



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: <https://doi.org/10.22214/ijraset.2025.71571>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Intrusion Detection Using Machine Learning

Anshul Kumar¹, Devansh², Arjun Chaudhary³, Prathu Dwivedi⁴, Mrs. Sristhi Vashisth⁵

Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh, India

Abstract: *The rapid growth of technologies not only formulates life easier but also exposes a lot of security issues. With the advancement of the Internet over years, the number of attacks over the Internet has been increased. Intrusion Detection System (IDS) is one of the supportive layers applicable to information security. IDS provides a salubrious environment for business and keeps away from suspicious network activities. Recently, Machine Learning (ML) algorithms are applied in IDS in order to identify and classify the security threats. This paper explores the comparative study of various ML algorithms used in IDS for several applications such as fog computing, Internet of Things (IoT), big data, smart city, and 5G network. In addition, this work also aims for classifying the intrusions using ML algorithms like Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART) and Random Forest.*

Keywords: *Intrusion Detection System (IDS); Machine Learning (ML) Algorithm; Classification; Random Forest; Support Vector Machine; Accuracy.*

I. INTRODUCTION

The world Internet statistics report informs that, the Internet growth (2000-2019) reached 1,114%, since more than 2 quintillion bytes of data are generated every day. This shows that, the rate of data growth from various sources are extremely very fast, and at the same time the development of hacking tools and methodologies also growing in the faster manner. Therefore, there is a need for information security and data analysis for protecting the data from the intrusion. Due to the huge volume and high speed of data, the traditional detection system is not able to detect intrusion in the faster manner. In order to handle intrusion efficiently, the big data techniques are employed. The big data is defined under 7v's such as (i) volume: size of the data, (ii) velocity: speed at which the data are generated, (iii) variety: different types of data, (iv) value: the worth of data, (v) veracity: trustworthiness of data, (vi) variability: constant change of data meaning, and (vii) visualization: easy accessible or readable of data. The exponential rate of data growth makes the traditional data handling system as complex due to consuming more time and resources. Big data are very complex in nature to handle such kind of data and they need powerful technologies and advanced intelligent algorithms. IDS plays an important role in detecting the attacks. The IDS is a system that will monitor the network traffic in the intent to find out any suspicious activity and known threats. It may also issue the alert to the admin when such activity is discovered. To handle and classify the attacks in the efficient manner, various ML algorithms can be used. This section focuses on various techniques that are used for identifying the intrusion. IDS can be a hardware system or software system that automatically monitors, identify the attack or intrusion, and alert the computer or network. This alert report helps the administrator or user to find and resolve the vulnerability present in the system or network. Some common ways of intrusion detections are: Anomaly-based detection, Signature-based detection and Hybrid-based detection. The anomaly-based intrusion detection is also known as behaviour-based detection, because this method models the behaviour of the users, network, and host systems and thus generates alarm or alert the admin whenever the behaviour is deviated from the usual behaviour. The signature-based IDS is also called as knowledge-based detection. This method is relying on the database which contains previous known attack signature and known system vulnerabilities. Hybrid based detection system is the combination of anomaly-based intrusion detection and signature-based intrusion detection. Most of the IDSs use any one of the intrusion detections namely anomaly or signature. Since both intrusion detections have their own drawbacks, hybrid IDS can be used.

II. METHODOLOGY

ML Algorithm for Intrusion Detection

ML is a subset of Artificial Intelligence (AI). ML makes the system to learn and improve their automatic ability from the experience without being explicitly programmed. For Intrusion Detection System (IDS), ML algorithm works more accurately in detecting the attacks for huge amount of data under less time. Typically, ML algorithms can be classified into three categories:

- Supervised
- Unsupervised
- Semi-supervised.

Supervised ML Algorithm The supervised algorithm deals with fully class labelled data, and finds the relationship between data and its class. This can be done by either classification or regression. The classification has two steps such as training and testing. The training data is done with the help of response variable. The common algorithms under classification category are Support Vector Machine (SVM), Discriminant Analysis, Naïve Bayes, Nearest Neighbour, Neural Network, and Logistic Regression. While some algorithms under regression category are Linear Regression, Support Vector Regression (SVR), Ensemble Methods, Decision Tree, and Random Forest. In this paper, Support Vector Machine, Logistic Regression, Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART), Random Forest (RF) and Ensemble methods discussed.

Support Vector Machine (SVM) SVM is one of the mostly used supervised ML algorithm. SVM can be used for both classification and regression. The algorithm can be trained with the labelled data, and it can output the separation of data into classes by the hyper plane that maximizes the margin among all attack classes. Mehmood et al.

[14] stated that SVM as a binary classifier, it will also perform multi-class classification by using cascaded manner. SVM is mainly depends on the types of kernel used and parameters.

Logistic Regression (LR) LR is a supervised ML classification algorithm used to observe the discrete set of classes. The logistic function makes use of cost function which is called as sigmoid function. This function maps predictions to probabilities. Belavagi et al.

[5] mentioned that by fitting data to the logistic function the probability of occurrence of event can be predicted.

Linear Discriminant Analysis (LDA) LDA is a simple linear supervised ML algorithm used for dimensionality reduction and prediction. Based on Bayes theorem, LDA estimates the probability. When LDA is used as a classification problem, the output variable should be categorical and supports binary as well as multi-class class.

Classification and Regression Tree (CART) CART is a simple nonlinear supervised ML algorithm used for classification and regression. In CART, the target variable should be categorical, whereas in regression tree the target variable should be continuous.

Random Forest (RF) RF is a complex nonlinear supervised algorithm used for classification and regression. This will construct many decision trees at training the model and the outcomes of predictions from all trees are pooled to make a result so, it is mentioned as Ensemble techniques. The RF classifiers works as follows: the higher the number of trees in the model will result in the higher accuracy and not over-fit the model. 2.1.6 Ensemble Methods In order to produce the optimal predictive model, this ML technique combines several models. The main idea behind ensemble method is to grouping of all weak learners to form a strong learner; thereby the accuracy of the model is increased. Some common types of ensemble methods are Bagging, Boosting, and Stacking. Gautam et al. [9] approached the bagging ensemble method and works out the trail with Naive Bayes, partial decision tree algorithm (PART) and Adaptive Boost. They showed that ensemble approach has the higher rate than PART, Naive Bayes and Adaptive Boost. **Unsupervised ML Algorithm** For intrusion detection, the unsupervised learning algorithm will try to find out the hidden structure in unlabelled data. There is no training data for unsupervised learning. This can be done by clustering or association analysis or dimensionality reduction. The clustering algorithms such as K-Means, K-Medoids, and C-Means can be used. The dimensionality reduction algorithms such as Singular Value Decomposition (SVD), Principle Component Analysis (PCA) can be used.

K-means K-means is one of the unsupervised ML algorithms. This algorithm works based on the finding groups in the data, and the number of groups can be represented by the variable. K-means algorithm is highly used in time series data for pattern matching. Sridevi et al. [17] proposed clustering based pattern matching algorithms for predicting the time series data. Varuna et al. [18] proposed K-means clustering, with the cluster of five types such as four types of attack and one normal traffic. These five features are then classified by using Naive Bayes classifier. The drawback of K-Means algorithm is it is not applicable for non-spherical form of data.

Principle Component Analysis (PCA) PCA is a technique which is used for dimensionality reduction. PCA provides new set of variables called principle components and can also be used as an input to any supervised ML algorithm. Aburomman et al. [1] proposed ensemble PCA-LDA method. The PCA is able to remove only linear feature information and LDA will remove the non-linear feature information.

Semi-Supervised ML algorithm The semi supervised ML algorithm lies between unsupervised learning and supervised learning. These learning techniques make use of unlabelled data for training and also a small amount of labelled data for large set of unlabelled data. Jarrah et al. [2] proposed semi-supervised multi layered clustering model for network intrusion detection. This algorithm provides a multiple layers of randomized K-Means clustering algorithm, which improves the diversity among classifiers and results in accurate intrusion detection.

III. LITERATURE REVIEW

Intrusion Detection Systems (IDS) is very essential technology to keep the people away from cyber-attack. Every transaction and information processing is take place through Internet which is very prone to more different types of malicious activity. Therefore, there is a need to provide more concentration for the information security. The application areas covered in this paper are:

- IDS for Internet of Things
- IDS for Smart City
- IDS for Big Data Environment
- IDS for Fog
- IDS for Mobile.

IDS for Internet of Things (IoT) Internet of Things (IoT) is a network of object or device with unique identification, which can sense, accumulate and transfer data over Internet without any human to human or human to computer intervention. IoT devices are powered with low power and it is developed with lightweight protocols. It is lightweight also. Ghasempouret al. [10] discussed the purpose of IoT device in smart grid. It can be highly vulnerable and even attackers can modify the sensors data. The major attack that take place in IoT devices are physical attack, side channel attack, environmental attack, cryptanalysis attacks, Black hole attack, Sybil attack and so on. Jan et al. [16] proposed lightweight intrusion detection by using supervised learning strategy. They developed SVM classifier to detect the attacks (target DDoS). Hasan et al. [11] discussed anomaly and attack detection. They implemented their work using ML algorithm such as LR, SVM, Decision Tree, RF and Artificial Neural Network.

IDS for Smart City Elsaedy et al. [7] discussed Intrusion detection on smart cities. The author used the data set collected from smart water distribution plant. The work is to detect the DDoS attack in smart city applications. The proposed method of this paper consists of two parts: Restricted Boltzmann Machines (RBM) model and classifier model. This RBM model is applied to learn high level features in an unsupervised manner. The classification is used to differentiate the normal and variety of DDOS attack. They used four types of classifiers such as Feed Forward Neural Network (FFNN), Automated FFNN, RF, and SVM. For the high level of features, K-Means algorithm is processed by RBM model and they developed up to 5 layers which provide 5 sub versions of each from clustering algorithm with different k value. For each 5 data set generated from the clustering, 4 types of classifiers are applied and totally 20 experiments have done.

IDS for Big Data Environment Big data consists of very large amount of structured, unstructured, and semi structured data in heterogeneous format. For such a huge amount of data, traditional intrusion handling system is not capable to solve the issues. IDS for big data environment can be only possible by employing ML algorithm. Othman et al. [15] used an Apache Spark big data platform for feature selection and SVM to find intrusion detection. Pre-processed model is standardized to unit variance in spark MLlib. ChiSq selector and SVM are used for feature selection and the feature selection model is based on the method of num Top Features. In order to reduce the effect of misclassification error, the soft SVM margin is used. The user defined variable called slack variable is used to trade between margin and misclassification error. Their result shows that the intrusion detection on big data is achieved with higher performance and speed.

IDS for Fog Computing Fog computing is a new technology of computing paradigm, which bring analytic service to the edge and improving the performance by placing the resources close to where they are needed. The fog computing has three types of layers such as cloud service layer, fog service layer, and user layer. The fog service layer has a geographically distributed fog node which composed of routers, gateway, server at the edge and offers a unique layer in fog computing. Fog nodes support heterogeneous computing which make the fog node more vulnerable to attack such as DDoS, Remote-to-Local (R2L), User-to-Root (U2R), PROBE and so on. An et al. [4] contributes the attack process of DDOS in fog computing, and explore the relationship between the fog node and DDOS based on hyper graph. The state of the fog node is computed by the load factor. To determine the state of the fog node, it is compared with the threshold load level of node. Their model is used to analyse the association of fog nodes suffering from DDOS attack.

IDS for Mobile Mobiles are becoming more predominant tool among the people for communication and for storing more sensitive information. The mobile vulnerabilities are application vulnerability, device vulnerability, networks vulnerability, web and content vulnerability. To solve these vulnerability and threats, the device should have IDS. Maimo et al. [12] proposed a 5G-oriented cyber defence architecture to identify cyber threats in 5G mobile networks by using self-adaptive deep learning based system. They design their architecture for classifying the intrusion by arranging the anomaly detection in two levels: ASD module (anomaly symptom detection) and NAD module (network anomaly detection).

The NAD is implemented by a supervised way of LSTM (long short term memory recurrent networks) and ASD module is implemented by atwo levels supervised or semi-supervised way of DBN (deep belief network) and SAE (stacked auto-encoders).

IV. RESULT & DISCUSSION

In order to evaluate the above literature work, this research work implements Linear Discriminant Analysis (LDA), Classification and Regression Tree (CART) and Random Forest (RF) algorithms for testing purpose. It is implemented on standard KDD'99 Cup data set. The data set has 42 features and 494021 instances with 25 predictors which was mapped to 5 types of classes such as DoS, probes, user to remote attack (U2R), remote to local (R2L), and normal. The work has three step processes such as Data pre-processing, Classifications and Evaluation.

```

Models: lda, cart
Number of resamples: 10

Accuracy
      min. 1st Qu.  Median    Mean 3rd Qu.    Max. NA's
lda  0.9818334 0.9825095 0.9835408 0.9833409 0.9841676 0.9845652  0
cart 0.9805162 0.9807505 0.9842619 0.9830170 0.9845534 0.9849194  0

kappa
      min. 1st Qu.  Median    Mean 3rd Qu.    Max. NA's
lda  0.9462902 0.9483059 0.9513086 0.9507276 0.9531384 0.9543379  0
cart 0.9410264 0.9418288 0.9520842 0.9484624 0.9530112 0.9540817  0
    
```

Fig.1. Classification result of LDA and CART algorithms

pred	dos	norma1	probe	r2l	u2r
dos	312282	31	24	2	0
norma1	49	87485	42	73	30
probe	0	15	3870	3	0
r2l	0	18	835	4	0
u2r	0	1	0	0	12

Fig. 2. Confusion Matrix Result of Random Forest Algorithm

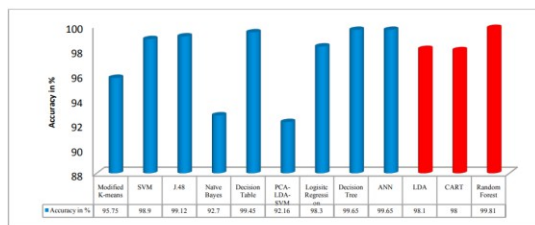


Fig. 3. Performance Comparisons of various ML algorithms in IDS

In data pre-processing steps, the features are mapped to their appropriate functions and features are selected based on filter method. From 42 variables, 20 variables were selected based on correlation attribute evaluation by choosing ranker search method. In classification phase, LDA algorithm, CART algorithm and Random Forest are used.

The data set is divided into training set and testing set based on 80-20 rule. In evaluation phase, the metrics like accuracy and kappa were used to measure the performance of LDA, RF and CART algorithms. The experimental results show that the RF algorithm yields better accuracy (99.65%) than LDA (98.1%) and CART (98%) algorithms. The work was implemented using RStudio. The classification result of KDD cup dataset using LDA, CART and RF are shown in Fig.1 and Fig.2. The different ML algorithms such as LDA, CART and RF used in this work as well as in above survey is compared in terms of accuracy is shown in Fig. 3. The graph shows that, the RF algorithm used in this research work also yields better accuracy among other algorithms. In general, the algorithms like RF, ANN and decision tree give better results for classifying the attacks. From the above comparisons, it is observed that the performance of the algorithms also depends on the size of the dataset and applications employed.

V. CONCLUSION

This paper provides an extensive review of the network intrusion detection mechanisms based on the ML and DL methods to provide the new researchers with the updated knowledge, recent trends, and progress of the field. A systematic approach is adopted for the selection of the relevant articles in the field of AI-based NIDS. Firstly, the concept of IDS and its different classification schemes is elaborated extensively based on the reviewed articles. Then the methodology of each article is discussed and the strengths and weaknesses of each are highlighted in terms of the intrusion detection capability and complexity of the model. Based on this study, the recent trend reveals the usage of DL-based methodologies to improve the performance and effectiveness of NIDS in terms of detection accuracy and reduction in FAR.

About 80% of the proposed solutions were based on the DL approaches with AE and DNN are the most frequently used algorithms. Although DL schemes have much superior performance than the ML-based methods in terms of their ability to learn features by itself and stronger model fitting abilities. But these schemes are quite complex and require extensive computing resources in terms of processing power and storage capabilities. These challenges need to be addressed to fulfill real-time requirements for NIDS and hence improves NIDS performance. The study also shows that 60% of the proposed methodologies were tested using KDD Cup'99 and NSL-KDD data sets mainly because of the availability of extensive results using these datasets. But these datasets are quite old to address modern network attacks, and hence limits the performance of the proposed methodologies in real-time environments. For AI-based NIDS methods, the model should be tested with the latest updated dataset like CSE-CIC-IDS2018 for better performance in terms of detection accuracy for intrusions. This article also highlights the research gaps in improving the model performance for low-frequency attacks in a real-world environment and to find efficient solutions to reduce complexity for the proposed models. Proposing an efficient NIDS framework using less complex DL algorithms and have an effective detection mechanism is a potential future scope of research in this area. For future research, we will use this knowledge to design a novel, lightweight, and efficient DL-based NIDS which will effectively detect the intruders within the network.

REFERENCES

- [1] Tarter A. Importance of cyber security. Community Policing-A European Perspective: Strategies, Best Practices and Guidelines. New York, NY: Springer; 2017:213-230.
- [2] Li J, Qu Y, Chao F, Shum HP, Ho ES, Yang L. Machine learning algorithms for network intrusion detection. AI in Cybersecurity. New York, NY: Springer; 2019:151-179.
- [3] Lunt TF. A survey of intrusion detection techniques. Comput Sec. 1993;12(4):405-418.
- [4] Anderson JP. Computer Security Threat Monitoring and Surveillance. Fort Washington, PA: James P Anderson Co; 1980.5. Debar H, Dacier M, Wespi A. Towards a taxonomy of intrusion-detection systems. Comput Netw. 1999;31(8):805-822.
- [5] Hoque MS, Mukit M, Bikas M, Naser A, An implementation of intrusion detection system using genetic algorithm; 2012. arXiv preprint arXiv:1204.1336.
- [6] Prasad R, Rohokale V. Artificial intelligence and machine learning in cyber security. Cyber Security: The Lifeline of Information and Communication Technology. New York, NY: Springer; 2020:231-247.
- [7] Lew J, Shah DA, Pati S, et al. Analyzing machine learning workloads using a detailed GPU simulator. Paper presented at: Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). Madison, WI, USA: IEEE; 2019:151-152.
- [8] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2(1):1
- [9] Dong B, Wang X. Comparison of deep learning method to traditional methods using for network intrusion detection. Paper presented at: Proceedings of the 8th IEEE International Conference on Communication Software and Networks (ICCSN). Beijing, China: IEEE; 2016:581-585.
- [10] Vasilomanolakis E, Karuppayah S, Mühlhäuser M, Fischer M. Taxonomy and survey of collaborative intrusion detection. ACM Comput Surv. 2015;47(4):1-33.
- [11] Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor. 2015;18(2):1153-1176.
- [12] Thomas R, Pavithran D. A survey of intrusion detection models based on NSL-KDD data set. Paper presented at: Proceedings of the 5th HCT Information Technology Trends (ITT). Dubai, United Arab Emirates: IEEE; 2018:286-291.
- [13] Liu H, Lang B. Machine learning and deep learning methods for intrusion detection systems: a survey. Appl Sci. 2019;9(20):4396.
- [14] Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity. 2019;2(1):20.
- [15] DKAC, Papa JP, Lisboa CO, Munoz R, DVHC A. Internet of Things: a survey on machine learning-based intrusion detection approaches. Comput Netw. 2019;151:147-157.3.
- [16] Keele S. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Technical Report, Technical Report, Ver. 2.3. EBSE Technical Report. vol. 5, EBSE; 2007.
- [17] Scopus Preview Welcome to Scopus Preview; 2020. <https://www.scopus.com/>. Accessed June 25, 2020.
- [18] Mukkamala S, Janoski G, Sung A. Intrusion detection using neural networks and support vector machines. Paper presented at: Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat.No.02CH37290). Honolulu, HI, USA: IEEE; vol. 2, 2002:1702-1707.
- [19] Garcia-Teodoro P, Diaz-Verdejo J, Maciá- Fernández G, Vázquez E. Anomaly-based network intrusion detection: techniques systems and challenges. Comput Secur. 2009;28(1-2):18-2



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)