Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Intrusion Detection System: A Survey

Hamza Nachan
Dept. of Computer Science Engineering School of Engineering, MIT ADT University
Pune, India

Dristi Poddar
Dept. of Computer Science Engineering School of Engineering, MIT ADT University
Pune, India

Sambhaji Sarode
Dept. of Computer Science Engineering School of Engineering, MIT ADT University
Pune, India

Pratik Kumhar
Dept. of Computer Science Engineering School of Engineering, MIT ADT University
Pune, India

Simran Birla
Dept. of Computer Science Engineering School of Engineering,
MIT ADT University
Pune, India

***Abstract*—Intrusion Detection System is one of the most important tools in cybersecurity. It helps detect if an attack is occurring on our system and also helps prevent it. It can detect attacks on the system's network and also detect host attacks. IDS segregates signature and anomaly attacks. While signature attacks can be easily detected by comparing the signatures already stored in a database. Whereas anomaly attacks are observed by the divergence of the traffic of the system from the normal traffic of the system. These anomaly attacks are new threats that are not contained in the signature database. Due to the rise in technical advancements, there is also a sudden spike in cyberattacks. To defend against these threats, the IDS is an effective method, but the standard IDS is not very clever and powerful to keep the user from encountering new attacks on an everyday basis. In order to enhance the classifiers and algorithms, machine learning classifiers and algorithms can be used. These machine learning models are very helpful and can train models to distinguish between normal traffic and bad traffic. Via the assistance of machine learning, IDS can then recognize anomaly attacks and avoid them. With standard IDS, when detecting anomaly attacks, the false positive rate is high, which means that it is inaccurate, in order to minimize the false positive rate ML algorithms can be used. Also, anomaly detection is possible using ML in intrusion detection which will give a high accuracy of attacks being detected. Our proposed system included training a data set with machine learning algorithms so that it would detect anomaly attacks and also segregate different types of attacks.**

Keywords*—Cyber Security, Machine Learning, Classifiers, Intrusion Detection System.

I. INTRODUCTION

Cybersecurity is becoming important for everybody. With the increase in cyber-attacks, our data is at risk, our systems are vulnerable to these attacks and our privacy does not exist. To protect our systems and ourselves from such attacks, there are many cybersecurity tools available. Among these tools, the Intrusion Detection System is an important tool that protects our data and tells us if someone is sniffing through our system. Intrusion Detection System is a tool that is used by people working in big companies to people who have personal computers. IDS can be defined as a tool or application software or a device that detects and reports malicious activities in a system. IDS works with the help of a SIEM (security information and event management). The intrusion in our systems and violation of the rules specified in our systems is typically collected and sent to the administrator or collected centrally using SIEM. The purpose of an IDS is to collect information and send it to an authority that will fire a signal. That signal will then indicate to the user that there is an intrusion in the system. Intrusion detection system identifies the intrusion and malicious activities by analyzing traffic patterns of the system. The overall function of an intrusion detection system is to analyze the traffic patterns, look for anything that is suspicious, collect the data regarding any malicious activities and then fire an alarm for the user to know that there is an intrusion in the system.

Intrusion Detection System is of types such as NIDS, IDS, PIDS, APIDS, and Hybrid.

NIDS (Network intrusion detection system): This IDS is placed at strategic points in a network. NIDS analyzes the traffic of all the systems present in one subnet. Depending on the rules that are configured by the network administrator or user, NIDS analyzes the traffic pattern.

NIDS also analyzes traffic patterns based on the signatures that are present in a database. Placing an IDS before a firewall is an example of Network IDS.

HIDS (Host intrusion detection system): HIDS analyzes and protects only the host to which it is attached to. It analyzes the packets incoming and outgoing packets and then alerts the administrator if there is any malicious activity found on them.

PIDS (Protocol based Intrusion detection system): It analyzes the protocol that is used for sharing data between the device and the server. It resides on the frontend of the server and consists of an agent.

APIDS (Application Protocol based Intrusion detection system): It analyzes the protocols that are related to a specific application. For e.g., If you want to analyze the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

transaction between the database and the system is secure then APIDS will analyze the SQL protocol.
Hybrid Intrusion Detection System: This IDS is made using two or more types of IDS. Here, the network data along with system data is combined and analyzed, which gives a complete view of the system. This is more effective than any other IDS. The directions through which the framework of intrusion detection identifies intrusions or malicious detection is based on signature and anomaly-based behaviors.
In Signature based-detection: The detection is just like the detection with the virus in anti-virus software which uses the database and matches signatures. In this, the signature can be bytes of the network traffic or malicious instructions that are known.
Anomaly based-detection: To identify new threats that are not present in the database, we use anomaly-based detection. For this, we use machine learning algorithms for analyzing the traffic.
In IDS, anomaly-based detection can be strengthened by using algorithms for machine learning. Machine learning can be used in order to build a prediction model to determine whether the incoming and outgoing packets have malicious code in them. It is possible to use multiple machine learning algorithms to identify an attack happening in the system. There are three types of Machine learning algorithms:

A. .Supervised Learning:
Here. the input data contains labels or results such as spam or not spam. The input data is called training data. The input data must have a class type or feature vector. Then the data is trained and it continues till the data achieves a certain level of accuracy.
B. Unsupervised Learning:
Here the data doesn't have any labels. The data is unknown. This is used to extract general rules. The data is deduced into structural elements and then the model is prepared. Problems like clustering, self organizing maps, associate learning are dealt with unsupervised learning. Though, clustering is told to avoid unsupervised learning because if the center is wrong the whole data will be wrong.
C. Semi-supervised Learning:
Some data is labeled and some are not labeled. In case of loads of incomplete data this can be applied. A prediction problem is expected, the model must evaluate the structures and make predictions in order to organize the data

II. LITERATURE SURVEY

[1] This paper states the use of supervised machine algorithms for anomaly detection. The algorithms used are SVM, J decision tree and decision table and naive Bayes. It is done to find out the anomaly in the normal traffic. The model has some data and with machine learning algorithms that data is trained into a prediction model. Then after that some data is taken to test that the predicted model is correct. The input taken for the algorithms must be understood by the algorithms. So, we take input which contains some examples, instances or records. These records have special attributes to differentiate them called as feature vector. Some instances have class types. The algorithm which requires each input to have feature vector or class type, it is known as supervised algorithm. Which require some are semi-supervised .Which requires none and then trains with the help of pattern or similarities in the input is known as unsupervised. The dataset used by the author is KDD dataset. Here 492021 instances are used in the training package, while 311029 instances are for checking the dataset. In the model proposed in this paper FPR for DOS attacks are very less, J48 has the highest accuracy, also SVM has a high accuracy with 0% FPR in u2r attacks.

[2] In this manuscript, the authors have developed a model that utilizes supervised machine algorithms for anomaly detection. The algorithms used in this paper are C4.5 decision tree and Naive Bayes classifier. It is shown that c4.5 has good accuracy while making decisions in this paper. The model proposed in this paper has mainly three parts which is pre-processing, processing and post processing
Pre-processing: With the help of C4.5 decisions are taken the training model according to the input given. Train the selected machine learning algorithm. Based on this training dataset classification of the records into the normal or abnormal activity is done. Two parameters are taken out X and Y. Where x is protocol used: TCP, IP, UDP and Y is the name of attack: U2R, probe, DOS etc.
Processing: Naïve Bayes classification algorithm requires k $(k-1)/2$ two-class Naïve Bayes classification techniques where each would be trained on data from two classes. Binary tree Naive Bayes is being used for IDS. Based on the characteristics of different intrusion detection types, the authors have developed four Naïve Bayes classifiers to identify the five states: normal state (NS) and the four intrusion state or attack state Denial of Service (DoS), Remote to User (R2L), User to Root (U2R), and Probing.
Post processing: Minimize False positive rates N voting algorithm is used; the authors grouped the attacks according to IP addresses both destination as well as source. After grouping checking is done for attack or normal if 3 out 5 are attacks then the IP addresses are termed as malicious. This model produced an accuracy up to 99.48%.

[3] It discusses SDN (software development network) based NIDS. The paper gives an overview on how different machine learning based algorithms are used in NIDS. It segregates the algorithms in 3 types:
Supervised: where data is labeled and
algorithms such as random forest and SVM is used for classification of data
Unsupervised: here the data is made in a structural element and then the unknown data is predicted. Algorithms like self-organizing maps and clustering are used. But clustering is told to avoid as it can give inaccuracies if the center is not proper.
Semi -supervised: If a lot of data is missing then these algorithms are used: Spectral Graph Transducer and Gaussian Fields approach, used to detect malicious behavior and one semi regulated clustering method MPCK-means used to boost the efficiency of the detection system. The approaches used here in this paper are as follows:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

1.Decision tree algorithm for feature selection and random forest for classification
2.PCA for feature selection and SVM as classifier. 3.Soft-max regression as classifier and spare encoder for further reduction and has accuracy of about 92%. According to this model the decision tree holds high accuracy and low false alarm.
Further in this paper, what is SDN and how it is used is described. SDN is used as a method to help strengthen security which can also be used with the help of a variety of tools. Most popular tool being open flow. SDN can be used in another simulation tools like ns2, ns3, omnet++. The challenges presented in this paper were, that SDN can be prone to attacks like DOS,DDOS etc. Most of the papers are on KDD dataset which is old, though contains a lot of data the data is outdated and new dataset must be used. Determination of the optimum number of model parameters was also a challenge

[4] In this review paper about Machine Learning Approach to IDS the authors start by explaining how IDS works: So, in signature packets detection, the ids collect the information or packets, preprocesses the data and then checks the signature and then matches the signature with the already available database of signature.
Then the working of anomaly detection is explained: it preprocesses the data and then does pattern building. After that outlier detection is done and then it generates reports Future the Supervised, unsupervised and semi-supervised techniques are explained
Then data reduction techniques are explained which is important. If not done it will not produce an outcome with good accuracy also will take up a lot of space.
It also defines feature classification/selection and feature extraction algorithms
Feature selection: helps in determining the subset of features that should be selected in order to improve overall outcome and also generate few errors. It reduces the computation time and storage utilization. E.g.: PCA, IG, GA
In the paper it is explained that feature selection is done by 2 strategies:
First being the wrapper technique, a classifier is used as a black box for optimum assessment of features. Such methods achieve great speculation, yet sometimes endure high dimensionality due to the computational expense of preparing the classifier.
Second being Filter methods which don't utilize any classifier for feature evaluation and are relatively powerful against over-fitting, yet it utilizes autonomous estimation techniques, for example, distance measures, consistency measures, and correlation measures.
Feature extraction: this is done by having the dataset into rows and columns where rows are samples and columns are features. Thus, the extraction helps in reducing the dimensions of the data without altering necessary data. Numerous feature extraction methods are available in the field. For example, self-organizing maps, principal component analysis etc. Then further Clustering is explained. In clustering data samples are grouped into sets of data where data samples are identical in one way or the other in each package. It provides with a formula to calculate the accuracy, false detection rate of the model True-Positive (TP): Correctly classify an anomalous sample as attack.
True-Negative (TN): Correctly classify a non-attack sample as an ordinary instance.
False-Positive (FP): Incorrectly identify a usual sample as an unusual case.
False-Negative (FN): Incorrectly label an attack experiment as an ordinary case.
Accuracy: TP+TN/TP+TN+FP+FN
Attack detection rate: TP/TP+FN
False alarm rate: FP/FP+TN
It then gives a comprehensive review on using different feature classifiers and extraction algorithms on the same dataset kDD and find out who had the most accuracy.

[5] In this paper various machine learning algorithms like Bayesian Network, Naive Bayes classifier, Decision Tree, Random Decision Forest, Random Tree, Decision Table, and Artificial Neural Network, have been implemented to detect intrusions and test the effectiveness of various experiments on cyber-security datasets having several categories of cyber-attacks had been done and the effectiveness of the performance metrics, precision, recall, f1-score, and accuracy has been evaluated. Out of the various machine learning classification algorithms employed by the authors Random Decision Forest algorithm had the best result in terms of accuracy, precision, recall and fscore. In this paper the author has implemented a data driven intrusion detection model which incorporates several steps such as dataset exploration, data processing, and machine learning-based security modeling.

[6] In this paper, the authors carried out an investigation and analysis of various ML techniques for finding the cause of problems associated with various ML techniques in detecting intrusive activities. In this paper grouping and mapping of attack features corresponding to each attack is given. In this paper all Issues related to detecting Low-frequency attacks using the data collection of network attacks are often mentioned and feasible approaches are mentioned. are suggested for improvement. The authors have analyzed and compared machine learning techniques in terms of their detection capability for detecting the various categories of attacks. In this model the comparison has been carried out with single classifier approaches and multiple classifier approaches. The influence of a classifier with another classifier is not only analyzed but also the influence of a feature subset with the classifier is analyzed. It has also been shown if an optimal feature set is sufficient for analyzing the
[7] Since Big Data is the data that is difficult to store, handle, and analyze using conventional database and software techniques, this paper introduced the Spark-Chi-SVM intrusion detection model that can deal with Big Data. They used ChiSqSelector for feature selection in this model and developed an intrusion detection model on Apache Spark Big Data classifier by using support vector machine (SVM) A contrast between the Chi-SVM classifier and the Chi-Logistic Regression classifier was implemented in the experiment. The experiment results showed that the model has high efficiency, decreases training time and is useful for Large Data. First to convert the categorical data to numerical data, a preprocessing method is used and then the dataset is standardized for the purpose of improving classification

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

performance. Secondly, in order to further increase classification efficiency and decrease computation time, the ChiSqSelector approach is used to decrease the dimensionality of the dataset. Thirdly, the classification of data uses SVM More precisely, in order to correct the optimization, they use SVMWithSGD. In addition, on the Apache Spark They introduce a comparison between SVM and Logistic Regression classifiers using a Big Data framework based on area under curve (AUROC), area under precision-recall curve (AUPR), and time metrics. They used KDD99 to train the model and to test it. The outcome of the experiment, based on the outcome of the method of data standardization, training time and prediction time, showed that the model has high performance and reduces the false positive rate. The Chi-SVM is the best classifier, based on the contrast between the Spark-Chi-SVM model and other research approaches based on training and time forecasting. Although to its disadvantage, the model is not a multi-class model that would detect types of attack.

[8] In this paper, the emphasis is on detecting real-time network-based intrusion where the incoming network information is recorded online and the result of detection is posted instantaneously or within a fraction of a minute, so that the network administrator is alerted and can stop the ongoing attack. The technique may also be implemented as a detection dependent on the host. Using a misuse detection technique, they developed an intrusion detection system (IDS). By comparison, the approach to anomaly detection can only distinguish between normal behavior and abnormal/attack activity. While there are several potential network data features that could serve as an input to an IDS, they considered only 12 network traffic data features derived from data packet headers. They demonstrate that these 12 characteristics are successful in defining standard network activity and dividing main attack activities into two types, namely Port Scanning (PS or probing) and Denial of Service (DoS) (DoS). By applying well-known machine learning algorithms, they have also created an IDS. This offered a basic but successful technique for detecting intrusion, so that it can be easily followed by all. To do that, several existing machine learning algorithms, such as Decision Tree, Ripper Rule, Back-Propagation Neural Network, Radial Basis Function Neural Network, Bayesian Network and Naive Bayesian, were considered and compared to classify incoming network data. The experimental results showed that the Decision Tree was superior in terms of detection precision to the other methods. The Decision Tree approach for the Real-Time Intrusion Detection System (RT-IDS) was therefore further developed, where data from the input network is collected on-line in a real-world environment. In addition, an optional post-processing technique for the real-time intrusion detection system was introduced in order to reduce and thereby increase the intrusion detection false alarm rate. The RT-IDS can achieve a total detection rate (TDR) of greater than 99 percent with a very low false alarm rate (less than 0.5 percent). When collecting network traffic at full load (100 Mbps), our RT-IDS uses less than 25% of CPU power and just 94.5 MB of memory. Our RT-IDS can detect malicious data packets in as little as 2-3 seconds, which is enough time to warn the computer user or network administrator. Our RT-IDS can detect malicious data packets in as little as 2-3 seconds, which is enough to signal the computer user or system administrators to system security. Ultimately, the results of the post-processing approach showed that the detection rates of regular network operations, DoS attacks and Probe attacks can be enhanced thus reducing false detection rates for all record real-time detection in this paper is powerful and more accurate.

[9] This model was made using NIDS especially for mobile. The main focus of this was detection of malware. The datasets used by the authors here are KDD99 and DARPA1.For traffic generation real time traffic was used with Wireshark. In Pre-processing, feature extraction feature selection is carried out (not auto because the author thought
that not all features would be considered) and then labelling it is done malicious or normal. Machine learning classifiers ML algorithms that were used in this research are J48, Random Forest, RIDOR, JRIP and PART. With the help of TP FP FN TN, the accuracy was determined. Various experiments were carried out on several datasets. On Dataset 1, Cross Validation on Dataset 1 was done, this experiment was to test the accuracy of the classifiers of machine learning. Cross Validation is one of the validation techniques
commonly used to test the efficacy of a classifier for machine learning. A 10-fold cross validation protocol was used in this experiment. As a training set/validation set for the classifiers, Dataset 1 was used. On Dataset 2 percentage split was done some part was for training rest to be utilized for testing purposes. Dataset 3 and 4 random forests were used, features were 5 and 10. Dataset 5 was used for detecting unknowns, dataset 1(train) and dataset2 (test). The model developed in this study was able to detect about 99.6 percent of malicious traffic with the TPR and it also performed well with an unspecified data with an accuracy of 97.5%

[10] In this paper the authors use supervised machine learning algorithms for anomaly detection. The algorithm used in this model is Random Forest in which k features are selected. Then these k features are split to get node d and then again split till you will reach the first node. Continuous repetition of this process is done for the formation of random forests. In this model the authors have also performed PCA which is essential for classification purposes. This method takes all the input as the dataset, which has a high number of attributes so the dimension of the dataset is very high. The authors have implemented this method as this method reduces the size of the dataset by taking the data points on the same axis. Then attributes such as eigenvector and value are evaluated which are then utilized to form a matrix. Further using this matrix, the principal component is obtained. This model as compared to SVM and Naive Bayes gives an accuracy of about 90%.

[11] In this paper the authors have proposed the usage of a supervised learning method called multi-layer perceptron (MLP) neural network instead of the traditional machine learning algorithms and neural network-based techniques like Decision Tree, Random Forest, or Support Vector Machine (SVM) for the purpose of intrusion detection. The dataset used by the authors for building the model is a recent

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

one which is CICIDS2017 which contains intrusion attacks and traffic that are representative of current network usage. The model proposed in this system multi-layer perceptron is a fully connected, feed-forward neural network classifier. The classifier has 15 outputs for the 14 types of attacks and the traffic. For the purpose of implementation and training of the model, python and TensorFlow as a deep learning framework has been utilized by the authors. Once training was done several key information such as True Positive (TP) which is the number of attacks correctly predicted as attacks, True Negative(TN) which is the number of normal instances classified as normal traffic while False Positive(FP),False Negative (FN) which is the number of normal instances classified as attacks and the number of attacks predicted as normal traffic was extracted from a confusion matrix using which other parameters such as TN rate, FP rate, precision, recall, accuracy, f1score, and MCC (Matthews Correlation Coefficient) were evaluated . Using this methodology, the authors were able to achieve better performances than traditional machine learning techniques with a detection of intrusion accuracy above 99% and a low false positive rate kept below 0.7%.

[12] In this paper the authors introduced the Spark-Chi SVM model for intrusion detection. In their proposed model the authors have used ChiSqSelector for feature selection, and built an intrusion detection model by using a support vector machine (SVM) classifier on Apache Spark Big Data platform. The dataset used by the authors to train and test the model is KDD99. In the methodology proposed in this system the first dataset is loaded and then exported into Resilient Distributed Datasets (RDD) and Dataframe in Apache Spark. Which is followed by data preprocessing, in this step preparation of data is done in which the categorical data in the dataset is converted to numerical data as the SVM algorithm deals with numerical data only. After this the authors proceed to feature selection which is done by ChiSqSelector and SVM combined. The feature selection that is applied to the dataset numTopFeatures method. Then the training on the training dataset, lastly followed by testing and evaluation is done with KDD dataset. Using this model the authors could process and analyze data with high speed; they achieved higher performance, reduced training time and were efficient for Big Data with an accuracy of 90-95%.
[13] In this paper, the authors have proposed a new hybrid intrusion detection method that hierarchically integrates a misuse detection model and an anomaly detection model in a decomposition structure which comprises the use of supervised machine algorithms such as 1 class SVM and j48 decision tree algorithms for anomaly detection. The dataset used here was the NSL-KDD data set, which is a modified version of the well-known KDD Cup 99 data set. The authors have implemented Decision tree (DT) on training models as it has comparatively low false positive rates. Once decomposed by DT then 1 class SVM is used on normal training datasets Here DT divides the dataset and then 1 class SVM is applied to each of the decomposed regions of the dataset. Throughout the integration, the anomaly detection model indirectly uses the known attack information to enhance its ability when building profiles of normal behavior. This methodology used by the authors to build their model demonstrated that the hybrid intrusion detection method could improve the IDS in terms of detection performance for unknown attacks and detection speed as the proposed system significantly reduced the high time complexity of the training and testing processes.

[14] The clustering approach under unsupervised machine algorithms for anomaly detection based on K-means for IDS has been well developed, but when directly using it in a big data environment it suffers from inappropriateness. Also, the efficiency of data clustering is low. Also, differ from the classification, as there is no unified evaluation indicator for clustering issues. So, in this paper, the authors in order to address these issues have proposed a clustering method for IDS based on Mini Batch K-means also combined with principal component analysis. The dataset used is the KDDCUP99 dataset. The Kmeans is used here for performing clustering of the records in the dataset. Also, the main idea of the PCA is to find a direction vector in order to project the original dataset onto it. The PCA method used reduces the dimension so as to improve the clustering efficiency. The Mini Batch Kmeans method is used for the clustering of the processed dataset; the evaluation of clustering results is based on the degree of in-cluster density and the degree of inter-cluster discretization. Compared with Kmeans (KM), the results of the methodology used here which is Kmeans with PCA (PKM), as well as Mini Batch Kmeans (MBKM), the results showed that their proposed model is more effective and efficient.

[15] In this paper, the authors used decision tree and SVM algorithms for building a machine learning model. The decision tree algorithm that is used is based on the divide and conquer strategy and it recursively divides the data till the conditions are satisfied. The other algorithm used is one-class SVM. It is used for outlier detection. The proposed systems use both the algorithm for developing an IDS with a low false-positive rate. The decision tree is used first on the data and then the data is segregated into smaller parts. Once the data is segregated into smaller parts or sections then one-class SVM is applied. SVM could be applied to the whole training data set but it is not advised as SVM is sensitive to that data and then in turn it decreases the accuracy. So, to avoid this problem SVM is applied to the subsets of the training data.

Once the DT is applied the known attacks are separated from the normal attack, present in the training data. After DT is applied then the training data is divided into normal data and then one class SVM is applied for finding out anomalies in the data. The NSL-KDD dataset is used for checking the accuracy of the model created. The parameter on which decision tree decision limit is dependent is flexible. The detection rate of DT is 99.1% for known attacks and 30.5% for unknown attacks.

[16] Here the authors proposed the model into 3 parts. The classification module, behavior pattern module, and final classifier module. In the classification module, the KDD cup 1999 is used and then the decision tree is used for separating into DOS, PROBE, and OTHER categories. The decision tree is then trained for the data and then the decision tree is generated. Then comes the next step behavior partition module, here the EFHCAM algorithm is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

used which is a clustering algorithm. The K value in the clustering is not fixed in the beginning. The OTHER categories go through the algorithm and then it gets segregated into Normal and attack data. After that it goes through the final classifier module where the attack data further gets labeled. It gets converted into a decision tree again with U2R and R2L attacks. This module is made up of classification as well as the clustering module. The training data is used and then an enhanced C4.5 decision tree is used on the data. When the Enhanced fast heuristic clustering is used the unsupervised data is then divided into normal and attack data. The data is usually separated by what data is closer to the center of the cluster. The dataset used is KDD cup 99 with 41 different attributes. Different categories have different misuse and anomaly detection rate. DOS has 99.19% misuse detection and 83.59% anomaly detection. PROBE has 99.71% misuse detection and 78.60% anomaly detection. U2R has 66.67% misuse detection and 58.06% anomaly detection. R2L has 89.50% misuse detection and 28.53% anomaly detection.

[17] This paper develops a hybrid intrusion detection system with a machine learning algorithm. In this paper, the dataset goes through ensemble feature selection classifier, fuzzy belief K-NN classification algorithm and data mining classifier. First the dataset goes through the ensemble feature selection classifier and each different classifier gives different output. Subset of features are done and then passed through the base classifier. The k-NN algorithm is used as an intrusion detection method. The fuzzy belief k-NN classification algorithm is used for assigning the training data in classes such as attack or normal data. After these steps are done, a data mining classifier is used. This is done to reduce the false positive rate and in turn increase the accuracy. It provides a strategy to find useful information from the large clusters of information and then decides if the data is normal or attack data. Here the C4.5 decision tree algorithm is used to extract patterns and data. The data mining technique is also used for extracting network data of the user's normal behavior by training and testing data. The KDD-99 dataset is used. It has 92.30% detection rate and 3.13% false-positive rates in detecting U2R attacks. In detecting R2L attacks, 71.66% averaged detection rates and 3.15% averaged false positive rates are obtained.

[18] In this paper probability of the occurrence of data is used for determining if the data is normal data or attack data. Here naive bayes and decision tree algorithms are used in combination with each other for determining normal and attack data. The author starts by cleaning out the data set. It removes extra or redundant data in the dataset. After that it groups or discretize the continuous attributes containing continuous values so that when it is done, proper interval sets are determined. Here authors focus greatly on proper attribute selection technique. Then finding the prior probabilities and conditional probabilities of the training data is done. Then all the data is segregated into different groups based on these probabilities. Once done each class value of the data in the dataset is then updated with Maximum Likelihood of posterior probability. This procedure is repeated again with updated class values. After this best attribute is selected from the dataset. Then the dataset is divided into subsets of data based on the attribute and then again, the procedure of calculating probabilities is done. Thus, the repetition of procedure is done for acquiring a decision tree and the naive bayes is used for the decision making process. The attacks are classified into Probe, DOS, R2L, U2R. The attacks are detected with 99% accuracy using the proposed algorithm.

[19] The authors used an improved Genetic Algorithm and Levenberg- Marquard backpropagation algorithm to present Hybrid Neural Network Algorithm. Here Genetic algorithm used to find global optimal point combined with fast local searching Levenberg- Marquard backpropagation algorithm to find global optimal point. Neural network classifier and multiple classifier combination technique is used in designing of IDS. This classifier helps in improving and optimizing the overall working of the intrusion system. KDD CUP 99 data set was used for training and testing. DOS, Probe, Remote to Local (R2L), User to Root (U2R) were some of the sub classifier sets in which only one type of attack sample with normal sample was used. Hybrid neural network algorithm had improved detection rate and less error rate compared to improved Genetic algorithm and Levenberg- Marquard backpropagation algorithm when tested individually against hybrid neural network algorithm. Thus, Hybrid neural network algorithm enhances the performance of Intrusion Detection System.

[20] The authors used KDD data set of Intrusion detection with machine learning classifiers like J48, Random Tree, MLP, Random Forest, Naive Bayes, Bayes Network and Decision Table. Machine learning tool used was Waikato Environment for knowledge analysis (WEKA). A deep understanding of the dataset was provided through statistical measurement to extract impartial experiments. Twenty-one types of different attacks were grouped into four groups (DOS, R2L, U2R and Probe). Where 79% dataset was of DOS, rest other types of attacks were 2% and 9% was normal traffic. The performance metrics of each classifier was computed considering True positive, True negative, False positive and False positive. Results were, the Decision Table classifier scored the lowest false negative value, the Bayes Network classifier scored the highest value in detecting normal traffic packets, the Random Forest classifier scored the highest accuracy rate with smallest root mean square error and false positive rate, and the Multi Layer Perceptron (MLP) classifier built its training model in acceptable period of time. In [21-27], the various different techniques are discussed for different sensor data sets in the real time environment. These methods are useful for gathering and processing the information from time bound application viewpoint.

III. METHODOLOGY

Machine learning is a data analysis tool that automates the creation of analytical models. It is a branch of artificial intelligence focused on the premise that, with minimal human interaction, systems can learn from data, recognize trends and make decisions. When adequate training data is available, machine learning-based IDSs can achieve satisfactory detection levels, and machine learning models have sufficient generalizability to detect attack variants and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

new attacks. The aim of this study is to identify and summarize the IDSs proposed to date based on machine learning, to abstract the key ideas of applying machine learning to security domain issues, and to analyze the present problems and possible changes.

A. Anomaly-based intrusion detection

AIDS has different advantages. They have the opportunity to discover internal malicious operations. If an attacker begins making purchases that are unidentified in the normal user behavior in a compromised account, it triggers an alert. Then, since the framework is built from personalized profiles, it is very difficult for a cybercriminal to know what is a common user activity without triggering an alarm. AIDS techniques can be divided into three major groups: knowledge-based statistics and machine learning-based statistics. In order to detect unknown threats, they were added. Complex pattern-matching abilities are obtained from training data through machine-learning methods. To construct a model simulating normal operation, IDS uses machine learning and then compares new behavior with
the current model. It uses data at various time intervals to form a baseline usage of the networks. Machine Learning Algorithms can be categorized generally into: Supervised machine learning algorithms, Unsupervised machine learning algorithm and Semi supervised machine learning algorithms.

1. Supervised machine learning algorithms

A labeled dataset is added to decrease the false positives and construct a supervised machine learning model by teaching it the difference between a usual and an attack packet in the network. The supervised model can expertly control the recognized attacks and can identify variants of those attacks as well. The most common role in supervised learning is classification; but it is costly and time-consuming to manually mark data. The lack of adequate labeled data therefore forms the key constraint for supervised learning. Typically, a supervised learning approach consists of two steps, i.e. Testing and preparation. Appropriate features and classes are defined in the training stage and then the algorithm learns from these samples of data. A supervised learning technique is then used to train a classifier using the training data for selected characteristics to learn the inherent relationship between the input data and the labeled output value. Labelled data and the desired output are given to the learning algorithm. Pictures of flowers labeled "flower" will, for example, help the algorithm define the rules for classifying flower images. Bayes Network, Random Forest, Random Tree, MLP, Decision Trees, are typical supervised algorithms. To construct a classification model, each of these techniques use a learning process. However, not only should a suitable classification method manage the training details, but it should also correctly classify the class of records it has never seen before.

a. Support Vector Machines (SVM)

Support Vector Machine (SVM) is a discriminatory classifier identified by a splitting hyperplane that falls under the supervised method of learning in which different types of data are trained from different subjects. The strategy in SVMs is to find a hyperplane of max-margin separation in the space of the n-dimension function. SVMs are well known for their ability to generalize and are primarily useful when the number of attributes is high and the number of data points is limited. SVMs, however, are vulnerable to noise near the hyperplane. Many features are redundant or less influential in dividing data points into correct classes in IDS datasets. Therefore, during SVM preparation, feature selection should be considered. SVM produces hyperplanes or multiple hyperplanes in a high-dimensional vacuum. The hyperplane, which divides the given data optimally into different classes with the main partition, is considered as the best hyperplane. A non-linear classifier applies various kernel functions to determine the margins between hyperplanes.

b. Naïve Bayes

The Naïve Bayes algorithm is based on the conditional possibility and the attribute independence hypothesis. They are able to predict the probability that a particular class would match the given model. The Naïve Bayes classifier calculates the conditional probabilities for various classes for any sample. This principle is known as class conditional independence.

Due to its ease of use and calculation performance, the Naïve Bayes classification model is one of the most prevalent IDS models. The method, however, does not work well if the presumption of conditional independence is not true.

Fig : Classification of AIDS

c. Genetic algorithms (GA)

The machine uses genetic algorithms (GA) to execute natural selection and evolution. This idea is derived from the 'adaptive survival of natural species.' The algorithm begins by creating a large population of candidate programs

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

randomly. Some type of fitness measure is used to determine the health of each person in a community. To create basic rules for network traffic, GA was used. Every rule is represented by a genome and a number of random rules are the primary population of genomes. Each genome consists of various genes that correspond to properties such as source of IP, destination of IP, source of port, destination of port and form of 1 protocol.

d. Artificial Neural Network (ANN)

The design principle of an ANN is to simulate the functioning of human brains. An ANN includes a layer of data, multiple hidden layers, and an output layer. The units in the neighboring layers are completely connected. ANN is one of the names of the Machine-learning techniques that have been implemented most widely and have been shown to be efficient in detecting various malware. The backpropagation (BP) algorithm is the most common learning technique used for supervised learning. First of all, at the beginning of training, random weights are given. The algorithm then performs weight tuning to define whatever hidden unit representation is most successful at minimizing the misclassification error. However, for ANN-based IDS, it is still important to improve detection accuracy, particularly for less frequent attacks, and detection accuracy. Compared to that of more frequent attacks, the training dataset for less frequent attacks is limited and this makes it difficult for the ANN to correctly learn the properties of these attacks.

e. Fuzzy logic

The word fuzzy means items that are not quite specific or ambiguous. Fuzzy logic is based on the idea of the sometimes-occurring fuzzy phenomena in the real world. Fuzzy logic is a form of reasoning that parallels human reasoning. The Fuzzy Logic approach is based on the degrees of uncertainty rather than the typical true or false Boolean logic on which contemporary PCs are created, imitating the way of decision-making in humans that includes all intermediate possibilities between digital values 1 and 0. It therefore offers an easy way to arrive at a final conclusion based on uncertain, ambiguous, noisy, incorrect or incomplete input data. Fuzzy logic takes truth degrees on the vagueness model as a mathematical basis, while probability is a mathematical model of ignorance. In various fields such as control system engineering, image processing, power engineering, industrial automation, robotics, consumer electronics, and optimization, Fuzzy logic has been successfully used. Fuzzy logic is a good classifier for IDS issues because vagueness is involved in the protection itself, and the borderline between usual and abnormal states is not well understood. Furthermore, the issue of intrusion detection involves different numeric features in the data obtained and many statistical metrics extracted.

f. Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is a Markov statistical model in which the modeling system is believed to be a Markov mechanism with unnoticed states. Hidden Markov models are known for their applications in thermodynamics, statistical mechanics, economics, signal processing, information theory, pattern recognition, such as speech, handwriting, gesture, part - of - speech tagging, partial discharges and bioinformatics. HMMs have been shown to be very useful for detecting a problem with a multi-step attack. In these attacks, within each stage, an attacker can use a different set of acts to mask and throw off conventional IDS systems that rely on the detection of attack fingerprints. These engines can become useless in the face of constantly changing attacks. Past authors applied HMMs to system call data and concluded that when mapping application system calls to a number of states, the best performance was achieved which means Markov method achieved. When trained on the same datasets, they perform better than neural networks. Just by drawing an intuitive image, HMM provides a conceptual toolkit for constructing models. They are at the center of a wide range of initiatives, including gene finding, profile searches, multiple synchronization of sequences and identification of regulatory sites.

g. Decision Trees

A decision tree is a decision support tool that uses a tree-like decision model and its potential implications, including outcomes of chance events, cost of resources, and utility. It is one way of showing an algorithm that only includes statements of conditional control. The algorithm of the decision tree will automatically remove irrelevant and redundant characteristics. Decision tree classification is a two-step method in machine learning, learning step and prediction step. Feature collection, tree generation, and tree pruning are included in the learning process. The algorithm selects the most relevant characteristics individually when training a decision tree model and generates child nodes from the root node. Unlike other supervised learning algorithms, the decision tree algorithm can also be used to solve regression and classification problems. In Decision Trees, we start from the root of the tree to predict a class label for a record. We compare the values of a root attribute with the attributes of the record. We follow the branch corresponding to that value on the basis of comparison, and leap to the next node. Decision trees use many algorithms to determine whether to divide a node into two or more sub nodes. Sub-node formation increases the homogeneity of the resulting sub-nodes. In other words, with respect to the target variable, we can assume that the purity of the node increases. The decision tree splits the nodes into all
available variables and then chooses the split that results in most of the sub-nodes being homogeneous.

h. K-Nearest Neighbors (KNN) classifier

K-nearest neighbor (k-NN) is one of the easiest and most traditional non-parametric approaches to classify samples used in machine learning that can be used to solve problems with classification and regression. The KNN algorithm assumes close proximity to similar objects. It calculates the approximate distances on the input vectors between different points, and then assigns the unlabeled point to its K-nearest neighbors' class. The k parameter significantly affects the efficiency of KNN models. The lower the k, the more complicated the model is and the greater the chance of overfitting. On the other hand, the larger k, the simpler the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

model is and the poorer the fitting power does not include the model training level, but only searches for input vector examples and categorizes new instances. The KNN function is only locally approximated and all computation is delayed until the evaluation of the function. As this algorithm relies on distance for classification, it can significantly improve its accuracy by normalizing the training data. For all other classifiers, k-NN can be properly implemented as a benchmark because it provides reasonable classification efficiency in most IDSs. As the amount of data increases, KNN's key drawback of being slightly slower makes it an impractical alternative in environments where predictions need to be made quickly. There are, moreover, faster algorithms that can yield more detailed results for classification and regression. However, KNN can also be useful in solving problems that have solutions that rely on recognizing similar artifacts if you have ample computational power to quickly handle the data you are using to make predictions. An example of this is the use of the KNN algorithm, a KNN-search application, in recommending systems.

B. Unsupervised machine learning algorithms

Unsupervised learning is a form of machine learning that searches for previously undetected patterns in a data set without pre-existing labels and with a minimum of human supervision. These are called unsupervised learning since there are no correct answers and there is no teacher, unlike supervised learning above. To discover and present the interesting structure in the data, algorithms are left to their own designs. Unsupervised learning algorithms allow users to perform more complex processing tasks. The aim of unsupervised learning is to find the underlying dataset structure, group the data according to similarities, and display the dataset in a compact format. Unsupervised learning algorithms would learn the standard network pattern and, without any labeled dataset, can report anomalies. It can recognize new kinds of intrusions, but it is very vulnerable to false positive warnings. The information provided to the learning algorithm is unidentifiable, and the algorithm is asked to classify input data patterns. For instance, an e-commerce website's recommendation system where the learning algorithm discovers comparable products frequently purchased together. Unsupervised learning generates useful feature data from unlabeled data, making it much easier to collect training data. However, the output of Unsupervised learning methods for detection is typically lower than that of supervised learning methods.

a. K-means

The K-means techniques are among the most common clustering methods for the separation of 'n' data objects to 'k' clusters in which each data object in the nearest mean is selected-means clustering is a vector calculation technique that is initially used as a prototype for signal processing. The K-means clustering algorithm utilizes the iterative refining method to achieve the final output. It aims to make the data points as close as possible within the intra-cluster, while holding the clusters as separate. Data points are given to a cluster such that the sum of the squared distance between the data points and the centroid of the cluster (arithmetic mean of all data points in that cluster) is as low as possible. The lower the variance we have in the clusters, the more homogenous the data points are. The number of clusters K and the data settings are the algorithms entries. The data set is a function array for each data point. Initial calculations for K centroids begin with the algorithm, which can be either randomly generated or altered from the data set. It's a clustering strategy based on distances and doesn't have to measure the distances of all record combinations. It is a measure of similarity with a Euclidean metric. The consumer decides the number of clusters beforehand.

b. Hierarchical Clustering

Hierarchy clusters is an algorithm that groups related objects into groups called clusters, also known as Hierarchical Cluster Analysis. The endpoint is a collection of clusters where each cluster is different, and each cluster's artifacts are mostly identical. Hierarchical clustering begins by treating each observation as a separate cluster. Then the two steps are repeatedly executed: (1) the two nearest clusters are found and (2) the two most similar clusters are fused together. This iterative process goes on until all of the clusters are combined.
This technique of clustering is split into two types:
Agglomerative Hierarchical Clustering
Divisive Hierarchical Clustering

Agglomerative Hierarchical Clustering

The most common form of hierarchical clustering used for grouping objects into clusters on the basis of their similarity is Agglomerative Hierarchical Clustering. It's a "bottom-up" technique in which each data point is viewed as an independent cluster initially. Similar clusters merge with other clusters at each iteration, until a single cluster or K cluster is created.

Divisive Hierarchical Clustering

We consider all the data points as a single cluster in Divisive Hierarchical Clustering and we separate the data points from the cluster in each iteration, which are not identical. As an independent cluster, each data point that is isolated is considered. Ultimately, we will be left with the n clusters. Divisive is a top-down method of clustering where all results are allocated to a single cluster and the cluster is then split into two less related clusters. This approach to clustering is also precisely the opposite of agglomerative clustering. There is evidence that, under certain circumstances, divisive algorithms generate more precise hierarchies than agglomerative algorithms, but are conceptually more complex. Users need to define the optimal number of clusters as a termination condition in both agglomerative and divisive hierarchical clustering (when to stop merging).

c. Apriori algorithm

The Apriori algorithm is designed to work on databases that contain transactions and uses frequent item sets to generate association rules. With the help of this association law, it determines how strongly or weakly two objects are

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

associated. This algorithm effectively measures the itemset associations by using a breadth-first search and a Hash Tree. It is an iterative method for identifying frequent item sets in a broad dataset. The Apriori algorithm was the first algorithm proposed for frequent object sets mining. It was later improved by R Agarwal and R Srikant and became known as Apriori. This algorithm employs two "join" and "prune" steps to reduce the search space. Identifying the most popular item groups. It can also be used for patients to find drug reactions in the healthcare sector. Frequent item sets are iteming whose support is greater than the upper bound or the minimum support defined by the user. This means that if the frequent item sets together are A & B, then the frequent itemset should also be A and B separately. A powerful algorithm that scans the database only once is the Apriori algorithm. It greatly decreases the size of the item sets providing a good result in the database. Data mining thus improves the decision-making process for customers and industries. The Priori Algorithm can be sluggish. The key drawback is the time taken to retain a large number of candidates sets with very frequent item sets, low minimum support or large item sets, i.e., for a large number of datasets it is not a successful approach.

C. Semi-supervised learning

This is an approach to machine learning that blends a large number of unlabeled data with a small amount of confidential data during planning. Among unsupervised learning and supervised learning, semi-supervised learning falls in between. There must be a relation to the underlying data distribution in order to allow some use of unlabeled data. The basic technique involved is that, first, using that, first, using an unsupervised learning algorithm, the programmer would cluster related data and then use the current labeled data to mark the rest of the unlabeled data. There is a common property among the typical use cases of such type of algorithm-the acquisition of unlabeled data is relatively inexpensive while it is very costly to mark the same data.

At least one of the following assumptions is used by semi supervised learning algorithms:

Consistency assumption: The algorithm assumes that the points closer to each other are more likely to have the same output label.

Cluster assumption: The data can be divided into different clusters and points in the same cluster are more likely to share an output label.

Manifold assumption: This hypothesis makes it possible to use distances and densities identified on a manifold.

IV. OBSERVATION TABLE

Table 1: This table shows the comparison between the different datasets

Dataset	Realistic Traffic	Label data	IoT traces	Zero day attacks	Full packet captured	Year
DARPA 98	+	+	-	-	+	1998
KDD CUP 99	+	+	-	-	+	199
CAIDA	+	-	-	-	-	2007
NSL-KDD	+	+	-	-	+	2009
ISCX 2012	+	+	-	-	+	2012
ADFA-WD	+	+	-	+	+	2014
ADFA-LD	+	+	-	+	+	2014
CICIDS2017	+	+	-	+	+	2017
Bot-IoT	+	+	+	+	+	2018

(+ = True, - = False)

Table 2: This table shows the different systems assessed and the datasets used by them.

	Systems Used	Dataset Used
[1]	SVM (Support Vector Machine), J Decision Tree and Decision Table and Naive Bayes.	KDD
[2]	C4.5 Decision Tree and Naive Bayes Classifier	KDDCUP99
[5]	Bayesian Network, Naive Bayes classifier, Decision Tree, Random Decision Forest, Random Tree, Decision Table, and Artificial Neural Network.	CICIDS, KDD
[7]	Apache Spark Big Data classifier by using support vector machine (SVM)	KDD99
[8]	Decision Tree, Ripper Rule, Back-Propagation Neural Network, Radial Basis Function Neural Network, Bayesian Network and Naive Bayesian	KDD99, CICIDS, DARPA

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

[9]	Network based Intrusion Detection Systems (NIDS) using J48 Decision Tree and Random Forest Algorithms	DARPA, KDD99
[10]	IDS by using the principal component analysis (PCA) and the random forest classification algorithm.	NSL-KDD
[11]	multi-layer perceptron (MLP) neural network	CICIDS
[12]	Support vector machine (SVM) classifier on Apache Spark Big Data platform.	KDD99
[13]	supervised machine algorithms such as 1 class SVM and j48 decision tree algorithms	NSL-KDD, KDDCUP99
[14]	unsupervised machine algorithms based on Mini Batch K-means combined with principal component analysis	KDDCUP99
[16]	Multiple Level Hybrid Classifier using Enhanced C4.5.	KDDCUP99
[17]	Fuzzy belief k-NN classification algorithm in combination with data mining technique	KDD99
[18]	Naive bayes and decision tree algorithms	KDD99
[19]	Improved Genetic Algorithm and Levenberg Marquard backpropagation algorithm	KDD CUP 99
[20]	J48, Random Tree, MLP, Random Forest, Naive Bayes, Bayes Network and Decision Table	KDD 99

V. CONCLUSION

It is possible to use an intrusion detection system to track the file system for changes. It is helpful in determining what improvements after an attack are made to the device. An intrusion detection system is used to identify many types of malicious actions that can undermine a computer system's protection and confidence. Both IDS and firewall are linked to network protection, but an IDS varies from a firewall as a firewall searches for intrusions outwardly to prevent them from occurring. Firewalls block access between networks to avoid interference and do not signal whether an attack is from inside the network. A suspected intrusion is described by an IDS until it has occurred and then signals an alarm. Cybercriminals commit several illegal acts using computers or other digital technologies. The criminal can use computer skills, human behavior intelligence, and a range of resources and services to accomplish his or her goal. Hacking, identity theft, online scams and fraud, developing and distributing malware, or attacks on computer systems and sites may be the types of crimes a cybercriminal can commit. The key aspect that makes a crime a cybercrime is that it is targeted at a computer or other devices and/or that the crime is committed using these technologies. In order to exchange malicious goods and services, such as hacking instruments and stolen data, cybercriminals are known to access the cybercriminal underground markets contained in the deep web. In such goods or services, cybercriminal underground markets are known to specialize. Cybercrime related laws continue to develop across different countries around the world. When it comes to identifying, arresting, prosecuting, and verifying cybercrimes, law enforcement authorities are still constantly questioned. Experts in cybersecurity say that cyber criminals are using increasingly brutal tactics to accomplish their targets, and as they continue to create new methods for cyber-attacks, the skill of attacks is expected to progress. A variety of different types of cyber criminals have been generated by the development of the global cybercrime network, which is primarily credited with the increased potential for financial rewards, many of which pose a major threat to governments and businesses.

In this paper, we presented, in depth, with their advantages and drawbacks, a survey of intrusion detection system methodologies, styles, and technologies. Several techniques of machine learning have been suggested to detect zero-day attacks. The Intrusion Detection System (IDS) helps to recognize the intrusion and misuse of computer systems by collecting and analyzing data. IDSs have historically been developed for wired devices and networks to detect intrusion and misuse. IDSs are developed more recently for use on wireless networks. These wireless IDSs can monitor and analyze the behavior of known attacks by users and devices, identify suspicious network activity and detect policy violations. Wireless IDSs collect all local wireless broadcasts and produce warnings based on either predefined signatures or traffic anomalies.
Lastly while quite a few individuals seem to think that IDSs must now be relegated to the past, we believe that the combination of in-depth study of conventional IDSs and the recognition of capabilities with the ability to block and secure new IDS technology goes a long way towards the future of intrusion detection.

REFERENCES

[1] Tahir Mehmood1 and Helmi B Md Rais2 “Machine Learning Algorithms in Context of Intrusion Detection” 3rd International Conference on Computer and Information Sciences (ICCOINS),2016.
[2] Kunal and Mohit Dua “Machine Learning Approach to IDS: A Comprehensive Review” Proceedings of the Third IEEE Conference Record # 45616,2019
[3] Preeti Mishra, Vijay Varadharajan, Uday Tupakula and Emmanuel S. Pilli, “A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection” IEEE Communications Surveys & Tutorials,2018.
[4] Mr. Subhash Waskle, Mr. Lokesh Parashar and Mr. Upendra Singh “Intrusion Detection System Using PCA with Random Forest Approach” Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020).
[5] Hamed Alqahtani, Iqbal H., Sarker Asra Kalim, Syed Md. Minhaz Hossain ,Sheikh Ikhlaq, Sohrab Hossain “Cyber Intrusion Detection Using Machine Learning Classification Techniques” International Conference on Computing Science, Communication and Security COMS2 2020.
[6] Arjunwadkar Narayan M. and Thaksen J. Parvat “An Intrusion Detection System, (IDS) with Machine Learning (ML) Model Combining Hybrid Classifiers” Journal of Multidisciplinary Engineering Science and Technology (JMEST) Vol. 2 Issue 4, April - 2015.
[7] Phurivit Sangkatsanee, Naruemon Wattanapongsakorn and Chalermpol Charnsripinyo “Practical real-time intrusion detection using machine learning approaches”Volume 34, Issue 18, 1 December 2011.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

[8] Suad Mohammed Othman, Fadl Mutaher Ba-Alwi, Nabeel T. Alsohybe and Amal Y. Al-Hashida "Intrusion detection model using machine learning algorithm on Big Data environment "Journal volume 5, Article number: 34 Published: 24 September 2018."

[9] Kumar, S., Hämäläinen, T., amd Viinikainen, A. "Machine Learning Classification Model for Network Based Intrusion Detection System" The 11th International Conference for Internet Technology and Secured Transactions, 2016.

[10] Subhash Waskle, Lokesh Parashar and Upendra Singh "Intrusion Detection System Using PCA with Random Forest Approach" International Conference on Electronics and Sustainable Communication Systems (ICESC),2020

[11] Waskle, S., Parashar, L., & Singh, U. (2020). Intrusion Detection System Using PCA with Random Forest Approach. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). doi:10.1109/icesc48915.2020.9155656.

[12] Gisung Kim ,Seungmin Lee and Sehun Kim "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection" Expert Systems with Applications 41(4):1690–1700 DOI: 10.1016/j.eswa.2013.08.066 March 2014.

[13] Ahmed I. Saleh, Fatma M. Talaat and Labib M. Labib "A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers" Artificial Intelligence Review volume 51, 2019.

[14] Peiying Tao, Zhe Sun and Zhixin Sun "An Improved Intrusion Detection Algorithm Based on GA and SVM" in IEEE Access, vol. 6, pp. 13624-13631, 2018, doi: 10.1109/ACCESS.2018.2810198, 2018.

[15] Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 41(4), 1690–1700. doi: 10.1016/j.eswa.2013.08.066

[16] Rajeswari, L. P., & Kannan, A. (2008). An Intrusion Detection System Based on Multiple Level Hybrid Classifier using Enhanced C4.5. 2008 International Conference on Signal Processing, Communications and Networking. doi:10.1109/icscn.2008.4447164

[17] Chou, T.-S., & Chou, T.-N. (2009). Hybrid Classifier Systems for Intrusion Detection. 2009 Seventh Annual Communication Networks and Services Research Conference. doi:10.1109/cnsr.2009.51

[18] Farid, D. M., N. Harbi and M. S. Rahman. "Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection." *ArXiv* abs/1005.4496 (2010): n. pag.

[19] L.Xiangmei and Q. Zhi, "The application of Hybrid Neural Network Algorithms in Intrusion Detection System," 2011 International Conference on E-Business and E-Government (ICEE), Shanghai, China, 2011, pp. 1-4, doi: 10.1109/ICEBEG.2011.5882041.

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

[21] Deshpande, Vivek, Prachi Sarode, and Sambhaji Sarode. "Root cause analysis of congestion in wireless sensor network." International Journal of Computer Applications 1, no. 1 (2010): 27-30.

[22] Bang, Raghav, Manish Patel, Vasu Garg, Vishal Kasa, Jyoti Malhotra, and Sambhaji Sarode. "Redefining smartness in township with Internet of Things & Artificial Intelligence: Dholera city." In E3S Web of Conferences, vol. 170, p. 06001. EDP Sciences, 2020.

[23] Sarode, Sambhaji. "VSRS: variable service rate scheduler for low rate wireless sensor networks." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 9, no. 4 (2017): 37-44.

[24] Sarode, Prachi, and R. Nandhini. "Intelligent query-based data aggregation model and optimized query ordering for efficient wireless sensor network." Wireless Personal Communications 100, no. 4 (2018): 1405-1425.

[25] Sarode, Prachi, and R. Nandhini. "APDA: Adaptive pruning & data aggregation algorithms for query based wireless sensor networks." In 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), pp. 219-224. IEEE, 2016.

[26] Parikh, Smit, Srikar Banka, Isha Gupta, and Sambhaji Sarode. "Dynamic Based face authentication using Video-Based Method." International Journal of Computing and Digital Systems 10 (2020): 1-9.

[27] Deshpande, Vivek, Prachi Sarode, and Sambhaji Sarode. "EDCAM-early detection congestion avoidance mechanism." International Journal of Computer Application 7, no. 2: 11-14.