International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

A Review on Machine Learning Based Approaches of Network Intrusion Detection Systems

Basmah Alsulami[1], Abdulmohsen Almalawi[2], Adil Fahad[3]
[1,2] School of Computer Science Information Technology, King Abdulaziz University, Jeddah 21589
[3] Department of Computer Science, College of Computer Science Information Technology, Al Baha University, Al Baha 65527, Saudi Arabia

ABSTRACT: The rapid growth of using the Internet raises the possibility of network attacks. In order to secure internal networks, intrusion detection systems are widely employed to address a major research challenge in network security, which aims to efficiently detect unusual access or attacks. To do so, various intrusion detection systems approaches based on the concepts of machine learning algorithms have been developed in the literature to tackle computer security threats. These IDs approaches can be broadly classified into Signature-based Intrusion Detection Systems and Anomaly-based Intrusion Detection Systems. This review paper presents a taxonomy of current intrusion detection systems (IDs), a comprehensive review of significant recent works, and a variety of recent attacks that can be detected in the network environment.

KEYWORDS: Intrusion Detection, Network traffic anomaly detection, Semi-Supervised learning, Supervised learning, Unsupervised learning.

I. INTRODUCTION

Due to the fast growth of communication, a variety of systems, including power grids, transportation networks, industrial control processes, and critical infrastructure, have made extensive use of networking. A trusted communication environment may expose such systems and their users to assaults and threats. Many governments are concerned about network security in order to secure their internet-connected systems, such as networks, computers, programs, and data, against assaults, damage, and unauthorized access [1],[2]. This has led to the development of a variety of security technologies and methodologies in order to resist increasing risks and the fast spread of attack patterns (IDS). One of the most important components of any cyber security system is an intrusion detection system (IDS). Detecting and identifying intrusions on computers and networks is the primary goal of this tool. Anomaly/behavior- based and misuse/signature based methods are used to identify attacks, respectively, based on a database of known attack patterns and normal-operation profiles. False positives are rare with the previous IDS strategy, which relies on constantly-updating databases, however this method cannot identify zero-day assaults. Most threats, including zero-day attacks, may be detected using the second method.

Recent years have seen a rise in the use of traffic intrusion detection technologies such as supervised, semi-supervised and unsupervised algorithms to monitor the normal and abnormal behavior of network traffic in order to correctly and effectively battle network assaults [1],[3]. Unpredictable network traffic makes it difficult to apply pre-defined labels to large data sets of network traffic, despite the fact that supervised methods are often thought of as quick and straightforward to use. To be successful in identifying traffic intrusions, however, the model must be free of noise and inaccurate data labelling. Anomaly detection models may be trained utilizing inexpensive and fast-to- collect unlabeled data to overcome the drawbacks of supervised techniques. As a result, unsupervised techniques are ineffectual and incorrect since they don't need subject matter experts to identify them. This has led to the development of semi-supervised techniques, which combine a small quantity labelled data with a large amount of unlabeled traffic information to overcome the constraints of both supervised and unsupervised IDs approaches. Such semi-supervised techniques [4],[3],[5],[2], however, have a high false detection rate because of the quick appearance of new threats and anomalies. Network intrusion data may be obscured by other traffic flows because of a lack of visibility into the intrusion data set [5],[3].

An IDS classification system based on network traffic data auditing and detection algorithms is the topic of this research. In this article, a comprehensive review of some machine learning based methods for IDS. The research also examines the current literature

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

from a holistic approach, encompassing detection architecture, detection method, auditing sources, and the practicality of integrating the suggested IDS to different real-world network systems.

II. INTRUSION DETECTION SYSTEM (IDS)

An anomaly or outlier can be defined as given by Hawkins [6]: "An outlier is an observation that varies so greatly from other observations as to raise suspicions that a different process has generated it.". Using algorithms to detect such anomalies in software is called Anomaly Detection. The Anomaly

Detection field has a very broad spectrum with applications in health care, fraud detection, intruder detection, industrial applications, and big sensor data systems. We can define intrusion as any kind of unauthorized activities that damage an information system which includes any attack that might pose a potential threat to information integrity, availability, or confidentiality. For example, activities that would make the services of the computer not responsive to authorized users are considered as an intrusion. An Intrusion Detection System (IDS) is a hardware or software system that recognizes malicious activities on computer systems in order to allow maintenance for system security. The main aim of an IDS is identifying distinct types of malicious computer usage and network data traffic, that cannot be recognized by a regular firewall. This is essential to achieving high protection toward actions that affect the confidentiality, availability, or integrity of computer systems.

A. Detection Method

An The IDS systems can be classified broadly according to detection method into two categories: Signature-based Intrusion Detection System (SIDS) and Anomaly-based Intrusion Detection System (AIDS) [7].

1) Signature-based IDS (SIDS),

SIDS [7] are based on the method of pattern matching to find a previous intrusion; this type of IDS is also known as Misuse Detection or Knowledge-based Detection. In SIDS, matching techniques are used to find a known attack. In other words, the main idea is to create a database of intrusion signatures and compare the current set of actions with known signatures, and to notify the user if a match is found. Figure 1 shown the general architecture of a SIDS. In this architecture the detector is used to find and compare the activity signatures recorded in the monitored environment with the existing signatures in the signature database. The detector will do nothing if there is no match while when a match is found, an alarm is issued. Although SIDS often provides a high detection accuracy for previously identified intrusions; it has a problem in detecting zero-day attacks because there will be no matching signature in the signature database until the signature of the new attack is extracted and stored. Although these tasks require knowledge of system protocols, operations, and particular issues of attacks, they usually tend to offer a low false alarm rate especially in comparison to the machine learning solution, which is more practical for industry. In addition, the signature based process needs that existing signature of attacks be kept updated and that new threats of intrusion cannot be guaranteed such as zero-day attacks. The increasing frequency of zero-day attacks has made SIDS approaches progressively less appropriate, as there is no prior information available for those kinds of attacks. The rising number of targeted attacks and polymorphic malware variants can further undermine the effectiveness of this traditional approach.

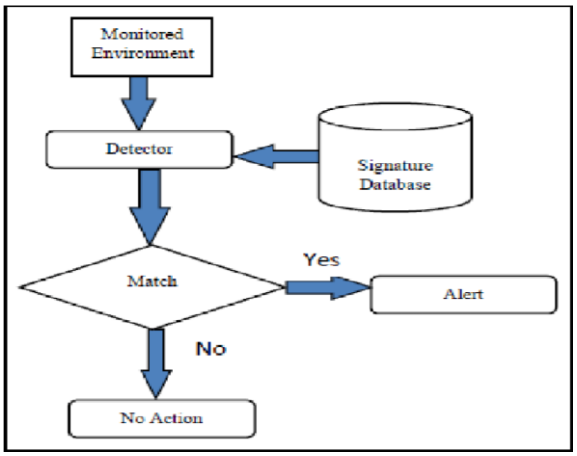

Figure 1: General architecture for SIDS.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

2) Anomaly-based IDS (AIDS),

AIDS [7] has captured the attention of many scholars because of its ability to overcome the limitation of SIDS. AIDS works by comparing the activity observed with the baseline profile. The baseline profile is the acquired normal behavior of the monitored environment and is established during the learning phase when AIDS learns about the system and builds up a normal profile of the monitored environment. Any significant deviation between baseline profile and the observed behavior is considered an anomaly which can be interpreted as an intrusion. The argument for that kind of approaches would be that malicious behavior deviates from normal user behavior. The behaviors of abnormal users which are different from the normal behaviors are identified as intrusions. The implementation of AIDS consists of two parts: the training phase and the testing phase. In the training phase, the normal traffic profile is used to recognize a normal behavior model, afterwards, within the testing phase, a new set of data will be used to determine the system's ability to generalize to previously unknown intrusions. The key advantage of AIDS is the ability to detect zero-day attacks considering the fact that the recognition of abnormal user behavior doesn't always depend on a signature database. AIDS triggers a warning signal when the behavior under examination differs from the normal behavior. In addition, AIDS has a lot of advantages. First, they have all the ability to identify malicious internal activities. If a hacker begins making transactions in a stolen account which are not reported in a normal user activity, an alert system is triggered. Second, it is really difficult for cyber attackers to realize what is normal user behavior without alerting the system as it is constructed from customized profiles.

Table I: Comparisons of Methods to Detect Intrusions

Detection methods	Advantages	Disadvantages
SIDS	-Very efficient in detecting intrusions with minimum false alarms (FA). -Identifies the intrusions promptly. -Superior to identify known attacks. -Simple design	-Need to be updated regularly with a new signature. -SIDS is designed to identify attacks for known signatures. -If the existing intrusion was changed slightly to a newer version, the system would be unable to identify the new intrusion of the similar attack. -Unable to identify a zero-day attack. -Little recognizing of the knowledge of the attacks -High false positive alarms.
AIDS	-Can be used to detect unknown attacks. -Can be used to create an intrusion signature database.	-Difficult to create a normal profile over a very dynamic computer system. -Initial training is required.

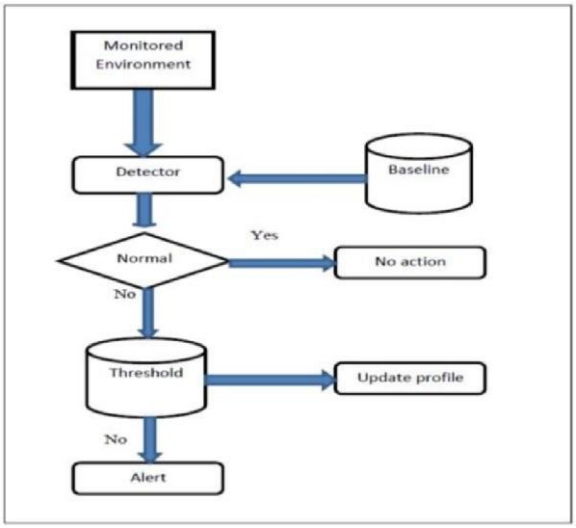

Figure 2: General architecture for AIDS.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

The general architecture of the AIDS system can be seen in the figure. The monitoring environment shall be monitored by a detector which examines the observed activities with the baseline profile. The profile is updated only if the observed activities do not match the baseline profile and are within the acceptable threshold range. Otherwise, no action will be taken. If the observed activities don't really match the baseline profile and fall outside of the threshold range, they will be marked as an anomaly and alert will be issued.

The differences between signature-based detection and anomaly-based detection are presented in Table 1. SIDS can only detect well-known attacks, while AIDS can identify zero-day attacks. Besides that, AIDS can lead to a high false-positive rate, since anomalies could be new normal actions instead of true intrusions.

B. Intrusion Data Sources

The two previous sections classified IDS based on intrusion detection methods. The IDS also can be categorized depending on the input sources of data used to detect abnormal behavior. In general, there are two kinds of IDS technologies regard to data sources, Host-based IDS (HIDS) and Networkbased IDS (NIDS) [7]. **HIDSs** are run only on individual hosts or devices on the network to monitor outbound and inbound packets from the device and notify the user or administrator if unusual activity is discovered. HIDS analyzes incoming and outgoing traffic and system settings such as software calls, local security policy, local log audits, and more. The HIDS should be located on every device and should be configured in that operating system. **NIDSs** are located at a strategic location or point inside the network to monitor traffic into and out of all network devices. Ideal world, one would check all network data traffic. NIDS analyzes network data traffic layer upon layer of the Open Systems Interconnection (OSI) model and makes the decisions about both the aim of the traffic, analyzing it for suspicious behavior. Most NIDSs are easy to install on a system and often can display data traffic from multiple devices at once.

Table Ii: Comparison of Ids Types Based on Their Positioning Within the Computer System

Technology	Advantages	Disadvantages	Data Source
HIDS	-No additional hardware needed. -HIDS can check the end-to-end encrypted communications behavior. -Each packet is reassembled. -Looks at the whole item, not just the streams. -Detects intrusions by scanning the hosts file system, the system calls or the network event.	-Only the machine where it is installed can be monitored for attacks. -Needs to be placed over each host. -Expends host resources. -Reporting attacks is delayed.	Audits records, Interface (API), Application Program log files, Interface (API), rule patterns, system calls.
NIDS	-Capable to detect the broad range of network protocols. -Can check different hosts at same time. -No need to be placed over each host. -Detects attacks by scanning the network packets.	-Allows only the detection of network attacks. -Hard to examine the highspeed network. -The challenge is to find attacks from encrypted network traffic. -The most dangerous challenge is the insider attack. -Specialized hardware is needed.	-Router NetFlow records. -Management Information Base (MIB). -Network packets (TCP/UDP/ICMP). -Simple Network -Management Protocol (SNMP).

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

NIDS installed in a set of locations inside a specific network topology, along with HIDS and firewalls, to provide concrete, flexible and multi-tier protection against both internal and external attacks. Table 2 presents a summary of comparisons between HIDS and NIDS.

C. Types of IDS Attacks

An attack is a set of operations that threaten the security of either a network or a computer system. Network attacks are classified as active and passive attacks [7]. Passive attacks provide an indirect influence. Intruders are tried to launch for two main goals: (i) traffic monitoring and recording, and (ii) gathering valuable information for the subsequent launch of an active attack. Because there is no active process, these attacks are not easy to detect. Traffic monitoring and analysis [7] is an example of a passive attack. On the other hand, active attacks are much more risky and thus are obviously more destructive to a network.

Active attacks are categorized into four groups according to the Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Assessment Plan. These four groups are: 1. Denial of Service (DoS) attacks

DoS attacks are intended to force the target to terminate the service(s) provided by flooding it with probes of illegitimate request. As a result, for DoS attack to be able to detect, packet-level features such as" percentage of error packets" and" source bytes", and traffic features such as" percentage of connections that have the same service and same host destination a" are significant. To detect DoS attacks, it is not necessary to know whether or not a user is" logged in".

2. Probe attacks

   The aim of the probe attacks is to acquire information about the target network from the source which is often external to the network. Hence, features such as" file number accessed" and" file creation number" are never expected to give information about probe detection while basic connection-level features such as" source bytes" and" connection duration" are significant.

3. Remote to Local (R2L) attacks

   R2L attacks are considered to be one of the most difficult to detect attacks since they affect both the host level and the network level features. Designers therefore select both the host level features such as the "number of failed login attempts" among others for R2L attacks detecting and the network level features such as the "service requested" and "duration of connection".

4. User to Root (U2R) attacks

   The U2R attacks include semantic details which are very hard to capture at an early phase. These attacks are mostly content-based and target an application. Consequently, for U2R attacks, features like "number of shell prompts invoked," and "number of file creations" are chosen whereas features like "source bytes" and "protocol" are ignored.

III. OVERVIEW OF MACHINE LEARNING TECHNIQUES

Machine Learning techniques have been widely used for the IDS due to its ability to classify normal/abnormal network traffic by learning patterns based on data collected. From ML perspective, anomaly detection algorithms work by learning what behavior is normal and what behavior is anomalous for the program. This learning can be done in three different ways: supervised, unsupervised and semi-supervised as shown in (Figure 5).

1. ***Supervised learning:*** in machine learning, the supervised techniques try to determine the mapping function based on labeled training data (both input and output in the training set are given) and then use this relationship or function to map a new unlabeled data. After that, when new input data set given, you can easily predict the output (Figure 3). There are a wide range of methods use supervised learning such as Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Artificial Neural Network (ANN) [1], and so on. The advantage is that this technique has great accuracy and speed compared to the other techniques. The downside is that labeled training data for both positive and negative behavior is often not available. This makes supervised learning techniques overall less generalizable than the other types of learning.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

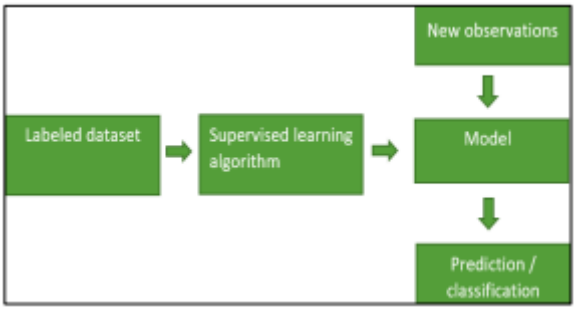

Figure 3: Supervised Learning

2. ***Unsupervised learning*** techniques deal with unlabeled data where you only have the input data, and the objective is modeling the hidden pattern to learn more about data based on the relationship between data themselves (Figure 4). Clustering algorithms such as K Means, DBSCAN, and Affinity Propagation (AP) [1] are examples of this class.

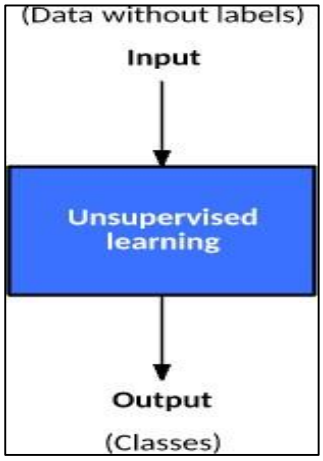

Figure 4: Unsupervised Learning

3. ***Semi-Supervised learning*** techniques is a blend of supervised and unsupervised that use a small amount of labeled data with a large amount of unlabeled data to train the model [3] By this the efficiency of supervised method is increased as the accuracy of anomaly detection rate also improved by using unsupervised method [3].

A. Classification Algorithms

Classification is the most popular task in supervised learning and it is also applied for most IDS systems; It refers to the process of classifying or categorizing given data items into one of several predefined classes, these classes are usually known as target, label, or category. It can be performed on both structured and unstructured data. There are many classification terminologies in machine learning as follows: -

- **Classifier** – It is a procedure that used to map input data to a particular class.
- **Classification Model** – The model that predicts the category or class of the input sample given for the training; it will predict the conclusion for the data item.
- **Feature** – A feature is an individual measurable attribute for the observed event.
- **Binary Classification** – It is a kind of classification with 2 classes only, whether true or false.
- **Multi-Class Classification** – Classification including more than two classes. In this type of classification, each sample shall be assigned to one and only one label or target.
- **Multi-label Classification** – This is a type of classification where a set of labels can be assigned to each sample.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

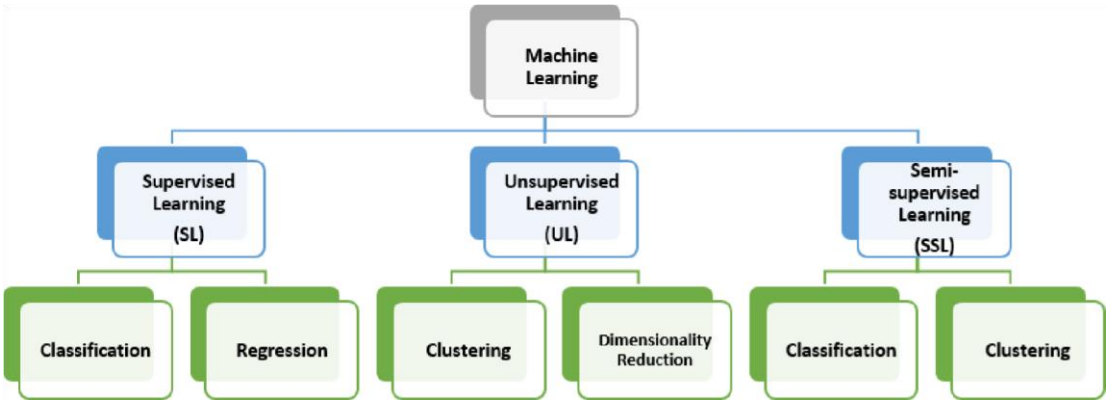

Figure 5: Machine Learning classes

In intrusion detection, we need to gather sufficient normal and anomalous data then use a classification method to teach the classifier how to classify or predict new, unseen data items as normal or anomalous. Despite different IDS processes have different methods, we will define here the generic process to provide a basic idea of how the detection model is trained and then used to detect attacks. Figure 6 shows the overview of the implementation of the IDS approach. In particular, such frameworks consist of five key processes [8]:

a. **Data collection:** Shows the measurement phase where input data has been collected, e.g., system status or traffic logs or event logs, and so on.

b. **Feature extraction and selection:** presents the extraction of features and the selection process, the discriminating features are extracted and selected in a form that can be used in the classification process. In some machine learning techniques, a feature selection process may not be required. In Figure 6(B.2), each data item (i.e., record) is viewed by a vector feature consisting of attribute values (e.g., behaviors or indicators). Sometimes feature values require normalization during this level to prevent large range features (e.g., payload size) from overweight relatively small range features (e.g., binary). For example, Min-Max normalization [9] can be used to transform the numeric value v for feature x to v' that ranges from [0, 1] as follows:

$$v' = \frac{v - \min_x}{\max_x - \min_x} \qquad (1)$$

where min_x *is* the minimum value and max_x *is* the maximum value of the feature x. Note that, each data item may not always be a tuple with a specific number of features, some instances of a sample could be a sequence of features.

c. **Tagging:** each training record should be tagged in order to identify the class to which the data item belongs (i.e., normal or attack, in regards to IDS) and this process is carried out either manually or automatically, using input from an expert or an analyst. In many cases, the attack incident is difficult to collect or simulate from its field. For such scenarios, only normal datasets are used by IDS designer to train the model.

d. **Training:** defines the process of model development. The dataset labeled is used to train and evaluate the classification model. In order to enhance detection accuracy, the model parameters could be adapted and the training/assessment process will be repeated until the anticipatory efficiency is satisfied. When the model training is completed, the model can be used to classify new data samples.

e. **Anomaly detection:** Describe the process of anomaly detection using the classification model.
Each unknown data sample must be labeled either normal or malicious (most systems will give lowerlevel detail of malicious points). At last, the outcome for this module will be submitted to the administrator to either notify or take a response action.

1) Taxonomy of Classification Approaches

In this section, we will discuss a structured view of the classification techniques based on the nine (9) groups mentioned in [8] (see Figure 7). Below, we're describing each group in detail.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

1. *Probabilistic Method:* This category uses probabilities to predict the class of unseen data items. Through the training phase, it will calculate the distribution of joint or conditional possibility [7], based on the existence of particular features of each class. This result reflects the probability that a given input belongs to one of the predefined classes of objects to be classified. Examples of methods in this group are Naïve Bayes (NB), Hidden Markov Model (HMM), Bayesian Network, and Conditional Random Field (CRF). NB [10] is a basic probabilistic method that discovers the conditional probability for each feature in the training data based on a hypothesis that any feature can be viewed independently (i.e., conditionally independent). NB predicts the input class based on the maximum joint probability. The Bayesian Network approach [11] expands the NB classifier by merging Bayesian variables with the directed acyclic graph representation. The graph vertex describes the variable (i.e., class and features), whilst the graph edge denotes the probabilistic relationship among these variables. The graph itself is used to calculate the posterior probability for each class given all evidences.

   Moreover, some probabilistic systems are designed to model a sequence of features rather than a fixed number of features like HMM [12] which inherits the NB concept in order to predict the class of data point based on a set of observable features over time, is based on assumption that an observable feature is produced by the particular hidden system states. The hidden system state at time t depends on its prior state $t-1$. When the joint probability of a future state is computed, the probability of an observable sequence will be estimated through using knowledge of the predicted hidden states.

   HMM is used to classify event labels depending on sequences of features which are traced over the monitoring period (e.g., communication log messages are considered to be a sequence of features).

   CRF is more complicated than the HMM. Rather than depending on a joint probability, CRF works with a conditional probability [13] which makes CRF more flexible. On the other hand, CFR includes a large diversity of overlapping features which make HMM more efficient compared to CRF in terms of computational complexity measurement. On the other hand, when newer data becomes available, the CRF does not support the model re-training. It is therefore not appropriate for attacks that developed over time.

2. *Divide & Conquer Method:* Divide and conquer method contains a broad class of techniques known as the decision tree which formulates a tree data structure from a set of features of each training example in such a way that the rule set became easily derived and can then be used to classify the input into a specific category. [14; 15; 16]. This category includes several algorithms such as Classification and Regression Tree (CART), Iterative Dichotomiser (ID3), C4.5, Frequent Pattern (FP) Growth, Supervised Learning in Quest (SLIQ), SPRINT, and Random Forest (RF). CART is a

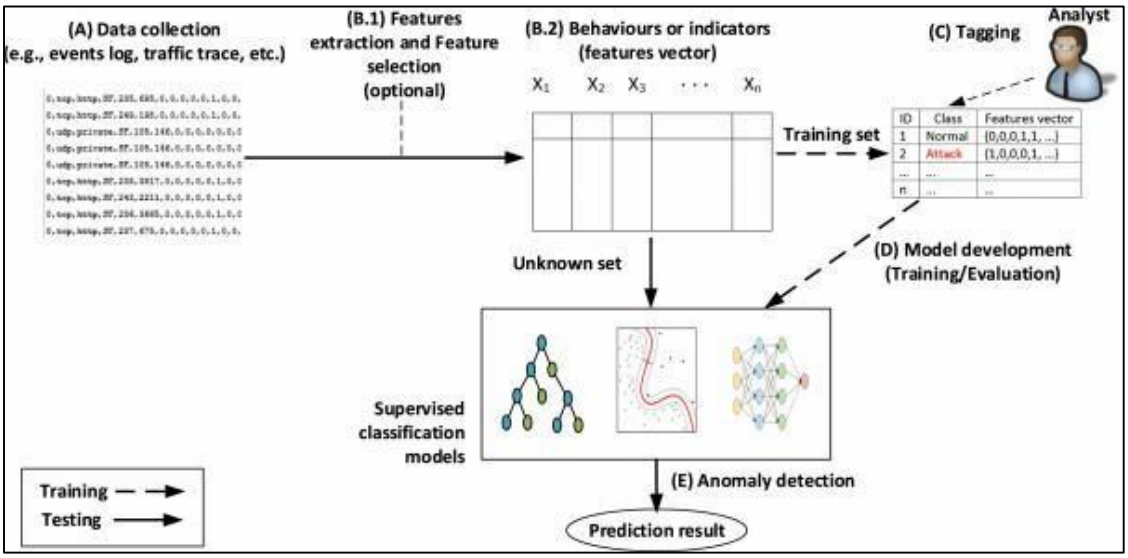

Figure 6: The process of the supervised learning-based IDS approach

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

binary tree consisting of a set of decision rules described using a set of if-else rules. The tree node shows a feature which is used to classify the data item, whereas the edge represents the decision path. The path connecting more than one class is called a sub-tree. The path that only yields one class is called the leaf node and represents the final output. Actually, the classification node provides a combination of instances from different groups; and thus, the feature with the best discriminates of the input data is chosen as the *splitter*. Then, the recursive partitioning is performed at the next level of the tree till the classification arrives at the leaf node. The process of obtaining the best splitter is based on a greedy (non-backtracking) method which makes the resulting decision tree might not be optimized. To reduce tree size, the tree pruning technique is the most commonly used [17].

ID3 and C4.5 consider as advanced decision tree approaches. Both techniques create a decision tree in the same way, however, C4.5 has been developed over ID3 in several aspects. For example, C4.5 deals with both continuous and discrete features. So, for the feature $F \in R$, the *splitting threshold* will be used to split the data into two subgroups instead of just the exact value of the discrete feature F. Besides that, C4.5 allows training data with some absent attribute values and supports tree pruning after formulation, which makes C4.5 outperform ID3 in terms of speed and flexibility [14]. Scalability is an important factor for tree-based methods since the size of the tree grows when training data is larger. Sophisticated decision tree algorithms are proposed to enhance scalability and accuracy compared to classical methods such as IBM's SLIQ [18] and SPRINT [19] algorithms, which are focused on a scalable method that divides parts of the tree from memory to the database on a hard drive to overcome the scalability issue. Other than decision tree strategies, the divide-and conquer method also includes a technique that worked based on the frequent pattern, such as the FP-Growth algorithm, which identifies a pattern that is infrequently realized as an anomaly. FP Growth uses the FP-Tree data structure to summarize patterns of events (i.e., nodes) that frequently occur together (i.e., edges of the tree). Even though the FP-Growth algorithm is effective, the data structure may be too large to fit into the main memory.

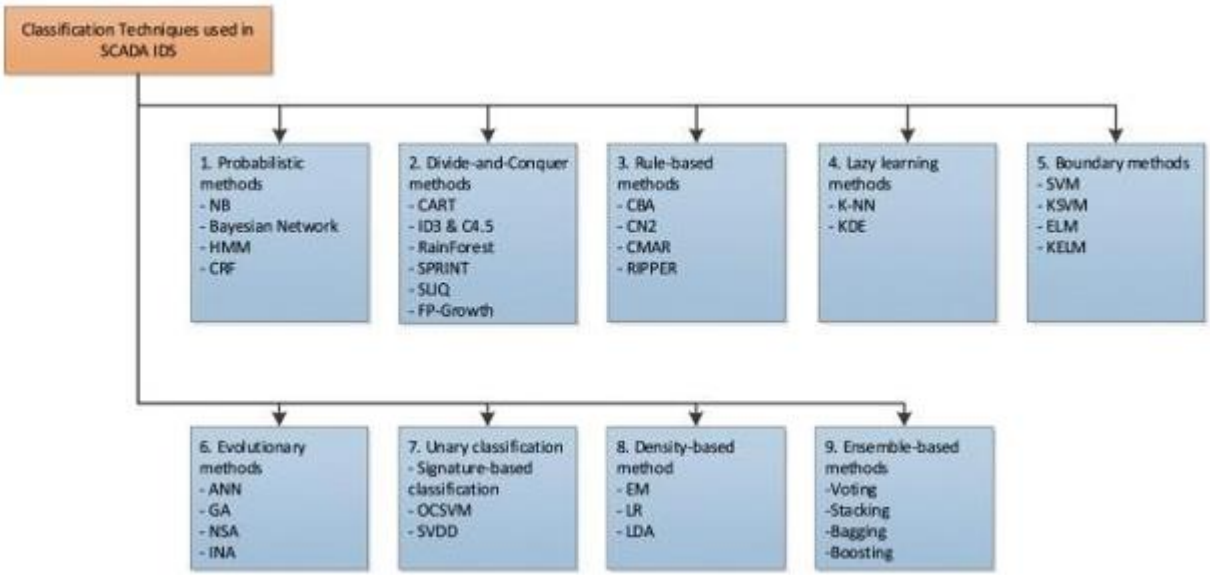

Figure 7: Groups of classification algorithms

3. *Rule-Based Method:* In this technique, a set of rules is used to decide the classes of data inputs. It is undeniably, the rule is easier for a human to realize the reason behind each decision, compared to the probabilistic and numerical models. Generally, the rule includes the condition and conclusion parts. The condition is the features of a particular object that are being classified, whereas the conclusion is the output of the classification. The coverage and accuracy. are used to measure the efficiency of rule-based method. Coverage represents how many items in the data set meets the condition and the accuracy defines how many data items that apply the conclusion. Concerning the automated rules discovery approach, training data can be used to obtain a set of rules. This task can be achieved by either decision tree methods or sequential covering algorithms [17]. The decision tree formulates multiple tree-like structures of decision

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

rules, whereas the sequential approach extracts rules in sequence, one-by-one, from training data. There are many different rules mining strategies, such as Classification Based on Associations (CBA)[20], Repeated Incremental Pruning to Produce Error Reduction (RIPPER)[21], Multiple Association Rules (CMAR)[22], and CN2[23]. A simple technique, such as CBA, picks the most accurate rules to identify the class of each data item in the training set. The CBA algorithm discovers the Class Association Rule (CAR) from the D training set, defined as $x_i \rightarrow C_i$, where the x_i feature implies the C_i class. The rule support $sup\% = |x_i|/|D|$ and confident $conf\% = |x_i \rightarrow C_i|/|D|$, where $|x_i|$ represents the total number of tuples with the attribute x_i ,$|x_i \rightarrow C_i|$ represents the total number of attribute value x_i which means the class value C_i, and $|D|$ indicates the total number of tuples in the training set. These are the minimum thresholds (i.e., *sup*% and *conf* %) that decide whether the rule must be selected for the classification. Even though the selected rules are the most effective, they may not be the best discrimination. Besides, the issue of scalability is an important limitation. Therefore, the rule-based method has a poor scale with huge training data, in particular, once outliers cannot be avoided compared to the decision tree (i.e., C4.5). On the other hand, more advanced techniques such as CMAR and RIPPER aim to improve the robustness of the algorithm by removing impurities from training data and minimizing the set of rules generated during the training stage. CMAR extends the FP-Growth technique [15] to build the Class Association Rules (CR)-Tree and reduces memory space constraints (e.g., use tree pruning techniques); therefore, CMAR has higher classification accuracy and scalability compared to the classical C4.5 and CBA methods. RIPPER [24] tries to address the issue of impurity by proposing an iterative pruning technique to reduce error and allow the use of large and noisy training datasets. Specifically, the training data are split up into growing and pruning sets. The growing step will be used to construct a rule-set, while the new set of rules is directly pruned with the pruning dataset until there is no improvement from the pruning step. Finally, the new set of rules and a rule-set from the previous iteration are integrated.

4. *Lazy Learning Method:* In terms of the learning process, this method differs from the others. The basic idea behind Lazy Learning is to update the training model as late as possible. For instance, training data is kept in your memory without constructing a prediction model as an eager learning technique. The prediction model is built only during the classification phase.

   In fact, it offers advantages and limitations. Although the lazy learning approach provides the best performance during the learning phase, this benefit is a trade-off of computational complexity when making a classification or prediction. Also, it is more flexible than the others because the training model can be improved incrementally [25; 26]. The most common lazy learning method is kNearest-Neighbor (k-NN) [27] that works by comparing input data to k data points by using Euclidean distance measure. Another advanced technique like Kernel Density Estimation (KDE), works by implementing k-NN-based methods (e.g., nearest neighbors, shared neighbors, and reverse neighbors[28]) as a kernel function which is used to formalize the density function (i.e. Multivariate Gaussian [26]) of the normal data points, hence, the boundary of the normal data can be defined by using the density function parameters given that one data point contains *n* features, the anomaly is detected by projecting the data point into the n dimensional space of the features, the data point beyond the normal boundary is classified as an anomaly.

5. *Boundary Method:* Apart from the lazy learning approach, the boundary method is considered to be an eager learner. Support Vector Machine (SVM) [29] became the most commonly used technique from this type. It translates the training data into some kind of decision boundary (so-called hyperplane). The idea of a hyperplane is that nonlinear mapping training data is converted into an appropriate higher dimension, in which two classes can be separated linearly by a hyperplane. The algorithm also tries to maximize the width of the separation plane. Although training time is too long, this technique tends to dominate other techniques by being less over fitting on the training data.

   The SVM-based method combined with other strategies (such as pruning and kernel-trick) can be used to produce higher detection accuracy [30] and lower false positive and negative alarms [31]. Furthermore, it may be used to minimize the offline learning phase [32], or implement for limited input features environments [33]. [34] is also an example of the boundary method proposed to the process of anomaly detection for proprietary communication (e.g., vehicle CAN bus protocol). As the manufacturer does not disclose the communication protocol, it will be difficult to detect attacks on these critical systems.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

6. *Evolutionary Method:* The development of artificial intelligent methods was being inspired by the rapid development of a living organism. There are two popular methods: the Artificial Neural Network (ANN) and the Genetic Algorithm (GA). ANN operates a brain-like structure (i.e., a neural cell network) for classification. The neural network consists of three main layers, *input layer*, *hidden layer(s)*, and *output layer*. The input layer may contain many nodes based on the input (e.g., number of features of each tuple). Each input node is connected to all other nodes inside the hidden layer. links or connections have different weights. These weights are fitted in the training phase to produce an output with the highest accuracy according to supervised learning, there could be more than one hidden layer of the neural network. Although ANN provides higher prediction accuracy in a different application (e.g., anomaly detection, misuse of behavior, and image/voice recognition), the explanation of how ANN works remains an issue [1]. Researchers do not understand exactly how the input data is classified using ANN, and the design of the topology of ANN (i.e., number of nodes and hidden layers) is still ambiguous. In the case of a binary classification problem, the output layer could be just one node. Such that will give output between 0 and 1, which can be represented as a probability of being a particular class. In the case of multi-class classification problems, the output number could be *k* nodes, indicating the prediction of k-classes. Each output node represents the probability of each class. One of the Basic challenges of the ANN approach is the demand for a large training data set as well as the trained model could be over-fitted to the training data set; therefore, ANN is not applicable in such contexts. On the other hand, the GA algorithm [35] simulates the selective survival process in evolutionary theory, assuming that the knowledge base (e.g., the rules of anomaly detection) can be viewed as the living organism chromosomes. These chromosomes will be optimized according to evaluation objectives. In the context of the IDS, the selection of the evaluation function is critical for optimizing the fitness of the output. GA is usually used to minimize the effects of faulty training sets and sometimes deals with the problem of multiple local optima. Other factors of nature are also applied to anomaly detection.
   The Negative Selection Algorithm (NSA) or Immune Network Algorithm (INA) [36] defines the detector module for classification between normal and abnormal data. The concept was derived initially from the organic process in which immune cells detect harmful agents in human bodies.
7. *Unary Classification:* Unary or on-class classification (OCC) solves the issue of binary or multi-class classification mainly by learning from a training set of one class only (i.e. normal system states); therefore, the anomaly can be identified as a class of others. This method helps other approaches whenever one class of training data can be identified clearly, whereas information about other classes is severely difficult to record.
   The most commonly used technique in this group is the One-Class Support Vector Machine (OCSVM)(also called *Support Vector Data Description: SVDD*). The basic principle of the SVDD algorithm is similar to the SVM algorithm, previously explained but rather than splitting between two classes with a linear hyperplane. In the high-dimensional feature space of the data, SVDD formulates the boundary function of a spherical shape that covers the entire population of the target class minimally [37]. Let the sphere be identified by its center a and its radius $R \geq 0$, SVDD seeks to minimize the sphere volume by minimizing R^2 value defined in the *error function* as:

$$minimize \quad F(R,a) = R^2 \qquad (2)$$

$$subject\ to \quad ||x_i - a||^2 \leq R^2, \forall\, i, \qquad (3)$$

   where x_i is the distance between i and a, in which i is the data point and a is the center. Nevertheless, the perfect shape of the sphere could involve outliers from the training data set (so, not optimized). As shown in [37], the separation plane is formulated by solving the following optimization problem:

$$minimize \quad \frac{1}{2}||w||^2 + \frac{1}{vr}\sum_{i=1}^{r}\xi_i - \rho \qquad (4)$$

$$subject\ to \quad (\omega.\phi(x_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots., r \qquad (5)$$

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

where $x_i \in X$ is a data point out of total r samples in the training set X, and Φ: $X \rightarrow H$ *denotes* the mapping function from raw to the high dimensional space. For the hyper plane, ω indicates the normal vector while ρ indicates the compensation parameter; v = (01) is a tradeoff parameter that controls the support vectors proportion in the training data set. Finally, ξi is a slack variable that allows some training set samples to be classified incorrectly. Such an optimization problem is handled by using the Lagrange multiplier, which can be found in detail with further reading. [37].

8. *Density-Based Method.:* This technique creates a statistical data estimation function depending on the training data set and can be used for both clustering and classification issues. Logistic
Regression (LR), Expectation- Maximization (EM), and Linear discriminant analysis (LDA) are examples of this category.
Logistic regression (LR) [38] is primarily used for binary classification problems. In the network traffic, the feature value e.g., payload size x = (1 100) is mapped to the likelihood of the predicting class (say *normal* or *abnormal*).Mainly, the algorithm aims to fit the whole data samples in the sigmoid curve. It shifts the line and then recalculates the likelihood until finds the maximum likelihood value. A primary drawback of the LR technique is in the case of having *complete separation* classes. i.e., some of the features may separate two classes; so, the binary function cannot be used to classify the data items.
EM is an iterative method that works very well with incomplete training data [39]. During the training phase, the parameters of the *likelihood function* are adjusted based on the data point of the previous iteration to optimize the likelihood function. This process will be repeated in the next iteration until the stopping conditions have been satisfied (e.g., the output difference is zero or stick in constant value). As the EM method enhances the accuracy of the classification model, the optimized IDS strategy [40] combines the EM method with a single class classifier to reduce the outlier of the training set in the pre-processing period in order to obtain more accurate classification results.
A more advanced approach called LDA [41] is useful for classification of dataset with multiple classes or features (*dimensions*). Data item x with f features can be plotted on space of n-dimensions in order to find a separation point, line, or plane, when n = 1, 2, 3 respectively.
Actually, the data could have more than three dimensions however calculating the separation plane immediately looks complicated. This issue was solved by LDA by reducing the datadimension complexity. In the case of binary classification, a new axis is created and all data points are projected onto the new axis. Regardless of the dimension of the feature, the data points of the two classes can be separated on the new axis. The new axis is formulated from the training data set by minimizing the combination within and between scatters, and maximizing the combination of data point means in each class. After all, the density-based only performs well when large size of training observations are available.
9. *Ensemble-Based Method.:* The ensemble classifiers idea is to learn a group of classifiers rather than only one, called an ensemble of classifiers, and then aggregate their predictions to classify a new input by using certain techniques, Voting, Stacking, Bagging, and Boosting are the most commonly used techniques. *Voting* is the simplest way to determine the final decision based on multiple voters. Votes are collected from a set of classifiers. The majority vote will be selected as a final result.
A more advanced version of voting is called *Stacking*. Rather than accepting a majority vote, a meta-learner is used to assert the best decision based on supervised knowledge. In other words, outputs from the first-level classifiers are loaded into the second-level learning process. The metalearner keeps training to optimize the final output.
[42].
Bagging and *Boosting*, on the other hand, depend on the distribution of the training dataset, because the combination of independent units can significantly enhance the effectiveness of the final prediction. *Bagging* or Bootstrap Aggregating acquires data from different baselines by employing the bootstrap sampling strategy [43]. Multiple base classifiers are integrated using the *voting* and *averaging* strategies to maximize the accuracy of the model.
Boosting [44] improves accuracy by creating a more powerful classifier using an existing weak single classifier. Assume that the data D distribution includes three parts $X1$, $X2$, and $X3$, and we only have a weak classifier that predicts only $X1$ and $X2$ correctly. Also, let $h1$ be the wrong classification of $X3$. To fix the error of $h1$, the boosting method derives

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

a new D' from D. For illustrations, the developer must focus mainly on the samples inside the $X3$ part and after that train the $h2$ classifier from D'. Assume that this new classifier classifies $X1$ and $X3$ correctly. Now, we can combine the $h1$ and $h2$ classifiers to provide a greater classifier. The procedure is iterated by adjusting the parameters of the distribution until no more enhancement is possible.

B. Clustering Algorithms

Clustering refers to the task of splitting the dataset into multiple clusters or groups, in which each cluster includes data patterns that are very relevant according to specific metrics. Clustering algorithms found in the previous research can be broadly classified into five distinct categories [45]:

- **Partitioning-based:** In these kinds of techniques, all clusters are immediately decided. Initial classes are defined and reallocated to a union. In other words, partitioning techniques divide data items into multiple partitions, where a certain partition shows a cluster. The following two conditions should be satisfied by these clusters: (1) each cluster should contain at least one object, and (2) each object should belong to exactly one cluster. There are various partitioning methods, such as PAM, CLARA, CLARANS, K-means, Kmedoids, K-modes, and FCM.
- **Hierarchical-based:** The data is arranged in a hierarchical structure, based on the intermediate nodes. The main output of the Hierarchical Clustering is called a dendrogram, where the individual data is presented by leaf nodes. The hierarchical clustering approach works by handling each data item as a separate cluster, and then it performs the following two steps iteratively: (1) Identify two clusters that are closest to each other, and (2) combine the two most similar clusters. This iterative step will performed until all clusters are combined. Some of the widely known methods in this category are CURE, ROCK, BIRCH, and Chameleon.
- **Density-based:** These techniques classify distinguishing data clusters/groups depending on the fact that a cluster in the data space is a contiguous area of high-density point, separated from other clusters by contiguous regions of a low density point. A cluster described here as dense, connected elements expands in any direction that leads to density. Objects in separating areas of a low-density point are generally considered to be outliers or noise. DBSCAN, DENCLUE, DBCLASD, and OPTICS are examples of methods that using the Density-based technique.

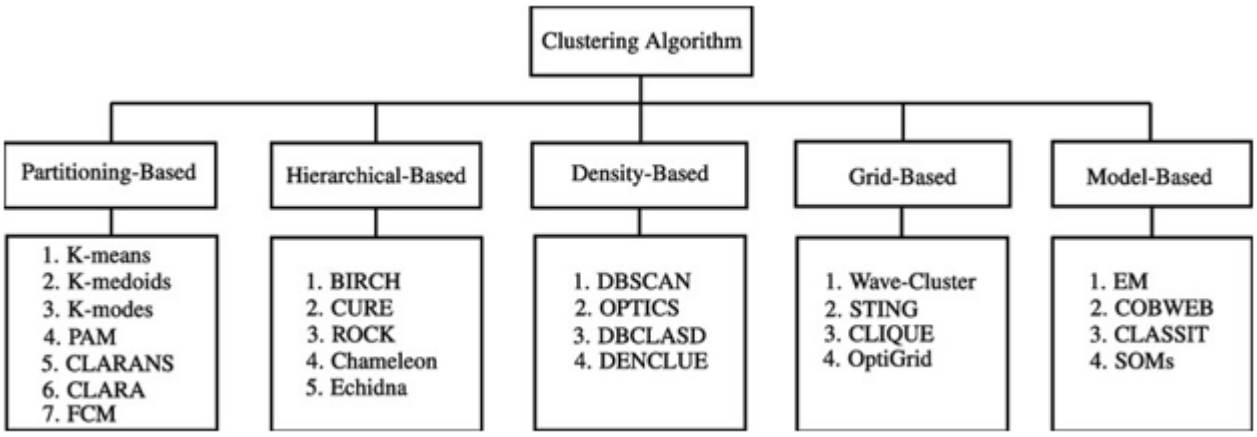

Figure 8: An overview of clustering algorithms

Grid-based: The data object space is split into grids. The key advantage of this technique is its low time complexity since it goes through the dataset once to calculate the statistical data of the grids. Cumulative grid-data helps to make grid-based clustering approaches independent of the number of data objects used for a uniform grid to capture local statistical values, and then execute clustering directly on the grid rather than the database. STING and Wave-Cluster are classic examples in this category.

- **Model-based:** This technique tries to optimize the fit between both the data and some mathematical (predefined) model. The data is represented as coming from a combination of probability distributions, each one shows a specific cluster. In other words, in model-based clustering, the data is assumed to be created by a combination of probability distributions

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

where each item shows a particular cluster. MCLUST, EM, COBWEB, and neural network approaches are some of the popular methods in this category. Figure 8 summarizes the clustering method classification regarding the five classes explained above.

IV. RELATED WORK

There are three types of algorithms in use today: supervised, unsupervised, and semi-supervised.

A. Techniques based on supervised learning

The set of all potential traffic flows and output data must be known in advance for supervised algorithms, such as [5],[60],[3]. Any unknown occurrence may be predicted based on this model's analysis of the usual values of these flows. Using supervised learning, new occurrences may be categorized into existing categories [3],[60]. We need to offer the machine with a large number of pre-selected instances that it can use to learn. A classification model is the outcome of the learning process. By analyzing a collection of inputs and then drawing conclusions, a categorization model is created. A model is built by mapping input characteristics to output classes, and this is a key goal of supervised learning. Categorization rules, decision trees, flowcharts, and other visual representations may all be used to represent newly obtained information. This information will be put to good use in the future when classifying fresh events. Training and evaluation are the two main components of supervised learning. An analysis of training data sets and a classification model are the initial steps in the training process. Testing, the second stage, is sometimes referred to as the classification phase. Using the model to analyze any unanticipated occurrences is the next step after training. In contrast to supervised machine learning, clustering algorithms employ internalized heuristics to generate clusters in a given dataset instead of example cases. Subheading Techniques that are overseen based on the KNN and DPC concepts of intrusion detection, this work proposes a new classifier called the DPNN. U2R attacks cannot be detected even if the experiment shows that DPNN is superior to other classifiers in terms of average accuracy and efficiency. KNN [48] has been demonstrated to be successful in overcoming difficulties with feature selection difficulty when employed as the evaluation function in [1]. Although their suggested technique outperformed the competition in terms of classification accuracy, they still have room to improve their initialization process. Many studies have shown that Support Vector Machines are an effective machine learning technique, such as the [46] which constructs a robust intrusion detection model that has high accuracy, training speed, false alarm rate, and detection rate for binary IDS problems but only for binary IDS. A new intrusion detection approach dubbed optimal allocation-based least square support vector machine was presented in [50]. The Least Square Support Vector Machine was used to sample this data (LS-SVM). [50] was tested using KDD 99, a standard database recognized as the de facto performance benchmark for machine learning systems. Static and incremental data are no problem for this method, which works well in both cases. The detection accuracy of a DoS attack was increased to 99.9998 percent using an intelligent detection and protection strategy reported in the journal [49] Additionally, this method's key drawback was that the suggested strategies' performance was only evaluated in a normal computer network simulation environment with a sorted dataset, rather than the real-world setting. Because of the time and money involved in gathering specialized expertise, categorizing and labelling large amounts of data is out of the question. However, if your model relies on love2002comparing, erroneous or noisy data labelling will have an impact on its performance.

B. Techniques based on unsupervised learning

To deal with the Irrelevant feature in [51] they used a hybrid strategy that incorporates Information Gain and Principal Component Analysis (Principal Component Analysis). MLP (multilayer perceptron) and SVM ensemble classifiers were employed in this study (Support Vector Machine). In terms of accuracy, detection rate, and detection time, this new approach is excellent. For certificate-free authentication in VANETs, [52] has been a successful and secure solution. They have a three-step process to ensure that their conversations are safe. When a vehicle registers with a trusted offline party, this is the beginning of the process. Authentication, the next step, verifies that the vehicle is really the one that was registered. Following this, a shared key is established via the Chinese Remainder Theorem (CRT). The results of the experiment show that their design is successful and able to meet the requirements for security. Unsupervised deep learning based on the Robust Software Modeling Tool (RSMT) was introduced by the creators of [53] which monitors and characterizes web application runtime behavior. Unsupervised and monitored approaches were used in the empirical investigation. Unsupervised ML tests fared poorly, however supervised RSMT highlighted concerns regarding computational cost (at the best case, around 0.728 for XSS using SVM). In addition, unsupervised

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

deep learning proved to be a useless endeavor (around 0.906). Despite the fact that unsupervised-based approaches have emerged as a promising method for learning anomaly detection systems from unlabeled data, they have low accuracy and efficiency.

C. Techniques Under Some Form of Supervision

Research on semi-supervised methods, despite the fact that most academics have concentrated on ML strategies to improve anomaly detection accuracy, is still restricted. [54] offers a multilevel semi-supervised intrusion detection approach based on the KDDcup99 dataset in order to solve issues about non-identical distribution and balance. There was a lack of flexibility in the selection of hyper-parameters, therefore they only used a single data set to test their MSML framework. A novel semi-supervised feature grouping strategy based on the linear correlation coefficient and the cuttlefish algorithm was presented in [55] A considerable increase in detection rates, accuracy, and false-positive rates was discovered. With a high detection rate of 95.23 percent, an accuracy of 95.03 percent and a low false- positive rate of 0.4The authors of [56] present a hybrid system for high dimensional data anomaly detection that combines a K-NNG ensemble with a deep autoencoder (DAE). With the help of the nonlinear mapping strategy, the DAE was able to reduce a large dataset into a more manageable amount of data by initially learning just the important unlabeled data characteristics in an unsupervised mode. After that, random subsets of the given dataset were used to build nonparametric K-NN classifier ensembles. Experiments on a variety of real world datasets show that the suggested strategy improves the accuracy of anomaly detection while simultaneously reducing computer complexity. Semi-supervised anomaly identification using a newly devised multivariate statistical network monitoring algorithm and Partial Least Squares is presented in this study. Supervised and unsupervised aspects of this system were combined to provide the best results. As with rule-based systems, it provides a machine learning system capable of updating harmful behavior patterns (zero-days). The findings of a traffic-based system experiment show that the proposed strategy can be implemented in a real-world setting. SemiSupervised Detection of Outliers (SSDO) was recently revealed by researchers in [58] which is used to monitor water consumption in supermarkets using time series data. SSDO offers two stages for semi-supervised anomaly detection. Calculation of an anomaly score begins with clustering. Enduser labels are collected using an active-learning mechanism. After receiving labels, the model enters a semi-supervised phase in which the labels guide clustering. The suggested strategy was put to the test on a variety of different datasets. Using just a limited number of labelled data points, our method was able to outperform both unsupervised and current semi-supervised methods. It was recommended by [59] to use a semi-supervised approach to identify DDoS assaults. The unsupervised mode reduces the number of false positives by eliminating irrelevant data via the use of information gain ratio, entropy estimation, and co- clustering. Extra Trees ensemble classifiers are used in the supervised mode to classify traffic flows. 98.23.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

Table III: Most recent studies of related works.

References		Techniques	Dataset used	Results	Demerit
[46]	Supervised	SVM ensemble with feature augmentation.	NSL-KDD	High accuracy. High detection rate. Low false alarm rate.	It works only with the binary case of IDS problems.
[47]		Combine k -nearest neighbors (KNN) with density peaks clustering (DPC).	KDD-Cup 99	High average accuracy. Testing time is reduced. The efficiency increased by 20.688%.	It cannot detect U2R attack.
[48]		1- Four effective Candidate Solution Generation Strategy (CSGS) are used in self-adaptive differential evolution (SaDE) 2- K-Nearest Neighbor (KNN) for feature selection.	KDDCUP99	Efficient to solve the IDS problems. About 57% of the features reduced. High classification accuracy.	We can make more improvements in the initialization phase.
[49]		1- BP Neural Networks. 2- Dynamic defense strategy based on game theory.	KDDCUP99	High DoS detection accuracy (99.998%). High DoS detection rate (99.998%).	High computational complexity
[50]		Least Square Support Vector Machine (LS-SVM).	KDD 99	High classification accuracy (CA).	NA
[51]	Unsupervised	1- IG (Information Gain) combined with PCA(Principle component analysis). 2- IBK(Instance-based learning algorithms). 3- MLP (multilayer perceptron). 4- SVM (Support Vector Machine).	ISCX 2012 NSL-KDD Kyoto 2006 +	(In the ISCX 2012 dataset): Accuracy rate (99.01%), DR (99.1%). FAR (0.01%). (In NSL-KDD and Kyoto 2006 + datasets): Accuracy rate (98.24% and 98.95%). DR (98.2% and 99.8%). FAR(0.017% and 0.021%).	Poor performance. Poor accuracy. Not efficient.
[52]		1- Certificate-less authentication technique 2- Chinese Remainder Theorem (CRT).	RSU, where the aggregated traffic data are unlabeled VANETs data.	Provide efficient authentication and can deliver the desired security properties.	
[53]		Robust Software Modeling Tool (RSMT).	They created several test applications and synthetic trace datasets.	RSMT produce overhead issues. Supervised ML procedure was poor (0.728). Unsupervised deep learning performance was not efficient (0.906).	
[54]	Semi-Supervised	Multilevel semi-supervised machine learning.	KDDcup99	High accuracy (99.3%). Improve the F1-score.	The choice of hyper-parameters was not flexible enough.
[55]		linear correlation coefficient combined by cuttlefish algorithm.	KDD-Cup 99	High accuracy (95.03%). High detection rate (95.23%). Small false-positive rate (1.65%).	The training time was high.
[56]		Hybrid semi-supervised anomaly detection model combined: Deep autoencoder (DAE). Ensemble k-nearest neighbor graph (K-NNG).	Opportunity activity recognition (OAR). Gas sensor array drift (GAS) MiniBooNE particle identification dataset (MPID). KDD 2008.	Enhances the anomaly detection accuracy. Reduces the computational complexity.	NA
[57]		Multivariate Statistical Network Monitoring (MSNM). Partial Least Squares (PLS).	UGR'16	Have the flexibility to update (zero-day) attack. Consider the first-time application of sparse methodologies in intrusion detection. Proved the practical applicability.	NA
[58]		Semi-Supervised Detection of Outliers (SSDO). Active-learning strategy.	Several datasets from four stores provided by the Belgian retailer Colruyt.	The approach outperformed the competing techniques.	NA
[59]		Entropy estimation. Co-clustering. Information gain. Extra Trees ensemble classifiers.	NSL-KDD UNB ISCX 12 UNSW-NB15	High accuracy reported (98.23%, 99.88%, and 93.74%). Low false-positive rate.	Complex analysis of unlabeled data. Single classification algorithm.

The existing semi-supervised-based intrusion detection algorithms are a cost-effective solution since domain experts are not required to label training data sets in their entirety. But the accuracy and efficacy of these procedures are not as high as they may be. We will build an IDS on top of existing solutions to reduce the requirement for domain experts to identify millions of data flows. The goal of this design is to decrease the number of false positives while also increasing the accuracy.

V. CONCLUSION

Cyber attackers have demonstrated their capability to obscure their identities, mask their activities, protect their identities isolated from illegal profits, and use infrastructure that is resistant to compromise. Thus, advanced intrusion detection systems design to detect modern attacks are becoming increasingly vital for computer systems to be secured. It is essential to have a comprehensive understanding of the strengths and limitations of existing IDS research in order to design and build robust IDS systems. To do so,

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

we have presented in this paper a comprehensive recent review of intrusion detection system approaches based on machine learning, including: IDs methods types, and technologies, along with their advantages and disadvantages.

REFERENCES

1. Clarence Chio and David Freeman. Machine Learning and Security: Protecting Systems with Data and Algorithms. " O'Reilly Media, Inc.", 2018.
2. Adil Al-Harthi. Designing an accurate and efficient classification approach for network traffic monitoring. PhD thesis, RMIT University, 2015.
3. Ayman Taha and Ali S Hadi. Anomaly detection methods for categorical data: A review. ACM Computing Surveys (CSUR), 52(2):38, 2019.
4. Dhruba Kumar Bhattacharyya and Jugal Kumar Kalita. Network anomaly detection: A machine learning perspective. Chapman and Hall/CRC, 2013.
5. Bradley C Love. Comparing supervised and unsupervised category learning. Psychonomic bulletin & review, 9(4):829–835, 2002.
6. Douglas M Hawkins. Identification of outliers, volume 11. Springer, 1980.
7. Stefan Axelsson. Intrusion detection systems: A survey and taxonomy. Technical report, Technical report, 2000.
8. Jakapan Suaboot, Adil Fahad, Zahir Tari, John Grundy, Abdun Naser Mahmood, Abdulmohsen Almalawi, Albert Y Zomaya, and Khalil Drira. A taxonomy of supervised learning for idss in scada environments. ACM Computing Surveys (CSUR), 53(2):1–37, 2020.
9. S Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462, 2015.
10. Wojciech Tylman. Scada intrusion detection based on modelling of allowed communication patterns. In New Results in Dependability and Computer Systems, pages 489–500. Springer, 2013.
11. Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. Machine learning, 29(23):131–163, 1997.
12. Kyriakos Stefanidis and Artemios G Voyiatzis. An hmmbased anomaly detection approach for scada systems. In IFIP International Conference on Information Security Theory and Practice, pages 85–99. Springer, 2016.
13. Roman Klinger and Katrin Tomanek. Classical probabilistic models and conditional random fields. Citeseer, 2007.
14. Daesung Moon, Hyungjin Im, Ikkyun Kim, and Jong Hyuk Park. Dtb-ids: an intrusion detection system based on decision tree using behavior analysis for preventing apt attacks. The Journal of supercomputing, 73(7):2881–2895, 2017.
15. Shengyi Pan, Thomas Morris, and Uttam Adhikari. Developing a hybrid intrusion detection system using data mining for power systems. IEEE Transactions on Smart Grid, 6(6):3104–3113, 2015.
16. Rishabh Samdarshi, Nidul Sinha, and Paritosh Tripathi. A triple layer intrusion detection system for scada security of electric utility. In 2015 Annual IEEE India Conference (INDICON), pages 1–5. IEEE, 2015.
17. Jiawei Han, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
18. Manish Mehta, Rakesh Agrawal, and Jorma Rissanen. Sliq: A fast scalable classifier for data mining. In International conference on extending database technology, pages 18–32. Springer, 1996.
19. John Shafer, Rakesh Agrawal, and Manish Mehta. Sprint: A scalable parallel classifier for data mining. In Vldb, volume 96, pages 544–555. Citeseer, 1996.
20. Bing Liu, Wynne Hsu, Yiming Ma, et al. Integrating classification and association rule mining. In KDD, volume 98, pages 80–86, 1998.
21. William W Cohen and Yoram Singer. A simple, fast, and effective rule learner. AAAI/IAAI, 99(335-342):3, 1999.
22. Wenmin Li, Jiawei Han, and Jian Pei. Cmar: Accurate and efficient classification based on multiple class-association rules. In Proceedings 2001 IEEE international conference on data mining, pages 369–376. IEEE, 2001.
23. Peter Clark and Tim Niblett. The cn2 induction algorithm. Machine learning, 3(4):261–283, 1989.
24. Zhiwen Pan, Salim Hariri, and Youssif Al-Nashif. Anomaly based intrusion detection for building automation and control networks. In 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), pages 72 –77. IEEE, 2014.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

25. Pedro Silva and Michael Schukat. On the use of k-nn in intrusion detection for industrial control systems. In Proceedings of The IT&T 13th International Conference on Information Technology and Telecommunication, Dublin, Ireland, pages 103 –106, 2014.
26. Bo Tang and Haibo He. A local density-based approach for outlier detection. Neurocomputing, 241:171 –180, 2017.
27. Zubair Shah, Abdun Naser Mahmood, Mehmet A Orgun, and M Hadi Mashinchi. Subset selection classifier (ssc): a training set reduction method. In 2013 IEEE 16th International Conference on Computational Science and Engineering, pages 862–869. IEEE, 2013.
28. MF Schilling. Mutual and shared neighbor probabilities: Finite-and infinite-dimensional results. Advances in Applied Probability, pages 388–405, 1986.
29. Corinna Cortes and Vladimir Vapnik. Support-vector networks. Machine learning, 20(3):273–297, 1995.
30. Bisyron Wahyudi Masduki, Kalamullah Ramli, Ferry Astika Saputra, and Dedy Sugiarto. Study on implementation of machine learning methods combination for improving attacks detection accuracy on intrusion detection system (ids). In 2015 International Conference on Quality in Research (QiR), pages 56–64. IEEE, 2015.
31. Ahmed Patel, Hitham Alhussian, Jens Myrup Pedersen, Bouchaib Bounabat, Joaquim Celestino J´unior, and Sokratis Katsikas. A nifty collaborative intrusion detection and prevention architecture for smart grid ecosystems. Computers & Security, 64:92–109, 2017.
32. Alecsandru Patrascu and Victor-Valeriu Patriciu. Cyber protection of critical infrastructures using supervised learning. In 2015 20th International Conference on Control Systems and Computer Science, pages 461–468. IEEE, 2015.
33. R Vijayanand, D Devaraj, and B Kannapiran. Support vector machine based intrusion detection system with reduced input features for advanced metering infrastructure of smart grid. In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), pages 1–7. IEEE, 2017.
34. Moti Markovitz and Avishai Wool. Field classification, modeling and anomaly detection in unknown can bus networks. Vehicular Communications, 9:43–52, 2017.
35. Wei Li. Using genetic algorithm for network intrusion detection. Proceedings of the United States department of energy cyber security group, 1:1–8, 2004.
36. Fernando PA Lima, Anna DP Lotufo, and Carlos R Minussi. Disturbance detection for optimal database storage in electrical distribution systems using artificial immune systems with negative selection. Electric power systems research, 109:54–62, 2014.
37. Xien Liu, Mengjun Li, Yuanlun Sun, Xiaoyan Deng, et al. Support vector data description for weed/corn image recognition. Journal of Food, Agriculture and Environment, 8(1):214–219, 2010.
38. David G Kleinbaum, Lawrence L Kupper, Azhar Nizam, and Eli S Rosenberg. Applied regression analysis and other multivariable methods. Nelson Education, 2013.
39. Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.
40. Hyunguk Yoo and Taeshik Shon. Novel approach for detecting network anomalies for substation automation based on iec 61850. Multimedia Tools and Applications, 74(1):303–318, 2015.
41. Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468), pages 41– 48. Ieee, 1999.
42. Joseph Sill, G´abor Taka´cs, Lester Mackey, and David Lin. Feature-weighted linear stacking. arXiv preprint arXiv:0911.0460, 2009.
43. Bradley Efron and Robert J Tibshirani. An introduction to the bootstrap. CRC press, 1994.
44. Zhi-Hua Zhou. Ensemble methods: foundations and algorithms. CRC press, 2012.
45. Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Sebti Foufou, and Abdelaziz Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3):267–279, 2014.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 05 Issue 06 June 2022
DOI: 10.47191/ijcsrr/V5-i6-47, Impact Factor: 5.995
IJCSRR @ 2022
www.ijcsrr.org

46. Jie Gu, Lihong Wang, Huiwen Wang, and Shanshan Wang. A novel approach to intrusion detection using svm ensemble with feature augmentation. Computers & Security, 2019.
47. Lixiang Li, Hao Zhang, Haipeng Peng, and Yixian Yang. Nearest neighbors based density peaks approach to intrusion detection. Chaos, Solitons & Fractals, 110:33–40, 2018.
48. Yu Xue, Weiwei Jia, Xuejian Zhao, and Wei Pang. An evolutionary computation based feature selection method for intrusion detection. Security and Communication Networks, 2018, 2018.
49. Lijun Gao, Yanting Li, Lu Zhang, Feng Lin, and Maode Ma. Research on detection and defense mechanisms of dos attacks based on bp neural network and game theory. IEEE Access, 7:43018–43030, 2019.
50. Enamul Kabir, Jiankun Hu, Hua Wang, and Guangping Zhuo. A novel statistical technique for intrusion detection systems. Future Generation Computer Systems, 79:303–318, 2018.
51. Fadi Salo, Ali Bou Nassif, and Aleksander Essex. Dimensionality reduction with ig-pca and ensemble classifier for network intrusion detection. Computer Networks, 148:164–175, 2019.
52. Haowen Tan, Ziyuan Gui, and Ilyong Chung. A secure and efficient certificateless authentication scheme with unsupervised anomaly detection in vanets. IEEE Access, 6:74260–74276, 2018.
53. Yao Pan, Fangzhou Sun, Zhongwei Teng, Jules White, Douglas C Schmidt, Jacob Staples, and Lee Krause. Detecting web attacks with end-to-end deep learning. Journal of Internet Services and Applications, 10(1):1– 22, 2019.
54. Haipeng Yao, Danyang Fu, Peiying Zhang, Maozhen Li, and Yunjie Liu. Msml: A novel multilevel semi-supervised machine learning framework for intrusion detection system. IEEE Internet of Things Journal, 6(2):1949– 1959, 2018.
55. Sara Mohammadi, Hamid Mirvaziri, Mostafa Ghazizadeh-Ahsaee, and Hadis Karimipour. Cyber intrusion detection by combined feature selection algorithm. Journal of information security and applications, 44:80–88, 2019.
56. Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang. A hybrid semi-supervised anomaly detection model for high-dimensional data. Computational intelligence and neuroscience, 2017, 2017.
57. Jos´e Camacho, Gabriel Maci´a-Fern´andez, Noem´ı Marta Fuentes-Garc´ıa, and Edoardo Saccenti. Semi-supervised multivariate statistical network monitoring for learning security threats. IEEE Transactions on Information Forensics and Security, 14(8):2179–2189, 2019.
58. Vincent Vercruyssen, Meert Wannes, Verbruggen Gust, Maes Koen, B¨aumer Ruben, and Davis Jesse. Semi-supervised anomaly detection with an application to water analytics. In Proceedings/IEEE International Conference on Data Mining. IEEE, 2018.
59. Mohamed Idhammad, Karim Afdel, and Mustapha Belouch. Semi-supervised machine learning approach for ddos detection. Applied Intelligence, 48(10):3193–3208, 2018.
60. Stephen Marsland. Machine learning: an algorithmic perspective. Chapman and Hall/CRC, 2014

Cite this Article: Basmah Alsulami, Abdulmohsen Almalawi, Adil Fahad (2022). A Review on Machine Learning Based Approaches of Network Intrusion Detection Systems. International Journal of Current Science Research and Review, 5(6), 2159-2177