JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

An Efficient Intrusion Detection System Using Machine Learning Model

Mrs. Priyavengatesh, Resarch Scholar,
Dr. R. Kannan, Associate Professor,
Sri Ramakrishna Mission Vidyalaya College of Arts and Science Coimbatore.

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10

Abstract

Evolving network technologies and increase in reliance on internet applications, cyber attacks are on the rise. More people and devices get connected across internet and generate large volume of traffic data. Utilizing such connected networks, exploiters attack using fraudulent activities for many reasons. To mitigate such malicious activities, an intelligent system capable of analyzing traffic and attacks becomes essential. Machine learning techniques are increasingly been utilized to detect anomalies and intrusions in network traffic. The performance of machine learning techniques largely depends on the dataset dimensions and the type of information present inside the data. This paper proposes a tree based ensemble model to detect intrusions in network traffic which addresses the feature dimension problem and model performance on detecting intrusions. The performance of the proposed model is compared against different machine learning techniques such as LDA, Logistic Regression, NN and SVM.

1. INTRODUCTION

Early network security rely on firewall and encryption however firewall and encryption security measures are easily evaded by the attackers with growing technology and communication development. Monitoring network traffic for intrusions has become essential to ensure the network is protected from intruders. Intrusion detection system (IDS) offers security to the network through continuous monitoring of network activities and alarm of malicious or anomalous network behavior (Othman et al., 2018). Intrusion detection system monitors network traffic flows

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

and alerts activities that breaks the network policies (Liao et al., 2013). An intrusion is an effort to escape the network security or an attempt to violate the security policies. Successful breach of security policies results in exploitations such as spreading malware, phishing, stealing of information, denial of service etc. Intrusion detection system is mainly divided into signature based and anomaly based intrusion system (Khraisat et al., 2019) based on detection mechanism. Signature based IDS monitor for known threats and the known threats are referred to signatures and also known as rules or patterns and the drawback of signature based system is that it cannot monitor for newer threats or signatures (Masdari & Khezri, 2020). Anomaly based IDS monitor for deviations in the traffic flow from previously detected traffic flows. The main drawback of this type of IDS is that it produces higher false positives for traffics that are not present earlier in the network. The quality of detection depends on the learning rate of machine learning models. The better learning of machine learning models yields better accuracy otherwise the models produce false detections. The learning rate of machine learning models is affected by noise or outliers, redundant data and large number of features (Chandrashekar & Sahin, 2014). Machine learning techniques are classified into supervised learning and unsupervised learning (Sarker, 2021). In supervised learning models, the models are trained using known labels and detect patterns for the corresponding labels present in the data, example includes classification and regression. In an unsupervised learning, the models are trained using unlabelled data and the detect patterns present inside the data example includes clustering and association (Soysal & Schmidt, 2010).

Network traffic generates large number of traffic data sent and received across request and reply. Traffic refers to the amount of data sent over a period of time in a network (Shafiq et al., 2020). Network flow data involves information of packets sent, packets received, ports, source and destination IPs, bytes, start time, end time, protocol information etc and these are stored as features in the network data. The number of features and the size of the data affect the performance and computation time of the machine learning models. To reduce the features and to improve machine learning model performance techniques such as feature selection, feature extraction and feature reduction are employed (Venkatesh, & Anuradha, 2019). Feature selection refers to

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

selecting features which contain more information about the class labels. Feature extraction is transforming features into new a feature without losing the information. Feature reduction refers to removing irrelevant features and irrelevant features are the features that do not carry information about the classes (Miao & Niu, 2016).

Feature selection is mainly divided into filter, wrapper and embedded methods. Filter method selects features based on the importance and the importance is calculated using statistical methods. Wrapper method selects the feature subsets in forward or backward direction based on the model performance and embedded methods select the features as a part of the model and utilize the selected features to complete the model. Feature selection is employed to reduce the computational time and complexity, and to improve the predictive power of the learning models. In this study an ensemble tree based classifier is proposed to classify the network traffic into benign and attack. Since network traffic data generates large volume of data for a given time, understanding the anomalies is difficult as attackers mimics the normal traffic patterns. The complexity for considering each network flow features becomes expensive in terms of computation and memory allocation and to reduce the complexity choosing the features that contribute to the classify benign and attack is required. The proposed model creates a subset of features that have higher relationship with the class types and utilize the subset features to classify the traffic into benign and attack. The efficiency of the proposed method is evaluated on KDD Cup 99 and compared against state-of-art machine learning techniques. This paper is organized into five sections, section 2 discusses the related works, section 3 discusses proposed methodology; section 4 discusses the experimental analysis and section 5 presents the results and discussion and finally concludes the paper.

2 RELATED WORK

Intrusions in the form of attacks are evolving rapidly and attack types include malware, phishing, password exploitation, Dos etc. Mitigating such attacks require sophisticated techniques to understand network resources such as machines, types of networks, communication processes and patterns of exploitation. Application of machine learning algorithms on network traffic data to detect network intrusions is extensively studied. This section details various literatures that utilize

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

machine learning algorithms for intrusion detection.

(Abrar et al., 2020) investigated different machine learning models for classifying traffic data on NSLKDD dataset. Random selection of features is used to reduce the features and four different feature sets are created. SVM, KNN, LR, MLP, RF, Extra trees, and DT models are used to evaluate the feature set created. Extra trees, RF and DT performed well over other models with 99% of accuracy.

(Ahanger et al., 2021) investigated four machine learning models for intrusion detection in traffic data. RF, DT, MLP and SVM models are used for classify network traffic data NLSKDD data. The traffic features are randomly selected into three sets containing 23 features in the first set, 15 features in the second set and 12 features in the third set. RF showed highest accuracy on first set (23 features). The study concluded that RF can be applied to detect intrusion in the network.

(Amaizu et al., 2020) employed different machine learning models and compared the performance on three different dataset for intrusion detection namely NSL-KDD, UNSW-NB15 and CSE-CIC-IDS2018. The study uses PCA for feature extraction and deep neural network model for classification. The performance of DNN showed highest accuracy on NSL-KDD dataset with 97.89%, 89.99% for UNSW-NB15 dataset and 76.47% on CSE-CIC-IDS2018.

(Serinelli et al., 2020) evaluated machine learning models on KDD CUP99 and NSL-KDD dataset. The proposed work illustrated the training of machine learning models (SVM, RF, XGBoost, Neural Network) and evaluated on the capabilities of the models to reduce the false alarm rate. The features that are highly correlated are removed and features are selected using RF feature selection. RF achieved highest accuracy of 98.99 on KDDCUP99 and 97.93% on NSL-KDD dataset. FAR of XGBoost is comparatively low on KDD Cup99.

(Bhati & Rai, 2020) investigated different SVM techniques such as linear SVM, Quadratic SVM, Fine Gaussian SVM and Medium Gaussian SVM for intrusion detection on NSL-KDD dataset. The performance of the SVM models is evaluated using accuracy and ROC. Fine Gaussian SVM achieved highest detection accuracy of 98.7% than other models.

(Krishnaveni et al., 2020) proposed an effective anomaly detection system using SVM. The proposed model is tested on NLS-KDD

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

dataset using info Gain feature selection. The selected features are trained and tested in RBF kernel and the performance is evaluated against different machine learning models. The highest ranking features of about ten features are selected and applied to the model. The performance of the proposed model is compared against linear SVM, Logistic Regression and KNN. RBF SVM achieved highest accuracy of 96.34% while LSVM achieved 92.65%, Logistic regression achieved 92.41% and KNN achieved 92.87%.

(Pradeep Mohan Kumar et al., 2021) proposed a hybrid IDS to improve the classification rate and reduce false alarm rate. The proposed model is based on a fuzzy classification which new rule sets are built using Genetic algorithm and PCA. Initially the features are selected using PCA and are optimized for new rules through GA and classified through Fuzzy classifier (GA-Fuzzy). The performance of the model showed 99.6% accuracy using six features on NSL-KDD dataset.

(Bhati, & Rai, 2020) proposed an ensemble model to detect intrusion in network. The performance of proposed model is evaluated on KDDcup99 and NSL KDD dataset for intrusion detection. The ensemble model is constructed using randomized extra trees and the model achieved 99.97% on KDD-Cup99 and 99.32% on NSL-KDD dataset. The study recommended ensemble models for intrusion detection.

(Ogundokun, et al., 2021) proposed two different techniques to classify intrusion in the network. The first technique was developed using PSO and decision tree while the second technique is developed using PSO and KNN. The two models are evaluated on KDDCup99 dataset. The evaluation result shows that PSO-KNN achieved an accuracy of 96.2% and PSO-DT achieved 98.6% of accuracy. The false positive rate for PSO-KNN achieved a lowest score of 0.004.

(Hindy et al., 2021) proposed an intrusion detection system based on similarity learning and the proposed model is tested on CICIDS2017, NSL-KDD and KDDCup99 dataset. The similarity based learning utilize Siasame network, a twin ANN networks using constructive loss function. The proposed model achieved accuracy of 84% for CICIDS2017, 88% for KDD Cup'99 and for NSL-KDD is 91%.

(Mohammad & Alsmadi, 2021) proposed a novel feature selection method for intrusion detection. The proposed feature

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

selection 'HW' is a statistical method and performance is compared with IG and chi-square method. The selected features are trained using decision tree and NB model. The performance of decision tree model achieved 99.36% on selected features and 99.56% on all features and NB achieved 89.59% on all features and 88.31% on selected features and finally the study concluded that decision tree was able to produce more decision rules while using proposed feature selection method.

(Dwivedi et al., 2021) proposed an intrusion detection system using grasshopper optimization algorithm (GOA). The proposed GOA is a hybrid of ensemble features selection and GOA. The selected are trained on SVM kernels and are compared. The performance of the feature selection is compared against CMIM, mRMR and JMI methods. On NSL-KDD dataset, the proposed feature selection trained with SVMR achieved highest accuracy of 96.08% outperforming other models and on KDDcup99 dataset the proposed model achieved 95.15% of accuracy. The proposed out performed GA-SVM, PSO-SVM on both the datasets.

3. PROPOSED METHOD

The main objective of the proposed method is to reduce the computational complexity and dimensional problems of having large number of features in the traffic data and to improve the classification performance of the model. Features with no information about the target classes add noise to the classification model and increase the error rate. To reduce the error rate, features that contain high information on target classes are selected and applied to the classification model. The proposed method involves creation of subset using feature selection in the first step and classifies traffic data using subset of selected features created. For dataset D with X features and target class y, a scoring function S is applied on each feature-class pair and the score C_i for each features is calculated. Features with high C_i are included in the subset D_{sub} while features with lower scores are eliminated.

The scoring function S computes the Chi-square value for each feature-class pair $((X_1,y),(X_2,y)\ldots X_n,y))$ for the given dataset D. The scoring function chi-square is a statistical test that measures the relationship between features and target. The relationship is interpreted as the degree of association between X features and y target, higher the chi-square score, higher the relationship between a feature and the target class. The scoring function is given by,

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

$$C_i = \sum \frac{(X_k - e_k)2}{e_k}$$

where X_k is the observed frequency, e_k is the expected frequency and sum refers to summation of each rows of a feature. Observed frequency is the number of instances for target class k and expected frequency is the expected number of instances for target class k when there is no relationship between feature and target class k. The lower chi-square score indicates the weaker association between the feature and the class and it describes the degree of independence of the feature to the target class. The probability of chi-square value for a feature is independent from the class is determined by the p-value. The p-value greater than 0.5 indicate the features are independent from the class while p-value less than 0.5 indicate that the features are dependent and have association with the target class. Features with high chi-square values are selected and added to the subset D_{sub} while other features are removed.

Ensemble tree models are powerful and use random method to build collection of trees. Several studies have shown that ensemble tree models have good prediction and classification accuracy. The proposed model inspired by (Breiman 2001) builds classification trees randomly using bagged samples BT and prediction are made by averaging the majority of vote of each ensemble of trees. The subset D_{sub} contains the selected features and D_{sub}, $X=(x_1, x_2, x_3 \ldots x_n)$ is a feature vector and y is the classification of traffic data into benign (y=0) and attack (y=1) drawn from bagged samples BT with replacement. Features in X predicts y using bagged ensembles of classifier bCL where $bCL= (bCL_1(X), bCL_2(X) \ldots bCL_n(X))$. Each ensemble of classifier $bCL_1(X)$ is a decision tree with hyperparameter P and is denoted by $P= (P_1, P_2, P_3 \ldots, P_n)$. The decision tree is denoted as $bCL_n(X)=bCL(X|P_n)$ and each decision tree with hyperparameter P_n votes y in the feature vector and the class with majority of vote is selected as prediction result. P_n determines the subset of each bagged ensemble trees and the corresponding class C_1 for each bagged ensemble classifier $bCL_n(X)$ is given by $bCL_n(1 \leq n \leq N)$. The corresponding class score Cs is calculated using votes and the number of trees in each bag which is denoted as,

$$Cs(X, C_1) = \frac{v(X, C_1)}{btrees}$$

and the majority of vote among btrees is given by,

$$C^i = mvote\{C^{btrees}\}bCL_n$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

The ensemble classifiers bCL accuracy is estimated using out-of-bag observations and each ensemble classifier $bCL_n(X)$ predicts the OOB samples. The OOB samples are the leftovers of bagged samples BT. OOB errors are the misclassification of OOB samples by the ensemble classifiers. The mean squared error (MSE) of the classifier bCL for OOB samples is given by,

$$\text{MSE}^{oob} = \frac{1}{oob_n}\sum_{n}(Y - Y'^{oob})^2$$

The major advantage of bagged trees is that all features are used for node split in a tree and the final prediction is based on the ensemble of decision trees vote aggregation. Since the results are averaged in bagged trees, which reduces the variance between features and the training set could not alter the prediction performance where as in trees that are deep and not pruned, the variance is high. The performance of the proposed model is tested against features selected with sequential forward selection (SFS) and machine learning models such as linear discriminant analysis (LDA), Logistic Regression (LR), Neural Network (NN) and SVM on KDD CUP 99 dataset.

4. Experiment and Analysis

4.1 Dataset

The proposed method is evaluated using NSL-KDD dataset. The KDD CUP 99 dataset was collected by DARPA using network traffic TCP dump (Stolfo et al., 2000) and it is an improved version of DARPA98 dataset. The data is collection of network traffic for three weeks with 48, 38,430 records as a part of IDS evaluation program (Lippmann et al., 200). The NSL-KDD dataset is modified dataset without any redundant data and it contains 41 features and labeled into DOS, R2L, U2R, probe, and normal. Attack type DOS refers to denial of services, R2L refers to unauthorized access from remote, U2R refers to unauthorized access to local superuser, and probe refers to surveillance and port scanning. The train set approximately contains 125975 records and the traffic data is grouped into basic feature, traffic feature and content feature. For this study only binary classification is considered and different attack types Dos, R2L, U2R, probe are renamed into attack.

4.2 Evaluation Metrics

The performance of classification model is summarized and visualized using confusion matrix. The confusion matrix for binary

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

classification is given in Table 4.2.1. It represents the actual values and predicted values. The performance is interpreted using TP, TN, FP and FN where TP refers to True Positives which represent the number of positive samples correctly classified as Positives, TN refers to True Negatives which represent the number of negative samples classified correctly as Negatives, FP refers to False Positives which represent the number of Negative samples incorrectly classified as Positive and FN refers to False Negatives which represent the number of Positive samples incorrectly classified as Negative. Accuracy, Sensitivity, specificity, Precision and F-Score are the metrics that explains the classification model's performance.

Table 4.2.1 Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	True Positive TP	False Positive FP
	Negative	False Negative FN	True Negative TN

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{F-score} = \frac{TP}{TP + ½ (FP+FN)}$$

5. Results and Discussion

To improve the classification of network traffic data and to reduce the computational complexity, tree based ensemble model is proposed. Traffic data consists of large number of features that eventually increase the training time and demand more memory space. The proposed model selection features based on the degree of association between features and the target class and removes features that have weak associations with the target class and classify traffic data using ensemble of tree classifiers. This section discusses the performance of the proposed method while using all features and selected features and the results are compared against LDA, LR, LSVM and NN models.

Feature selection remove unwanted features and reduce the dataset dimension, the reduction of feature dimension eventually lowers the computation time and improve prediction results. The sequential forward selection method is wrapper method that selects

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

the features with respect to the model performance. Using random forest as the base classifier, the SFS select best features in forward direction. Five best features are selected by the SFS wrapper method to classify network intrusion. The selected features include 'src_bytes', 'dst_bytes', 'dst_host_same_srv_rate', 'dst_host_serror_rate' and 'dst_host_rerror_rate' with a model performance of 99.7% average score and std error of 0.0002 (Table 5.1). The proposed model using Chi-square estimation selected five best features which include 'src_bytes', 'dst_bytes', 'count', 'dst_host_count' and 'dst_host_srv_count'. The chi-square value and corresponding p-value is given in Table 5.2 and it is noted that the selected features have p-value of less than 0.000 which shows a strong association between selected features and the target class.

Table 5.1 SFS score for features selected (k=5)

Feature names	Avg score	Std error
Src_bytes	0.963	0.0004
Src_bytes, dst_host_same_srv_rate	0.994	0.0004
Src_bytes, dst_host_same_srv_rate, dst_bytes	0.994	0.0002
Src_bytes, dst_host_same_srv_rate, dst_bytes, dst_host_serror_rate	0.996	0.0002
Src_bytes, dst_host_same_srv_rate, dst_bytes, dst_host_serror_rate, dst_host_rerror_rate	0.997	0.002

Table 5.2 Chi-square score for feature selected (k=5)

Feature	Chi Score	p-value
Src_bytes	3.340	0.000
Dst_bytes	1.746	0.000
Count	6.525	0.000
Dst_host_count	9.574	0.000
Dst_host_srv_count	9.574	0.000

The proposed model achieved highest accuracy of 99.65%, while NN achieved 98.73%, LSVM achieved 95.28%, LR achieved 94.53 and LDA achieved 94.51% on all features Table 5.3. The proposed model show higher performances with respect to sensitivity

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

(99.80%), specificity (99.48%), precision (99.54%), F1 score (99.67%) and NPV (99.77%) over other models. Next to the proposed model, neural network achieved sensitivity (98.79%), specificity (98.66%), precision (98.85%), F1 score (98.82%) and NPV (98.59%) while LR, LDA and LSVM perform less than 98% on all features.

Table 5.3 Performance of machine learning models using all features

Models	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**NPV**
LDA	94.51	94.51	94.51	95.27	94.89	93.64
LR	94.53	94.39	94.69	95.43	94.91	93.48
LSVM	95.28	94.77	95.89	96.50	95.63	93.88
NN	98.73	98.79	98.66	98.85	98.82	98.59
Proposed	**99.65**	**99.80**	**99.48**	**99.54**	**99.67**	**99.77**

The proposed model on selected features for k=5 using SFS achieved highest accuracy of 94.91%, while NN achieved 90.80%, LSVM achieved 87.54%, LR achieved 88.38% and LDA achieved 88.27%, Table 5.4. The sensitivity (99.62%), F1score (95.02%) and NPV (99.6%) of the proposed model also outperformed other models while specificity (90.43%), precision (90.83%) of the proposed model is decreased. Neural network achieved sensitivity (98.41%), specificity (81.96%), precision (86.23%), F1 score (91.92%) and NPV (97.82%) and the sensitivity and specificity of LDA has higher score of 92.37% and 94.13% over other models. The decreased score of specificity and precision of the proposed model on features selected with SFS shows that the model is losing its ability to classify as the information on the selected features do lower association with the target class.

Table 5.4 Performance of machine learning models using selected features (SFS)

Models	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**NPV**
LDA	88.27	85.42	**92.37**	**94.13**	89.56	81.54
LR	88.38	85.63	82.28	94.04	89.64	81.88

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

LSVM	87.54	84.43	92.13	94.04	88.98	80.08
NN	90.80	98.41	81.96	86.23	91.92	97.82
Proposed	**94.91**	**99.62**	90.43	90.83	**95.02**	**99.6**

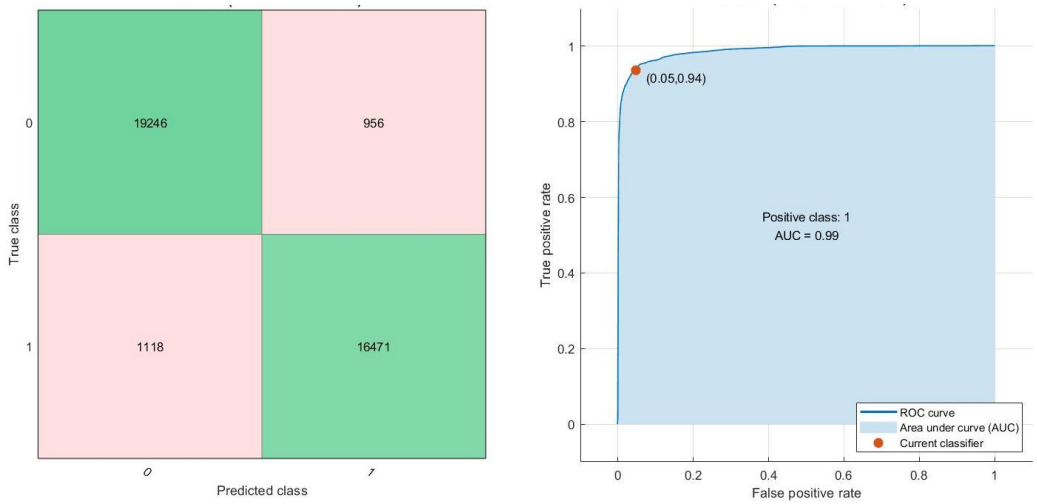

Figure 5.1 confusion matrix and ROC for LDA (all features)

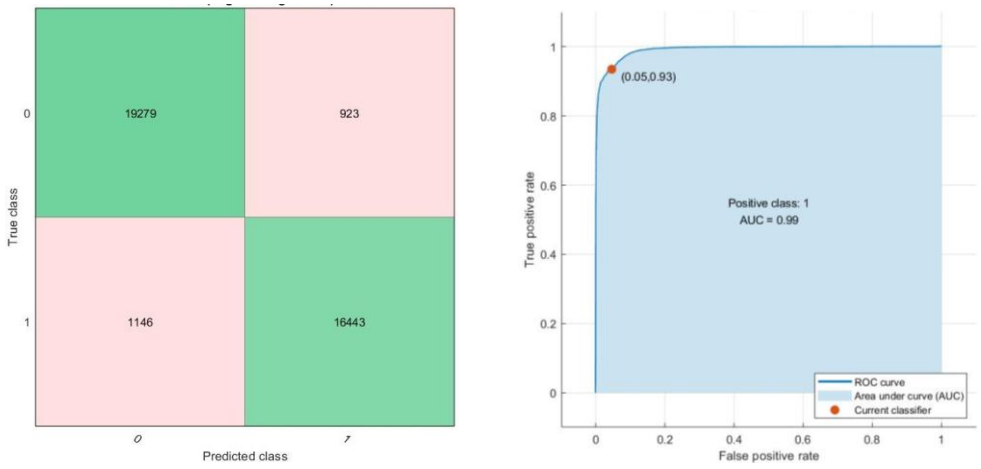

Figure 5.2 confusion matrix and ROC for LR (all features)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

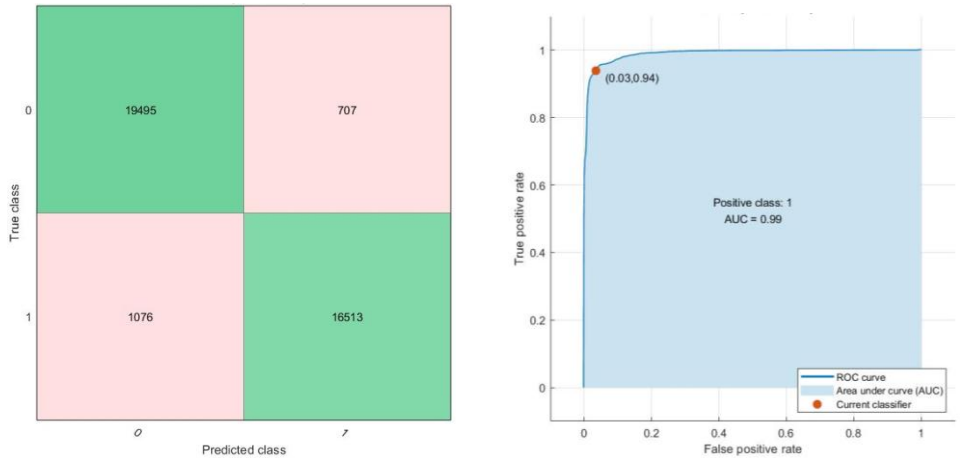

Figure 5.3 confusion matrix and ROC for LSVM (all features)

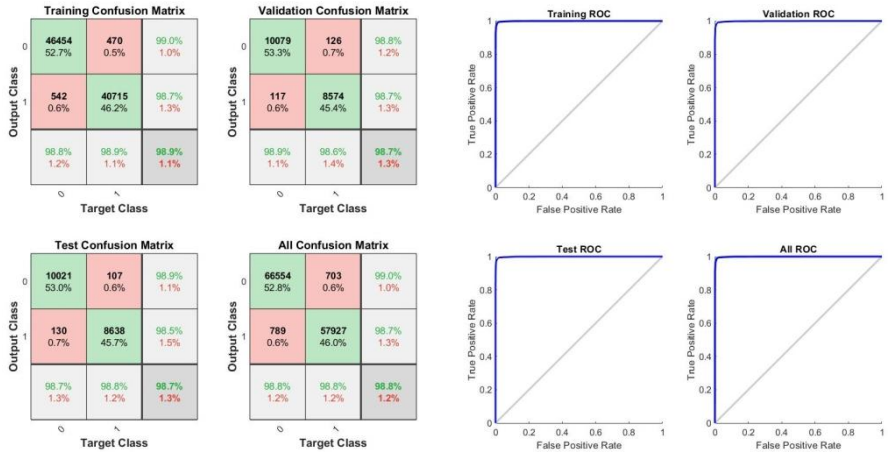

Figure 5.4 confusion matrix and ROC for NN (all features)

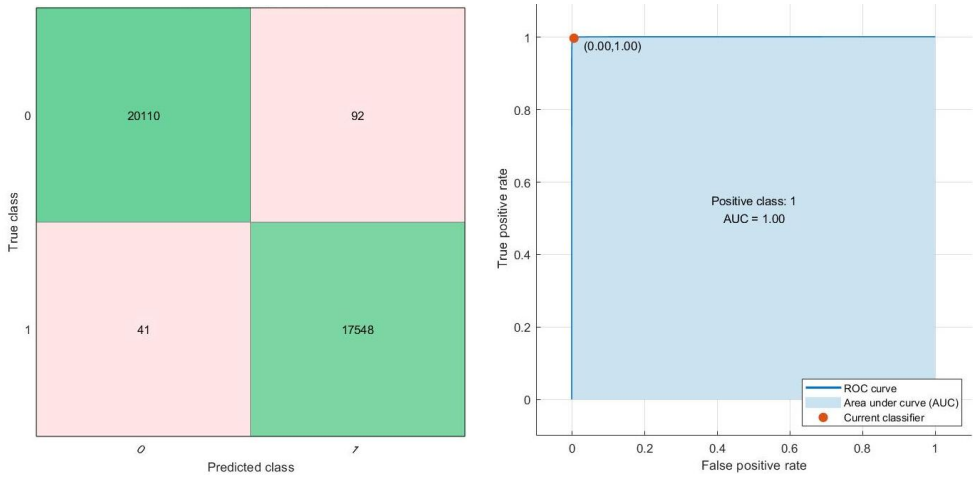

Figure 5.5 confusion matrix and ROC for proposed model (all features)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

Table 5.5 Performance of machine learning models using selected features (Chi-square)

Models	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**NPV**
LDA	86.21	90.67	81.96	82.71	86.51	90.23
LR	86.32	89.04	83.49	84.85	86.90	88.01
LSVM	87.39	88.43	86.21	87.91	88.17	86.79
NN	91.97	93.26	90.50	91.82	92.54	92.15
Proposed	**99.63**	**99.68**	**99.58**	**99.63**	**99.66**	**99.63**

The performance of the proposed model on selected features for k=5 using chi-square achieved highest accuracy of 99.63%, while NN achieved 91.97%, LSVM achieved 87.39%, LR achieved 86.32% and LDA achieved 86.21%, Table 5.5. Also the proposed model on using selected features show higher sensitivity (99.68%), specificity (99.58%), precision (99.63%), F1 (99.66%) and NPV (99.63%) outperforming other models. The improved accuracy, sensitivity, specificity, precision shows that the proposed model gains better classification ability to classify normal and attack network traffic. The improvement in classification shows that the features selected through chi-square have higher association with the target class.

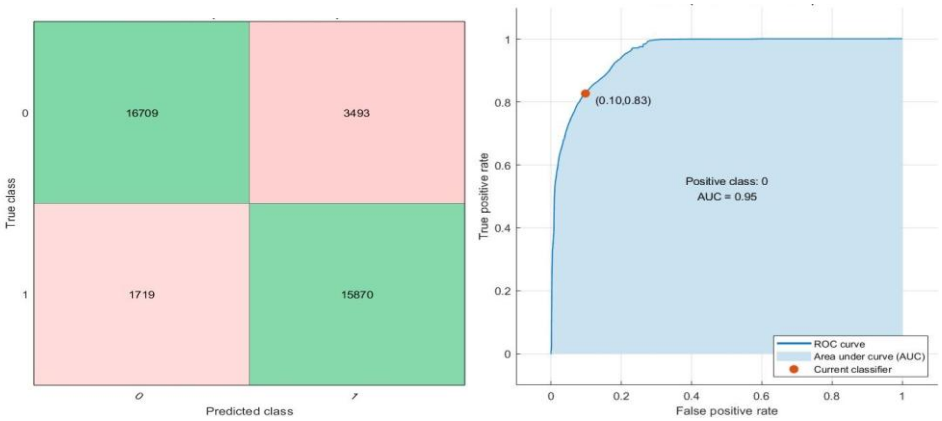

Figure 5.6 confusion matrix and ROC for LDA (Chi-square)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

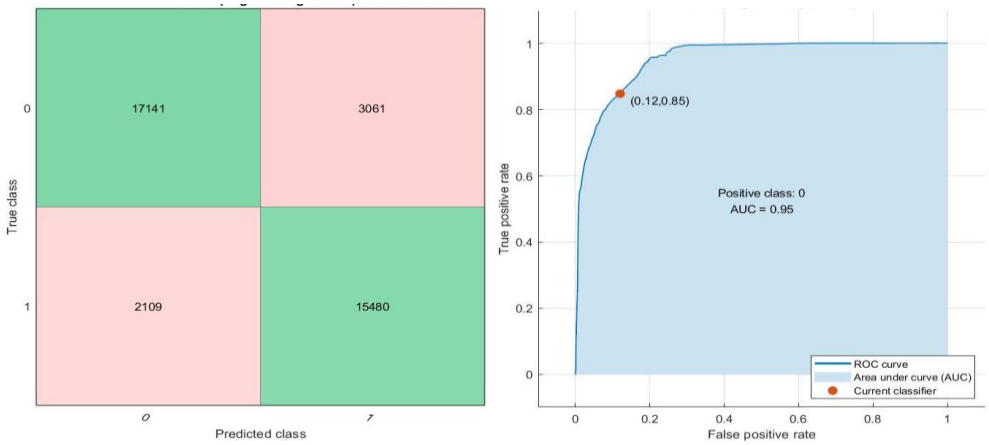

Figure 5.7 confusion matrix and ROC for LR (Chi-square)

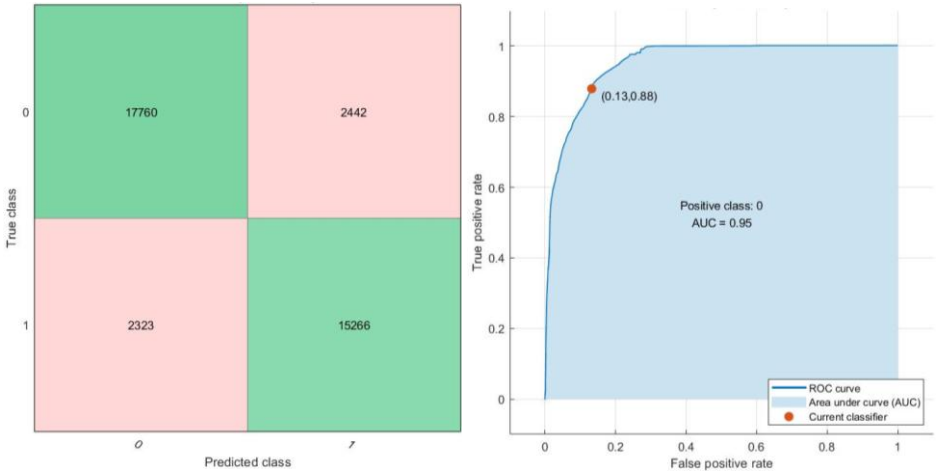

Figure 5.8 confusion matrix and ROC for SVM (Chi-square)

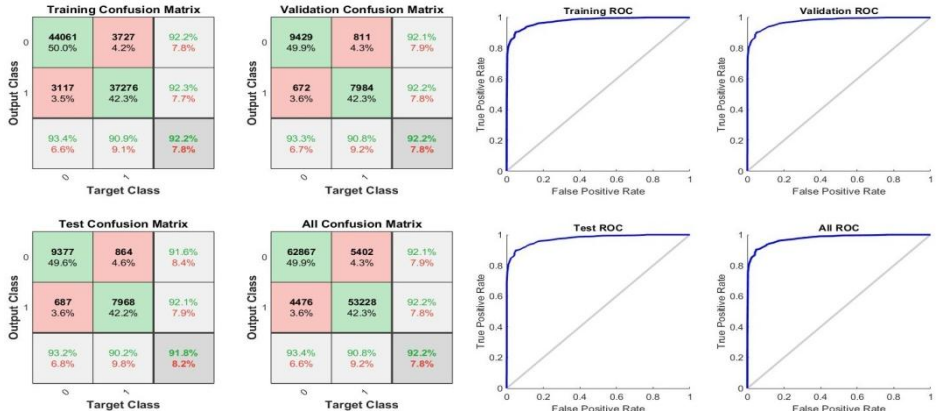

Figure 5.9 confusion matrix and ROC for NN (Chi-square)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

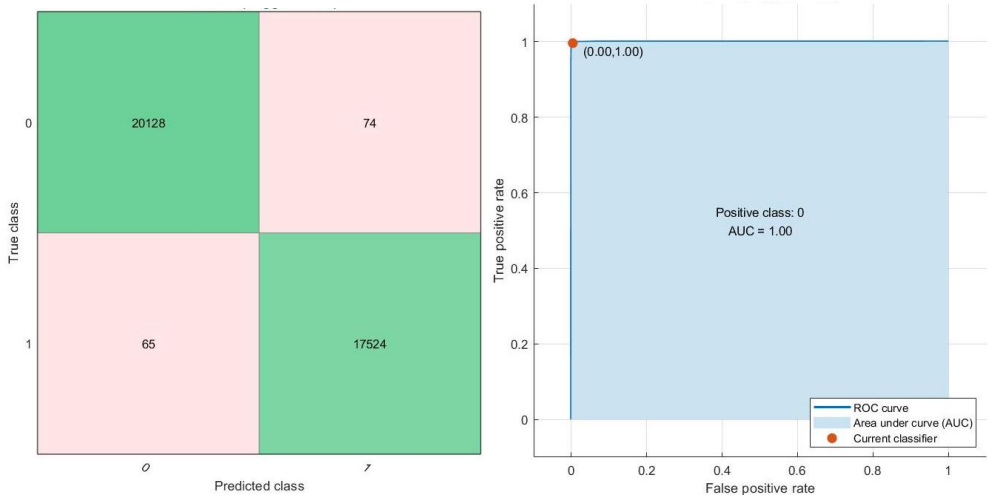

Figure 5.10 confusion matrix and ROC for proposed model (Chi-square)

Table 5.6 Model evaluation

Measure	**LDA**	**LR**	**LSVM**	**NN**	**Proposed**
Accuracy	86.21	86.32	87.39	91.97	**99.63**
FPR	0.180	0.165	0.137	0.095	**0.004**
DR	0.906	0.890	0.884	0.932	**0.996**

According to the table 5.6, the detection rate of the proposed model is 0.996 and the false positive rate is 0.004 which is and an accuracy of 99.63%. The proposed model outperforms other models such as LDA, LR, LSVM and NN in terms of Accuracy, detection rate and false positive rate. The ROC represents the discrimination power of the classification models and ROC for the proposed model shows AUC =1.00 and ROC in the top left corner which indicate that the proposed model is efficient in discriminating normal and attack traffic data in the network. The proposed model performance show a highest accuracy of 99.63% outperformed (Dwivedi et al., 2021), (Hindy et al., 2021), (Ogundokun, et al., 2021), (Bhati, & Rai, 2020), (Krishnaveni et al., 2020), (Serinelli et al., 2020) and (Amaizu et al., 2020) and consistent with (Pradeep Mohan Kumar et al., 2021), however (Pradeep Mohan Kumar et al., 2021) involved six features and ours involved five features and using five features the proposed model achieved highest accuracy score.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4984-5002
https://publishoa.com
ISSN: 1309-3452

CONCLUSION

The internet traffic attacks are increasing and becoming capable of evading intrusion detection systems. The exploitation of machines and networks can be secured with robust intrusion detection system. The present study proposed a network intrusion detection system which is capable of capturing network anomaly and attacks. The proposed model using ensemble trees improves detection rate and reduces feature dimensions effectively. The performance of the proposed tree based ensemble model is demonstrated on NSL-KDD dataset which achieved higher accuracy of 99.63% and FPR 0.004. Also, understanding every attack types is required to capture the attack patterns to efficiently capture different attacks in the network. As a future, the present work will extended to address the classification of different types of attacks in the network, which improve the Intrusion detection capability for capturing newer threats.

References:

1. Abrar, I., Ayub, Z., Masoodi, F., & Bamhdi, A. M. (2020, September). A machine learning approach for intrusion detection system on NSL-KDD dataset. In 2020 international conference on smart electronics and communication (ICOSEC) (pp. 919-924).
2. Ahanger, A. S., Khan, S. M., & Masoodi, F. (2021, April). An effective intrusion detection system using supervised machine learning techniques. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1639-1644)
3. Amaizu, G. C., Nwakanma, C. I., Lee, J. M., & Kim, D. S. (2020, October). Investigating network intrusion detection datasets using machine learning. In 2020 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1325-1328)
4. Bhati, B. S., & Rai, C. S. (2020)[1]. Analysis of support vector machine-based intrusion detection techniques. Arabian Journal for Science and Engineering, 45(4), 2371-2383.
5. Bhati, B. S., & Rai, C. S. (2020)[2]. Ensemble based approach for intrusion detection using extra tree classifier. In Intelligent computing in engineering (pp. 213-220).
6. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

7. Dwivedi, S., Vardhan, M., & Tripathi, S. (2021). Building an efficient intrusion detection system using grasshopper optimization algorithm for anomaly detection. Cluster Computing, 24(3), 1881-1900.
8. Hindy, H., Tachtatzis, C., Atkinson, R., Bayne, E., & Bellekens, X. (2021, April). Developing a Siamese network for intrusion detection systems. In Proceedings of the 1st Workshop on Machine Learning and Systems (pp. 120-126).
9. Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity, 2(1), 1-22.
10. Krishnaveni, S., Vigneshwar, P., Kishore, S., Jothi, B., & Sivamohan, S. (2020). Anomaly-based intrusion detection system using support vector machine. In Artificial intelligence and evolutionary computations in engineering systems (pp. 723-731).
11. Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16-24.
12. Lippmann, R. P., Fried, D. J., Graf, I., Haines, J. W., Kendall, K. R., McClung, D., ... & Zissman, M. A. (2000, January). Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00 (Vol. 2, pp. 12-26).
13. Masdari, M., & Khezri, H. (2020). A survey and taxonomy of the fuzzy signature-based intrusion detection systems. Applied Soft Computing, 92, 106301.
14. Miao, J., & Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91, 919-926.
15. Mohammad, R. M. A., & Alsmadi, M. K. (2021). Intrusion detection using Highest Wins feature selection algorithm. Neural Computing and Applications, 33(16), 9805-9816.
16. Ogundokun, R. O., Awotunde, J. B., Sadiku, P., Adeniyi, E. A., Abiodun, M., & Dauda, O. I. (2021). An enhanced intrusion detection system using particle swarm optimization feature extraction technique. Procedia Computer Science, 193, 504-512.
17. Othman, S. M., Alsohybe, N. T., Ba-Alwi, F. M., & Zahary, A. T. (2018). Survey on intrusion detection system types.

International Journal of Cyber-Security and Digital Forensics, 7(4), 444-463.

18. Pradeep Mohan Kumar, K., Saravanan, M., Thenmozhi, M., & Vijayakumar, K. (2021). Intrusion detection system based on GA-fuzzy classifier for detecting malicious attacks. Concurrency and Computation: Practice and Experience, 33(3),
19. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1-21.
20. Serinelli, B. M., Collen, A., & Nijdam, N. A. (2020). Training guidance with kdd cup 1999 and nsl-kdd data sets of anidinr: Anomaly-based network intrusion detection system. Procedia Computer Science, 175, 560-565.
21. Shafiq, M., Tian, Z., Bashir, A. K., Du, X., & Guizani, M. (2020). IoT malicious traffic identification using wrapper-based feature selection mechanisms. Computers & Security, 94, 101863.
22. Soysal, M., & Schmidt, E. G. (2010). Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. Performance Evaluation, 67(6), 451-467.
23. Stolfo, S. J., Fan, W., Lee, W., Prodromidis, A., & Chan, P. K. (2000, January). Cost-based modeling for fraud and intrusion detection: Results from the JAM project. In Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00 (Vol. 2, pp. 130-144).
24. Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics and Information Technologies, 19(1), 3-26.