IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

ENHANCING INTRUSION DETECTION SYSTEM PERFORMANCE LEVERAGING MACHINE LEARNING MODEL AND FEATURE SELECTION

A.Priyavengatesh
Ph.D. Scholar, Department of Computer Science
Dr.R.Kannan
Associate Professor , Department of Computer Science
Sri Ramakrishna Vidyalaya Arts &Science College, Coimbatore-641020.

ABSTRACT

Intrusion detection system (IDS) plays a vital role in mitigating cyber-attacks and exploitations. The internet and communication technology growth have witnessed the attacks and threats are evolving continuously. Intrusion detection systems are designed to capture cyber-attacks efficiently and various types of IDS are studied extensively to capture different forms of exploitation. The performance of any IDS depends on detection and prevention mechanism involved. Though many machine learning techniques are employed, there is still a need to enhance the performance of IDS with respect to rise of newer vulnerabilities in the network. This paper focus on enhancing IDS performance through building machine learning models that leverage feature selection to capture intrusions and anomalies. The proposed model utilizes feature selection to remove redundant and irrelevant features which helps machine learning models to gain better learning. The performance of the proposed model is evaluated on two datasets, namely NSL-KDD and HIKARI-2021 and compared against machine learning models such as LDA, Logistic Regression, NN and SVM.

1. INTRODUCTION

The main requirement of IDS is to detect and capture any forms of intrusions in the network with high detection accuracy. In the past decade many machine learning models are studied and introduced to classify intrusions. Intrusions detection systems are of two types anomaly based and signature based. Signature based IDS use patterns that are known to capture intrusions and suffer from poor detection of newer patterns while anomaly-based IDS use network activity to capture intrusions where the traffic flow that deviate from the normal network activity are treated as intrusions (Javed et al., 2022). The main drawback of present IDS is that it does not capture patterns and network activity that arises in future and end-up in higher false alarm rate (Agrawal et al., 2022).

Many machine learning models are utilized for IDS due to its capability of detection of new intrusions and threats in the network (Rahman et al., 2022). The performance and accuracy of machine learning models depends on the learning ability on the dataset. Also, the detection accuracy of machine learning models depends on the type of algorithm utilized since each algorithm has its own merits and merits (Saranya et al., 2020).

The machine learning algorithms can be grouped into three categories such as supervised, unsupervised and semi supervised methods. Supervised methods have the ability to learn the relationship between known classes and features which include algorithms such as SVM, LDA, NB and DT. Unsupervised methods learn the patterns hidden inside the data which includes methods like K-means, K-mediods, PCA etc. Semi supervised methods utilize both labeled and unlabelled data uncover information which combines both clustering and classification (Van Engelen & Hoos, 2020). The presence of large number of traffic data and traffic features which cause machine learning models to suffer from high dimensional problem and computational complexity (Belavagi, & Muniyal, 2016). Feature selection is a process of identifying useful and relevant features present in the dataset while the features that are irrelevant are removed. Feature selection process helps to reduce the feature number and to remove redundant features. The advantage of utilizing feature selection is that it reduces the model complexity & model learning time and improves the model accuracy & performance (Thakkar & Lohiya., 2022).

The main purpose of this study is to propose an ensemble tree based classifier 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 intruders imitate the normal traffic patterns. Considering every traffic flow features to classify into benign and attack, it requires high memory and produce computational complexity. In order to reduce the memory consumption and computational complexity, feature selection is utilized to reduce the data dimension. The proposed model creates a subset of features that have high relationship with the target class and classifies the traffic into benign and attack. The performance of the proposed method is evaluated on NSL-KDD Cup 99 and HIKARI-2021 using machine learning algorithms such as LDA, Logistic Regression, NN and SVM. 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 concludes the paper.

2 RELATED WORK

(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.

(Fernandes & Lopes., 2022) investigated the Hikari-2021 dataset on different machine learning algorithms. Using feature selection, the data set is reduced to 22 features from 83 features. The dataset is balanced with equal number of target class instances and Ml achieved accuracy of 99% however the accuracy of the ML models reduced to 80% when feature selection is applied.

(Rajak et al., 2022) proposed a CNN and LSTM based IDS for securing precision farming. The proposed model is trained using Hikari-2021 dataset and the proposed model achieved accuracy of 93.27% without applying feature selection.

(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 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 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 Chi-Square Feature Selection Ensemble Tree Model (CFSETM) uses Chi-square based Feature Selection and Ensemble Tree Model to classify the traffic into benign and attack. CFSETM creates subset using feature selection in the first step and classifies traffic data using subset of selected features. 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,

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

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,

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

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,

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

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 CFSETM is tested against features selected using sequential forward selection (SFS) and machine learning models such as Linear Discriminant Analysis (LDA), Logistic Regression (LR), Neural Network (NN) and SVM on NSL-KDD CUP 99 & HIKARI- 2021 datasets.

Step 1 for CFSETM Pseudo Code

```
Input: $D=(X_1,X_2,...,X_d : Y_d)$
Output: $D_{sub}$
$D_{sub}$ returns a subset with K features where K < X
Compute scoring function $C_i$ for each feature target pair ($C_i = \sum \frac{(X_k - e_k)2}{e_k}$)
$D_{sub}$ =select features with high chi-square score (p-value $\leq 0.05$)
```

Step 2 for CFSETM Pseudo Code

```
Input: $D_{sub}=(X_1,X_2,...,X_d : Y_d)$
Output: bagged ensemble trees $bCL_n$
For n to 1 $\longrightarrow$ n do
Construct bagged subset of samples BT
Build decision tree classifier $bCL_n(D_{sub})=bCL(D_{sub} |P_n)$ with hyperparameter $P_n$
For each bagged ensemble classifier $bCL_n(D_{sub})$
Determine corresponding class $C_1$ for $bCL_n(1\leq n \leq N)$
Compute class score Cs ($Cs(X, C_1) = \frac{v(X,C_1)}{btrees}$) for each trees in bag
Return the majority of vote for each btrees (mvote{$C^{btrees}$}$bCL_n$ )
```

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. The HIKARI-2021 dataset was introduced by (Ferriyan et al., 2021) and the dataset has a total of 555278 instances with 84 features. Of the total instances 347431 instances are benign and 207847 instances are attack. The details of attack types are given in Table 4.1.1. The Hikari-2021 data is an unbalanced dataset with more benign instances. For this study a balanced dataset set is prepared with total instances of 20,000 with 10000 instances of benign and 10000 instances of attack. After preprocessing and removing nine null features, final 75 features are used to evaluate the proposed model CFSETM.

Table 4.1.1 Hikari-2021 Details

Traffic Category	**Type**	**Instances**
Background	Benign	1,70,151
Benign	Benign	3,47,431
Bruteforce	Attack	5,884
Bruteforce-XML	Attack	5,145
Probing	Attack	23,388
XMRIGCC Cryptominer	Attack	3,279
Total		**555278**

4.2 EVALUATION METRICS

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

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

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

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

$$\text{F-score} = \frac{TP}{TP + \frac{1}{2}(FP+FN)}$$

5. RESULTS AND DISCUSSION

To improve classification of traffic data and to reduce the computational complexity, tree based ensemble model CFSETM is proposed. The proposed CFSETM model selects 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 CFSETM using all features and selected features using sequential feature selection and the results are compared against LDA, LR, LSVM and NN models on NSL-KDD and Hikari-2021 datasets. The sequential forward selection method is wrapper method that selects 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 sequential forward selection method is wrapper method that selects 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 on each dataset. The selected features for NSL-KDD dataset 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 selected features for Hikari-2021 by SFS are Fwd_pkts_payload.avg, Bwd_pkts_payload.max, Fwd_subflow_bytes, Bwd_init_window_size and Fwd_last_window_size with model performance of 97.9% average score and std error of 0.00012 (Table 5.3). The proposed model using Chi-square estimation on NSL-KDD dataset selected five best features which include ‘src_bytes’, ‘dst_bytes’, ‘count’, ‘dst_host_count’ and ‘dst_host_srv_count’. The proposed model using Chi-square estimation on Hikari-2021 dataset selected five best features which include ‘Bwd_init_window_size’, ‘Bwd_header_size_min’, ‘Bwd_header_size_max’, ‘Bwd_bulk_rate’ and ‘Down_up_ratio’. The chi-square value and corresponding p-value is given in Table 5.4 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) for NSKL-KDD dataset

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	0.0002

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

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

Table 5.3 SFS score for features selected (k=5) for Hikari-2021 dataset

Feature names	Avg score	Std error
Fwd_pkts_payload.avg	0.950	0.0014
Fwd_pkts_payload.avg , Bwd_pkts_payload.max	0.989	0.0011
Fwd_pkts_payload.avg , Bwd_pkts_payload.max , Fwd_subflow_bytes	0.989	0.0012
Fwd_pkts_payload.avg , Bwd_pkts_payload.max , Fwd_subflow_bytes , Bwd_init_window_size	0.966	0.0012
Fwd_pkts_payload.avg , Bwd_pkts_payload.max , Fwd_subflow_bytes , Bwd_init_window_size , Fwd_last_window_size	0.979	0.0012

Table 5.4 Chi-square score for feature selected (k=5) for Hikari-2021 dataset

Feature	Chi Score	p-value
Bwd_init_window_size	55.88	0.000
Bwd_header_size_min	19.07	0.000
Bwd_header_size_max	17.63	0.000
Bwd_bulk_rate	14.51	0.000
Down_up_ratio	9.80	0.000

The proposed model CFSETM for NSL-KDD 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 5.5. The proposed model CFSETM show higher performances with respect to sensitivity (99.80%), specificity (99.48%), precision (99.54%), F1 score (99.67%) and NPV (99.77%) over other models. Next to the proposed model CFSETM, 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 for NSL-KDD dataset.

Table 5.5 Performance of machine learning models using all features for NSL-KDD dataset

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
CFSETM	**99.65**	**99.80**	**99.48**	**99.54**	**99.67**	**99.77**

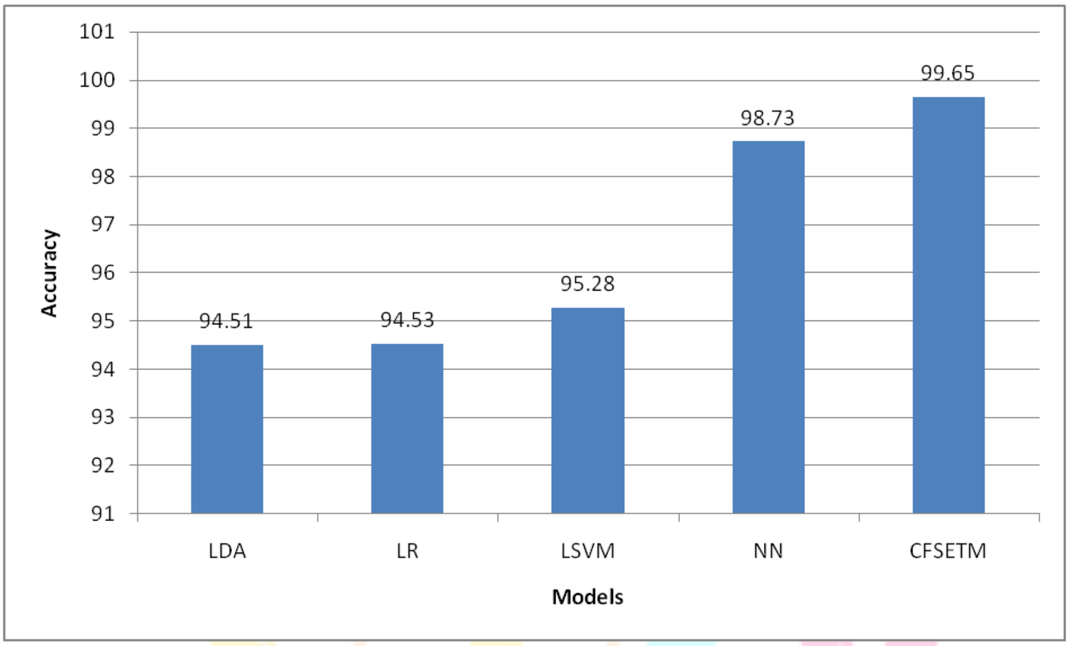


Figure 5.1 Accuracy of the models using all features on NSL-KDD dataset

The proposed model CFSETM on selected features for k=5 using SFS for NSL-KDD dataset 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 5.6. The sensitivity (99.62%), F1score (95.02%) and NPV (99.6%) of the proposed model CFSETM 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 CFSETM on features selected with SFS shows that the model is losing its ability to classify as the information on the selected features have lower association with the target class.

Table 5.6 Performance of machine learning models using selected features (SFS) for NSL-KDD Dataset

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
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
CFSETM	**94.91**	**99.62**	90.43	90.83	**95.02**	**99.6**

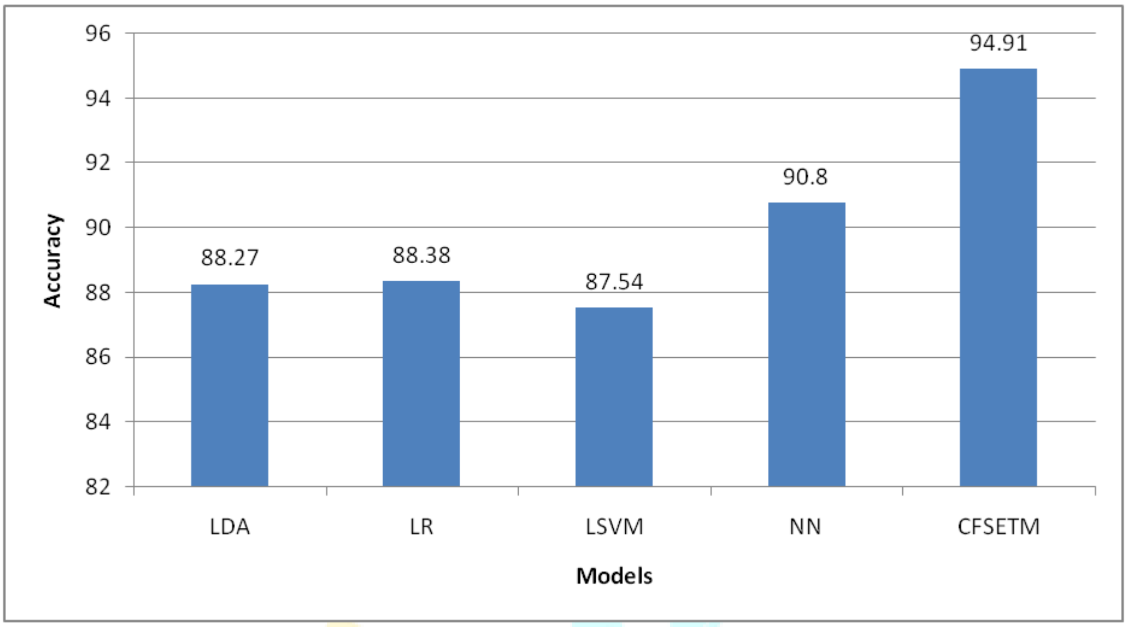


Figure 5.2 Accuracy of the models using selected features (SFS) on NSL-KDD dataset

Table 5.7 Performance of machine learning models using selected features (Chi-square) for NSL-KDD dataset

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
CFSETM	**99.63**	**99.68**	**99.58**	**99.63**	**99.66**	**99.63**

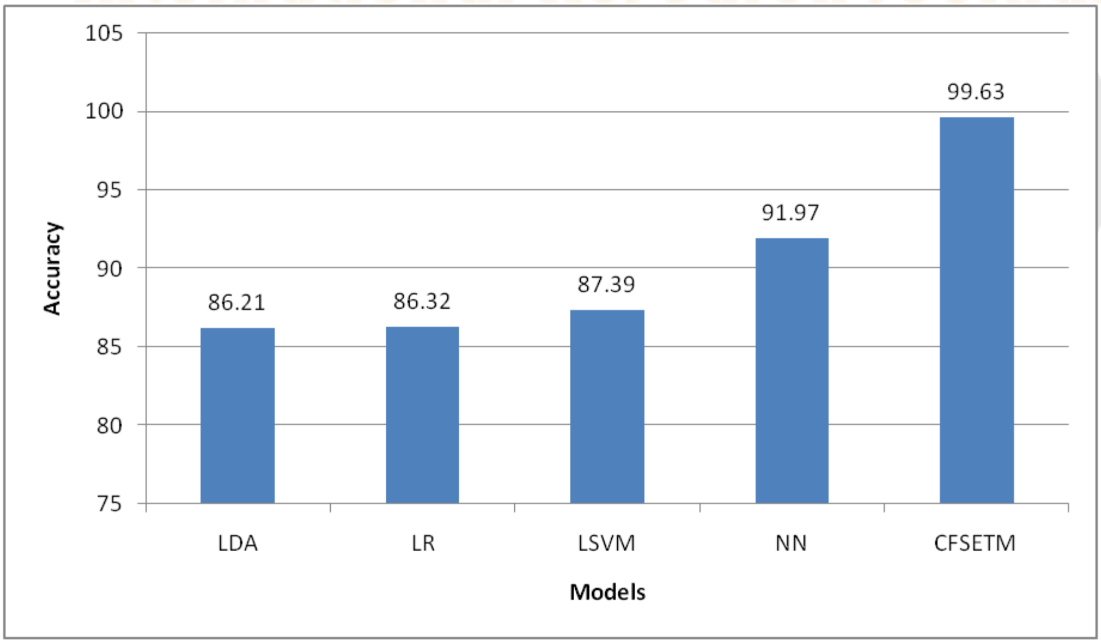


Figure 5.3 Accuracy of the models using selected features (Chi-square) on NSL-KDD dataset

The performance of the proposed model CFSETM 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 5.7. 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.

Table 5.8 Performance of machine learning models using all features for Hikari-2021 dataset

Models	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**NPV**
LDA	99.22	99.29	99.18	99.16	99.22	99.31
LR	99.35	99.26	99.44	99.43	99.34	99.28
LSVM	97.82	98.56	97.11	97	97.77	98.62
NN	**99.98**	**99.97**	**1**	**1**	99.98	**99.97**
CFSETM	92.73	87.18	1	1	93.15	85.63

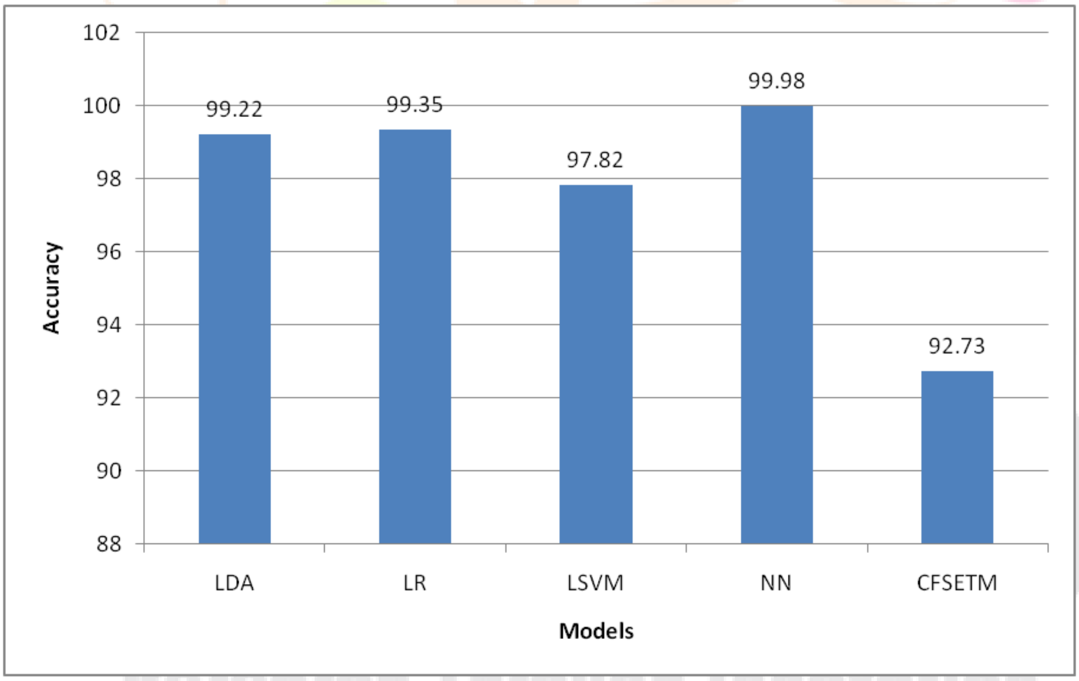


Figure 5.4 Accuracy of the models using all selected features on Hikari-2021 dataset

The performance of the proposed model CFSETM on Hikari-2021 dataset using all features achieved an accuracy of 92.73% and neural network achieved a highest accuracy of 99.98% with sensitivity of 99.97%, specificity of 100%, and precision of 100% (Table 5.8). Similarly LDA and LR models achieved accuracy of more than 99%. The decrease in performance of the proposed model CFSETM over other models is a result of poor generalization and the model becomes underfitting, which indicates that the model requires more number of data. Proposed model CFSETM achieved precision of 100% which indicates that the model correctly predicts the positive cases as positive while the precision score for LDA, LR and LSVM is less than 100%.

Table 5.9 Performance of machine learning models using selected features (SFS) for Hikari-2021 Dataset

Models	Accuracy	Sensitivity	Specificity	Precision	F1	NPV
LDA	92.55	1	87.05	85.09	91.94	1
LR	94.63	1	90.31	89.24	94.31	1
LSVM	95.2	**99.94**	91.29	90.44	94.96	99.95
NN	**98.78**	98.27	**99.29**	99.3	**98.78**	**98.25**
CFSETM	49.89	49.91	0	**99.99**	66.57	0

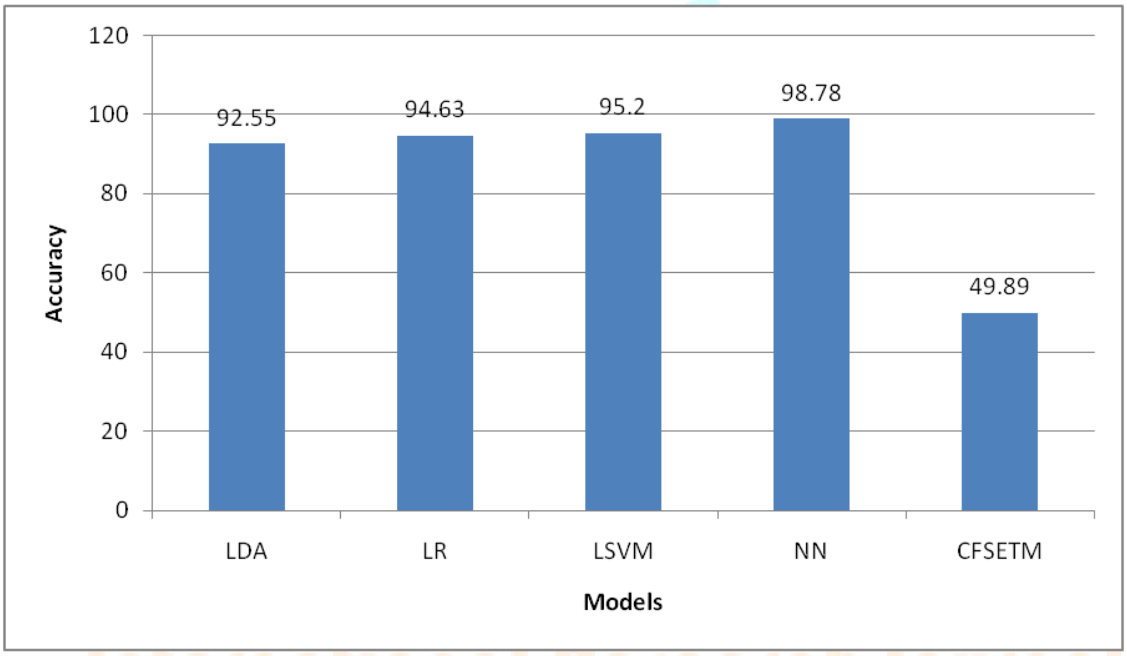


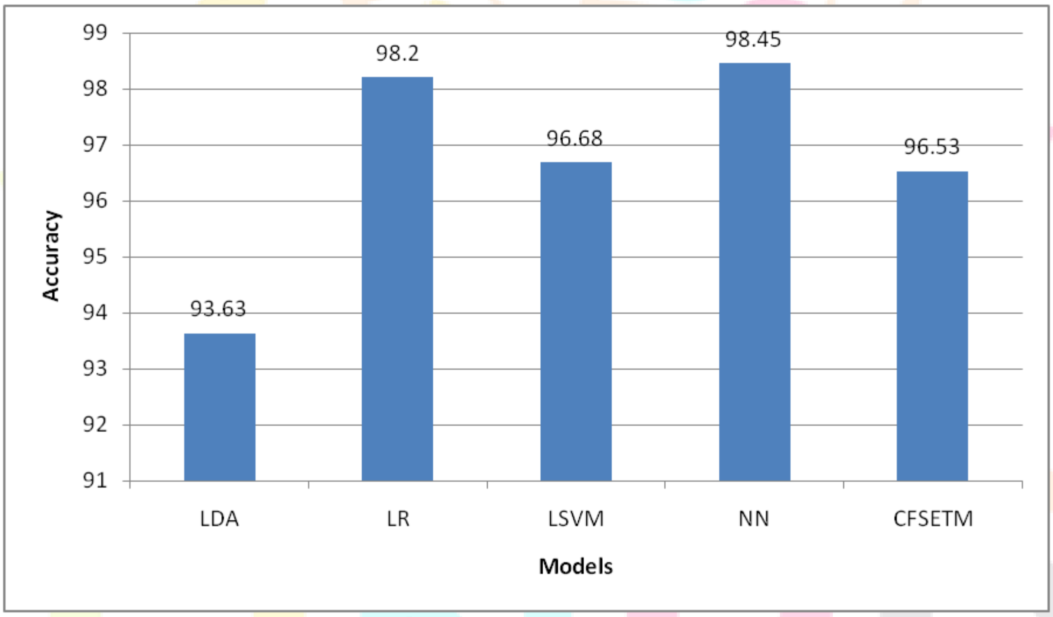
Figure 5.5 Accuracy of the models using selected features (SFS) on Hikari-2021 dataset

The proposed model CFSETM on selected features for k=5 using SFS for Hikari-2021 dataset achieved low accuracy of 49.89%, while NN achieved highest accuracy of 98.78%, LSVM achieved 95.20%%, LR achieved 94.63% and LDA achieved 92.55% (Table 5.9). The neural network model with specificity of 99.29%, precision of 99.3%, F1score of 98.78% and NPV of 98.25% outperformed other models while SVM achieved a highest accuracy of 99.94% than other models which indicates that the model is classifying positive cases correctly as positive. The decreased performance of the proposed model CFSETM on features selected with SFS shows that the model is losing its ability to classify as the information on the selected features have poor association with the target class. It is noted that the performance of LDA, LR, SVM and NN models using selected features gets reduced with feature selection. The accuracy of LDA decreased from 99.22% (full features) to 92.55%, for LR the accuracy decreased from 99.35% to 94.63%, for SVM the accuracy decreased from 97.82% to 95.20% and for NN the accuracy decreased from 99.98% to 98.78% and for proposed model CFSETM the accuracy decreased from 92.73% to 49.89% which indicate that the features selected through SFS does not have enough information to make good predictions and the models becomes underfitting and in simple terms, the features selected by SFS are not useful.

Table 5.10 Performance of machine learning models using selected features (Chi-square) for Hikari-2021 dataset

Models	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**NPV**
LDA	93.63	1	88.71	87.24	93.18	1
LR	98.2	98.73	97.68	97.65	**98.19**	**98.75**
LSVM	96.68	**99.57**	94.1	93.74	96.57	99.6
NN	**98.45**	98.74	98.16	98.15	98.44	**98.75**
CFSETM	96.53	94.16	**99.16**	**99.20**	96.61	93.65

Figure 5.5 Accuracy of the models using selected features (Chi-square) on Hikari-2021 dataset



The performance of the proposed model CFSETM on selected features for k=5 using chi-square on Hikari-2021 dataset achieved accuracy of 96.53%, sensitivity of 94.16%, specificity of 99.16% and precision of 99.20%. NN achieved highest accuracy of 98.45%, LSVM achieved 96.68%, LR achieved 98.20% and LDA achieved 93.63% (Table 5.10). Also the proposed model on using selected features show higher specificity (99.16%) and precision (99.20%) outperforming other models. The improved accuracy, 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 high information on the target class. It is noted that the performance of LDA, LR, SVM and NN models using selected features gets reduced with feature selection. The accuracy of LDA decreased from 99.22% (full features) to 93.63%, for LR the accuracy decreased from 99.35% to 98.20%, for SVM the accuracy decreased from 97.82% to 96.68% and for NN the accuracy decreased from 99.98% to 98.45% and for proposed model CFSETM the accuracy improves from 92.73% to 96.53% (3.8% increase) which indicate that the features selected are useful features with information on target class. Further the decrease in model performance of LDA, LR, SVM and NN is due to poor generalization that causes the models to become underfit.

Table 5.11 Model evaluation

Measure	**LDA**	**LR**	**LSVM**	**NN**	**Proposed**	**Dataset**
Accuracy	86.21	86.32	87.39	91.97	**99.63**	NSL-KDD
FPR	0.180	0.165	0.137	0.095	**0.004**	NSL-KDD
DR	90.67	89.04	88.43	93.26	**99.68**	NSL-KDD
Accuracy	93.63	98.20	96.68	98.45	**96.53**	Hikari-2021
FPR	0.112	0.023	0.059	0.018	**0.008**	Hikari-2021
DR	**100**	98.73	99.57	98.74	94.16	Hikari-2021

The proposed model CFSETM for NSL-KDD dataset achieved detection rate of 99.68%, false positive rate of 0.004 and 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 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 this study involved only five features and using five features the proposed model achieved highest accuracy score for NSL-KDD dataset. The proposed model CFSETM for Hikari-2021 dataset achieved detection rate of 94.16%, false positive rate of 0.008 and accuracy of 96.53%. The proposed model achieved a 2% less than other models such as LR (98.20%) and NN (98.45%) in terms of accuracy; however the accuracy of the proposed model CFSETM improved from 92.73% to 96.53% with detection rate of 94.16% while other models achieved high detection rate. As pointed out by (Fernandes & Lopes., 2022) the accuracy of the model on Hikari-2021 dataset is reduced when feature selection is applied. On Hikari-2021 dataset, the proposed model achieved a lowest FPR of 0.008% which indicate that the model correctly classifies positive cases as positive. On Hikari-2021 dataset, the proposed model outperforms (Fernandes & Lopes., 2022) and (Rajak et al., 2022) with 96.53% of accuracy.

CONCLUSION

In this paper, enhancing the performance of IDS leveraging feature selection has been demonstrated with NSL-KDD and Hikari-2021 dataset. The experimental study establishes suitable evidences that proposed feature selection model improves intrusion detection rate. The proposed model CFSETM achieved a highest accuracy of 99.63% with 0.004 of FPR on NKSL-KDD dataset and also for Hikari-2021 dataset, the accuracy improved from 92.73% to 96.53% with FPR of 0.008. The ability to effectively capture the intrusions by machine learning models depends on the information that can be used by the machine learning models but the learning rate is largely affected by noises and redundant information in the dataset. Identifying relevant and informative features contribute to better learning of the models which eventually improve the model performances. In the future work, the capabilities of the proposed model will be improved further to use adaptive parameter optimization on larger datasets.

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. Agrawal, S., Sarkar, S., Aouedi, O., Yenduri, G., Piamrat, K., Alazab, M., ... & Gadekallu, T. R. (2022). Federated learning for intrusion detection system: Concepts, challenges and future directions. Computer Communications.2022, 195, 346–361.

3. 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)

4. 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)

5. Belavagi, M. C., & Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion detection. Procedia Computer Science, 89, 117-123.

6. 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.

7. 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).

8. 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.

9. Fernandes, R., & Lopes, N. (2022, June). Network intrusion detection packet classification with the hikari-2021 dataset: a study on ml algorithms. In 2022 10th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-5). IEEE.

10. Ferriyan, A., Thamrin, A. H., Takeda, K., & Murai, J. (2021). Generating network intrusion detection dataset based on real and encrypted synthetic attack traffic. Applied Sciences, 11(17), 7868.

11. 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)

12. Javed, A. R., Ahmed, W., Alazab, M., Jalil, Z., Kifayat, K., & Gadekallu, T. R. (2022). A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions. IEEE Access, 10, 11065-11089.

13. 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).

14. Mohammad, R. M. A., & Alsmadi, M. K. (2021). Intrusion detection using Highest Wins feature selection algorithm. Neural Computing and Applications, 33(16), 9805-9816.

15. 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.

16. 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),

17. Rahman, A. U., Mahmud, M., Iqbal, T., Saraireh, L., Kholidy, H., Gollapalli, M., ... & Ahmed, M. I. B. (2022). Network Anomaly Detection in 5G Networks. Mathematical Modelling of Engineering Problems, 9(2).

18. Rajak, P., Lachure, J., & Doriya, R. (2022, October). CNN-LSTM-based IDS on Precision Farming for IIoT data. In 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA) (pp. 99-103). IEEE.

19. Saranya, T., Sridevi, S., Deisy, C., Chung, T. D., & Khan, M. A. (2020). Performance analysis of machine learning algorithms in intrusion detection system: A review. Procedia Computer Science, 171, 1251-1260.

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. 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).Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

22. Thakkar, A., & Lohiya, R. (2022). A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. Artificial Intelligence Review, 55(1), 453-563..

23. Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine learning, 109(2), 373-440.