Ensemble of Statistically Based Methods for Identifying Features in Intrusion Detection System

Dr. Rahul Adhao
School of Computer Engineering MIT Academy of Engineering, Alandi, Pune 412105.

Mr. Nikhil Kamble
Department of Computer Engineering & IT, COEP Technological University, Shivajinagar, Pune, 411005.

Dr. Vinod Pachghare
Department of Computer Engineering & IT, COEP Technological University, Shivajinagar, Pune, 411005.

Abstract: *An intrusion detection system is an element or part of software that monitors and evaluates network traffic for suspicious activity. Improving the accuracy of an intrusion detection system is an essential goal. Additionally, analyzing the massive volume of data takes a long time. It is necessary to figure out how to efficiently minimize the amount of data in a dataset without sacrificing any crucial or pertinent information. A feature selection model for efficient IDS is presented in the work. CICIDS2017 is the data set used in the research. The approach suggested uses analysis of variance (ANOVA) and chi-square to select features. There are a total of 79 features in the CICIDS2017 data set. In this paper, out of 79 features, the proposed model focuses on 22 relevant features, which are found by a combination of the chi-square and ANOVA methods, which gives improved accuracy and less processing time.*

***Keywords*: Network Security, Intrusion Detection System, ANOVA, CICIDS2017, Statistics.**

INTRODUCTION

Information security has become an important issue as the threats to the data over the network are increasing day by day. An intrusion detection system assists with recognizing host- and network-based system threats. Nowadays, with the rapid growth of computer and database technologies, users have been given more access to high-dimensional data. From this, it can be inferred that machine learning models and data mining applications have improved by a large margin [1]. On the other hand, there are many approaches that are irrelevant, and they often neglect the correlation between a set of selected features. So, in order to overcome this, the feature selection process comes into the picture.

Feature selection is an important data preprocessing stage in data mining. In the proposed system, a model is designed to preserve a small subset of possible features by excluding features with no predictive information, which are irrelevant, useless, and noisy.

Given a large number of data sets such as CICIDS2017, the proposed approach can help to find out which data are different or correlated to each other. An irrelevant feature results in overfitting and can affect the modelling power of the classification algorithm; hence, combining feature selection techniques such as Chi-square and Analysis of Variance (ANOVA), as it requires less processing time, helps in the removal of data that is irrelevant and redundant data from the dataset to increase classification power of algorithms.

The proposed model has two main objectives, which always conflict with each other: one is the maximization of classification accuracy, and the second one is the minimization of a number of selected features.

The proposed model has reduced the number of features, reducing features also reduced the computation cost, and as a result, they can have faster prediction time. The combination of chi-square and ANOVA methods gives 22 relevant features out of the 79 features from the CICIDS 2017 dataset.

This paper's remaining content is as follows: The literature review will be covered in section two. The research gaps and problem statement are explained in section three. Section four gives a detailed description of data preprocessing and an explanation of the algorithms used. System requirements are defined in section five. Section six explains the proposed model, followed by results, discussion, conclusion and further work.

I. LITERATURE REVIEW

Research has been carried out in the feature selection process along with several datasets, and feature selection approaches with the classifiers to give the best intrusion detection system for keeping the system safe. Here, most of the research on the CICIDS2017 dataset-based detection system is taken into consideration.

In the approach of Kurniabudi et al. [3], the CICIDS2017 dataset, which consists of eight traffic monitoring sessions, is used. In this method, the feature selection method is used as a filter, where chi-square is used along with the classifiers random forest and J48, which gives an accuracy of 99.99% and 99.91%, respectively. This approach tries to identify the most important feature of the intrusion detection system. Here, the relevant features selected by the random forest, which is the classifier is 22 in this case, and it was found to be the most reliable classifier as compared to J48, which gives the accuracy lesser than it and significantly affects the accuracy with a total of 52 selected features and also has the longer execution time comparatively to the random forest classifier.

This paper has proposed the method in which the CICIDS2017 dataset with 78 features after processing it with a total of is used with the Random forest as the classifier along with the other four classifiers in the ensemble technique, which gives the accuracy of 99.93% when the random forest is applied and where the intrusion detection is done using the filter feature selection method where symbolic features have been converted into the numeric values with label encoding technique and the methods which are used for feature selection of filter method is chi-square and the optimal features that have been obtained used in the high detection rate of the intrusion[4].

Bayu Adhi Tama et al. [5] have given the approach in which the dataset used is CICIDS2017, in which the filter feature selection method is used. Anova is used as well, as the random forest is used as the classifier, which picks the features that produce the trees and builds the tree models, which are built using a small number of variables in the trees, which also reduces the overfitting problem in this case and gives the accuracy of 99.92%.

CICIDS2017 dataset is used in this approach by Shaik Razia et al. [6] with the filter-based feature selection method in which the chi-square, Pearson's correlation method and information gain methods are used with the random forest and decision tree as the classifier, which gives the accuracy of 99.84% with the random forest. In this process, sets of all features selected are carried out, along with selecting the best subset out of it by preprocessing and using the classifiers mentioned, which gives the performance of the given system.

Murtaza Ahmad Siddiqi, Wooguil Pak et al. [7] have proposed a method in which the CICIDS2017 dataset has been used along with the filter-based feature selection method in, here basic data cleaning is done after this power transformation technique, which includes methods like Min Max scalar is used then the features are reduced using preprocessing of the dataset. Here, the filter-based feature selection method is Pearson's correlation method, where the variance of the features is calculated using the respective formula. Here, the classifier used is a Decision tree, which gives an accuracy of 99.58%.

In this, Yonghao Gu and Kaiyue Li et al.[8] proposed a model which focuses on methods for detecting DDOS attacks from data. They implemented a hybrid feature selection technique based on Hadoop. This discovers the best feature sets and also uses a Semi-supervised K-means algorithm with the hybrid feature selection, which is used to detect attacks from the "CICIDS- DDoS attack-2017" dataset. The semi-supervised k-means algorithm with a hybrid feature selection algorithm helps to achieve better detection performance. As concluded, other feature selection algorithms are not better than the hybrid feature selection method, but the provided process does not verify the generalization and robustness of the algorithms used.

A multimodal-sequential intrusion detection approach can extract the different level features, and this can process the feature selection information separately and more efficiently. Based on the approach, it investigated the performance of detecting attacks. In this experiment, results showed that this approach, based on the method of multi-viewing features, helped improve performance, providing a new point of view for researchers to understand the nature of the network behaviour. They got to know the experimental result accuracy of this method, 94%. Method, where the CICIDS2017 dataset used the result, has not provided a sufficient number of attacks—moreover, the view which improves the accuracy in the intrusion detection system [9].

In this approach, Murizah Kassi et al. [10] used the CICIDS2017 dataset for intrusion detection evaluation. The proposed method developed an intrusion detection system that improves security response for network systems. Classification is done based on the Deep Neural Network and the K-Nearest Neighbors, which are used with machine learning for K-Nearest Neighbors, and the Deep Learning method is used for deep neural networks. A deep neural network has higher correctness, which shows an accuracy of 92.93% as compared with the performance of K-Nearest Neighbors at 88.24%. These results are according to the accuracy, precision, time, and confusion matrix. In the provided approach, K-Nearest Neighbors lacks the accuracy of the detection scheme.

In this paper Mohammad Noor Injadat etal. [11], proposed a framework approach that was used to reduce computational complexity in the detection of the performance on the CICIDS2017 dataset. In this approach, a comparison between information gain and the feature selection techniques, which are based on the correlation, are used, which affects the performance of detection and the time complexity. They also compared the K-Nearest Neighbors and random Forests classifiers based on that concluded that K-Nearest Neighbors and Random Forest have an accuracy of 99%, but while choosing a subset of features which is most suitable and the optimizing parameters to enhance the performance, it failed.

The proposed approach is to build an intrusion detection system using the convolutional neural network method, which provides security to the internet. It is evaluated using terms such as overall accuracy, attacks, and training overhead, which has the potential to detect innovative attacks. This model has an accuracy of 99.78%. However, this model has not given perfection and results in the class imbalance issue of the training dataset [12].

In this approach, Deepak Kshirsagaret al. [13] proposed a model that detects the Dos attacks on the CICIDS2017 dataset using machine learning and neural network algorithms such as Random Forest and Multi-Layer Perceptron. The result of this research concluded that Multi-Layer Perceptron gave an accuracy of 98.87% with 30% of training data. The random forest gives 99.95% accuracy when 50% of training data is given. The research concluded that random forest gives higher accuracy in detecting the Dos attacks than the Multi-Layer Perceptron algorithm. The proposed model does not classify multiple DOS attacks.

Feature reduction using filter-based feature reduction techniques such as Information Gain Ratio, Correlation, and Relief are proposed in this paper. From the CICIDS 2017DoS dataset, the proposed method reduced features to 24 out of 77 features presented in the dataset. The experiment concluded the reduced features with 99.95 % accuracy [14].

The above literature study shows a problem of overfitting in the existing systems, which can be removed from the proposed system using a combination of chi-square and ANOVA. Working on classification problems requires the best accuracy; the proposed system provides better accuracy with less processing

time. Many existing systems use chi-square, and its advantage is the robustness of data. Random forest used in the proposed model is suitable as it uses the Ensemble Learning technique, reduces overfitting, and improves accuracy.

Research develops an Intrusion Detection System (IDS) based on a deep learning model called Pearson-Correlation Coefficient - Convolutional Neural Networks (PCC-CNN) to detect network irregularities. Important characteristics from linear-based extractions and convolutional neural networks are combined in the PCC-CNN model [2]. It carries out multiclass classification for many kinds of assaults in addition to binary classification for anomaly detection. Three publicly accessible datasets—NSL-KDD, CICIDS-2017, and IOTID20—are used to assess the model. To assess model performance, we first train and test five distinct PCC-based machine learning models (support vector machine, logistic regression, linear discriminant analysis, K nearest neighbour, classification and regression tree, and K nearest neighbour).

II. COMPONENTS OF PROPOSED MODEL

This section presents the dataset, feature selection methods and the classifiers. The section also briefly describes the proposed methodology.

A. Dataset Used

The proposed model is built using the CICIDS2017 dataset for data preprocessing. The CICIDS2017 dataset contains eight different files, and each file contains traffic data of attacks. The details of this dataset are given in [15]. The CICIDS2017 dataset contains up-to-date network attacks. The dataset contains records (2827876) and columns (79), and it fulfils all the criteria of real-world attacks. In addition, some new types of threats that were not present in any older datasets are addressed here.

B. Feature Selection

Filter feature selection is taken to design the intrusion detection system. The dataset is given to the search strategy, which selects the quality features [16].In feature evaluation, it evaluates and selects the set of features that are best for the model. The best features are further given to the classification model, which gives the result in the form of accuracy, better accuracy, and less processing time.

C. Chi-square test

The chi-square test is used to find out whether the variables are independent or related to each other. It is best suited for categorical variables as it is not based on parameters like mean but on frequencies. It can be applied to complex contingency tables, which have many classes. It is useful for testing the hypothesis [18].

D. ANOVA test

It is used to find the features among the categorical variable and numerical variable; thus, on the alpha value, the features are further selected [19-20].

E. Random Forest as a classifier

Machine learning classifiers are used to analyze data and categorize it into one or more sets of classes. They allow users to update the model from time to time with new learning data [19]. The top five classification models are decision trees, naive Bayes, k-nearest neighbours, support vector machines, and artificial neural networks.

In the random forest, each individual tree spits out a class prediction, and the class with the most votes becomes our model's prediction[19]. Random forest is an ensemble machine learning method used for classification. It is used to rank the importance of variables in a natural way. It tends to give more accurate results as it uses an ensemble algorithm.

III. PROPOSED MODEL

The dataset is input, and then data preprocessing is done. After that feature selection, Chi-square and Anova are applied to it, getting the reduced combined features. Classifier like Random Forest is applied to the reduced features, giving accuracy and time as the results from the model. Fig 1 shows the flowchart of the proposed model.

This section provides a brief explanation of the algorithms used and illustrates the flow chart of the proposed model. The next section gives the system requirements.

IV. EXPERIMENTATION AND RESULT ANALYSIS

The experiment was performed on a 32-GB RAM Windows 10 operating system. Python3 IDE, such as Kaggle, PyCharm, or JupyterNotebook. Machine learning libraries and Scikit Learn libraries NumPy, SciPy, and Pandas are used.

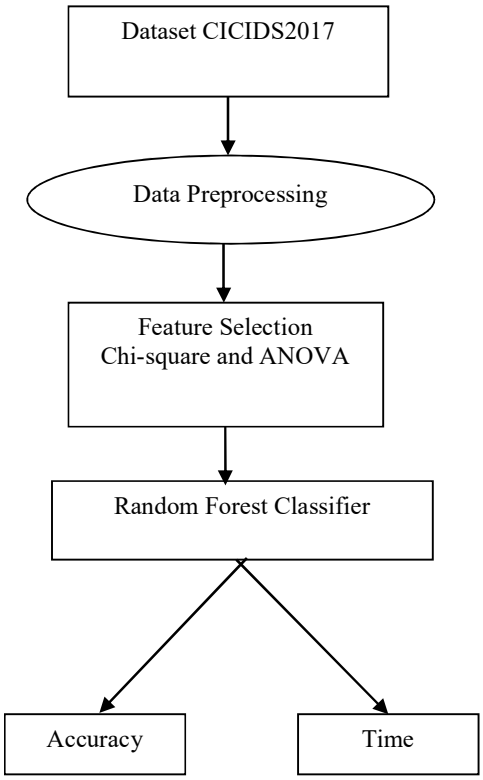

Fig. 1 Flowchart of the Proposed Model

A. Data Preprocessing

The eight CSV files of the CICIDS2017 dataset, as shown in Table 4.1, are combined into one CSV file, giving the dimensions (2827876, 79). Only 20% of the dataset is used in this experiment. Firstly, the columns having infinity, NAN, missing values or string values are dropped, and identical rows are removed, finally reducing it to 67 columns, which are all in numerical data type. The dataset has an imbalanced class distribution of attacks, which has more majority class than the minority. This oversampling or undersampling can duplicate, delete or merge the classes. To overcome this class imbalance, the resampling methods of Python documentation using Random Under Sampler and SMOTETomek are used. After this dataset is split into two portions: train data (70%) and test data (30%). Next, this data is analyzed using feature selection methods.

B. Feature Selection and Classifications

Chi-square feature selection and ANOVA are used for model building. Applying the preprocessed dataset on Chi-square and ANOVA methods, the best 20 features from both

Table 1 Comparison of our model with other existing models and classifiers

ATTACK NAME	Our Model	Naïve Bayes	J48	Ada boost	Random forest	One-Class SVM [22]	Ensemble [4]	MLP[21]
BENIGN	99.07	72.36	99.87	88.63	99.99	84.84	95.83	99.51
SSH PATATOR	99.96	99.64	99.99	99.79	100	80.26	99.99	99.89
PORTSCAN	99.99	96.76	99.95	94.31	100	60.04	99.99	99.72
DOS SLOWLORIS	99.99	98.43	99.96	99.79	100	8.38	99.99	99.95
DOS HULK	99.16	96.32	99.97	94.05	99.99	91.35	99.99	91.82
DOS SLOWHTTPTEST	99.99	98.27	99.99	99.77	100	98.66	99.97	99.94
BOT	100	75.30	99.98	99.92	100	46.15	99.98	99.95
FTP PATATOR	99.98	99.85	99.99	99.70	100	5.34	99.97	100
INFILTRATION	100	99.88	99.99	99.99	100	93.8	99.68	99.99
WEB ATTACK-XSS	100	98.61	99.98	99.97	100	9.76	99.99	99.97
DDOS	99.96	98.38	99.98	95.41	99.99	39.94	95.93	99.91
DOS GOLDENEYE	99.99	99.01	99.98	99.63	100	72.39	99.67	99.20
WEB ATTACK-SQL INJECTION	99.99	99.72	100	99.99	100	6.31	99.99	99.77
HEARTBLEED	99.99	99.99	100	99.99	100	99.54	99.99	99.49
WEB ATTACK-BRUTE FORCE	99.99	99.03	99.98	99.94	100	80.26	99.98	99.91

Individual methods are found to give combined output as the 40 most relevant features. After this combination, the intersection of features is done from these two methods and gets 22 features. The output of 22 features is given to the classification algorithm used here, which is Random Forest. Accordingly, the accuracy of the model, along with the accuracies of 15 individual attacks present in the CICIDS2017 dataset, is found. Table 1 Comparison of our model with other existing models and classifiers. As shown in Table 1 above, getting the 22 features, Random Forest is applied to it and gives an accuracy of 99.07%. From the confusion matrix, the accuracies of individual 15 attacks are found. A confusion matrix was generated from each feature selection technique. To evaluate our model, we take into consideration the accuracy and time required, along with the detection rate (DR) and false alarm rate (FAR).

For evaluation purposes of our model, we compare it with some of the existing models of One class SVM [22], Ensemble model [4], Multi-Layer Perceptron (MLP) [21] and different classifiers, namely NaiveBayes, J48, Adaboost, Random Forest. Here, for classifiers, the results of accuracies are taken using WEKA by applying the inbuilt algorithms of the classifiers taken on the preprocessed dataset of CICIDS2017.

Table 1 summarizes the performance of our model with the existing models and classifiers for the 15 different attacks of the CICIDS2017 dataset. It shows that our model gives 99.07% accuracy for Benign attacks with almost the highest accuracies for the remaining 11 attacks and 100% accuracies for Bot, Infiltration and Web Attack-XSS. Overall, our model performance is best for most of the different attack types.
Table 2 summarizes the overall performance of our model as compared to other models. Our model gives a false alarm rate(FAR) of 0.359%, a detection rate(DR) of 99.07% and an accuracy of 99.07%. The model building time is 24.72 seconds, which is the lowest when compared with all the models except the Ensemble model, but it has less accuracy than our proposed model. So, it can be concluded that our model performs at its best with lower processing time.

Table 2 Overall performance of our model relative to other models and classifiers

	Our Model	Naïve Bayes	J48	Ada boost	Random Forest	One class svm[22]	Ensemble [4]	MLP [21]
FAR	0.359	58.33	0.036	0.11	0.002	0.15	5.57	0.35
DR(Overall)	99.07	72.36	99.87	88.63	99.99	98	99.99	99.51
ACCURACY (%)	99.07	72.36	99.87	88.63	99.99	98	95.83	99.51
TIME (Sec)	24.72	53.83	847.86	151.89	64.74	50	17	115

Table 2 shows the results of our model compared to other existing models and classifiers, which include false alarm rate(FAR), detection rate(DR), accuracy and the time taken for the model building.

VI. CONCLUSIONS

Since intrusion detection systems are the first line of defence against attacks, developing them is essential. This study proposes an intrusion detection model based on random forest ANOVA and feature selection using Chi-square. Using the CICIDS2017 dataset, the model was verified and assessed. It achieved up to 99.07 % accuracy and took 24.72 seconds to complete. Because two distinct feature selection strategies were used throughout the training process, the model demonstrated a significant potential for identifying assaults. The model is compared with the other existing models and classifiers, and the results are obtained using WEKA software. The goal of future research is to integrate additional models with other classifiers and a probabilistic approach.

REFERENCES

[1] Surbhi Solanki, Chetan Gupta and Kalpana Rai, "A Survey on Machine Learning based Intrusion Detection System on NSL-KDD Dataset", International Journal of Computer Applications, vol. 27, no. 3, pp. 36-39, June 2020.

[2] Isabelle Guyon, AndréElisseef, "An Introduction to Variable and Feature Selection", The Journal of Machine Learning Research, vol.3, pp. 1157-1182, March 2003.

[3] Kurniabudi, Deris Stiawan, Darmawijoyo, Mohd Yazid Bin Idris, Alwi M Bahmdi, Rahmat Budiarto, "CICIDS-2017 dataset feature analysis for information gain for anomaly detection", IEEE Access, vol.8, pp. 132911-132921, July 2020.

[4] Adel Binbusayyis, Thavavel vaiyyapuri, "Identifying and benchmarking key features for cyber intrusion detection system", vol.7, pp. 106495-106513, August 2019.

[5] Bayu Adhi Tama, Lewis Knereyye, S M Riazul Islam, Kyung Sup Kwak, "Enhanced anomaly based detection in web traffic using stack of ensemble classifier", vol.8, pp. 24120-24134, February 2020.

[6] Vainkat Ramani Varanasi, Shaik Razia, "Intrusion Detection using Machine Learning and Deep Learning", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.

[7] Murtaza Ahmad Siddiqi, Wooguil Pak, "Efficient Filter Based Feature Selection Flow for Intrusion Detection System", International Workshop on Emerging ICT, November 2020.

[8] Yonghao Gu, Kaiyue Li, Zhenyang Guo, Yongfei Wang, "Semi-Supervised K-Means DDoS Detection Method Using Hybrid Feature Selection Algorithm", vol.7, pp. 64351-64365,2019.

[9] Haitao He, Ligang He, Jiadong Ren, "A Novel Multimodal-Sequential Approach Based on Multi-View Features for Network Intrusion Detection", vol.7, pp. 183207-183221, 2019.

[10] Kayvan Atefi, Habibah Hashim, Murizah Kassim, "Anomaly Analysis for the Classification Purpose of Intrusion Detection System with K-Nearest Neighbors and Deep Neural Network", IEEE 7th Conference on Systems, Process and Control (ICSPC), pp. 269-274,2019.

[11] Mohammad Noor Injadat, Abdallah Moubayed, Ali Bou Nassif, Abdallah Shami, "Multi-Stage Optimized Machine Learning Framework for Network", IEEE Transactions on Network and Service Management, 2020.

[12] Samson Ho, Saleh Al Jufout, Khalil Dajani, Mohammad Mozumdar, "A Novel Intrusion Detection Model for Detecting Known and Innovative Cyberattacks Using Convolutional Neural Network", IEEE Open Journal of the Computer Society, vol.2, pp. 14-25, January 2021.

[13] Shreekhand Wankhede and Deepak Kshirsagar, "DoS Attack Detection Using Machine Learning and Neural Network", Fourth International Conference on Computing Communication Control and Automation, pp. 1-5,2018.

[14] Deepak Kshirsagar, Sandeep Kumar, "An efficient feature reduction method for the detection of DoS attack", January 2021.

[15] Samarjeet Borah and Ranjit Panigrahi, "A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems", International Journal of Engineering and Technology, vol.7, pp. 479-482, January 2018.

[16] Sumaiya Thaseen Ikram, Aswani Kumar Cherukuri, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM", Journal of King Saud University - Computer and Information Sciences, vol.29, Issue 4, October 2017.

[17] Arowolo, M.O., Abdulsalam, S.O., Saheed, Y.K. and Salawu, M.D, “A Feature Selection Based on One-Way ANOVA for Microarray Data Classification”,Journal of Pure & Applied Sciences,vol.3, pp. 30-35 December 2016.

[18] Abbas Salami, Farnaz Ghassemi, Mohammad Hasan Moradi, “A Criterion to Evaluate Feature Vectors Based on ANOVA Statistical Analysis”, 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME), pp. 14-15,2017.

[19] Tony Yiu, “Understanding Random Forest: How the Algorithm Works and Why it is So Effective.” [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.[Accessed: 27-April-2021].

[20] Jason Brownlee, “How to Combine Oversampling and Under sampling for Imbalanced Classification."[Online]. Available:https://machinelearningmastery.com/combine-oversampling-and-undersampling-for-imbalanced-classification. [Accessed: 27-April-2021].

[21] M.H Abdulraheem, N.B Ibraheem, “A Detailed Analysis of New Intrusion Detection Dataset", Journal of Theoretical and Applied Information Technology, vol.97, pp. 4519-4537, September 2019.

[22] Hanan Hindy, Robert Atkinson, “Towards an Effective Zero-Day Attack Detection Using Outlier-Based Deep Learning Techniques”, vol 1, June 2020.

[23] Bhavsar, M., Roy, K., Kelly, J. et al. Anomaly-based intrusion detection system for IoT application. Discov Internet Things 3, 5 (2023). https://doi.org/10.1007/s43926-023-00034-5