Intrusion Detection System Using Machine Learning: An Overview

Nidhi Kumari[1], Vimmi Malhotra[2]

[1]Student,Master of Technology, Computer Science Engineering
[2]Assistant Professor, Computer Science Department, Dronacharya college of Engineering, Gurugram

---***---

Abstract - *Today's wireless networks are faced with rapid expansions in errors, flaws, and attacks that threaten to undermine their security. Since computer networks and applications are built on multiple platforms, network security is becoming increasingly important. Both complex and expensive operating programs may have security vulnerabilities. The term "intrusion" refers to attempts to break security, completeness, and availability. Network security vulnerabilities and abnormalities can be identified using an IDS. The development of intrusion detection technology has been a burgeoning field, despite being often regarded as premature and not as an ultimately comprehensive method of fighting intrusions. Security experts and network administrators have also made it a priority task. This means that more secure systems cannot replace it completely. Using data mining to detect intrusion, IDS is able to predict future intrusions based on detected intrusions. An extensive review of literature on the use of data mining methods for IDS is presented in this paper. First, we will review data mining approaches for detecting intrusions using real-time and benchmark datasets. This paper presents a comparison of methods of detecting intrusions in the network with their merits and demerits. In this paper, we propose approaches to improve network intrusion detection.*

Key Words: Intrusion detection, Security, Machine Learning, Data mining.

1. INTRODUCTION

Due to the rapid increase in the number of applications and organizations using computer networks, security is becoming increasingly important. Most companies use network security tools like antivirus and anti-spam software to protect themselves from network attacks. These tools can't detect complex or new attacks, however.

An IDS [1] enables computer networks and computers to detect and eliminate unwanted intrusions. Identifier systems can collect and process information from various sources within a network or computer, identifying threats that can make people vulnerable, such as misuse and intrusion. IDSs (Intrusion Detection Systems) [2] are systems that continuously monitor and analyze events occurring on a network to detect malicious activity. IDS are now regarded as an important element of the security infrastructure in most companies. By detecting intrusions, companies can deter attacks on their networks. Security professionals could use this method to reduce current network security risks and the complexity of current threats.

The procedure of gaining extra approval by gaining access to a database is how attackers to compromise databases, approved users who abuse their assent are how approved users gain access to databases. An IDS identifies assaults that appear to be unusual or harmful in purpose [3]. The existence of various types of intrusions has been identified using different techniques, but there are no heuristics to confirm their accuracy. The majority of traditional IDS rely on human analysts to distinguish between invasive and noninvasive network data. Because of the considerable time frame necessary to notice an assault, fast attacks are not practical. For network owners and operators, access to the internet is a particularly delicate topic. As the online world offers different hazards, several solutions are designed to avoid internet assaults.

The data mining method is used to derive models from massive data sets. The technology behind machine learning and deep learning has enabled a wide range of data mining techniques in recent years. Intrusion detection research uses a variety of techniques, including classifiers, link assessments, and sequence analysis.

Data mining is essential for detecting intrusions by machine learning. It can provide insights into possible behaviors based on prior experiences. The most common data mining techniques are a hybrid association, clustering, and classification. Grouping data based on results is called clustering. Clustering is most commonly done using K-means.

The most popular technique used by mining analysts is classification and prediction, which creates models, characterizes data and projects the future to extract important insights. Extends the IDS by categorizing outcomes as regular or abdominal using metric-based categorization. The techniques used to mine audit data were questioned for consistency, which resulted to several proposals for improvements to the present data mining technologies.

A variety of data mining approaches have been described in the literature as being useful in detecting network breaches. This paper delves into how data mining algorithms may be utilized for intrusion detection in great depth. Advantages, restrictions, and effectiveness are also discussed in the paper. With this extensive investigation, IDS functionality may be enhanced even further in the future.

As far as content is concerned, Section 2 of the paper summarizes the previous literature studies on intrusion detection based on data mining. Identifying the strengths, and weaknesses, and evaluating their performance efficiency, Section 3 summarizes these techniques, Section 4 reviews prior studies, while Section 5 wraps up the debate by outlining potential future improvements.

2. LITERATURE SURVEY

Based on fuzzy entropy, Varma et al. [4] proposed a features technique for real-time intrusion detection datasets using Ant Colony Optimization (ACO) techniques. Both discrete and continuous traffic characteristics could be extracted using this technique. In order to determine the most valuable characteristic among the detected characteristics, ACO uses the fuzzy entropy heuristic. Classifiers are therefore more accurate at detecting intrusions. A real-time intrusion threat detection technology provided the best solution for this task.

Using a support vector machine, Thaseenn and Kumar [5] describe an intrusion detection method which is based on chi square properties. By calculating the largest variance for each feature, we improved the parameters of SVM. Reverse variance considerably reduces variance, which improves kernel parameters. Using variance balancing, the SVM parameters were improved in this intrusion detection model. The results improved classification accuracy.

Khammassi & Krichen [6] derived the best sample of characteristics from IDS using a featured selection method. To reduce the dataset size, the pre-processed dataset was first re-sampled, and then a wrapper method was used. Genetic algorithms and logistic regression were used in the method. The wrapping technique allows network intrusions to be identified using NBTree, Random Forest (RF), and C4.5 classifiers.

An anomaly-based IDS has been proposed by Aljawarneh et al.[7]. In order to determine which characteristics were most important, a voting method led to an information gain. By doing so, basic learners' probabilities could be integrated. Several classifiers, including REPTree, AdaBoostM1, Meta Paging, Na*ve Bayes, and Random Tree, were implemented to identify network intrusions with the given attributes.

Kabir et al. [8] developed LS-SVM (Least Square Support Vector Machine). There were two stages to this method. The dataset was divided based on arbitrary criteria into preset subgroups. The characteristics that distinguished these groups were then analyzed. They were listed together in the same order. For determining the most efficient allocation method, the variability of the data within subgroups was examined. To extract samples from a network, we used LS-SVM in phase two.

According to Khan et al. [9], intrusion detection should be performed in two stages. First, network traffic was categorised using likely score values. When determining if the intrusion was a routine or an assault, deep learning used this likelihood score value as a second measure. The probability score in step two was applied to avoid overfitting. By using this two-stage technique, it is possible to handle large volumes of unlabeled data effectively and automatically.

Based on Convolutional Neural Networks (CNNs) and feature reduction techniques, Xiao et al. [10] developed an intrusion detection model. A step in the process of intrusion detection is the reduction of dimensionality by eliminating irrelevant or redundant characteristics. A CNN algorithm was used to extract features from the reduced data. A supervised learning approach was used to obtain the data that are more successful in detecting intrusions.

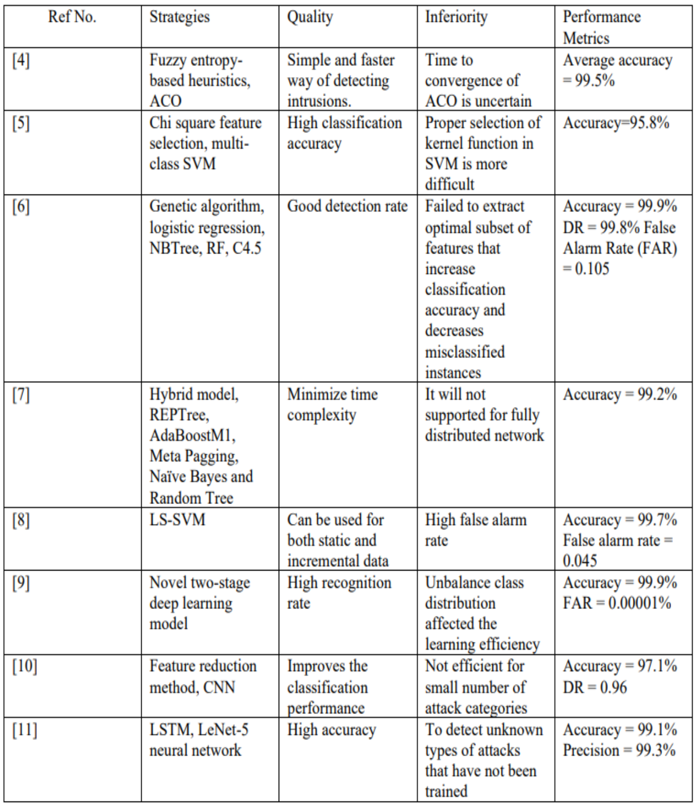
Among the approaches presented by Zhang et al.[11] to detect network intrusion is a deep hierarchical network. Spatial and temporal aspects of flow were studied using LeNet-5 and LSTM. Various network cascade mechanisms were used to train the deep hierarchical network instead of two. It was also examined how the flow of information in the network varies.

3. OBSERVATION

Comparative study of strengths and limitations of the intrusion detection technologies discussed in the preceding section. Each approach has advantages and downsides, as illustrated in Table 1. The tabulated data makes it simple to determine which strategy works best and provides the most advantages. In this observation table, we can see how current methods are flawed, and we can come up with new ideas to solve them.

This observation table gives an idea about the study of IDS using machine learning methods. Fuzzy entropy based which gives 99.5% accuracy with time convergence of ACO. Chi-square feature selection methods give 95.8% accuracy but the selection of functions in SVM is difficult. The genetic Algorithm with 99.9%, failed to obtain the optimal subset.

Hybrid model,99.2 % accuracy .and supported fully distributed network. These types of algorithms used especially in network-based IDS have higher accuracy obtained during the learning process. Several approaches to data mining were tested on two sets of intrusion data: NSL-KDD and UNSV-NB15 to determine the effectiveness of intrusion detection. The study provides information on unbalanced data distributions, convergence times, normal and abnormal traffic distributions, classification, and detection rates. The combination of network intrusion algorithms and deep learning algorithms such as OSS, SMOTE, and BiLSTM, combined with a deep hierarchical network and learning algorithm, provides superior performance in terms of accuracy and precision.



Ref No.	Strategies	Quality	Inferiority	Performance Metrics
[4]	Fuzzy entropy-based heuristics, ACO	Simple and faster way of detecting intrusions.	Time to convergence of ACO is uncertain	Average accuracy = 99.5%
[5]	Chi square feature selection, multi-class SVM	High classification accuracy	Proper selection of kernel function in SVM is more difficult	Accuracy=95.8%
[6]	Genetic algorithm, logistic regression, NBTree, RF, C4.5	Good detection rate	Failed to extract optimal subset of features that increase classification accuracy and decreases misclassified instances	Accuracy = 99.9% DR = 99.8% False Alarm Rate (FAR) = 0.105
[7]	Hybrid model, REPTree, AdaBoostM1, Meta Pagging, Naïve Bayes and Random Tree	Minimize time complexity	It will not supported for fully distributed network	Accuracy = 99.2%
[8]	LS-SVM	Can be used for both static and incremental data	High false alarm rate	Accuracy = 99.7% False alarm rate = 0.045
[9]	Novel two-stage deep learning model	High recognition rate	Unbalance class distribution affected the learning efficiency	Accuracy = 99.9% FAR = 0.00001%
[10]	Feature reduction method, CNN	Improves the classification performance	Not efficient for small number of attack categories	Accuracy = 97.1% DR = 0.96
[11]	LSTM, LeNet-5 neural network	High accuracy	To detect unknown types of attacks that have not been trained	Accuracy = 99.1% Precision = 99.3%

4. METHODOLOGY

1.1 Machine Learning

The creation of analytical models is automated through the use of machine learning, a data analysis tool. A type of artificial intelligence that uses data analysis to detect patterns, recognize trends, and take action based on minimal human involvement. When machines learn from sufficient data and develop models capable of detecting attack variants and new attacks, intelligent intrusion detection systems are able to achieve satisfactory detection levels. Our study is primarily focused on identifying and summarizing IDSs that have utilized machine learning in the past.

1.1.1 Decision Tree

Decision Trees are a class of Supervised Learning algorithms that can be used for predicting categorical or continuous variables. It works by breaking data from the root node into smaller and smaller subsets while incrementally building an associated decision tree. Decision nodes create a rule and leaf nodes deliver a result. In addition to CART, C4.5, and ID3, there are numerous other DT models available. Multi-decision tree methods such as RF and XGBoost are used to build advanced learning algorithms.

1.1.2 K-Nearest Neighbors

In KNN, "feature similarity" is used to predict a data sample's class based on its features. By calculating the distance between it and its neighbors, it identifies samples based on their neighbors. A parameter called k affects KNN algorithm performance. The model can overfit with small values of k. Karatas et al[12]. CSE-CIC-IDS2018 has been used as a benchmark dataset to compare the performance of ML algorithms. Selecting very large k values led to incorrect classification of the sample instances. An improvement in detection rate for minority class attacks resulted from using Synthetic Minority Oversampling Technique (SMOTE) to resolve dataset imbalance.

1.1.3 Support vector machine (SVM)

A supervised machine learning algorithm that consists of an n-dimensional hyperplane whose elements are spaced closer than the distance between them. Both linear and nonlinear problems can be solved with SVM algorithms. A kernel function is typically applied to nonlinear problems. With the kernel function, an input vector is transformed into a high-dimensional feature space first. After that, the support vectors are used as a decision boundary to determine the maximal marginal hyperplane. NIDS can be improved by using the SVM algorithm to correctly identify normally occurring and malicious traffic.

1.1.4 K- mean clustering

A cluster is a group of similar data that is grouped together. K-Mean clustering is an unsupervised system for dividing data into meaningful clusters using centroid based iterative learning. Datasets consist of centroids (cluster centers), and K is their number. To assign data points to clusters, a distance calculation is usually used. During clustering, reducing the distance between data points is the main objective.

The clustering concept is used in the RF model in a multilevel intrusion detection model framework, Yao et al. [13]. Four modules were combined into the proposed solution: clustering, pattern discovery, fine-grained classification, and model updating. A potential attack that isn't detected in one module will then be passed to the next one. KDD Cup'99 dataset has been used in testing this proposed methodology. The model still showed superiority despite fewer attacks in our dataset.

1.1.5 Artificial neural network

A neural network simulates the way the brain works. As part of the ANN output, there are hidden layers and data layers. Layers within a layer are entirely connected to one another. An ANN is one of the most popular Machine-learning techniques and has proven to be an efficient method to detect different types of malware. The most frequently used learning algorithm for supervised learning is backpropagation (BP). To start with, weights are randomly assigned at the start of training. A weight tuning algorithm is then implemented to specify which hidden unit representation minimizes the misclassification error the best. IDS based on ANN still requires improvement, especially for less frequent attacks. Less frequent attacks have a smaller training dataset than more common attacks, meaning ANN has a harder time learning their characteristics correctly.

1.1.6 Ensemble method

Ensemble methods' principal advantage is that they allow you to take advantage of the different classifiers by using them together. Classifiers differ in their behaviors, so their use may not always be optimal. It may be that some types of detection programs work well for detecting some types of attacks but fail to detect others. Combining weak classifiers is an ensemble approach, which involves training many classifiers while selecting the best one through a voting algorithm.

An ensemble method by Shen et al.[14] to propose an IDS that utilized most of the ELM features. During the ensemble pruning phase, our proposed methodology is optimized using a BAT optimization algorithm. Data from the KDD Cup'99, NSL-KDD, and Kyoto datasets were used to test the model. Results of the experiments indicated that combining multiple ELs in an ensemble manner outperformed each EL individually.

Using the deep neural network (DNN) and DT as a base classifier, Gao et al.[15] proposed an adaptive ensemble model using and adaptive voting algorithm to pick the best classifier. In experiments performed with the NSL-KDD dataset, the proposed methodology has been verified. Other models were compared to demonstrate its efficiency. For weaker attack classes, it didn't perform well.

CONCLUSION AND RECOMMENDATIONS

A detailed overview of data mining strategies based on IDS mostly in network is offered in this study. The advantages and disadvantages of these strategies are also examined in order to offer future options for improving intrusion detection performance and thereby improving IDS. The findings of the comparison investigation revealed that insider threat detection utilising deep hierarchical networks had greater accuracy, clarity, and recall. However, the intrusion detection system algorithm's training period is lengthy. Because the efficiency of the system of wireless intrusion detection systems are poorly quantified in the preceding comparisons, a machine learning-based network detection model is presented. Machine learning capabilities that automatically extract and select features reduce the difficulty of calculating domain-specific, manually generated features and allow you to skip the traditional attribute selection phase. Deep learning (DL) is also widely used in a variety of fields and has proven to be effective. Therefore, for the next few years, we will use machines and deep learning algorithms to prevent overfitting with zero elements, address model training issues with a limited percentage of attack classifications, and avoid DNNs. Increases the effectiveness of intrusion detection and prevention. Misunderstandings due to controversial input formation and ultimately solving the problem of instability in cyber attacks.

REFERENCES

[1] iMohit iS iD, iGayatri iB iK, iVrushali iG iM, iArchana iL iG iand iNamrata iR. iB i(2015). Using lArtificial iNeural iNetwork iClassification iand iInvention iof iIntrusion iin iNetwork iIntrusion iDetection iSystem. iInternational iJournal iof iInnovative iResearch iin iComputer iand iCommunication iEngineering, i3(2). i

[2] iZaman iS, iEl-Abed iM iand iKarray iF i(2013 iJanuary). ,Features iselection iapproaches ifor iintrusion idetection isystems ibased ion ievolution ialgorithms.

[3] iNazir iA i(2013). iA iComparative iStudy iof idifferent iArtificial iNeural iNetworks ibased iIntrusion iDetection iSystems. iInternational iJournal iof iScientific iand iResearch iPublications i.

[4] iVarma iP iR iK, iKumari iV iand iKumar iS iS i(2016). iFeature iselection iusing irelative ifuzzy ientropy iand iant icolony ioptimization iapplied ito ireal-time iintrusion idetection isystem. iProcedia iComputer iScience, i85, i503-510. i

[5] iThaseen iI iS iand iKumar iC iA i(2017). Intrusion idetection imodel iusing ifusion iof ichi-square ifeature iselection iand imulti iclass iSVM. iJournal iof iKing iSaud iUniversity-Computer iand iInformation iSciences, i29(4), i462-472. i

[6] iKhammassi iC iand iKrichen iS i(2017). iA iGA-LR iWrapper iApproach ifor iFeature iSelection iin iNetwork iIntrusion iDetection. iComputers i& iSecurity, i70, i255-277. i

[7] iAljawarneh iS, iAldwairi iM iand iYassein iM iB i(2018). iAnomaly-based iintrusion idetection isystem ithrough ifeature iselection ianalysis iand ibuilding ihybrid iefficient imodel. iJournal iof iComputational iScience, i25, i152-1613] iKabir iE, iHu iJ, iWang iH iand iZhuo iG i(2018). iA inovel istatistical itechnique ifor iintrusion idetection isystems. iFuture iGeneration iComputer iSystems, i79, i303-318. i

[8] iKabir iE, iHu iJ, iWang iH iand iZhuo iG i(2018). iA inovel istatistical itechnique ifor iintrusion idetection isystems. iFuture iGeneration iComputer iSystems, i79, i303-318. i

[9] iKhan iF iA, iGumaei iA, iDerhab iA iand iHussain iA i(2019). iA inovel itwo-stage ideep ilearning imodel ifor iefficient inetwork iintrusion idetection. iIEEE iAccess, i7, i30373-30385. i

[10] iXiao iY, iXing iC, iZhang iT iand iZhao iZ i(2019). iAn iintrusion idetection imodel ibased ion ifeature ireduction iand iconvolutional ineural inetworks. iIEEE iAccess, i7, i42210-42219. i

[11] iZhang iY, iChen iX, iJin iL, iWang iX iand iGuo iD i(2019). iNetwork iintrusion idetection: iBased ion ideep ihierarchical inetwork iand ioriginal iflow idata. iIEEE iAccess, i7, i37004-37016. i

[12] iKaratas iG, iDemir iO, iSahingoz iOK. iIncreasing ithe iperformance iof imachine ilearning-based iIDSs ion ian iimbalanced iand iup-to-date idataset. iIEEE iAccess. i2020;8:32150-32162. ihttps://doi.org/10.1109/ACCESS.2020.2973219. i

[13] iYao iH, iFu iD, iZhang iP, iLi iM, iLiu iY. iMSML: ia inovel imultilevel isemi-supervised imachine ilearning iframework ifor iintrusion idetection isystem. iIEEE iIoT iJ. i2018;6(2):1949-1959. ihttps://doi.org/10.1109/JIOT.2018.2873125.

[14] iShen iY, iZheng iK, iWu iC, iZhang iM, iNiu iX, iYang iY. iAn iensemble imethod ibased ion iselection iusing ibat ialgorithm ifor iintrusion idetection. iComput iJ. i2018;61(4):526-538. ihttps://doi.org/10.1093/comjnl/bxx101.

[15] iGao iX, iShan iC, iHu iC, iNiu iZ, iLiu iZ. iAn iadaptive iensemble imachine ilearning imodel ifor iintrusion idetection. iIEEE iAccess. i2019;7:82512-82521. ihttps://doi.org/10.1109/ACCESS.2019.2923640.