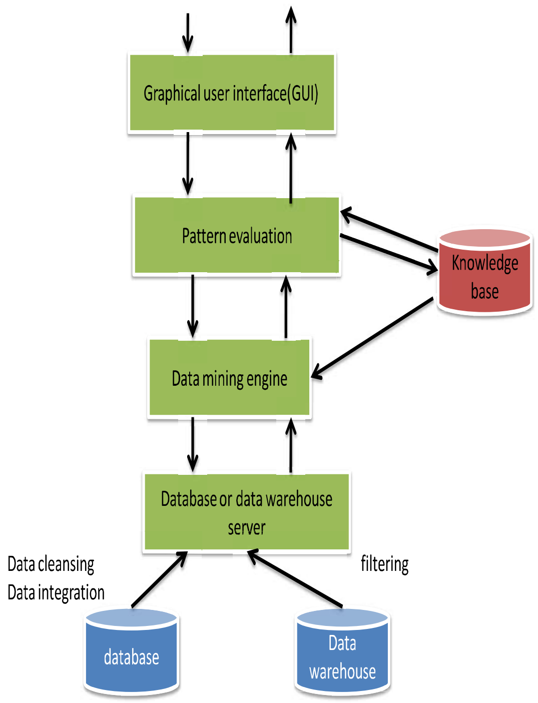

Figure 1.1 Architecture of a Typical Data Mining System.

There are data mining methods which invent in the artificial intelligence and the machine learning. In modern years the utility of the methods has been proven also in network attack. Data mining aims at describing specific patterns which may be present in data. These patterns, exposed in historical data, may be used to support future decisions relating to attack. Such knowledge may also have a huge value for decision making in action planning, hazard analysis and other predictions. Earlier to the mining procedure it is essential to grow sufficient amount of data. This may possibly require integrating data from multiple varied information sources and transforming it into a form which is specific to a target decision support application. Afterwards the data has to be organized for knowledge extraction. The next step comprises induction of rules which may be encouraging in the fining attacks on network The Figure 1.2 shows the process of discovering knowledge from data.

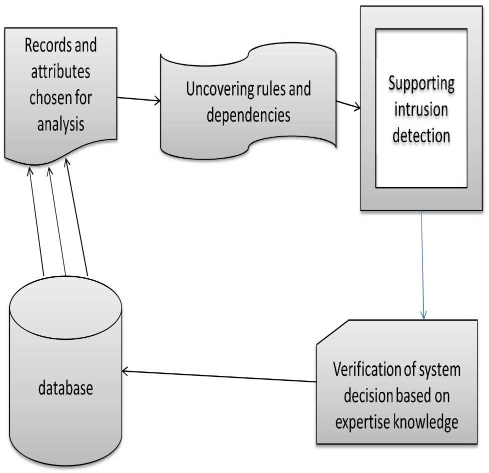
Records and attributes chosen for analysis
Uncovering rules and dependencies
Supporting intrusion detection
database
Verification of system decision based on expertise knowledge

Figure 1.2 Discovering Knowledge from Data

The data mining process is a complex process and it has been divided into several steps:

- Domain analysis and data accepting
- Data collection
- Data analysis and pre-processing
- Data reduction and transformation
- Attribute selection
- Reduction in number of dimensions
- Normalization
- Aggregation
- Selection of data mining
- Visualization
- Evaluation
- Knowledge utilization

1.2 Intrusion Detection System

Now a day's community security infrastructure depends on to network Intrusion Detection Gadget (NIDS). NIDS gives protection from acknowledged intrusion attacks. It's impossible to forestall intrusion attacks, so organizations need to be prepared to deal with them [13]. Intrusion detection machine (IDS) Intrusion detection system (IDS) is a defensive mechanism whose elementary purpose is to keep work going on considering all possible attacks on a computer system. Intrusion detection is a one type of process used for to detect suspicious activity and malicious activity both at network level and host level. Two main Intrusion Detection techniques available are anomaly detection and misuse detection. In anomaly based totally detection device, audit facts is used to differentiate atypical facts from ordinary statistics. On the other hand, in misuse detection gadget, additionally known as signature primarily based IDS, using patterns of well-known attacks to match with audit data and identify them as intrusions. Functioning of misuse detection models is very much similar to that of antivirus applications. Misuse IDS can Examine network.

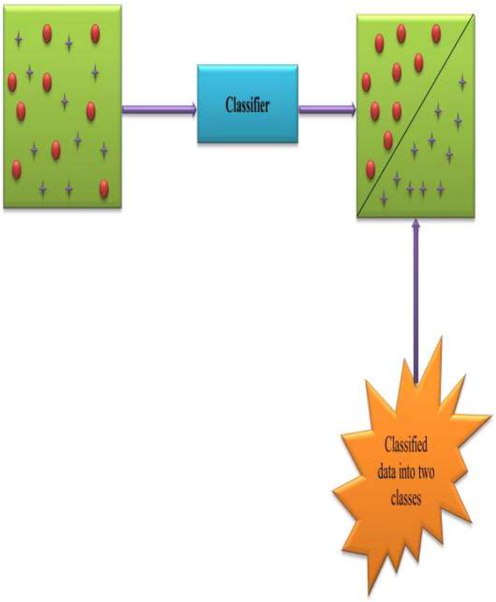

Figure 1.3 Two Class Classifier of IDS

S.NO.	INTRUSION ATTACK	DESCRIPTION WITH EXAMPLE
1	**Worms**	**It is a section of code or program which replicates or folded themselves from computer to computer without any use of host file. The whole document such as word or excel vary from one system to another system should consider as worm. E.g. PrettyPark**
2	**Trojans**	**Trojans having a malicious code or malevolent code, when triggered it may causes loss or even theft of data. E. g Mail Bomb**
3	**Physical Attack**	**It is an attempt to damage the physical components or hardware components of networks or computers. E. g Cold boot**
4	**Password Attack**	**In this type of attack, password is obtained within a small or short period of time which is indicated by a sequence of login failure. E. g SQL injection attack**
5	**Information gathering attack**	**It is Able to gather or gain information to find out intrusion or vulnerabilities. These activities are generally done by scanning the whole computers or networks. E. g XMAS scan**
6	**Malware**	**Malware is a program that is basically written intentionally to harm or attack system. It is not the buggy software or any programs.**
7	**Vulnerability**	**The Program that is written by humans. Programmers are sometimes forget to or unaware of cross t's and dot i's and those mistakes create strange behaviour in the programs, it create a hole or make a secret path that malware or attacker could use to access system more easily and effectively, that is known as a vulnerability.**

Table 1.1 Types of Attack in Network

1.3 Motivation

Over the past decades, Internet and computer systems have raised numerous security issues due to the explosive use of networks. In the present day intrusion detection gadget is most useful because we are now use any time computer internet and the intruder easily attacks on our network. In this decade many modern techniques and computational system have been emerged in order to facilitate their operations .Nowadays many types of attacks are found for this we have data in databases and facts Warehouse. The essential data include just primary information about attack such as call, kind etc. In order to facilitate the storage and maintenance of this huge data a new type of system has been emerged that is known as intrusion detection system. Its goal is to carry out early detection of malicious pastime and in all likelihood save you greater extreme damage to the included system by using IDS. It originates in the trade intelligence in order to support attack decisions. The research may improve the procedure of finding attacks and stop that attack as well as decrease the risk of finding attack mistake or the time of a finding attack. This may revolve out to the critical especially in data is very important. This may also turn out to be a time saving effort. The research area which tries to find for practice of knowledge wrenching from data is described as know-how discovery or data mining.

1.4 Objectives

The main purpose of this paper is to discover the maximum promising data mining techniques which yield great effects in terms of all performance measurements like accuracy, sensitivity and specificity with reduced number of features. The evaluation is performed on NSL-KDD Dataset from UCI repository site (UCI Repository).

In order to reach the main goal of paper following objectives to be fulfilled:

a) Analysis of NSL-KDD dataset.

b) Identification of the most common data mining algorithms implemented within

Various datasets.

c) Identification of the most promising feature choice method for reducing features.

d) Improvement of algorithm for type of IDS assaults.

e) Evaluation of the specific strategies used on different datasets for comparing performance.

1.5 Research Methodology

Research is a way of knowledge improvement. The main motivation of research is to extend human knowledge rather than to create or discover new things. To conduct master's paper first literature survey is done for analysis of the existing facts and figures. In this literature survey assessment is performed not most effective at the papers associated with the intrusion detection device however also diverse assaults like u2r,r2l,everyday,prob and so on that has been discussed in bankruptcy 3. Subsequent evaluation of intrusion detection gadget used in community. there are numerous IDS which might be utilized in community that has discussed later on this chapter. the following step is to become aware of the common records mining algorithms and feature choice method which has been carried out and sooner or later to pick a aggregate of information mining approach and feature choice technique for assessment.

2. RELATED WORK

Intrusion detection gadget (IDS) is one of the maximum crucial research vicinity for community and facts safety with fast improvement of internet in all around the world . IDS are a classifier that can classify the network information as ordinary or assault. Researchers are generally used characteristic choice approach with statistics mining because it assists to lessen time for result and locating and some of author used combination of a couple of technique for

higher effects. There are many authors who use KDD99 information set and many of them makes use of NSL-KDD data set because both are bunch mark data available in UCI Repository site after a decade review we find an ensemble model with feature selection. There are lots

] of research work already done by the various authors as explained below with some recent publications:

Gang Kou et al.,(2009) have compared various classification algorithms like C4.5,SVM,Naive

Bayes, Logistics, CART, See5 with proposed model Multi-class multi-criteria mathematical programming (MCMP) model to develop IDS. MCMP model is best classifier for KDD99 data set in case of multi class classification problem .

V. Balon Canedo et al., (2011) proposed a new KDD winner method consisting of discretizations, filters and various classifiers like Naive Bayes (NB) and C4.5 to develop a robust

IDS. The proposed classifier gives high accuracy i.e. 99.45% compare to others.

Mohammad Saniee Abadeh et al., (2011) have suggested three kinds of genetic fuzzy system

based on Michigan, Pittsburgh and iterative rule learning (IRL) to deal with intrusion detection

as high dimensional classification, in which genetic fuzzy Michigan approach is better to deal intrusion detection problem compare to others.

M. Govindarajan et al., (2011) have proposed a brand new technique, that is ensemble of Multilayer Perceptron (MLP) and Radial foundation characteristic (RBF). The proposed ensemble version gives better improvement compare to its man or woman model

Saurabh Mukharjee et al. have discussed new feature reduction method known as feature validity based reduction method (FVBRM) applied on one of the efficient classifier Naive Bayes

on reduced data set with 24 features for intrusion detection. Result obtained in this case is better as compare to case based feature selection (CFS), gain ratio (GR), info gain ratio (IGR) to design efficient and effective network intrusion detection system.

Y. Li et al., (2012) have recommended regularly characteristic decreased (GFR) approach implemented on green classier assist Vector gadget (SVM) with KDD99 records set. SVM classifier offers high category accuracy as ninety eight.62% for intrusion detection in case of 10-fold pass validation.

L. Koc et al., (2012) have delivered Hidden Naive Bayes (HNB) model with promotional k-c programming language discretization and have interaction feature choice technique. They have compared their proposed model with conventional Naive Bayes method. The proposed models offers exceptional end result as ninety three.72% accuracy and zero.66% error charge in multiclass class for intrusion detection.

Mrutyunjaya Panda et al., (2012) have suggested hybrid method with aggregate of random wooded area, dichotomies, and ensembles of balanced nested dichotomies (cease) for binary class trouble, which gives detection fee 99.50% and occasional false alarm rate of zero.1%. They evaluated the overall performance of model with different measures like F-fee, precision and don't forget.

Nagaraju Devarakonda et al., (2012) have presented Hidden Markov Model(HMM) is a simplest kind of dynamic Bayesian network model for intrusion detection system. They have taken only five features (protocol_type, flag, src_bytes, dst_bytes, and, count) out of 41 to develop robust IDS to classify the attacks.

Hesham Altwaijry et al.,(2012) have suggested Bayesian network to improve the accuracy of R2L type of assault. Experiment accomplished with distinct function subset of KDD99 statistics set and offers better effects for R2L attack with detection price eighty five.35% using three features.

Muamer N. Mohammed et al., (2012) have proposed a new method focus on improving intrusion system in wireless area network by using support vector machine. This technique produced a better result in terms of detection rate and eliminating false positives and false negatives.

John Zhong Lei et al., (2012) have used various techniques like K-means, SOM, Improve Competitive Learning Network (ICLN), and Supervised Improve Competitive Learning Network (SICLN) to develop intrusion detection system. They have suggested SICLN technique and achieved high accuracy like 99.66% compare to others.

Y. Y. Chung et al.,(2012) have proposed a new hybrid network intrusion detection system using intelligent dynamic swarm based rough set (IDS-RS) and simplified swarm optimization with weighted local search (SSO-WLS) strategy for intrusion data classification. Proposed hybrid model SSO-WLS improve the overall performance of the network intrusion detection system with 99.3% accuracy.

B. Kavitha et al.,(2012) have expand a brand new rising Neurosophic common sense classifier (IDS), which is extension/aggregate of fuzzy logic ,intuitionist good judgment and three-valued logics that offers a excessive detection rate and coffee false alarm rate compare to others.

Shih-Wei Lin et al., (2012) have proposed a new technique based on feature selection and decision tree rule which is combination of support vector machine (SVM), decision tree (DT), and simulated annealing (SA).The proposed technique achieved 99.96% of accuracy as best model in case of 23 features.

3. Problem Description

According to the analysis of literature survey many authors have developed novel methods not only for the classification of attack but also for other attacks like u2r,r2l,probe,DoS etc.with the help of intrusion detection system. Development of novel techniques for classifying different attacks in network is a current research topic and is very popular among the researchers; it was found that many popular journals are publishing many research articles frequently. Problems related to classification have identified its applicability in real sense. The main problem is that all attacks share the same features. This may mystify a practicener. As a result we can state that one of the important problem in this research is decision making regarding a attack in which one attempts to match patterns and attack but on the other hand it does not be attached to typical behavior. From literature survey we conclude that each author proposed individual data mining techniques for calculating performance measures and from individual model we observe that accuracy does not increase. There is one more problem related to the number of feature which is used for classification. There are some authors who have applied feature selection methods like ranker method, best first search method etc. With the help of feature selection data quality gets improved as redundant and noisy get removed. And one more advantage is that the dimensionality of the feature space, to limit storage requirements and increase algorithm speed.

- ➢ Analysis of various previous researches comprises following problems:

a) Decision making process does not conform to standard behaviour.

b) Huge number of features due to which quality of data gets affected.

4. Proposed Approach for Classification

After analyzing the previous work of classification of intrusion it found that there is a need of

More reduction in features so that noisy and redundant data get remove and quality of data may get improved for better results. In this paper we have proposed an ensemble model approach of a feature selection for the classification of intrusions. In this proposed approach first model is trained using machine learning techniques and then features get reduced with the feature selection method and after testing of model it is evaluated using various performance measurements. The work flow of the proposed work is shown below with the help of diagram

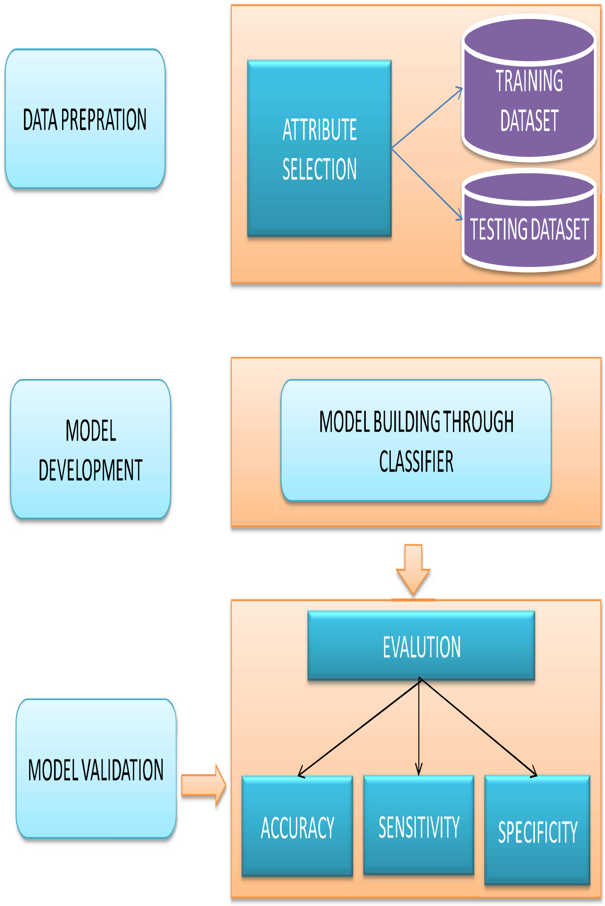

Fig 1.4 Work Flow of Proposed Approach

The work flow diagram show that for finding attacks first we prepared the data with training and testing set and then develop a model for classification and at the last stages a model for Classification and at the last stage we evaluate accuracy of that model.

5. Experiment and results analysis

This paper presents the results conducted with a hybrid technique of ANN and Bayes net with Feature selection method. The technique has been applied to NSL-KDD dataset which consists of 41 features. The experiments are conducted in CLEMENTINE 12.0 version.

This experiment used CLEMENTINE open source data mining tool to analysis of data. In these Experiments we have used NSL-KDD data which is collected from UCI repository data source. These facts set implemented in diverse records mining techniques for categorization of assault and ordinary. The test divided into sections: First evaluation the person and its ensemble model and then practice the feature selection technique on satisfactory version.

5.1. An Ensemble Model for Classification of IDS Data

In this experiment we've implemented different walls of records situate in distinct information Mining strategies like C5.0, SVM, ANN and Bayes net for category of NSL-KDD statistics. Number one we've got applied this data position into diverse individual's facts mining techniques and calculated the accurateness of models. Second we've hybrid the two fashions for category of NSL- KDD information. We have also ensemble ANN and Bayes net for classification of this data which gives higher accuracy compared to each individual's models. Data partitions acting very significant role for accuracy of model. From one partition to other partitions accuracy is changeable and our proposed ensemble ANN and Bayes net gives high classification testing accuracy as 98.74% in case of 65-35% as training-testing partitions.

Table 1.2 Accuracy of Different Model with Different Partitions of Dataset

Model	**60/40 % partition testing**	**80/20% partition testing**	**65/35% partition testing**
ANN	98.23%	97.77%	96.81%
Bayes Net	97.46%	98.68%	97.16%
ANN+ Bayes Net	98.73%	98.47%	98.74%

By using graphical representation we show that ensemble model is give high accuracy as Compare to individual.

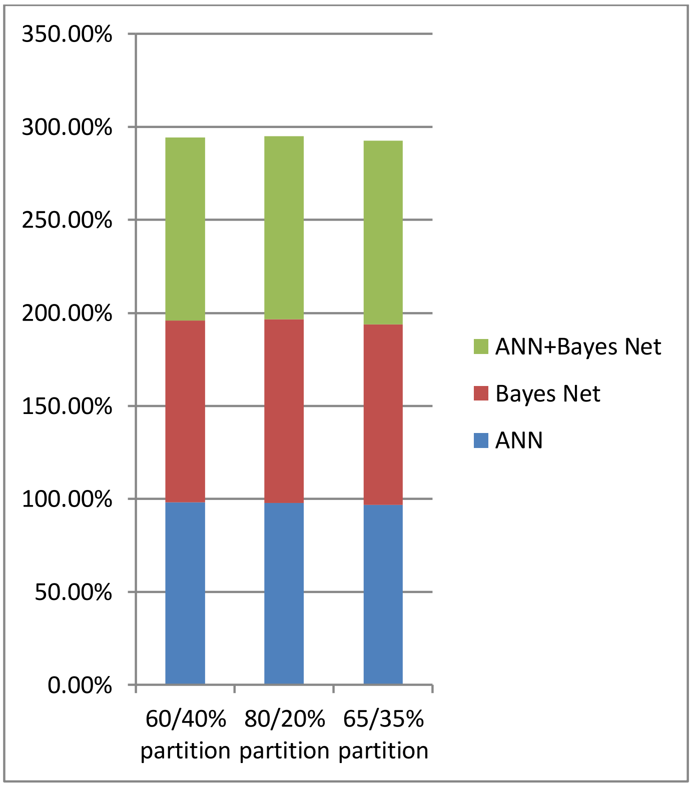

Figure 1.5 Classification Accuracy of Individual and Ensemble Model

6. CONCLUSIONS AND FUTURE WORKS

Due to huge amount of data transmission over public network it is mandatory to protect data and information from the intruders for individual as well as for any organization. ANN artificial neural network were very popular among the researchers, many models with many variations of ANN are developed and integrated with other techniques due its internal good characteristics on the other hand relatively new technique bayes net is widely accepted for development of intrusion detection system (IDS) .This study involves with a comparison of new and old techniques for intrusion related data classification based on two different partitions. An experimental result proves that bayes net is outperforming than ANN at both training and testing stages. After that we are ensemble ANN and bayse net and get highest accuracy with reduced feature. Since outcome of Research work as follows:

1. This research work have developed ensemble model (ANN+ Bayes net) for Classification of attacks. This model gives 98.74% accuracy in case of 65-35% data partition as robust model.

2. Feature selection is also very important role to improve the classification of data set. In this work, rank based feature selection techniques used on NSL-KDD data set and reduce the features. An ensemble of ANN and Bayes net model gives 98.74% of accuracy in case of rank based feature selection technique.

The main goal of this research work is to get higher accuracy with minimum number of features. In this study we have used NSL-KDD dataset which has been downloaded from UCI machine learning repository site. The proposed model can be used in future in following ways-

1. This ensemble model can also used for reduce feature, improve accuracy.

2. This proposed model can be applied to real time data.

3. This proposed model can be expanded with extra results.

In future, it is proposed to reduce features and make an Ensemble model for intrusion detection system having a better accuracy rate.

References

- Ahmadi, M., Sami, A., Rahimi, H., & Yadegari, B. (2013). Malware detection by behavioural sequential patterns. *Computer Fraud & Security*, *2013*(8), 11-19.

- Alazab, M., Huda, S., Abawajy, J., Islam, R., Yearwood, J., Venkatraman, S., & Broadhurst, R. (2014). A hybrid wrapper-filter approach for malware detection. *Journalof networks*, *9*(11), 2878-2891.

- Ali, M. A. M., & Maarof, M. A. (2013, December). Dynamic innate immune system model for malware detection. In *2013 International conference on IT convergence and security (ICITCS)* (pp. 1-4). IEEE analysis. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(4), 1377-128.

- Cai, M., Jiang, Y., Gao, C., Li, H., & Yuan, W. (2021). Learning features from enhanced function call graphs for Android malware detection. *Neurocomputing*, *414*, 301-307.

- Dhammi, A., & Singh, M. (2015, August). Behavior analysis of malware using machine learning. In *2015 eighth international conference on contemporary computing (IC3)* (pp. 481-486). IEEE.

- Gadhiya, S., Bhavsar, K., & Student, P. D. (2013). Techniques for malware
- Gao, Y., Lu, Z., & Luo, Y. (2014, August). Survey on malware anti-analysis. In *Fifth International Conference on Intelligent Control and Information Processing* (pp. 270-275). IEEE.
- Imtiaz, S. I., ur Rehman, S., Javed, A. R., Jalil, Z., Liu, X., & Alnumay, W. S. (2021). DeepAMD: Detection and identification of Android malware using high-efficient Deep Artificial Neural Network. Future Generation Computer Systems, 115, 844-856.
- Kumar, S., Krishna, C. R., Aggarwal, N., Sehgal, R., & Chamotra, S. (2014, March). Malicious data classification using structural information and behavioral specifications in executables. In *2014 Recent Advances in Engineering and Computational Sciences (RAECS)* (pp. 1-6). IEEE.
- Liang, G., Pang, J., & Dai, C. (2016). A behavior-based malware variant classification technique. *International Journal of Information and Education Technology*, *6*(4), 291.
- Pye, J., Issac, B., Aslam, N., & Rafiq, H. (2020, October). Android Malware Classification Using Machine Learning and Bio-Inspired Optimisation Algorithms. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (pp. 1777-1782). IEEE.
- Qiu, H., & Osorio, F. C. C. (2013, October). Static malware detection with Segmented Sandboxing. In *2013 8th International Conference on Malicious and Unwanted Software:" The Americas"(MALWARE)* (pp. 132-141). IEEE.
- Rad, B. B., Masrom, M., & Ibrahim, S. (2012). Camouflage in malware: from encryption to metamorphism. *International Journal of Computer Science and Network*
- Ranveer, S., & Hiray, S. (2015). Comparative analysis of feature extraction methods of malware detection. *International Journal of Computer Applications*, *120*(5).
- Roy, A., Jas, D. S., Jaggi, G., & Sharma, K. (2020). Android Malware Detection based on Vulnerable Feature Aggregation. Procedia Computer Science, 173, 345-316.*Sciences*, *176*, 420-435.*Security*, *12*(8), 74-83.
- Sikorski, M., & Honig, A. (2012). *Practical malware analysis: the hands-on guide to dissecting malicious software*. no starch press.
- Singh, A. K., Wadhwa, G., Ahuja, M., Soni, K., & Sharma, K. (2020). Android Malware Detection using LSI-based Reduced Opcode Feature Vector. Procedia Computer Science, 173, 291-298.
- Yoni Birman ∗, Shaked Hindi, Gilad Katz, Asaf Shabtai(2022). Cost-effective ensemble models selection using deep reinforcement learning.

Information Fusion, Elsevier1-16.

- Yoo, S., Kim, S., Kim, S., & Kang, B. B. (2021). AI-HydRa: Advanced hybrid approach using random forest and deep learning for malware classification. *Information*