A Review on Intrusion Detection System using Machine Learning Techniques

Usman Shuaibu Musa
Department of Computer Science & Engineering
School of Engineering & Technology
Sharda University
Gr. Noida, UP, India
usmanmusa04@gmail.com

Sudeshna Chakraborty
Department of Computer Science & Engineering
School of Engineering & Technology
Sharda University
Gr. Noida, UP, India
sudeshna.chakraborty@sharda.ac.in

Muhammad M. Abdullahi
Department of Computer Science & Engineering
School of Engineering & Technology
Sharda University
Gr. Noida, UP, India
mmabdullahi10@gmail.com

Tarun Maini
Department of Computer Science & Engineering
School of Engineering & Technology
Sharda University
Gr. Noida, UP, India
tarunmaini77@gmail.com

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397121

***Abstract*— Computer networks are exposed to cyber related attacks due to the common usage of internet, as the result of such, several intrusion detection systems (IDSs) were proposed by several researchers. Among key research issues in securing network is detecting intrusions. It helps to recognize unauthorized usage and attacks as a measure to ensure the secure the network's security. Various approaches have been proposed to determine the most effective features and hence enhance the efficiency of intrusion detection systems, the methods include, machine learning-based (ML), Bayesian based algorithm, nature inspired meta-heuristic techniques, swarm smart algorithm, and Markov neural network. Over years, the various works being carried out were evaluated on different datasets. This paper presents a thorough review on various research articles that employed single, hybrid and ensemble classification algorithms. The results metrics, shortcomings and datasets used by the studied articles in the development of IDS were compared. A future direction for potential researches is also given.**

Keywords—Machine learning, Single classifiers, Hybrid, Ensemble, Misuse detection, Intrusion Detection System

I. INTRODUCTION (HEADING 1)

For detecting illicit or abnormal behavior, IDS is used. Attack is launched in a network in a state of an anomaly behavior. Attackers use the opportunity of network weaknesses like poor security measures and practices, program bugs such as buffer overflows, yielding the breaches of the network. The attackers may be less privileged device operators who aim to claim more access control or black hat-hackers who are normal users of internet that intend to hijack sensitive information [1]. The Techniques for detecting intrusion can be centered on misuse detection or based on anomaly detection. Misuse based IDS tracks the flow of network traffic and compares them to the predefined malicious activities signatures in a database. Whereas in the technique of anomaly detection, attacks are detected when they are compared with actions which deviate from normal user operations [2].

The IDS could be Network-based IDS (NIDS) or Host-based IDS (HIDS). Computer network administrators utilize the host-based intrusion detection method to track and evaluate activities on a specific machine [1]-[3]. HIDS has a benefit that when moving over a network, encrypted information can be accessed. The downside is that HIDS is very difficult to handle, as every host needs to configure and manage data. In addition, some forms of denial-of-service attacks could disable HIDS. NIDS, is intelligently distributed software or hardware-based IDS in network which track packets passing through the network. NIDS is doubled interfaced, first being for listening network conversation and the second for monitoring [4][3]. The NIDS has the benefit that it needs a few well-fit NIDS to monitor a wide network and often NIDS is hidden to various intruders, so it is safe against invasions. However, during a time of heavy traffic, NIDS has the downside of finding it hard to discover an attack launch.

The rampant usage of internet makes it difficult to protect network resources from the mischievous action of attackers. According to Cybersecurity ventures, the damage related to cybersecurity is predicted to reach $6 trillion yearly by 2021[1]. Gartner reports that, in taking steps to counter the damage, Global expenses on cybersecurity could reaches $133.7 billion in 2022. Multiple measures have been taken in which various security tools such as IDS were developed [1], [2]. The various previously works done on building IDS showed effectiveness to some extent. However, several issues need to be addressed to build an efficient IDS that could detect and report malicious traffic with very high detection accuracy.

Most IDS were developed and evaluated using outdated and old dataset like KDD Cup '99, NSL KDD and so on, which lack the most recent and up to date attack labels. Slow detection rate is experienced in the existing works. This happens due to inability to get rid of all redundant and irrelevant columns. High false positive rate. This happens when a legit traffic is incorrectly detected and classified as an attack. The false positive rate increases complexity of IDS, hence, reducing its performance.

The rest of the paper is presented as follows; section two of the paper discusses the overview of ML techniques. The comparison of the studied works was covered by section three. The comparison was based on the classier used, the performance of the algorithms as well as the dataset applied to evaluate the algorithms. The last section of this paper

discusses, concludes and provides future scope in developing IDS using ML techniques.

II. OVERVIEW OF MACHINE LEARNING (ML) ALGORITHMS

A. Machine Learning (ML)

ML could be described as an approach whereby models undergone training for the purpose of learning and

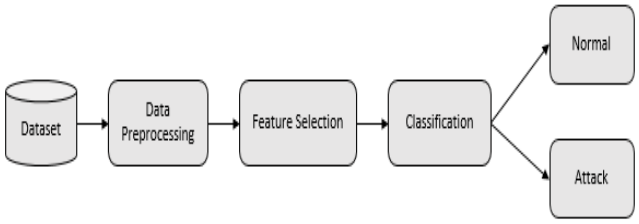

Fig 1: Block Diagram of Intrusion Detection System

enhancing performance parameters automatically so that they don't have to be solely programmed [2],[10],[11],[12] using previous experience or example data. In accordance with attributes, the ML model focuses on training data sets to predict various class labels [4],[13]. ML is typically divided into three groups:

1.Supervised Learning

In supervised ML the dataset to be trained is made up of examples of the input vector, each with their equivalent desired output vectors [1], [3], [4]. Algorithms in this type of learning include: Naïve Bayes, KNN, ANN, Decision Tree (C4.5, ID3, CART, RF, and J48), SVM, Ensemble methods (Bagging, Voting Classifier, Adaboost, Gradient Boosting), logistic regression [3]

2.Unsupervised Learning

In unsupervised ML, the learning algorithm is not given labels and as such it must by itself find structure in its input[1]. This is also known as learning without a teacher. Self-Organizing Map (SOM), Apriori algorithm, Éclat algorithm and outlier detection, Hierarchical clustering, and Cluster Analysis (K-Means clustering, Fuzzy clustering) are various unsupervised learning algorithms.

3. Reinforcement Learning

In reinforcement learning, the model is trained to make a sequence of decisions. The goal is achieved in an uncertain and potentially complex manner [1], [4]. The model performs trial and error to bring up a solution to the problem. Deep Q Network (DQN), Q-Learning, State-Action-Reward-State-Action (SARSA), Deep Deterministic Policy Gradient (DDPG) are various reinforcement learning algorithms.

A. Single Machine Learning Classifiers

ML Classifiers are classified as single classifiers when they only contain just one classification algorithm [1],[16],[17]. The single machine classification models have been adopted by many intrusion detection systems. SVM, ANN, DT, K-Nearest Neighbor, NB are made up of one ML algorithm and were adopted in various IDS studied in this work.[3], [5]–[9].

1. Support Vector Machine (SVM)

SVM could be used for both classification and regression cases. There is a provision of separating hyper-plane in SVM which defines the various classes to be predicted. The classification always depends upon the nature of the problem and the adopted dataset. The dataset could be one dimensional and, in such case, the hyper-plane is a point on one dimensional line. In a situation whereby the dataset is two dimensional, then the hyper-plane is a separating line and for three dimensional the hyper-plane is a plane and lastly for higher dimensional dataset it is a hyper-plane. SVM is widely used in most intrusion detection systems due its popularity in making accurate predictions [13],[5], [10], [11].

2. Artificial Neural Network (ANN)

ANNs are a category of ML algorithms motivated by the behavior and amount of computing performed by human brain in biological nervous system. The model of ANN is made up of an input layer, one or more hidden layer(s) and an output layer [6], [10]. The hidden layer(s) weigh and process the inputs fed to the artificial neurons so that the output to the next layer can be decided [8]. In ANN, a learning rule known as gradient-descent back propagation of error is used to adaptively adjust the various weights and biases of the hidden and output layer neurons so that the desired or required output is achieved [23],[7], [8].

3. Decision Tree (DT)

Decision tree is a type of machine learning algorithms that is applied for both categorical and numeric classifications. The decision tree is made up of three nodes, namely; the tree's topmost node, called the root node (root), the intermediate node also called internal node (node) and leaf nodes also called leaves. In decision tree the flow of learning rule is top to down [1], [12]. The leaves are the outcomes of decisions. In decision tree data sample is split into two homogeneous sets (subsample) based on most significant splitter [4], [13], [14]. The decision tree is widely used in classification problems due to its popularity in data exploration and less data cleaning requirement [17][1].

4. K-Nearest Neighbor (KNN)

KNN is a distance-based ML model, employed to solve classification problems. When it used and combined with prior knowledge, KNN produces a very good result. The classification in KNN is done by classifying each unlabeled example by the majority label among its K-nearest neighbors in the training set [3],[15]. The nearest neighbors are determined by the KNN's performance on distance metrics. Various techniques of measuring distance are used to identify the nearest neighbors, the most popular among which is the Euclidean distance. KNN is time efficient and can easily be interpreted, hence it's widely usage for classification problems in IDS and other applications [16].

C. Hybrid Classifiers

Hybrid classifier is an approach in which multiple ML models are combined in order to improve the efficiency of

the aggregate classifier in the IDS [26][10]. The purpose for the use of the IDS hybrid method is to improve the IDS performance as it is well known that hybrid systems work much more efficiently than the IDS classification of single machine learning [27],[28]. Either supervised or unsupervised ML models can be set as the initial hybrid classifier level [29],[30].

D. Ensemble Classifier

Ensemble Classifier is a combination of more than one ML classifiers sometimes referred to as poor learners, whose individual choices are combined as a consensus decision in some way to provide better effective predictive performance [31],[32]. Therefore, by aggregating different results of poor learners, the ensemble classifier provides enhanced efficiency. The work of many researchers who implemented ensemble models demonstrates a high precision and efficiency. Approaches for building ensembles include: Infusion of randomness, plurality polling, ensemble of function collection, bagging and random trees and performance coding error correction [14],[23].

III.REVIEW AND COMPARISON OF RELATED WORKS

In the work done by [9], multiple ML techniques have been adopted with the aim of overcoming the challenges of lacking accuracy when dealing with low frequent attacks often faced by the previous IDS when ANN with fuzzy clustering is used [8], [17]. They did this effectively by separating the heterogeneous collection dataset for training into homogeneous one, thus decreasing the size of each training set. In the study, J48 trees, MLP and BN classifiers were applied, offering the highest precision for J48 trees [33]. One of their work's big disadvantage is their failure to implement the feature extraction to discard all irrelevant, obsolete and unnecessary attributes.

An ensemble-based ML approach was applied in [1], in such the outputs of several models, both supervised and unsupervised ML were combined through voting classification. The work increases the accuracy and reliability of the IDS. Their work was evaluated on Kyoto2006+ dataset since it is more appealing in comparison with the most adopted datasets that are outdated. The accuracy of their work is quite good, though the false positive rate is high.

A real-time approach to hybrid IDS [11] was suggested in such approach the signature-based detection was utilized to discover well-known intrusions and the method of anomaly to discover new threats. A good rate of detection value was obtained in this work because the attacks that avoided the signature-based technique could be classified as an intrusion by the technique of anomaly detection. On the last day of the trial, the precision of the algorithm improved incrementally every day to a substantial value of 92.65 percent, and the rate of false negative declines sharply as the algorithm improves and train the machine every day. However, whenever the model is extended to a very broad data size, the problem of slow detection rate is observed.

A study conducted by [34] reveals that the performance of anomaly-based IDS could be improved, especially in the FPR. The NSL-KDD dataset was applied to evaluate the extreme-gradient boosting (XGBoost) and AdaBoost models. While a relatively high accuracy is achieved, the implementation of hybrid or ensemble ML classifiers is required to boost the effectiveness of the IDS.

Many works done failed to address the issues high execution time and detection rate as their work lack feature extraction. A study conducted by [13] evaluated various ML models on the NSL-KDD with various ML algorithms and attribute extraction methods. Because of the high FPR of the model and the work focusing solely on signature-based threats, a significant limitation in zero-day attacks remains unexplained, leaving novel attacks uncaptured. Many of the previous works failed to evaluate their adopted model on different datasets. A work proposed by [18] suggested a novel IDS in which a feature extraction was applied. The work gives the advantage of integrating the ensemble classifier with selecting features that provides increased intrusion detection performance and accuracy. Three separate datasets were utilized for the work; the popular NSL-KDD dataset and two newly released data sets, i.e. IDS2017-CIC. And AWID. The CFS-BA-based technique was used for feature collection. The ensemble-based method improves the efficiency of multi-class categorization. The model showed the best value of accuracy when evaluated on AWID dataset.

Both the FFANN and PRANN were applied in [15], which utilized scaled conjugate gradient and Bayesian regularization techniques in training the ANN based IDS. To assess the quality and capability of the work, various result metrics were used. In various output tests on different attack detections from the yielded result, the 2 models have been shown to bettered each other in performance. Overall, the FFANN provided 98,0742 percent improved precision. By checking the model on multiple datasets, the reliability of the work needs to be increased.

Four different algorithms were combined in the work of [16] in an ensemble model which includes; bay classifier, decision tree, random forest, and RNN-LSTM. the work provides contribution in which imbalanced dataset was handled through selecting the most effective intrusion detection features needed for detecting intrusions and signaling to system administrators whether the traffic is a legit or illicit behavior. Although the approach performs on NSL-KDD to some degree of precision, an experimental study on the most up-to-date datasets is required.

The work of [17] developed an IDS using single machine learning classifier. They applied RF and DT algorithms, evaluated on the NSL-KDD dataset. Having edged the decision tree in accuracy, the random classifier gives the superior results. The study has not addressed both the issue of detection rate and the FPR.

A work proposed by [18] on the IDS in which two separate datasets, NSL-KDD and UNSWN B-15, were evaluated on KNN and Random Committee. In this work, a feature extraction has been implemented that produces and

uses only the most appropriate attribute subsets for the applied datasets. The study findings show that the Random Committee algorithm works better than KNN. In future studies, it is important to further resolve the problem of large data scale, Data imbalance, and normal efficiency of IDS algorithms.

In Ponthapalli et al's proposed work, single classifiers were applied to detect network intrusion. The algorithms used are: SVM, LR, RF, and DT [19]. The work was evaluated on NSL-KDD dataset. The research showed that with the random forest classifier, the intrusion detection system performs the best. They have also found that there is the least execution time for the RF algorithm. The study has the drawback of only being evaluated efficiently with one dataset only.

In [20] A work was carried out on a stacking ensemble technique using heterogeneous datasets. LR, KNN, SVM and RF constitute the ensemble approach. The study uses two of the recent datasets, UGR'16 and UNSW NB-15 datasets. In a simulated machine, UNSW NB-15 was generated, while UGR '16 is generated in an actual data traffic scenario [20]. The method increased the IDS' estimation accuracy and detection speed and returned the best accuracy. Moreover, further studies need to be performed on multiple datasets that contain the latest attack types.

A hybrid NIDS was proposed in the work of [8], where several hybrid models were evaluated on the KDD-NSL dataset. A combination of Neural Network and K-Means clustering with Attribute extraction was made. Also, SVM was combined with K-means. The findings showed vividly that the hybridization of the types of ML complements each other, boosting IDS's performance. The highest accuracy is gotten in the integration of support vector machine and K-means with Attribute extraction. To decrease the FPR, further works need to be carried out using improved hybrid ML techniques.

A. Comparison of Related Work

In the review of the previous work, various papers have been studied. At least two research papers have been studied in each of the years within the stated range. Figure 2 below illustrated the overall distribution of the studied research papers.As it can be clearly observed in figure 3, ensemble classifier returns with highest accuracy whenever it is employed over the years.

B. Datasets employed in the previous Works

Dataset can be defined as the collection of records. A record is the word that is used to describe a single data row. Each record consists of many features, referred to as a data instance attribute. The KDD-NSL is the most common dataset used in the work being studied. In general, KDD'99, NSL-KDD, Kyoto2006, UGR2006, CICIDS'17, and UNSW-NB'15 are the dataset used to evaluate various algorithms applied in the studied works.

KDDCup dataset is the dataset applied in the 3rd International Knowledge Discovery and Data Mining Tools Competition. It is made up of 41 columns of attributes. The attributes are constituted in each input pattern instance. The second most used dataset is the NSL-KDD dataset. The NSL-KDD data set is an upgrade over KDD'99 data set, redundant records have been removed from KDD Cup'99 dataset to get rid of bias effects of classification. This dataset consists of 38 numeric features and 3 nominal features taking the total number of features stands at 41 [12].

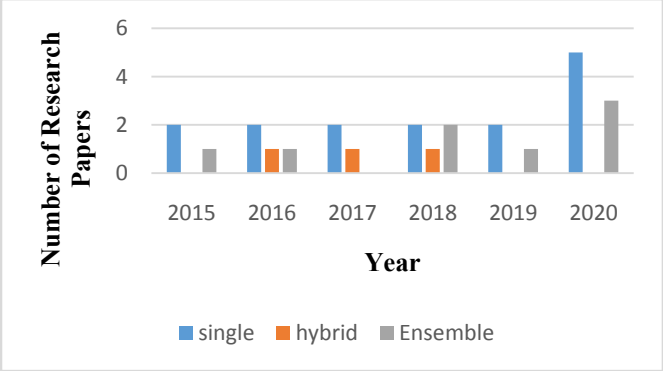

Fig 2: Grouping of Research papers based on type of classifier used.

The Kyoto2006+ is developed based on real traffic data gathered for three years at Kyoto University using 348 honeypots. In these, 24 features are used in this dataset; its 14 features are similar to that of KDD Cup'99 dataset. The remaining 10 columns containing six characteristics relevant to knowledge bring light to certain problems frequently encountered while using the KDD Cup '99 [10].

AWID is made up of real traffic of benign and attacks which was collected from real network environments. It was made publicly published in 2015. Standard and recent typical attacks are given in the CIC-IDS2017 dataset. It was founded in 2017 by the Canadian Institute for Cyber Security, it is an updated dataset for IDS. It consists of 3,119,345, 84 rows, with 84 distinct labelled features [14]. The UNSW NB-15 dataset was established in 2015. Its acronym means New South Wales University. There is a total of 47 attributes in the dataset with two class marks [20].

TABLE 1: Distribution of Dataset Usage over the Years

Year	KDD Cup 99	NSL-KDD	CIC-IDS 2017	UNSW NB-15	UGR '16
2015	1	1	0	0	0
2016	1	2	0	0	0
2017	0	3	0	0	0
2018	2	1	1	0	0
2019	0	3	0	1	0
2020	0	4	0	0	2

Table 2 and Figure 3 explain the distribution of datasets usage over the years from the review papers reviewed. KDD-NSL was applied 14 times, covering 58.33 percent of the total use of data sets. It was accompanied by KDD Cup '99 which was applied in four occasions. UNSW NB-15 is applied in two occasions, while Kyoto-2006+, AWID, CICIDS-2017, and UGR 2016 were applied each in one occasion.

TABLE 2: Comparison of the related works

TITLES	ALGORITHMS	DATASETS	RESULTS (ACCURACY)	FINDINGS	DRAWBACKS
IDS using bagging with partial decision tree base classifier[14]	Genetic Algorithm (GA) based feature selection. Bagged Classifier with partial decision tree	KDD-NSL	Bagged PART=99.7166%	High detection rate	High execution time
IDS based on combining cluster centers and nearest neighbors[15]	KNN CANN SVM	KDD'99	CANN=99.76% KNN=93.87% SVM=80.65%	They applied attribute selection for effective classification of intrusions and normal traffic	Some malicious traffic managed to escape detection
Comparison of classification techniques applied for network intrusion detection and classification[13]	BFTree NBTree J48 RFT MLP NB	KDD-NSL	BFTree=98.24% NBTree=98.44% J48=97.68% RFT=98.34% MLP=98.53% NB=84.75%	High decrease in FP	The study needs to be evaluated on updated datasets
Random Forest Modeling for Network IDS[18]	RF	KDD-NSL	99.67%	Relatively high detection rate and low false alarm was achieved	Attribute selection technique have to be evaluated to reduce the complexity experienced during execution like EFS.
Anomaly Detection Based on Profile Signature in Network using Machine Learning Techniques[19]	GA SVM	KDD'99	GA=84.0333% SVM=94.8000% Hybrid (GA+SVM) =98.333%	Decrease in FPR	There is need to evaluate the algorithm on different datasets
Fast KNN Classifiers for Network Intrusion Detection System[20] **Machine Learning Based Network Intrusion Detection[21]**	KNN C-ELMs	KDD-NSL KDD-NSL	99.95% 98.82%	Increase in accuracy compared to the previous works High detection rate	High Execution time The work has to be applied on different datasets
Intrusion detection in computer networks using hybrid machine learning techniques[22] **Machine Learning Methods for Network Intrusions[23]**	Hybrid NN, SVM and K-Means. J48 Tress MLP BN	KDD-NSL KDD '99	SVM+K-Means=96.81% NN+K-Means=95.55% J48=93.1083% MLP=91.9017% BN=90.7317%	Integration of supervise and unsupervised ML models complement each other in boosting IDS efficiency Provides solution to the low accuracy often faced in the detection of low frequent attacks	Have to evaluate the work on up to date datasets Failed to select the required attributes only
Evaluation of Machine Learning Techniques for Network Intrusion Detection[24]	K-Means KNN FCM SVM NB RBF	Kyoto2006+	RBF=97.54% KNN=97.54 Ensemble=96.72% NB=96.72% SVM=94.26% FCM=83.60% K-Means=83.60%	The work was evaluated using kyoto2006+	Low Recall
Anomaly Detection Based on Profile Signature in Network using Machine Learning Techniques[19]	GA SVM	KDD'99	GA=84.0333% SVM=94.8000% Hybrid (GA+SVM) =98.333%	Decrease in FPR	There is need to evaluate the algorithm on different datasets

Fast KNN Classifiers for Network Intrusion Detection System[20]	KNN	KDD-NSL	99.95%	Increase in accuracy compared to the previous works	High Execution time
Real Time Hybrid Intrusion Detection System[25]	FEC CSA	KDD'99	TP=92.65%	high detection rate is achieved	Failed to show consistency when the work is evaluated using high dimensional datasets
Network Intrusion Detection using Clustering and Gradient boosting[12]	XGBoost AdaBoost	KDD-NSL	XGBoost + Clustering=84.253% XGBoost =80.238 AdaBoost + Clustering=82.011% AdaBoost =80.731%	Improvement in FPR	Need to evaluate the model using recent datasets
Network Intrusion Detection using Supervised Machine Learning Technique with feature selection[4]	ANN SVM	KDD-NSL	ANN=94.02%	Attribute selection technique greatly helped in attaining good accuracy	High FP
Building an Efficient Intrusion Detection System[26]	CFS-BA Combination of C4.5, RF and Forest PA	KDD-NSL AWID CIC-IDS2017	NSL-KDD=99.80% AWID=99.50% CIC-IDS2017=99.90%	Various datasets were used to evaluate the work; hence, the work was very effective	FPR is recorded in one of the applied datasets
A Feed-Forward ANN and Pattern Recognition ANN Model for Network Intrusion Detection[27]	FFANN PRANN	KDD-NSL	FFANN=98.0792% PRANN=96.6225%	Hybridization of more than one ML classifiers often assists in getting a good accuracy	Need to evaluate the model using recent datasets
Anomaly Based Network Intrusion Detection using Ensemble Machine Learning Technique[28]	DT BC RNN-LSTM RF Ensemble of the 4 classifiers	KDD-NSL	85.20%	Addressed the issue of high dimensionality of dataset	Need to evaluate the model using recent datasets
Network Intrusion Detection System using Random Forest and Decision Tree Machine Learning Techniques [17]	RF DT	KDD-NSL	RF=95.323% DT=81.868%	Was easier to implement	Low detection rate
Detecting Intrusions in Computer Network Traffic with Machine Learning Approaches[29]	KNN RC	KDD-NSL UNSW NB-15	NSL-KDD using 1)KNN=98.727% 2) NSL-KDD using RC=99.696% 3) UNSW NB-15 using KNN=97.3346% 4) UNSW NB-15 using RC=98.955%	Ensemble returns with better accuracy than its Single counterpart	Data imbalanced issue
Implementation of Machine Learning Algorithms for Detection of Network Intrusion[30]	DT LR RF SVM	NSL-KDD	1) RF=73.784% 2) DT=72=303% 3) SVM=71.779% 4) LR=68.674%	RF classifier helps in reducing execution time	The model fails to replicate its performance when other classifiers were used
A Stacking Ensemble for NIDS using Heterogeneous Datasets[31]	KNN, LR, RF and SVM	UNSW NB-15 UGR '16	1) UNSW NB-15=94.00% 2) UGR '16=98.71%	High accuracy and detection rate	Need to evaluate the model using recent datasets
A study of IDS using advanced GA[32]	GA	NSL-KDD	96%	From the obtained results, it is possible to update and upload new rules to a system whenever new intrusions are	High execution time

				discovered	
Anomaly network based using a reliable hybrid artificial bee colony and adaboost algorithm[33]	Adabooost for classification and ABC for feature selection	NSL-KDD ISCXIDS2012	Adaboost=98.9%	As the result of applying feature selection using ABC, a good performance on different datasets was demonstrated	The approach needs to be evaluated on recent datasets
Improved off-line IDS using GA[34]	GA	NSL-KDD	GA=98.90%	The approach managed to reduce the false negative by building up aggregate solution sets of all compatible intrusions found	It is needed that future IDS development has to consider the standardization of audit files
An implementation of IDs using GA[35]	GA	KDD'99	99.40%	A reasonable level of detection rate was achieved	Need to evaluate the approach on more recent datasets
Improving adaboost-based IDS performance on CICIDS2017 dataset[36]	Adaboost	CICIDS2017	81.83%	The result of this work shows that the work's performance outperforms the previous works of similar technique	Machine learning based approach needs to be adopted

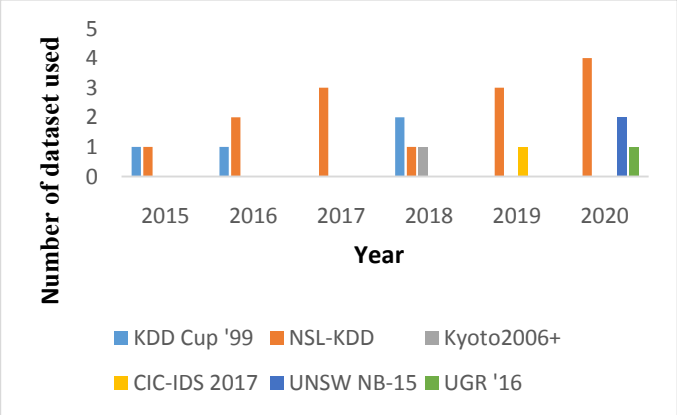

Fig 3: Distribution of dataset usage over the years

II. DISCUSSIONS AND FUTURE WORK

As seen in fig 4, the predictive accuracy and detection rate of ensemble and hybrid classifiers is higher than that of single classifiers. The following problems have been established for future research work and have to be addressed to enhance the efficiency of IDSs:

When they are mixed in a particular way, single machine learning classifiers perform so well, hence classification methods of hybrid and ensemble ML need to be adopted more frequently. In future experiments, more models need to be built in such a way that they can work successfully on different datasets. Certain classifiers perform best on particular datasets.

Some of the research papers studied did not apply attribute extraction prior to the classification phase, while others have done so. All irrelevant, unnecessary and redundant features have to be removed to increase the reliability and the detection rate of intrusion detection systems; the attribute selection must be considered in future study.

As mentioned in section 3, 58.33 percent of the studied works used KDD-NSL to evaluate their work's performance. In future studies, more recent and updated datasets must be used to evaluate employed algorithms in order to cope with the more current malicious intrusions and threats.

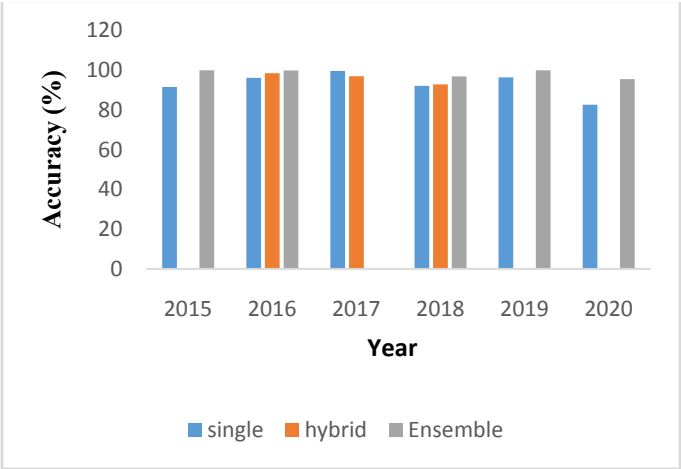

Fig 4: Comparison of Classifiers in terms of Accuracy

III. CONCLUSION

The introduction of ML invents new approaches for IDS whereby various researchers and academics have implemented diverse forms of classifications in the development of models of IDSs. The paper addressed numerous research papers written from 2015 to 2020 on the use of machine learning classifiers in intrusion detection

systems. Ensemble and hybrid classifiers have been able to outperform their single classifier equivalent among the different models implemented in the various works being done, and thus have the highest predictive accuracy and detection rate.

REFERENCES

[1] P. Sangkatsanee, N. Wattanapongsakorn, and C. Charnsripinyo, "Practical real-time intrusion detection using machine learning approaches," Comput. Commun., vol. 34, no. 18, pp. 2227–2235, 2011, doi: 10.1016/j.comcom.2011.07.001.

[2] P. Nader, P. Honeine, and P. Beauseroy, "Detection of cyberattacks in a water distribution system using machine learning techniques," 2016 6th Int. Conf. Digit. Inf. Process. Commun. ICDIPC 2016, pp. 25–30, 2016, doi: 10.1109/ICDIPC.2016.7470786.

[3] M. A. Jabbar, R. Aluvalu, and S. Sai Satyanarayana Reddy, "Cluster based ensemble classification for intrusion detection system," ACM Int. Conf. Proceeding Ser., vol. Part F1283, pp. 253–257, 2017, doi: 10.1145/3055635.3056595.

[4] N. S. Naganhalli and S. Terdal, "Network intrusion detection using supervised machine learning technique," Int. J. Sci. Technol. Res., vol. 8, no. 9, pp. 345–350, 2019.

[5] S. Peddabachigari, A. Abraham, and J. Thomas, "Intrusion Detection Systems Using Decision Trees and Support Vector Machines," Int. J. Appl. Sci. Comput., vol. 11, no. 3, pp. 118–134, 2004.

[6] L. P. Dias, J. J. F. Cerqueira, K. D. R. Assis, and R. C. Almeida, "Using artificial neural network in intrusion detection systems to computer networks," 2017 9th Comput. Sci. Electron. Eng. Conf. CEEC 2017 - Proc., pp. 145–150, 2017, doi: 10.1109/CEEC.2017.8101615.

[7] M. A. Manzoor and Y. Morgan, "Real-time Support Vector Machine based Network Intrusion Detection system using Apache Storm," 7th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEEE IEMCON 2016, pp. 1–5, 2016, doi: 10.1109/IEMCON.2016.7746264.

[8] Akashdeep, I. Manzoor, and N. Kumar, "A feature reduced intrusion detection system using ANN classifier," Expert Syst. Appl., vol. 88, pp. 249–257, 2017, doi: 10.1016/j.eswa.2017.07.005.

[9] D. G. Mogal, S. R. Ghungrad, and B. B. Bhusare, "NIDS using Machine Learning Classifiers on UNSW-NB15 and KDDCUP99 Datasets," Ijarcce, vol. 6, no. 4, pp. 533–537, 2017, doi: 10.17148/ijarcce.2017.64102.

[10] J. Hrabovsky, P. Segec, M. Moravcik, and J. Papan, Trends in application of machine learning to network-based intrusion detection systems, vol. 863. Springer International Publishing, 2018.

[11] S. Gupta, "ANALYZING THE MACHINE LEARNING ALGORITHMS- NAÏVE BAYES , RANDOM TREE , AND SUPPORT VECTOR MACHINES SVM USING THE KDD99 DATA SET TO PREDICT AND CLASSIFY THE," no. 2, pp. 452–459, 2016.

[12] P. Verma, S. Anwar, S. Khan, and S. B. Mane, "Network Intrusion Detection Using Clustering and Gradient Boosting," 2018 9th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2018, pp. 1–7, 2018, doi: 10.1109/ICCCNT.2018.8494186.

[13] A. S. Amira, S. E. O. Hanafi, and A. E. Hassanien, "Comparison of classification techniques applied for network intrusion detection and classification," J. Appl. Log., vol. 24, pp. 109–118, 2017, doi: 10.1016/j.jal.2016.11.018.

[14] D. P. Gaikwad and R. C. Thool, "Intrusion detection system using Bagging with Partial Decision Tree base classifier," Procedia Comput. Sci., vol. 49, no. 1, pp. 92–98, 2015, doi: 10.1016/j.procs.2015.04.231.

[15] W. C. Lin, S. W. Ke, and C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," Knowledge-Based Syst., vol. 78, no. 1, pp. 13–21, 2015, doi: 10.1016/j.knosys.2015.01.009.

[16] A. R. Syarif and W. Gata, "Intrusion detection system using hybrid binary PSO and K-nearest neighborhood algorithm," Proc. 11th Int. Conf. Inf. Commun. Technol. Syst. ICTS 2017, vol. 2018-Janua, pp. 181–186, 2018, doi: 10.1109/ICTS.2017.8265667.

[17] V. Hajisalem and S. Babaie, "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection," Comput. Networks, vol. 136, pp. 37–50, 2018, doi: 10.1016/j.comnet.2018.02.028.

[18] N. Farnaaz and M. A. Jabbar, "Random Forest Modeling for Network Intrusion Detection System," Procedia Comput. Sci., vol. 89, pp. 213–217, 2016, doi: 10.1016/j.procs.2016.06.047.

[19] K. Atefi, S. Yahya, A. Rezaei, and S. H. B. M. Hashim, "Anomaly detection based on profile signature in network using machine learning technique," Proc. - 2016 IEEE Reg. 10 Symp. TENSYMP 2016, pp. 71–76, 2016, doi: 10.1109/TENCONSpring.2016.7519380.

[20] B. Brao and K. Swathi, "Fast kNN Classifiers for Network Intrusion Detection System," no. April, 2017, doi: 10.17485/ijst/2017/v10i14/93690.

[21] C. H. Lee, Y. Y. Su, Y. C. Lin, and S. J. Lee, "Machine learning based network intrusion detection," 2017 2nd IEEE Int. Conf. Comput. Intell. Appl. ICCIA 2017, vol. 2017-Janua, pp. 79–83, 2017, doi: 10.1109/CIAPP.2017.8167184.

[22] D. Perez, M. A. Astor, D. P. Abreu, and E. Scalise, "Intrusion Detection in Computer Networks Using Hybrid Machine Learning Techniques," 2017.

[23] M. Alkasassbeh and M. Almseidin, "Machine Learning Methods for Network Intrusion Detection."

[24] M. Zaman and C. H. Lung, "Evaluation of machine learning techniques for network intrusion detection," IEEE/IFIP Netw. Oper. Manag. Symp. Cogn. Manag. a Cyber World, NOMS 2018, pp. 1–5, 2018, doi: 10.1109/NOMS.2018.8406212.

[25] A. Meryem and B. EL Ouahidi, "Hybrid intrusion detection system using machine learning," Netw. Secur., vol. 2020, no. 5, pp. 8–19, 2020, doi: 10.1016/S1353-4858(20)30056-8.

[26] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, "Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier."

[27] A. Iqbal and S. Aftab, "A Feed-Forward and Pattern Recognition ANN Model for Network Intrusion Detection," no. April, 2019, doi: 10.5815/ijcnis.2019.04.03.

[28] Y. V. Kumar and K. Kamatchi, "Anomaly Based Network Intrusion Detection Using Ensemble Machine Learning Technique," no. 4, 2020.

[29] P. Maniriho, "Detecting Intrusions in Computer Network Traffic with Machine Learning Approaches Detecting Intrusions in Computer Network Traffic with Machine Learning Approaches," no. April, 2020, doi: 10.22266/ijies2020.0630.39.

[30] P. Raviteja et al., "Implementation Of Machine Learning Algorithms For Detection Of Network Intrusion," vol. 8, no. 2, pp. 163–169, 2020.

[31] S. Rajagopal, P. P. Kundapur, and K. S. Hareesha, "A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets," vol. 2020, 2020.

[32] T. C. Ravekar, "A Study of Intrusion Detection System using Advanced Genetic Algorithm," vol. 3, no. 11, pp. 7–12, 2016.

[33] M. Mazini, B. Shirazi, and I. Mahdavi, "Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms," J. King Saud Univ. - Comput. Inf. Sci., vol. 31, no. 4, pp. 541–553, 2019, doi: 10.1016/j.jksuci.2018.03.011.

[34] P. A. Diaz-Gomez and D. F. Hougen, "Improved off-line intrusion detection using a Genetic Algorithm," ICEIS 2005 - Proc. 7th Int. Conf. Enterp. Inf. Syst., no. Section 2, pp. 66–73, 2005, doi: 10.5220/0002553100660073.

[35] M. Sazzadul Hoque, "An Implementation of Intrusion Detection System Using Genetic Algorithm," Int. J. Netw. Secur. Its Appl., vol. 4, no. 2, pp. 109–120, 2012, doi: 10.5121/ijnsa.2012.4208.

[36] A. Yulianto, P. Sukarno, and N. A. Suwastika, "Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset," J. Phys. Conf. Ser., vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012018.