# Evaluation of Efficient Classification Algorithm for Intrusion Detection System

**V. Priyalakshmi[1] and Dr. R. Devi[2]**
Assistant Professor, Department of Computer Science[1]
SRM Arts & Science College, Chennai, Tamil Nadu, India[1]
Associate Professor, Department of Computer Science[2]
Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India[2]
kumaran.durairaj.kd@gmail.com[1] and devi.scs@velsuniv.ac.in[2]

**Abstract:** *Intrusion detection system is one of the most significant network security problems in the technology world. To improve the Intrusion Detection System (IDS) many machine learning methods are implemented. In order to develop the performance of IDS, different classification algorithms are applied to detect different types of attacks. For building efficient IDS is not an easy task and choosing a suitable classification algorithm. The best method is to test the Performance of the different classification algorithms and select best method from them. This paper aim is to assemble an IDS model in terms of confusion matrix, accuracy, recall, precision, f-score, specificity and sensitivity. It also provides a detailed comparison with the dataset, data preprocessing, number of features selected, feature selection technique, classification algorithms, and evaluation performance of algorithms described in the intrusion detection system.*

**Keywords:** IDS, NSL-KDD, SVM, Confusion Matrix, Feature Selection, Classification Algorithm

## I. INTRODUCTION

Currently, intrusion detection system has gained a necessary role in computer and network security. IDS monitoring and analyzing network traffic is used to classify different types of attacks [1]. The network traffic action consists of many features collected in the form of a dataset to detect different types of attacks [2]. The enhance of the massive amount of data being generate the internet has caused the world of technology to look a big challenge [3]. Datasets represent instances that consist of several features and are related to the intrusion detection system [4]. It is essential to recognize the variety of data containing different types of attacks and features [5]. The trendiest data set that is being used for the intrusion detection system is a KDD'99 cup to develop predictive models for distinctive the relationship between intrusions or numerous attacks [6]. The intrusion detection system builds the model based on security data sets such as KDD99 and NSL-KDD [7]. It contains different types of features similar to an interpreter to distinguish the normal attacks from the abnormal ones as a features target [8]. The data set of classification model is splits into stage training and testing phase [9]. Therefore, it needs to select some useful and relevant features from the whole range of features to develop the performance of the model in the testing phase [10]. The significant stage to improve a classification model's quality is data preprocessing machine learning algorithms [11]. To solve a numerous types of big data set is one of crucial step to be performed.

Machine Learning (ML) techniques generally used in computer security data sets have newly become a trend in security tools [13]. It contribute to analysis and handling the massive amount of data and extract the essential features that are used in various techniques for feature selection [14]. IDS are a commonly used machine learning classifier to distinguish between various attacks as a class [15].For IDS many supervised classification algorithms such as Decision Trees, Naïve Bayes, K-Nearest Neighbor, Tree C4.5, Random Forest, Support Vector Machine, and Logistic Regression [16] are used. To classify and predict different types of threats especially using confusion matrix by evaluation of classification algorithms depends on different statistical metrics [17].

This paper is explained as follows: Section 2 explain the classification model, Section 3 describes effective of dimensionality reduction for feature selection, Section 4 describes performance evaluation appropriate metrics and Other sections are related works reviewed and compared with discussion finally give an conclusion.

DOI: 10.48175/IJARSCT-7751

## II. CLASSIFICATION MODEL

Supervised Machine learning model is used for intrusion detection systems based on binary or multi classes [18]. In supervised learning, data is always labeled, which takes each record in a dataset assigned to a particular class [19]. A classification model-based IDS classifies all the network traffic into either normal or abnormal classifier algorithms. [20]. Classification algorithms facing many problems in building an efficient model and need data preprocessing stage especially in high data dimensionality [21]. Choosing the best classification algorithm depends on the performance evaluation metrics in terms of confusion matrix and accuracy [22].

The two stages of training and testing are included in data classification process in the dataset [23]. A building model is used to predict the class labels for a given data during training and testing phase and learn a classified as a target[24].Preprocessing of the data helps the classification model decrease time and complexity by removing inappropriate data to improve the classification algorithm efficiency [25]. Two groups are divided into network traffic dataset classification in which one group for testing and training model [26], [27]. The most popular classifiers are used Decision Tree (DT), Random Forest, SVM (Support Machine Learning), KNN (K-Nearest Neighbor), Naïve Bayes, and Logistic Regression [28].

## III. EFFECTIVENESS OF DIMENSIONALITY REDUCTION FOR FEATURE SELECTION

The redundant and irrelevant data are removed by using feature selection process for dimensionality reduction. At same time it speeds up the training and testing time and also enhances the accuracy of classification model [29]. Reducing the high dimensionality of data improves the process of feature selection. To detect and prevent different cruel attacks, each dataset covers various kinds of features [30]. Reducing inadequate features by using dimension reduction techniques of the high number of features [31]. Feature selection and feature extraction are two main techniques to prevail over high dimensionality. The feature selection requires finding a subset of relevant features of the original data set. The feature extraction reduces the data in the original high-dimensional data set space to a lower dimension space [32].

There are many techniques used for dimension reduction, such as Principal Component Analysis (PCA) as a linear method, Linear Discriminant Analysis (LDA), Generalized Discriminant Analysis (GDA), and Support Vector Machine (SVM). In general, the IDS feature space faces the curse of dimensionality on a large scale. The curse dimension happens when big data set contains extra dimension space that does not occur in low dimensions [33]. By overcome from irritation of high dimensionality features problems using selection process [34]. Increase machine learning algorithm performance by removing unwanted features and uses the feature ranking [35]. Also, this process provides the model with the elimination of the unnecessary information and improvement in the generalization. Many techniques are used for feature selection such as Gain Ratio (GR), Symmetrical uncertainty, Chi-Square analysis, Information Gain (IG), and Practical Swarm Optimization (PSO) .

## IV. EVALUATION PERFORMANCE APPROPRIATE METRICS

Several metrics have been designed to measure the effectiveness of IDS. These metrics can be divided into three classes namely threshold, ranking and probability metrics. Ranking metrics include False Positive Rate (FPR), Detection Rate (DR), Precision (PR), Area under ROC curve (AUC) and Intrusion Detection Capability (CID). The value of ranking metrics lies in the range from 0 to 1. These metrics depend on the ordering of the cases, not the actual predicted value.

Confusion matrix is a matrix that represents result of classification. It represents true and false classification results. The followings are the possibilities to classify events and depicted is as follows,

- True positive (TP): Intrusions that are successfully detected by the IDS.
- False positive ( FP): Normal/non-intrusive behavior that is wrongly classified as intrusive by the IDS.
- True Negative (TN): Normal/non-intrusive behavior that is successfully labeled as normal/non-intrusive by the IDS.
- False Negative ( FN): Intrusions that are missed by the IDS, and classified as normal/non-intrusive.

To solve this problem, different performance metrics are defined in terms of the confusion matrix variables. These metrics produce some numeric values that are easily comparable and are briefly explained in subsequent paragraphs.

1. Classification rate (CR): It is defined as the ratio of correctly classified instances and the total number of instances.

DOI: 10.48175/IJARSCT-7751

2. Detection rate (DR): It is computed as the ratio between the number of correctly detected attacks and the total number of attacks.
3. False positive rate (FPR): It is defined as the ratio between the number of normal instances detected as attack and the total number of normal instances.
4. Precision (PR): It is the fraction of data instances predicted as positive that are actually positive
5. Recall: This metrics measures the missing part from the Precision; namely, the percentage from the real attack instances covered by the classifier.
6. F-measure (FM): For a given threshold, the FM is the harmonic mean of the precision and
7. Recall at that threshold.

The impact of splitting data set ratio into training and testing phase affects the result of a confusion matrix.

## V. REVIEW OF CLASSIFICATION ALGORITHMS FOR IDS

During the last decade, many works have been presented to improve the IDS to detect and prevent different malicious attacks from accessing computer information. Classification algorithms are one of the main concepts in the machine learning approach. Many researchers have analyzed and studied the Intrusion Detection System field, and recently more and more machine learning approaches are aligning with it to provide a better solution against intrusion.

Ashraf & Ahmed (2018)proposed comparing some of the most efficient machine learning algorithms - J48, Naïve Bayes, and Random Forest. This research aimed to deduct a better detection rate and accuracy of the Intrusion Detection System. The comparison was used to draw new patterns and procedures to overcome vast amount of audit data.

Colas & Bradzil (2006) proposed another comparative study to determine the best-suited algorithm in case of classification problems. The comparison was between Support Vector Machine (SVM) to K-Nearest Neighbor (KNN) and Naïve Bayes algorithms. Based on the performance analysis and processing time, they depicted a feature comparison between the above algorithms.

Jiang et al. (2012) developed a text categorization model using an improved and better KNN text categorization (INNTC) and one pass clustering algorithm, which shows that the combination of the two-classification algorithm improves text categorization by reducing text redundancy better than typical KNN, Naïve Bayes and SVM algorithms.

Aljawarneh,Aldwairi & Yassein (2018) provide insight into the network traffic meta-heuristic properties to create a better intrusion detection system. While large amount of data needs to be captured for performing an efficient analysis, a better model needs to be developed to feed the data into it.A new hybrid model was developed to estimate the intrusion level threshold on the network transaction data for training. From the results, it was found that the hybrid approach has a remarkable effect on the reduction of the computational and time complexity involved. The hybrid model's accuracy was 99.81% and 98.56% for binary and multi-class datasets, respectively. While obtaining low false and high false negative rates, the issues were handled using data filtration through Vote algorithm with information gain. The hybrid algorithm consists of several classifiers – J48, Random Tree, Naïve Bayes etc.

Elejla et al. (2019) showed a comparison between several classification algorithms such as KNN, SVM, Decision tree, Naïve Bayes, and Neural network to predict ICMPv6 based DDoS attacks by monitoring the network traffic and behavior of the attacks. The comparison showed that KNN was the fastest among the other algorithms to detect the attacks where neural network achieved the slowest rate. The classifiers detected most of the attacks ranging from 73% to 85% of the attacks.

In 2020, Kachavimath et al. proposed a Distributed Denial of Service (DDoS) detection model to improve network security cases by using machine learning techniques. The K- Nearest Neighbour and Naïve Bayes algorithms were used for classification, and for feature extraction, the correlation was utilized. The proposed model was compared with the conventional learning models being applied on NSL-KDD and KDD Cup 99 datasets. The experimental performance showed that the KNN algorithm with eight features obtained the best results than Naïve Bayes. The different parameters are used for measurement of classification algorithm performance accuracy 98.51%, precision 98.9%, recall 97.8% f-measure 1.005%, sensitivity 97.8%, specificity 99.12%, efficiency 98.48%, error rate 1.50% and ROC 0.99%.

In 2020, Bhosale and Nenova [68] proposed a new method for attack classification, which is Modified Naïve Bayes Algorithm (MNBIDS) with hybrid feature selection to improve the system accuracy of detecting attacks. The hybrid

DOI: 10.48175/IJARSCT-7751

feature selection method that ranks features according to the value of G_corrof for each selected feature. Additionally, compared the CNN, ANN, KNN, SVM algorithms and proposed (MNBIDS). The performance of MNBIDS is measured with the highest accuracy of 97%, precision 98%, and recall of 99%. The IDS performs data preprocessing, data normalization, and features extraction implemented in real-time data to KDD cup 99.

## VI. COMPARISON AND DISCUSSION

Machine learning techniques have been applied to the field of network security to improve intrusion detection systems. Previous sections reviewed some researches about classification algorithms applied to build the IDS model and evaluated the performance by different metrics in terms of accuracy, recall, precision, f-score, specificity, sensitivity, error rate, and dependable tool confusion matrix. The dimension reduction and feature selection had a good effect on the classification model performance because it reduces training and testing time via removing the irrelevant features, making the classification process more accurate and less complicated.

The best results for most reviewed studies showed that the Random Forest algorithm achieved the best accuracy of classification because it combines many decision trees that then decide the type of attack, leading to the decrease of the risk of over fitting. The random forest can deal with various big types of features that do not require data scaling. Moreover, the Practical Swarm Optimization (PSO) gets the best result for feature selection.

In this paper, the comparison is performed in terms of data set are data-preprocessing techniques, a number of features selected, feature selection techniques, classification algorithms, and evaluation metrics. This study aims to show different classification algorithms' performance by using different measurements to select a suitable classifier best model to gain speed and accuracy.

| Ref | Data set | Data preprocessing techniques | No of features selected | Feature selection techniques | Classification algorithm | Evaluation metrics |
|---|---|---|---|---|---|---|
| 2018 | AWID | Transformation values into integer Normalization scale | 32 set,10set 7 set,5 set | ZeroR | AdaBoost, Random Forest, Random Tree, J48, logit Boost, MLP | Best performance Random Forest with 32 features accuracy 99.64%, precision 0.995, recall 0.966 |
| 2018 | UNSWNB15 | Apache Spark processing tools | 42 features out of 49 | | SVM, Naïve Bayes, Decision Tree and Random Forest | best results Random Forest accuracy 97.49, Sensitivity 93.53, specificity 97.75 |
| 2019 | NSL-KDD | convert nominal attribute to binary attribute non-numeric, dimension reduction, Normalization | 24 | CfsSubsetEval | SVM Naïve Bayes | SVM best accuracy of 93.95 |
| 2019 | NSL-KDD | Reduce features | 17,35 | Correlation Chi-Square | ANN SVM | Highest ANN with Wrapper(correlation) 17 features, accuracy 94.02% |
| 2020 | UNSWNB15 | categorical features remove redundant and irrelevant features | 13 | Random Forest | Classification and Regression Trees (CART) | accuracy 87.74 |
| 2020 | NSL-KDD | outlier detection | when 34 Accepted features | Boruta Algorithm | Random Forest | accuracy 0.99892798 Sensitivity 0.99852158 Specificity 0.99939955 |
| 2020 | NSL-KDD | without need preprocessing | FeaturesF1 F2, F5, F6, F23, F24 | Software SDN | KNN, ELM, H-ELM | accuracy 84.29, False alarm rate 6.3 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2020 | KDD'99 | Dimension reduction | 9 | PSO | decision trees J48, SVM | 99.1 %, detection rate 99.6 %, FAR 0.9 % |
| 2020 | NSL-KDD | Data Normalization | - | CART Tree | Hybrid three decision tree | accuracy 83.1485, Precision 97.2193, recall 72.4694, F-score 83.0394 |
| 2020 | KDD | reduction high dimension using Python | - | Feature reduction PCA | Random forest decision tree, naïve bayes and SVM | Best result Random Forest accuracy 96.78% and error rate 0.21%. |
| 2020 | SE-CICIDS2018 | missing values small sample of data | 23 | Chi-square Correlation | Decision Tree, Logistics regression, , and gradient boosting ensemble | accuracy 98.8%, recall 97.1%, precision 98.8%, F1 97.9% |

## VII. CONCLUSION

Different machine learning methods depends on performance of IDS. Classification algorithms have a considerable role in helping IDS to discriminate different types of attacks. This paper aims to analysis different classifier algorithms and find the evaluation performance by using different metrics. By applying various metric measurements to evaluate classifiers' performance, noticed that the random forest algorithm achieved sufficient results and the highest accuracy to classify different types of attacks. The effectiveness of dimension reduction to reduce big data sets' complexity leads to select most favorable features to obtain better performance in classification in terms of accuracy and speed.

## REFERENCES

**[1].** Bace, R. (1998). An Introduction to Intrusion Detection & Assessment. Infidel, Inc. for ICSA, Inc

**[2].** Karthikeyan, K. R. & Indra, A. (2010). Intrusion Detection Tools and Techniques a Survey.International Journal of Computer Theory and Engineering, 2(6): 1793-8201

**[3].** Khraisat, A.; Gondal, I.; Vamplew, P.; Kamruzzaman, J. Survey of intrusion detection systems: Techniques,datasets and challenges. Cybersecurity 2019, 2, 20.

**[4].** Khraisat, A.; Gondal, I.; Vamplew, P.; Kamruzzaman, J.; Alazab, A. A Novel Ensemble of Hybrid IntrusionDetection System for Detecting Internet of Things Attacks. Electronics 2019, 8, 1210.

**[5].** Alazab, A.; Hobbs, M.; Abawajy, J.; Khraisat, A.; Alazab, M. Using response action with intelligent intrusiondetection and prevention system against web application malware. Inf. Manag. Comput. Secur. 2014, 22,431–449.

**[6].** Alazab, A.; Hobbs, M.; Abawajy, J.; Alazab, M. Using feature selection for intrusion detection system.In Proceedings of the 2012 International Symposium on Communications and Information Technologies(ISCIT), Gold Cost, Australia, 2–5 October 2012; pp. 296–301.

**[7].** Alazab, A.; Abawajy, J.; Hobbs, M.; Khraisat, A. Crime toolkits: The current threats to web applications. J. Inf.Priv. Secur. 2013, 9, 21–39

**[8].** Saranyaa, T.,Sridevib, S.,Deisyc, C., Tran Duc Chungd, Ahamed Khane, M. K. A. (2020).Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review.Third International Conference on Computing and Network Communications (CoCoNet'19),Procedia Computer Science 171 (2020) 1251–1260.

**[9].** Nguyen, H. A. and Choi, D. (2008). Application of data mining to network intrusion detection: classifier selection model.Asia-Pacific Network Operations and Management Symposium. Springer, 2008, pp. 399–408.

**[10].** Lahre, M. K., Dhar, M. T.,Suresh, D., Kashyap, K. and Agrawal, P. (2013). Analyze different approaches for ids using kdd 99 data set," International Journal on Recent and Innovation Trends in Computing and Communication, 1(8): 645–651.

[11]. Haddadi, F.,Khanchi, S., Shetabi, M. and Derhami, V. (2010). Intrusion detection and attack classification using feed-forward neural network.Computer and Network Technology (ICCNT), 2010 Second International Conference on. IEEE, 2010, pp. 262–266.

[12]. Alsharafat, W. (2013). Applying artificial neural network and extended classifier system for network intrusion detection. International Arab Journal of Information Technology (IAJIT), vol. 10, no. 3, 2013.

[13]. Bhargava, N., Sharma, G., Bhargava, R. and Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining, Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, 2013.

[14]. C. Fleizach and S. Fukushima, "A naive bayes classifier on 1998 kdd cup," 1998.

[15]. Alkasassbeh,M.,Al-Naymat, G.,Hassanat, A. B. and Almseidin, M. (2019). Detecting distributed denial of service attacks using data mining techniques, International Journal of Advanced Computer Science & Applications, 1(7): 436–445.

[16]. Mulay, Snehal&Devale, P.R. &Garje, Goraksh. (2010). Intrusion Detection System Using Support Vector Machine and Decision Tree. International Journal of Computer Applications. 3. 10.5120/758-993.

[17]. Khan, Latifur&Awad, M. &Thuraisingham, Bhavani. (2007). A new intrusion detection system using support vector machines and hierarchical clustering. VLDB Journal. 16. 507-521. 10.1007/s00778-006-0002-5.

[18]. AnsamKhraisat, Iqbal Gondal, Peter Vamplew, JoarderKamruzzaman and Ammar Alazab (2020). Hybrid Intrusion Detection System Based on theStacking Ensemble of C5 Decision Tree Classifier andOne Class Support Vector Machine. Electronics. 2020, 9, 173; doi:10.3390/electronics9010173, www.mdpi.com/journal/

[19]. Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T. et al. Intrusion detection model using machine learning algorithm on Big Data environment. J Big Data5, 34 (2018). https://doi.org/10.1186/s40537-018-0145-4

[20]. Mohamed El Boujnouni and Mohamed Jedra (2018).New Intrusion Detection System Based onSupport Vector Domain Description withInformation Gain Metric" International Journal of Network Security, 20(1): 25 – 34.

[21]. Vipin, Das & Vijaya, Pathak &Sattvik, Sharma &Sreevathsan, & Srikanth M. V. N. N. S., &Gireesh. T. (2010). Network Intrusion Detection System Based On Machine Learning Algorithms. International Journal of Computer Science & Information Technology. 2. 10.5121/ijcsit.2010.2613

[22]. Samra Zafar, Muhammad Kamran, Xiaopeng Hu (2019). Intrusion-Miner: A Hybrid Classifier for IntrusionDetection using Data Mining. International Journal of Advanced Computer Science and Applications, Vol. 10, No. 4, 2019.

[23]. Chung, Yuk & Wahid, Noorhaniza. (2012). A hybrid network intrusion detection system using simplified swarm optimization (SSO). Applied Soft Computing. 12. 3014–3022. 10.1016/j.asoc.2012.04.020.

[24]. Almseidin, M., Alzubi, M., Kovacs, S. and Alkasassbeh, M. (2017). Evaluation of machine learning algorithms for intrusion detection system, 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 2017, pp. 000277-000282, doi: 10.1109/SISY.2017.8080566.

[25]. Jayshree Jha and Leena Ragha (2013). Intrusion Detection System using Support Vector Machine. IJAIS Proceedings on International Conference and workshop on Advanced Computing 2013 ICWAC(3):25-30, June 2013

[26]. Enache, A. and Patriciu, V. V. (2014).Intrusion's detection based on Support Vector Machine optimized with swarm intelligence. 2014 IEEE 9th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 2014, pp. 153-158, doi: 10.1109/SACI.2014.6840052.

[27]. Manickam, M., Rajagopalan, S.P. A hybrid multi-layer intrusion detection system in cloud. Cluster Comput22, 3961–3969 (2019). https://doi.org/10.1007/s10586-018-2557-5

[28]. Goel, L. (2020). An extensive review of computational intelligence-based optimization algorithms: trends and applications. Soft Computing.24,16519–16549 (2020). https://doi.org/10.1007/s00500-020-04958-w

**[29].** P. Amudha,1 S. Karthik,2 and S. Sivakumari, “A Hybrid Swarm Intelligence Algorithm for Intrusion DetectionUsing Significant Features”, Hindawi Publishing Corporation Scientific World Journal Volume 2015, Article ID 574589, 15 pagehttp://dx.doi.org/10.1155/2015/574589.

**[30].** Adeel Hashmi and Tanvir Ahmad , “FAAD: A Self-Optimizing Algorithm for Anomaly Detection”, The International Arab Journal of Information Technology, Vol. 17, No. 2, March 2020

**[31].** Thiruvenkadam, Kalaiselvi& Perumal, Nagaraja& Z, Abdul. (2017). A Review on Glowworm Swarm Optimization. International Journal of Information Technology. 3. 49-56.

**[32].** Kaipa, Krishnanand& Ghose, Debasish. (2009). Glowworm Swarm Optimization: A New Method for Optimising Multi-Modal Functions. International Journal of Computational Intelligence Studies. 1. 10.1504/IJCISTUDIES.2009.515637.

**[33].** Kumar, A. S. A., Arpitha, K., Latha, M. N. and Sahana, M. (2017). A novel approach for intrusion detection system using feature selection algorithm. International Journal of Computational Intelligence Research. 13(8): 1963 – 1976.

**[34].** Kuang, F., Xu, W.& Zhang, S. (2014). A novel hybrid KPCA and SVM with GA model for intrusiondetection. Applied Soft Computing Journal, 18, 178–184, 2014.

**[35].** Wathiq, L. A.,Zulaiha, A. O. &Mohd. Z. A (2017). Multi-Level Hybrid Support Vector Machine andExtreme Learning Machine based on Modified K-means for Intrusion Detection System, ExpertSystems with Applications, 67, 296-303, 2017.

DOI: 10.48175/IJARSCT-7751