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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Intrusion Detection Using Machine Learning

Anshul Kumar[1], Devansh[2], Arjun Chaudhary[3], Prathu Dwivedi[4], Mrs. Sristhi Vashisth[5]
Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh, India

Abstract: ***Therapidgrowthoftechnologiesnotonly formulates life easier but also exposes a lotof securityissues. With the advancement of theInternetoveryears,thenumberofattacks overtheInternethasbeenincreased.Intrusion Detection System (IDS)isoneofthesupportive layers applicable to information security. IDS provideasalubriousenvironmentforbusiness and keeps away from suspicious network activities. Recently, Machine Learning (ML) algorithmsareappliedinIDSinorderto identify and classify the security threats. This paper explores the comparative study of various ML algorithms used in IDS for several applicationssuchasfogcomputing,Internetof 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 RegressionTrees (CART)and Random Forest.***

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

I. INTRODUCTION

The world Internet statistics report informs that,theInternetgrowth(2000-2019)reached 1,114%, since more than 2 quintillion bytes of dataaregeneratedeveryday.Thisshowsthat, 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 informationsecurityanddataanalysisfor 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 inthe faster manner. Inorder 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:theworthof 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.Bigdataarevery complex innature to handle such kind of data and they need powerful technologies and advanced intelligent algorithms.IDS plays an important roleindetectingtheattacks.TheIDSisasystem that will monitor the network traffic in the intent to find out any suspicious activity and knownthreats.Itmayalsoissuethealertsto theadminwhensuchactivityisdiscovered.To handle and classify the attacks in the efficient manner, various ML algorithms can be used. Thissectionfocusesonvarioustechniquesthat are used for identifying the intrusion. IDS can beahardwaresystemorsoftwaresystemthat 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 knownasbehaviour-baseddetection,because thismethodmodelsthebehaviouroftheusers, network,andhostsystemsandthusgenerates alarm or alert the admin whenever the behaviour is deviated from the usual behaviour.Thesignature-basedIDSsalsocalledasknowledge-baseddetection.Thismethodis relyingonthedatabasewhichcontains previous known attack signature and known systemvulnerabilities. Hybridbased 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

MLAlgorithmforIntrusionDetection

ML is a subsetofArtificialIntelligence (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 hugeamountofdataunderlesstime.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, andfindstherelationshipbetweendataandits class. This can be done by either classification or regression. The classification has two stepssuchastraining andtesting . Thetrainingdataisdonewiththehelpofresponsevariable.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,DecisionTree,andRandomForest.In this paper, Support Vector Machine, Logistic Regression,LinearDiscriminantAnalysis(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 andregression. Thealgorithmcan betrainedwiththelabelleddata,anditcan output the separation of data into classes by the hyper plane that maximizes the margin among all attack classes. Mehmood et al.

[14]statedthatSVMasabinaryclassifier,itwill also performmulti-classclassificationbyusing cascademanner.SVMismainlydependsonthe types of kernel used and parameters.

LogisticRegression(LR)LRisasupervised MLclassificationalgorithmusedtoobservethe 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]mentionedthatbyfittingdatatothelogistic functiontheprobabilityofoccurrenceofevent can be predicted.

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

ClassificationandRegressionTree(CART) CART is a simple nonlinear supervised ML algorithm used for classification and regression.InCART,thetargetvariableshould 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 willconstruct many decision trees at training the model and the outcomes of predictions from alltrees 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 themodelwillresultinthehigheraccuracyand not over-fit the model. 2.1.6 Ensemble Methods In order to produce the optimal predictivemodel,thisMLtechniquecombines 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, andStacking.Gautametal.[9]approachedthebaggingensemblemethod and works out the trail with Naive Bayes, partialdecisiontreealgorithm(PART)and Adaptive Boost. They showed that ensemble approachhasthehigherratethanPART,Naive Bayes and Adaptive BoostUnsupervised ML Algorithm For intrusion detection,theunsupervisedlearningalgorithm 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- Mediods, and C-Means can be used. The dimensionality reductions 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 worksbasedonthefindinggroupsinthedata, and the numberof groupscan 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 patternmatchingalgorithmsforpredictingthe timeseriesdata.Varunaetal.[18]proposedK-meansclustering,withtheclusteroffivetypes such as four types of attack and one normal traffic.Thesefivefeaturesarethenclassifiedby using Naive Bayes classifier. The drawbackof K-Means algorithm is it is not applicable for non-spherical form of data.

PrincipleComponentAnalysis(PCA)PCA isatechniquewhichisusedfordimensionality 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 ensemblePCA-LDAmethod.ThePCAisableto 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 supervisedlearning. These learning techniques make use ofunlabelleddatafortrainingandalsoa small amount of labelled data for large set of unlabelleddata.Jarrahetal.[2]proposedsemi- supervised multi layered clustering model for network intrusion detection. This algorithm provides a multiple layersofrandomizedKMeansclusteringalgorithm, whichimprovesthediversityamongclassifie randresultin accurate intrusion detection.

III. LITERATUREREVIEW

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 whichis very prone to more different typesofmaliciousactivity.Therefore,thereisa need to provide more concentration for the information security. The application areas covered in this paper are:

- IDSforInternetofThings
- IDSforSmartCity
- IDSforBigData Environment
- IDSfor Fog
- IDSforMobile.

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 modifythesensorsdata.Themajorattackthat 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).Hasanetal.[11]discussedanomalyand attackdetection.Theyimplementedtheirwork using ML algorithm such as LR, SVM, Decision Tree, RF and Artificial Neural Network.

IDS for Smart City Elsaeidyet 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 classifiermodel.ThisRBMmodelisappliedto learn high level features in an unsupervised manner. The classification is used to differentiate the normal and variety of DDOS attack.They usedfourtypesofclassifierssuch 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 eachfromclusteringalgorithmwithdifferentk value. For each 5 data set generated from the clustering,4typesofclassifiersareappliedand 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 byemployingMLalgorithm.Othmanetal.[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. Chisqselector 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 definedvariablecalledslackvariableisusedto 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 newtechnologyofcomputingparadigm,which bring analytic service to the edge and improving the performance by placing the resourcesclosertowheretheyareneeded.The fogcomputinghasthreetypesoflayerssuchas cloud service layer, fog service layer, and user layer.Thefogservicelayerhasageographically distributed fog node which composed of routers,gateway,serverattheedgeandoffers auniquelayerinfogcomputing.Fognodes support heterogeneous computing which makesthefognode morevulnerabletoattack suchasDDoS,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 computedbytheloadfactor.Todeterminethe 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.

IDSforMobileMobilesarebecomingmore predominant tool among the people for communication andforstoringmore sensitive information. The mobile vulnerabilities are application vulnerability, device vulnerability, networks vulnerability, web and content vulnerability.Toresolvethesevulnerabilityand threats, the device shouldhave IDS.Maimo et al. [12] proposeda5G-oriented cyberdefence 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 arrangingtheanomalydetectionintwolevels: ASDmodule (anomalysymptomdetection)and NAD module (network anomaly detection).

The NAD is implemented by a supervised way of LSTM (long short term memory recurrent networks)andASDmoduleisimplementedby atwolevelsupervisedorsemi-supervisedway ofDBN(deepbeliefnetwork)andSAE(stacked auto-encoders).

IV. RESULT & DISCUSSION

Inordertoevaluatetheaboveliteraturework, 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,usertoremoteattack(U2R),remoteto 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.ClassificationresultofLDAandCARTalgorithms

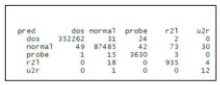


Fig. 2. Confusion Matrix Result of Random Forest Algorithm

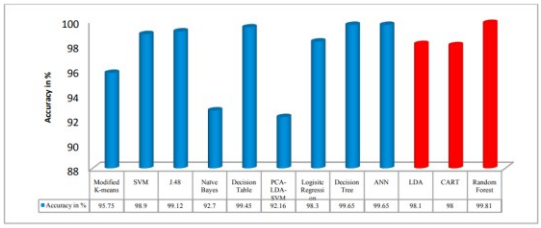


Fig. 3.Performane Comparisons of various ML algorithms in IDS

In data pre-processing steps, the features are mapped to their appropriate functions and features are selected based on filter method. From42variables,20variableswhereselected 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 testingsetbasedon80–20rule.Inevaluation phase, the metrics like accuracy and kappa wereusedtomeasuretheperformanceofLDA, RF and CART algorithms. The experimental resultshowsthattheRFalgorithmyieldsbetter accuracy(99.65%)thanLDA(98.1%)andCART (98%) algorithms. The work was implemented usingRStudio.TheclassificationresultsofKDD cup dataset using LDA, CART an RF are shown in Fig.1 and Fig.2.The different ML algorithms such as LDA, CART and RF used in this work as wellasinabovesurveyiscomparedintermsof 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

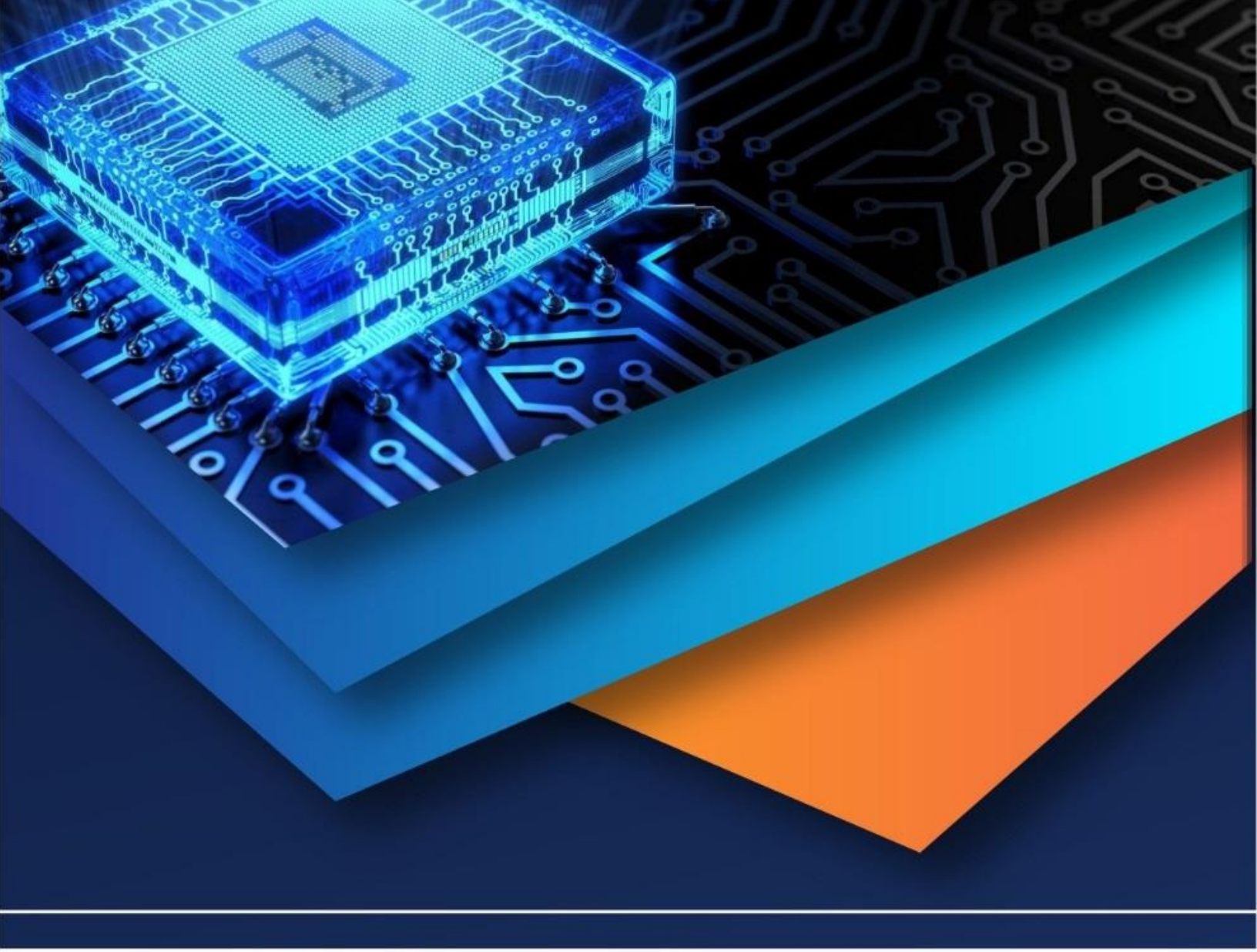
Thispaperprovidesanextensivereviewofthe network intrusion detection mechanisms based on the ML and DL meth-ods to provide the new researchers with the updated knowledge, recenttrends,andprogress ofthe field.Asystematicapproachisadoptedforthe selection ofthe relevant articles in the field of AI-basedNIDS.Firstly,theconceptofIDSandits different classification schemes is elaborated extensively based on the reviewed articles. Then the methodology of each article is discussedandthestrengthsandweaknessesof 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 improvetheperformanceandeffectivenessof NIDS in terms of detection accuracy and reduction in FAR.

About 80%of the proposed solutions were based on the DL approaches withAEandDNNarethemostfrequentlyused 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.Buttheseschemesarequitecomplex 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. Thestudyalsoshowsthat60%oftheproposed 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-basedNIDS methods, the model should be tested with the latest updated dataset like CSE-CIC-IDS2018 for better performance in termsofdetectionaccuracyforintrusions.This article also highlights the research gaps in improving the model performance for low- frequencyattacksinareal-worldenvironment 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 effectivedetectionmechanismisapotential futurescopeofresearchinthisarea.Forfuture research,wewillusethisknowledgetodesign a novel, lightweight, and efficient DL-based NIDSwhichwilleffectivelydetecttheintruders within the network.

REFERENCES

[1] Tarter A. Importance of cyber security. Community Policing-A European Perspective: Strategies, BestPractices andGuidelines. 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. NewYork, 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 MonitoringandSurveillance.FortWashington, 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 preprintarXiv: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 ofthe IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). Madison, WI, USA: IEEE; 2019:151-152.
[8] NajafabadiMM,VillanustreF,Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learningapplicationsandchallengesinbigdata analytics. J Big Data. 2015;2(1):1
[9] DongB,WangX.Comparisondeeplearning method to traditional methods using for network intrusion detection. Paper presentedat: Proceedings of the 8th IEEE International Conference on Communication SoftwareandNetworks(ICCSN).Beijing,China: IEEE;2016:581-585.
[10] Vasilomanolakis E, Karuppayah S, Mühlhäuser M, Fischer M. Taxonomy and survey of collaborative intrusion detection. ACM ComputSurv. 2015;47(4):1-33.
[11] BuczakAL,GuvenE.Asurveyofdatamining and machine learning methods for cyber security intrusion detection. IEEE Commun SurvTutor. 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 5thHCT 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,PapaJP,LisboaCO,MunozR,DVHC A. Internet of Things: a survey on machine learning-based intrusion detectionapproaches. 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/.AccessedJune25, 2020.
[18] Mukkamala S, Janoski G, Sung A. Intrusion detection using neural networks and support vector machines. Paper presented at: Proceecings 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