Survey on Machine Learning Approaches for Intrusion Detection System

Zina Garcia.R[1], Dr.C.Kavitha[2]
{20mz35@psgtech.ac.in[1], ckk.cse@psgtech.ac.in[2]}

PG scholar[1], [2]Assistant Professor, Department of Computer Science and Engineering, PSG College of technology, Coimbatore, India

Abstract*:* Wireless connectivity is easy available and low cost. Wireless networks are widely regarded as the most convenient and unavoidable in modern life. So, massive boom of information's are exchanged between every systems which causes intrusion events. Addressing, processing, and storing these massive amounts of data has become a challenge. As a result, attackers can easily gain access to the network and target private data transmission. There are ongoing efforts to implement and create a unique intelligent security system capable of managing and resisting intrusion events. The Intrusion Detection System (IDS) is one of among the unique security system. First, a single classifier may not be capable of detecting all types of attacks. Second, many models are designed to work with stale data and they'll be less adaptable in finding and detecting new attacks. Thus, it's unclear and incomplete with some of the factors, due to the lack in identification or addressing the issues of an intrusion detection system. Here it's easy to increase the anomaly detection speed and also can reduce the imbalance dataset using Machine learning approaches. The major and important contribution for this work includes more of the comprehensive research of papers among different authors that targeted on reducing the bias in the dataset. In addition, analyzing the previously done work and investigating on the Machine Learning techniques used across the referred papers. To efficiently balance the imbalanced data in the dataset and compare the classification process the proposed work makes an optimal use of K-Nearest Neighbor, Random Forest and Logistic Regression.

Keywords : Balance, Dataset, Intrusion Detection System, Machine Learning, K-NN, RF, LG

1 Introduction

Wireless networks are widely regarded as the most practical and unavoidable in modern life. The 802.11 networks, generally known as Wi-Fi, are the most popular low-cost wireless connectivity option. It enables for a rapid set-up in an enterprise context for data interchange with security standards. The Wireless Network Intrusion Detection paper contains the security specification (Wi-Fi)[1]. It is well known that the widespread use of the internet has resulted in a massive information explosion. Because of the open standards that are available, the network has become susceptible as a result of this huge development. Mainly for Addressing the respective data, processing it accurately , and storing these massive amounts of data from the datasets has become a very big challenge. As a result, attackers can easily gain access to the network and target private data transmission. Regardless, the use of earlier security programs could become a victim as well. Many organizations and specialists are working in this development to safeguard sensitive data. The pace of change in cyber security is

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2315107

quickening, and it's attracting a lot of attention throughout the world. There are ongoing efforts to implement and create a unique intelligent security system capable of managing and resisting intrusion events. The Intrusion Detection System (IDS) is one of among the unique security system. This technology assists in detecting and responding to intrusion events in a timely manner [2]. The traffic occurring and happening in the network is being monitored to see if it is normal traffic or malicious one. The wide range of security assaults on Wi-Fi makes it a hot topic and a popular research topic. As a result, wireless networks have been subjected to common intrusion detection approaches [3]. With the introduction of Machine Learning (ML), this has become a burgeoning field of study. The creation of datasets that rely on machine learning techniques has been given constant attention [4]. Despite the work in progress, there are still many fundamental issues with dataset imbalance. There have been studies that have used balancing techniques as their main conclusions. The paper [6] examines the impact and the judgment of the imbalance class distribution and also introduces a computational framework that includes data level as well as algorithm-level arrangements. A review of available approaches for categorization of unbalanced datasets is also shown [7]. As a result, dealing with issues of class imbalance is more crucial from the standpoint of comprehension. It is well known that, there are literatures addressing the issues of an intrusion detection system in machine learning. However, due to a lack of identification or expertise, several of the aspects in these works are ambiguous and incomplete. As a result, it's critical to assist and contribute to the development and research group's primary findings in cyber security using the imbalance dataset.

2 Literature Review

Literature surveys assist in determining what statistical knowledge is available about the research topic. It aids in the discovery of gaps in previous research, allowing for the generation of new, original ideas. The relevance to the proposed idea can be justified.A few references, and the paper's content is described here.

As per [1] which depends on the Intrusion Detection System is quite possibly the regularly utilized ways to deal with secure the respectability and accessibility of basic resources in ensured system. In this exploration, another interruption discovery framework dependent on include choice and gathering learning procedures is presented. For dimensionality decrease, a heuristic methodology called CFS-BA is proposed in the principal stage, which chooses the best subset dependent on include relationship Then, at that point, presenting a gathering approach that joins C4.5, Random Forest (RF), and Forest by Penalizing Attributes (Forest PA) approaches. At last, for assault acknowledgment and the voting mechanism are basically used to consolidate the probability distributions for the learners who learn from the scratch. Utilizing the NSL-KDD and CIC-IDS2017 datasets, the exploratory outcomes shows that the CFS-BA-Gathering technique beats other related and cutting edge approaches on various measures. So the advantage of the methodology apparently reduces the dimensionality radically and dispense with the insignificant elements of the dataset and the restrictions is without a doubt, they couldn't keep away from a high false positive (FP) rate. Albeit this sort of IDSs is solid for distinguishing known assaults, it can't recognize obscure assaults or varieties of known ones. The ensemble technique utilized in this work shows a prevalent exhibition and later on, its ability could be additionally improved to manage uncommon assaults from the monstrous network traffic.

As the occurrence of cyber threats increases day to day this [2], makes use of significantly skewed real-world benchmark network traffic statistics from a variety of assaults. After using an oversampling strategy to resolve the issue of class imbalance in our datasets, the research takes on a new dimension. To begin with, making distinct prediction models for each kind of attack and tuning the model for gaining the exact accuracy. The created model can precisely anticipate the threat or malicious activities and also the different types of assault. So the advantage of this methodology is by distinguishing and initiating alarms, IDS can support the counteraction of information loss because of safety breaks. This is refined by following and dissecting approaching traffic information. The limit of this technique is the powerlessness to distinguish new arising assaults. Basic Traditional IDS and methods say for example, signature-based detection frameworks are in every case useful for distinguishing old-known assaults, however new assaults won't be recognized by these frameworks. So later on, this work can explore more on new and ongoing sorts of assaults.

In [3], Big data is used to analyze the state-of-the-art in resolving unfavorable consequences owing to class imbalance, this work has concentrates in on imbalance class which is high (i.e., a larger part to minority class proportion somewhere in the range of 100:1 and 10,000:1). Information Level which is basically called data level (e.g., information testing (Sampling of data)) and Calculation Level that is basically called Algorithm level (e.g., cost-touchy and half and half/gathering). In order to alleviate class imbalance, data sampling methods are popular, with Random Over-Sampling approaches often producing overall a decent result. There are some performers which are really good enough to use at the algorithmic level. However, the outcomes of published investigations are varied and contradictory where this report presents a comprehensive review of studies. This work is aided by the use of a flexible computing framework for large data analysis, such as Apache Spark, which makes it easier to solve the problem of high-class imbalance. The principle restriction of this review identifies with the significance of the acquired outcomes. Another overall constraint is the restricted extent of assessed methods for each study. Likewise suggest that assessments ought to be performed on an assorted scope of classifiers introduced in the previous works. Varieties of a unique trial might be done on the equivalent dataset. In any case, giving these performance scores might be significant for future near research.

[4] has expressed that six AI based IDSs are presented that is the K Nearest Neighbor (KNN), Random Forest (RF), Gradient Boosting, Adaboost, Decision Tree (DT), and Linear Discriminant Analysis. A prevailing security dataset, CSE-CIC-IDS2018, is utilized rather than more established and generally worked datasets to make a more practical IDS. Thus, the proportion in the ratio is brought down by utilizing a manufactured information age model called has stated that six machine-learning-based IDSs are introduced that is the K Nearest Neighbor, Random Forest, Gradient Boosting, Adaboost, Decision Tree, and Linear Discriminant Analysis. An up-to-date security dataset, CSE-CIC-IDS2018, is used instead of older and mostly worked datasets to create a more realistic IDS. As a result, the imbalance ratio is lowered by employing a synthetic data generation model called has stated that six machine-learning-based IDSs are introduced that is the K Nearest Neighbor, Random Forest, Gradient Boosting, Adaboost, Decision Tree, and Linear Discriminant Analysis. An up-to-date security dataset, CSE-CIC-IDS2018, is used instead of older and mostly worked datasets to create a more realistic IDS. As a result, the imbalance ratio is lowered by employing a synthetic data generation model called Synthetic Minority Oversampling Technique to boost

system efficiency depending on attack types and to reduce missed incursions and false alerts (SMOTE). Minor classes' data is generated, and their numbers are grown to the average data size using this procedure. Oversampling Method to helps this system proficiency relying upon assault types and to lessen missed attacks and false alarms (SMOTE). Minor classes' information is created, and their numbers are developed to the normal information size utilizing this approach. This approach significantly boosts the detection rate for rarely occurring incursions, according to experimental results. The main advantage in it is the deep learning algorithms are used in detecting small sample attacks in up to date datasets. The disadvantage in this paper is when the distribution of classes in a dataset is not uniform, it is said to be unbalanced. This is a major issue in this work , as many of the classification issues are caused by the datasets already used. Due to the imbalanced dataset, the utilized classifier favors the majority class; the goal of this work is to discover the minority class. So the minority class samples can have a high classification error, and important objectives may be ignored. So for the future enhancements, deep learning algorithms are to be applied. It is expected that by employing a different design process, the system's efficiency will improve.

As proposed [5], An effective Intrusion Detection System is required to combat the increasing amount of wireless network infiltration activities. The method called Deep learning is based on a ladder network was proposed in this paper, which learns by itself the features required for the detection of network anomalies and effectively classify attacks. Furthermore, utilizing focal loss as a loss function helps to improve the model's discriminative capacity to categorize challenging samples. The network recordings were categorized into four classes in studies uses this the Aegean Wi-Fi Intrusion Dataset which is a public data-set which holds normal record, injection attack, impersonation attack, flooding attack. This work has accomplished the classification correctness/accuracies of these four sorts of records, and accomplished a general exactness of 98.54%. The value of this methodology is that, this is the first of the utilizations of ladder network with focal loss in interruption location task. However, a significant drawback is the requirement for extensive feature engineering and a very high level of participation and contribution from the human experts are required in order to obtain decent attack detection accuracy. Improving the detection of impersonation attacks is a future direction. It's also worth thinking about expanding the model structure to include more layers.

In [6], for the emerging cyber security threats, Network Intrusion Detection Systems are very essential so accordingly, this work offers a novel mode that can be prepared on an assortment of capabilities and applied to two unique intrusion detection applications: standard expertise networks and 802.11 remote networks. The model is made out of various layered sub-outfits and depends on a one-versus-all parallel structure. Each subset ensemble has an assortment of sub-learners who learn from the scratch, with just a piece of the sub-learners carrying out boosting to offer solid generalization capacity. The sub-ensemble of each class are given a class weight dependent on the affectability measure (true positive rate), which is just gained from the trained data of the respective dataset. Pruning sub-learners that don't add to or contrarily affect by and large framework impact was likely investigated. The outcomes show that the proposed framework can perform very well in both customary endeavor intrusion detection and more up to date undertaking intrusion detection applications and 802.11 remote interruption identification that is the wireless intrusion detection. The suggested work has the primary benefit of lowering the FPR for regular traffic. To put it another way, more attacks can be quickly identified. The fundamental hindrance is Oversampling adds

critical computational complexity, and along these lines this methodology can't be scaled productively for an application to intrusion detection datasets with many thousands to millions of tests. A Semi-Boosted Nested Model with Sensitivity Based Weighted Binarization isn't given, which tried altogether more modest UCI datasets. Accordingly, future work will concentrate and focus more on extending the application to different types of intrusion detection and presenting more sorts of sub-learners into the ensemble.

According to [7], the Interruption Recognition System (IRS) is used in this work so as a result, the IRS must be kept up to date on the most current gatecrasher attacks in order to protect administrations' privacy, trustworthiness, and accessibility. The CART and RBFN were chosen for this inquiry because of their expertise in Knowledge Disclosure and Data Mining (or Knowledge Discovery in Databases). The CART classifier has been proven to be effective for distinguishing and arranging all KDD dataset assaults, which are of sort DOS, R2C, C2R, and Test. So the benefit in using this classifier are CART can deal with missing qualities naturally utilizing surrogate parts. Uses any blend of consistent/discrete factors. It consequently performs variable choice and can set up communications among factors but does not fluctuate as indicated by the monotonic change of prescient variable. The main disadvantage is that it cannot work if there are un-correlated variables. More classifiers will be tried in the future, as well as element determination, to see the most important highlights.

In [8], An intrusion detection system plays a key role in ensuring that various networks are secure and protected from threats. There are many different types of IDS systems; however, for this work mainly, concentrated on those that use Machine Learning (ML) and Deep Learning (DL). In this paper, mainly used Feed Forward Deep Neural Networks (FFDNNs), Deep Long-Short Term Memory Recurrent Neural Networks (DLSTM RNNs), and Deep Gated Recurrent Unit Recurrent Recurrent Neural Networks to design and implement DL-based IDS systems (DGRU RNNs). Also implement an Information Gain (IG) based feature extraction method that is combined with the FFDNNs to address the issue of highly dimensional input spaces. Two wrapper-based feature selection methods were also created and implemented. One is inspired by the Random Forest and the other is based on the Extra-Trees (ET) classifier (RF). The ET is used in conjunction with the DLSTM and DGRU RNNs. FFDNNs are used in conjunction with the RF. The NSL-Knowledge Discovery and Data Mining (NSL-KDD) dataset, the University of New South Wales-NB15 (UNSW-NB15) dataset, and the Aegean Wi-Fi Intrusion Dataset were utilized to evaluate the performance of our frameworks (AWID). The main advantage is This work focuses on the usage of Intrusion Detection Systems to defend wireless networks. IDSs, on the other hand, are reactive measures in the sense that they are generated by events. They respond to attempted intrusions or attacks. As a result, it would be helpful to look into the use of prevention methods. This introduces the concept of an Intrusion Prevention System. So the main disadvantage is, it isn't capable of providing the protection against, the threats like brute force attacks, network discovery attacks etc. Consequently, in the future, the intended plan is to examine the use of the frameworks on real-time network traffic. It is well recognized that in most cases, different datasets do not provide a complete picture of real-time traffic flow. Furthermore, actual traffic flow is dynamic, and datasets are frequently recorded during a certain time period. As a result, events that occurred before or after the traffic flow was captured may be missing from the datasets. As a result, need to intend on the research applicability of the frameworks to real network traffic in future work.

In [9], A wireless Intrusion Detection Framework is utilized to recognize network-based assaults like flooding, assaults service denial, malware, and twin-abhorrent intruders. For identifying interruption or assaults in 5G and IoT organizations, this work utilizes a profound auto encoded dense neural network methodology . The method was tried utilizing the Aegean Wi-Fi Interruption dataset as a benchmark. For Flooding, Impersonation , and Injection assaults, our outcomes showed great execution, with a general discovery accuracy of 95.9%. The work likewise gives an examination of the previously used approaches in the research, which uncovered that the recommended algorithms and the techniques gave a huge expansion as far as accuracy detected and also the speed of detection. The fundamental benefit is that the algorithms utilized in this work are trained in the offline on superior PC which would help the degree of the accuracy with extraordinary and best performance. Subsequently, future models will actually have to deal with a more extensive scope of threats, react progressively, and redesign themselves over the long upcoming period. To help in boosting the performance of the framework, this work must be extended in the future to for further development and enhancements in recognition of intrusion detection and limit the ration of false negatives and also the true negatives in assault discovery.

According to [10], cyber-attacks has risen tremendously so, Software-based machine learning is one technique to creating an intrusion detection system. Threats can be predicted and detected using this method before they become big security incidents. Furthermore, an effective performance metric that can compare several multi-class and binary-class systems in terms of class distribution is required. Furthermore, anticipatory detection approaches must be able to distinguish true attacks from random flaws, inherent design flaws, system device misconfigurations, system failures, human errors, and software implementation flaws. Furthermore, a lightweight IDS that is tiny, real-time, versatile, and reconfigurable enough to be employed as permanent security infrastructure components is required. Three publicly available datasets are used to represent various networking setups, as well as realistically imbalanced class distributions and updated attack strategies. Three key elements make up the presented intrusion detection framework: feature selection and dimensionality reduction, handling imbalanced class distributions, and classification. The feature selection mechanism makes use of searching algorithms and correlation-based subset assessment techniques, whilst the feature dimensionality reduction part makes use of deep learning techniques such as principal component analysis and auto-encoder. are compared to determine which one(s) is/are the most efficient and accurate for the proposed intrusion detection framework. In addition, a hardware-based approach for detecting malicious behaviors of sensors and actuators integrated in medical equipment is proposed, in which the patient's safety is crucial and paramount. The objective is to create a Behavior Specification Rules Monitoring (BSRM) tool for four medical devices using a methodology that converts a device's behavior rules into a state machine. The BSRM tool's simulation and synthesis findings show that it efficiently identifies the device's predicted for normal behavior and also detects the deviations from it. For the identical problem, the performance of the BSRM strategy was compared to that of a machine learning-based approach. The BSRM's FPGA module can be used as an IDS in medical devices and can be combined with a machine learning-based method. The FPGA chip's programmable nature gives the developed model an added advantage: the behavior rules may be quickly changed and customized to the needs of the device, patient, treatment algorithm, and/or pervasive healthcare application. An additional benefit of the work is that the behavior rules may be readily modified and adjusted to the device, patient, treatment algorithm, and/or pervasive healthcare application's needs. Security threats, and new vulnerabilities as well as handling

older attacks that have not disappeared or have not evolved are not solved in this work. However, more research on this topic is needed, especially in light of the embedded class labels (Normal, Attacker, Victim, Suspicious, and Unknown). One of the most important outcomes of this research is that malicious behaviour can be detected using a hardware-based specification rules method. When compared to software-based machine learning, the hardware-based behavior specification rules method produced better outcomes in future.

[11] has stated modern intrusion detection has expanded to encloses the wireless 802.11 networks and Industrial Control Systems and Supervisory Control and Data Acquisition (ICS/SCADA) systems. The first model is a hybrid ensemble that employs complementary-based diversity metrics in a greedy search pruning technique which is considered to be more efficient and one of the effective metric. On an 802.11 network, the recommended hybrid ensemble is built from a diversified combination of decision tree and Naive Bayes classifiers and which these two classifiers are used for evaluating the performance of the Intrusion Detection. The subsequent model depends on a parallel system that comprises of many layered sub-ensembles and depends on a one-versus-all (OVA) twofold structure. Each sub-ensemble has an assortment of sub-learners, with just a fraction of the sub-learners utilizing boosting to offer strong generalization capacity. The sub-ensembles of each class are given a class weight contingent upon the sensitivity measure (genuine positive), which is just gained from the previously attained information. The subsequent model is utilized to recognize interruptions in both ordinary and 802.11 wireless organizations. When compared to state-of-the-art approaches for effective multiple domain Intrusion Detection, the proposed models gains higher rates of detection and the overall false positive performance was with a decent score. This work's adaptability is one of its primary advantages; it effectively recognized intrusions in three separate intrusion–detection domains that were trained on completely different features. The limitation is the traffic characteristics found in the dataset were just for a single or specific network implementation, not for a full network implementation. To effectively accomplish the work, the models must be trained on actual network traffic for a specific implementation that varies accordingly. Other efficient base classifiers could be included into the hybrid ensemble in the future, and new complementary-based diversity metrics could be developed using a global-optimal search method that maximizes the weights and/or events in complementary-based diversity measures.

According to [12], the complex and ever-growing network threats are addressed by the Intrusion Detection Systems (IDS). For effective network anomaly detection, an IDSs framework was created, with experimental results to demonstrate the benefits of the suggested framework. The proposed approach addresses one of the most pressing issues that SMEs' IDSs face: scalability and autonomic self-adaptation. This work discusses how to train, test, and evaluate IDSs using various machine learning techniques. Results of experiments show that using feature selection approaches can lead to better classification accuracy and improved computational speed. Machine Learning techniques are a major benefit to this work where SMEs are used for daily monitoring of security events, but they should not be employed blindly to avoid increased IDS complexity and system failures. So it would be ease to test the suggested framework with more real datasets in the future, including mobile and Internet of Things security platforms, to see if it can handle a wider spectrum of assaults. Experiments should also be carried out employing machine learning approaches capable of real-time intrusion detection.

In [13], the Deep learning methods are used finding out the effective classification with imbalanced data in the respective dataset. This poll delves into the specifics of each study's implementation and experimental outcomes, as well as providing additional insight into its strengths and drawbacks. Data complexity, frameworks that were explored, performance interpretation, easy usage, big data related application, and generalization to other respected domains were been investigated. This work has found that there is generally little concentrate in this field, that most existing work centers around PC vision assignments with CNN, and that the ramifications of enormous information are seldom considered. A few customary procedures for class imbalance, such as data sampling and cost-sensitive learning, have been displayed to work in deep learning, while more refined strategies that exploit neural network feature leaning include mastering abilities have demonstrated to be promising. The capacity to gain a effective cost matrix all through training period is the main advantage of this work, as this is regularly an intense and tedious methodology yet learns. The method involved with creating an ideal expense grid is probable the main requirement, and the CC strategy is vigorously dependent on the DNN's ability to produce discriminative components that group well. The significant objective is to utilize a big data approach to recognize information from quick organization cameras, however their test was bound to a particular number of pictures. The survey closes with a conversation that recognizes a few weaknesses in deep learning because of information that is slanted by class to direct future examination.

According to [14], in order to identify sophisticated attacks in the Wi-Fi networks, there is a necessary for effective algorithms to be introduced. So this paper, introduces an ensemble learning method which is used to develop an intrusion detection system for Wi-Fi networks. The AWID Wi-Fi intrusion dataset is used to identify the elements that are required for a successful IDS implementation. For the dataset being used in this work, multiple ensemble learning methods are applied before deciding on the best one for the proposed IDS implementation. This Intrusion Detection System's work and it's performance are measured with a good and known factors of metrics like as accuracy, precision, recall, and f-measure. The benefit of this work is that it can execute each algorithm with different random states and use the mean accuracy as a selection criteria for the best model, which aids in achieving respectable accuracy when compared to earlier ones. The main disadvantage is that time series features are not considered and they are blindly removed or eliminated without any considerations. So for the future , it can be examined that the features for impersonation and injection attacks more close, and also as previously discovered that many of the records for impersonation assaults were classed as injection attacks.

In [15], due to the traffic getting increased day to day in the network, The UNSW-NB15 and KDD99 datasets feature characteristics are compared in this work, and the UNSW-NB15 features are duplicated to the KDD99 data set to assess their efficiency. To make the most grounded highlights from the two datasets , there use of an Association Rule Mining algorithm as feature selection technique. To evaluate the intricacy as far as accuracy and FAR, some current classifiers are utilized. The test results uncover that the first KDD99 ascribes are less productive than the KDD99 dataset's repeated. UNSW-NB15 attributes. When comparing the two datasets, the KDD99 dataset has superior accuracy than the UNSW-NB 15 dataset, and the FAR of the KDD99 dataset is lower than the UNSWNB 15 dataset. The main benefit in this approach is that the Nave Bayes and EM clustering models are used in the NIDS decision engine to assess the complexity of the two data sets in terms of accuracy and FAR so as an output the replicated UNSW-NB15 features of the KDD99 data set have superior assessment

criteria than the original KDD99 features which is an extra advantage for this work and the disadvantage is due to the similarity of their values, the Decision engine's algorithms are unable to discriminate between the normal and attack rows. So for the future, this work can be intended to create a new NIDS algorithm to discover new patterns that can distinguish between comparable record values of each characteristic.

According to [16] a system for accurately detecting possible attacks has been built using decision-free, Random forest, and KNN approaches. A new solution is offered to address the limitations of the prior system, which was unable to identify IPV6 attacks. With the future in mind, the created system produces an amazing and efficient result in recognizing IPV4-based attacks. The efficacy of several algorithms is assessed. The accuracy, precision, and recall of the detection were all measured. The major advantage of this work is it has to improved intrusion detection rates and lower the number of false negatives. The main disadvantage is that, this work does not, include prevention of intrusions rather finds the threats alone . To increase performance, this work can be expanded in the future by integrating alternative data mining algorithms and data reduction approaches. An intrusion detection system that uses hybrid classification approaches to identify new and unique threats would be quick and reliable.

So in [17], a Survey of Data Mining and Machine Learning for CSID in the field of cyber security is given. To ensure cyber security, instruction detection is carried out. The instruction detection system uses packet headers and net flow packet headers to reach networks and kernel level data. The future scope is that data mining and machine learning cannot be done without representative data, and it is also extremely time-consuming. The paper discusses the difficulty of several machine learning and data mining algorithms, as well as a set of comparison criteria for machine learning and data mining approaches. And Intrusion Detection Systems aid in the detection, determination, and identification of illegal access, duplication, alteration, and destruction of data systems.

3 Intrusion Detection System

The main challenge is the high amount of data streams, which necessitates speedy reactions in real-time processing. Stream data is available in practically in all that the clients utilize and communicate with consistently on day to day, for example, online clicks streams and wireless network traffic [8]. Since stream information is regularly not kept in any type of data storehouse, efficient examination and the board of stream data is a major trouble. A popular query model in a stream data management system is the continuous query model, in which specified queries continuously analyze incoming streams and accumulate aggregate data. Individuals are keen on distinguishing interruptions dependent on irregularities in message stream that can be recognized by effectively making stream models and clustering stream data, or by contrasting current incessant examples with those seen at explicit time period previously. Although most of stream data is at a low degree of abstraction, experts are normally more excited and eager by higher and various degrees of reflection. As a result, stream data should also be subjected to multidimensional and multi-level live analysis and mining [9]. Concept evolution refers to the creation of new classes, whereas concept drift refers to data changes through time. Stream data of indefinite length necessitates infinite length storage and training time [10]. The velocity component of huge stream data introduces

concept drift in the learning model; particularly, concept drift shows that statistical features of the target variable predicted by a model vary in an unexpected way over time. This is a significant issue because the prediction will become less accurate over time [11]. Because of the vast amount of data involved, real-time intrusion detection is a time-consuming operation. A key stumbling block is data imbalance [12]. If there is a lot of imbalance in the data, classifiers will be less accurate and reliable. Because of the huge size and low recurrence of specific exchanges, this data imbalance of the dataset is an unavoidable issue with continuous data. The utilization of examining methods is to lessen the impact of irregularity that is imbalance on classifiers is more often [13]. malicious assaults or anomalies and intrusions are dynamic, consequently intrusion detection on data streams should be done continuously in the real time.. An event may be harmless on its own, but it becomes evil when it is part of a chain of occurrences. In this type of situation, stream data analysis is performed to aid in the detection of intrusions. The AWID-CLS-R dataset has been used for multi-class classification in an ensemble learning approaches/ techniques in the paper [14]. This effort resulted in an accuracy of 95.88 percent for the author. Another observation was made when the assault classes were unified into one. According to thework of calculation in gaining the accuracy, this resulted in a 99.11 percent accuracy. Both the impersonation and injection attacks, resulted in this work with a very low accuracy. To distinguish each of these assault classes, it is for sure a separate or an individual machine learning model were used. In [15], a classification framework was presented to distinguish a complex sample from a simpler one. Machine learning techniques played a key part in the majority of the projects listed above. As a result of their application of these methods, they have noticed a difference in their outcome cycles. As a result, here are some related works based on the AWID dataset.

4 Conclusion

The Intrusion Detection System is one of the most important aspects of cyber-security since it can detect intrusion before and/or after an attack. It serves as a critical defence mechanism for networks and systems. To detect any threat, an ID monitors and analyses data in a system. Due to newly enhanced technologies, intrusion detection has improved substantially over time, particularly in the last several years. The dataset is based on the Wi-Fi that is Wireless Intrusion system. This paper provides an detailed overview of intrusion detection system approaches and technologies, along with their benefits and drawbacks, in their works. A number of machine learning algorithms for detecting abnormalities are also been mentioned in the survey. However, such methods may have difficulty generating and updating information about new assaults, resulting in a high number of false alarms or poor accuracy. This paper also summarizes and displays many recent paper research techniques used for attaining solution to overcome IDS issues which indicates that this usage of methods can provide more benefits than using simply one method in terms of time efficiency and accuracy. For the future, the supervised machine learning techniques/approaches can be used to reduce the impact of bias in classification or prediction due to imbalance datasets and also it would aid the researchers to handle the issues in class imbalance which is more important.

References

[1]. Y. Zhou, G. Cheng, S. Jiang, and M. Dai, “Building an efficient intrusion detection system based on feature selection and ensemble classifier,” Comput. Networks, vol. 174, no. October 2019, 2020, doi: 10.1016/j.comnet.2020.107247.

[2]. H. Al Najada, I. Mahgoub, and I. Mohammed, “Cyber Intrusion Prediction and Taxonomy System Using Deep Learning and Distributed Big Data Processing,” Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018, pp. 631–638, 2019, doi: 10.1109/SSCI.2018.8628685

[3]. J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, “A survey on addressing high-class imbalance in big data,” J. Big Data, vol. 5, no. 1, 2018, doi: 10.1186/s40537- 018-0151-6

[4]. G. Karatas, O. Demir, and O. K. Sahingoz, “Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset,” IEEE Access, vol. 8, pp. 32150–32162, 2020, doi: 10.1109/ACCESS.2020.2973219.

[5]. J. Ran, Y. Ji, and B. Tang, “A semi-supervised learning approach to IEEE 802.11 network anomaly detection,” IEEE Veh. Technol. Conf., vol. 2019-April, 2019, doi: 10.1109/VTCSpring.2019.8746576.

[6]. J. W. Mikhail, J. M. Fossaceca, and R. Iammartino, “A semi-boosted nested model with sensitivity-based weighted binarization for multi-domain network intrusion detection,” ACM Trans. Intell. Syst. Technol., vol. 10, no. 3, 2019, doi: 10.1145/3313778.

[7]. P. Lavanya, A. Sangeetha, and S. Krishnan, “Intrusion detection using machine learning,” Int. J. Recent Technol. Eng., vol. 8, no. 2 Special Issue 6, pp. 832–837, 2019, doi: 10.35940/ijrte.B1154.0782S619.

[8]. “Development and Evaluation of a Deep Learning Based,” vol. 0002, no. August, 2017.

[9]. S. Rezvy, Y. Luo, M. Petridis, A. Lasebae, and T. Zebin, “An efficient deep learning model for intrusion classification and prediction in 5G and IoT networks,” 2019 53rd Annu. Conf. Inf. Sci. Syst. CISS 2019, 2019, doi: 10.1109/CISS.2019.8693059.

[10]. R. Abdulhammed, “Intrusion Detection: Embedded Software Machine Learning and Hardware Rules Based Co-Designs,” p. 176, 2019.

[11]. Joseph W . Mikhail, “(PhD" Good intro for Feature Selection) An Investigation of Anomaly-based Ensemble Models for Multi-Domain Intrusion Detection ///Hybrid Naive Bayes Decision Tree Ensemble,” no. January 2010, 2019.

[12]. O. Elezaj, S. Y. Yayilgan, M. Abomhara, P. Yeng, and J.Ahmed, “Data-driven intrusion detection system for small and medium enterprises,” IEEE Int. Work. Comput. Aided Model. Des. Commun. Links Networks, CAMAD, vol. 2019-Septe, 2019, doi: 10.1109/CAMAD.2019.8858166.

[13]. J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0192-5.

[14]. F. D. Vaca and Q. Niyaz, “An ensemble learning based Wi-Fi network intrusion detection system (WNIDS),” NCA 2018 - 2018 IEEE 17th Int. Symp. Netw. Comput. Appl., 2018, doi: 10.1109/NCA.2018.8548315

[15]. N. Moustafa and J. Slay, “The significant features of the UNSW-NB15 and the KDD99 data sets for Network Intrusion Detection Systems,” Proc. - 2015 4th Int. Work. Build. Anal. Datasets Gather. Exp. Returns Secur. BADGERS 2015, pp. 25–31, 2017, doi: 10.1109/BADGERS.2015.14.

[16]. Mohammed Anbar, Rosni Abdulah, Izan H. Hasbullah, YungWey Chong; Omar E. Elejla, “Comparative Performance Analysis of classification algorithm for Internal Intrusion Detection “, 2016 14th Annual Conference on Privacy Security and Trust (PCT), Dec 12-14,2016, Penang, Malaysia.

[17]. Anna L. Buczak, Erha n Guven, "A Survey of Data Mining and Machine Learning methods for cybersecurity intrusion detection", IEEE communication surveys and tutorials, vol. 18, Issue 2,2016