International Journal of Computer Sciences and Engineering
Vol.**11**, Issue **1**, pp.**01-10**, January **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Review Paper

A Review on Machine Learning Intrusion Detection Systems (MLIDS) in Encrypted Traffic

K.R. Harinath[1*], G. Kishore Kumar [2]

[1]Dept. of Computer Science and Engineering, JNTUA, Andhra Pradesh, India
[2]Dept. of Computer Science and Engineering, RGM College of Engineering and Technology, Andhra Pradesh, India

[*]*Corresponding Author: harirooba007@gmail.com, Tel.: +91-9966458007*

Received: 20/Dec/2022; **Accepted**: 02/Jan/2023; **Published**: 31/Jan/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i1.110

Abstract— Global connection depends on the internet and protecting it is a top priority for organizations and governments. As technology advances, so does the number of different types of network attacks. These attacks can be considered as intrusions. Due to deficiency of protection the information protection becomes onerous. To detect intrusions, a well defined intrusion detection system was utilized. It is one among the tools towards building secure system. To combat with advanced attacks and to protect the data and network, MLIDS (Machine Learning Based Intrusion Detection systems) is an advanced technology among best solutions. When accesses are encrypted, however, IDS is ineffective. Although encryption increases sender and receiver privacy, it causes an issue with inaccurate traffic categorization. There are Several ID approaches to analyse encrypted traffic interchange using data range, data similarity and data time without decryption. In this survey, paper presents a different techniques, datasets and challenges of detection over cipher text and comparative survey on machine learning algorithms from recent work.

Keywords— IDS, Encryption, Encrypted traffic, datasets, intrusions.

1. Introduction

IDS is generally used to detect intrusions (or) any other malicious attacks in the network and also used to test the traffic in networks. Intrusion detection system [18] exists as local for singular systems or exists at definite point to protect network in group of systems. The Intrusion detection system exists as local is called as Host-based intrusion detection system (HIDS) and Intrusion detection system exists in network called as Network based intrusion detection system (NIDS).

In the case of encrypted traffic IDS, it have been able to analyze decipher data is easier for HIDS. But HIDS is partial to a confined limited vision. It is not able to identify threats that expand over the network. Encryption is a problem that frequently solved by providing decryption keys.

Another method to classify IDS is by how it detects attacks. Commonly IDS classified into Signature based and Anomaly detection based. [15] The payload is examined and the findings are compared to known assaults in Signature based detections. It is generic method, since the findings are compared with familiar attacks, signature based technique is effectual for similar notorious attacks, even though it is not capable of dealing with new attacks. Anomaly detection based approach [2] analyses network traffic to express the known typical network traffic patterns and if a significant difference was found, that could be labeled as malevolent. The detection technique is little general since it adjusts the standard traffic representation to the learned connection, resulting in the specified adoptive model that accurately represents definite behaviour. Anomaly detection based approach was little precise than signature based techniques, however they can identify zero-day attacks, which are novel forms of assaults.

Payload is becoming less relevant as encrypted traffic becomes increasingly frequent. Because for the most part of intrusion detection method [16] depends upon scrutinizing the load of communications, decrypting of encrypted data becomes a common option. As a result, the issue with encrypted traffic IDS are reduced to the common obstacles that every IDS faces. Different procedures can be used to obtain decrypted data.

The paper planned as follows section 2 discusses about classification of IDS, section 3 conveys about performance metrics, section 4 depicts about MLIDS encrypted traffic, section 5 conveys about IDS datasets followed by conclusion and reference.

2. Related Work

In a network, [19] an Intrusion Detection System (IDS) is created to identify vulnerabilities by analyzing packets passed via it. IDS's are used by both inside and outside intruders to detect odd and hazardous activities. Large network traffic levels and unequal distribution of data are difficulties that IDS must cope with. The primary responsibility of IDS is to monitor data sources such as computers and networks for intrusion attempts. IDS's gather data from a wide variety of systems and network providers then analyse it for dangers. Network intrusion detection systems (NIDS) and host-based intrusion detection systems are two types of IDSs (HIDS).

Classification of Intrusion Detection Systems based on Information Source

2.1 Network Based IDS:

An Intrusion detection system identifies the vulnerabilities like message spreading and security issues. This intrusion detection detects the main source of the intrusion, which could be the machine or unauthorized individuals and their attempts. Basically an Intrusion is a Network based activity. NIDS is one of the classifications of Intrusion Detection systems [14] which acts as the detection tool to detect the intrusions through network Based method. Here network based means it has a server with an inbuilt system to detect and removes intrusions if possible; otherwise it verifies the data and sends it to destination node securely. The Network based system which has inbuilt server system has the benefit of installing the system over different hosts in the network with less complexity. With NIDS, the server installed with the system is sufficient other than installed in other hosts of the network. But, regular updates are necessary to the server.

Network intrusion detection system [14] observes and understands network connection by interpretation of individual packets from end to end network layer and transport layer. The system conducts search for any suspicious activity or network based attack such as Denial of Service (DoS) attack, port scans etc. An alert can be sent to the system administrator, when anomalous network traffic behavior is identified.

To identify intrusion on a network [24], increase the confidentiality and integrity of both receiver and sender. The packets are evaluated in this step, and several features are used as categorization parameters. The computer is trained and evaluated with various data in order to differentiate between harmful, legitimate encrypted and unencrypted packets. Any unnecessary information for any parameter may be quickly discovered and an intrusion warning issued. Thus, classifying data based on packet properties rather than content allows for easier and faster intrusion detection while simultaneously preserving secrecy and security.

2.2 Host Based IDS:

For more than two decades, the host-based intrusion detection systems (HIDS) have been gathered steam in cyber security field. HIDS features finer precision and the capacity to identify vulnerabilities when contrasted to network-based intrusion detection systems (NIDS). HIDS examines operating system monitoring data, whereas NIDS examines traffic on a network. [13]. HIDS detects illegal behavior by monitoring host operations such as framework or shell logs. A HIDS's execution speed is usually calculated by adding all of the training and testing times together. The HIDS's performance [28] may be inhibited while it must manage a high number of fine-grained system call paths. On host accounting data, HIDS [12] may use different data mining methods, such as artificial neural networks, to detect threats. HIDS scans and analyses the individual host to record the changes in the activities which are performed in the host. Activities like it include incoming and outgoing packets, system calls, files consisting of different applications, logs in the host. If any vulnerabilities are observed it alerts the system administrator to provide security from those attacks.

According to [12], HIDS can be classified into two types. They are abused HIDS and Anomaly-based IDS. The activities against predefined signatures of the host based systems are identified by misuse HIDS. Most of the Anti-Virus software's comes under this category. Signature-based HIDS can identify intrusions based on known patterns, but they cannot actively detect new attacks that occur regularly. To overcome this abnormal patterns detection it requires Anomaly HIDS.

3. Methodology

Classification of IDS based on Analysis Strategy:

Hybrid intrusion detection system implemented by using Meta heuristic algorithms. It is a network-based technique that organizes network protocols into sixty-four clusters using the K-means algorithm. It then divides these clusters into two groups: a. normal occurrences and b. abnormal occurrences. [19] The clustering process helps us to create meaningful network event classifiers. The second step of this method takes the network events at this point, which are referred to as data points, and educates a supervised detection model using the SVM algorithm. The SVM predictive model [10] is a detection model for finding abnormalities in fresh events. UNSW-NB15 is a labelled data set that is publicly available and used to assess a specific detection system.

In figure 1 broad classification of IDS is portrayed. Generally IDS systems classified into two groups based on Analysis Strategy. They are Signature-based Intrusion Detection System (SIDS) and Anomaly-based Intrusion Detection System (AIDS).

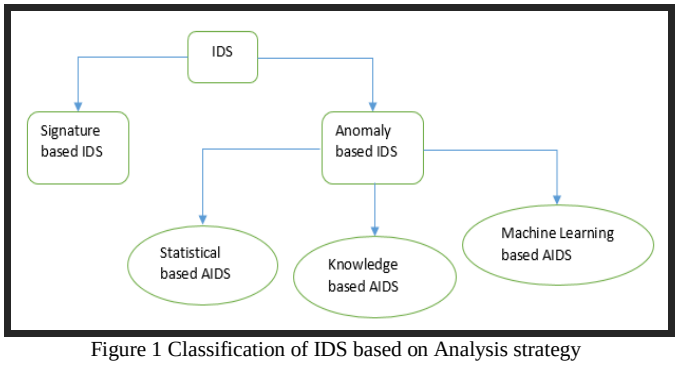

Figure 1 Classification of IDS based on Analysis strategy

3.1 Signature-Based intrusion detection system (SIDS):
Using pattern matching techniques, these systems can recognize known vulnerabilities. As a result, we might refer to them as Knowledge-based Detection Systems [1].The SIDS used pattern matching methods which means when an intrusion occurs, that can be compared with earlier intrusions which can be recorded in the signature database. If the signature of new intrusion matches with the earlier intrusion which exists previously, then the system alerts with an alarm signal.

SIDS has very good accuracy in detecting the intrusions based on previous signatures, but it is poor in detecting zero day attacks because these attacks have no matching signatures and considered as new attacks. SIDS are less effective due to lack of matching signatures, where the rate of zero-day attacks are increased [1].

3.2Anomaly-Based intrusion detection system (AIDS):
Using various methods, the system can be created as a normal model [5] of behaviour. If the system's behaviour differs significantly from what is expected, this malevolent conduct will be classified as intrusion. To develop an AIDS, training phase and testing phase was used. To learn normal performance of the system, a regular traffic contour was used in training phase, and then to find invisible intrusions and to establish the capacity of system, a new dataset is used in the testing phase [32].
According to the type of dispensation connected to the ‘‘behavioural’’ model of the aiming system, anomaly detection techniques can be classified into three main categories [8] statistical based, knowledge-based, and machine learning-based.

3.2.1 Statistical-based Anomaly based IDS techniques:
The action of network traffic was recorded and an outline was created which represents the dynamic behaviour of these techniques [33]. The outline was based on different criteria which includes traffic rate, the quantity of packets for every protocol, the rate of associations, the number of dissimilar IP addresses, etc. In anomaly detection, it can be considered two datasets of network traffic. One is to represent the profile which is currently observed and second one is statistical profile which is formerly trained. When an event occurred in the network, the behaviours of anomaly estimated prescribed score and present profile will be compared. Generally the attainments indicate the grade for deformity for precise event. The IDS will alert the incongruity system when the score reaches or cross the specified threshold value. The Statistical-Based AIDS differentiated as Univariate, Multivariate and Unauthorized user models. Here Univariate model considers both Network and Host based IDS as combined IDS where the parameters as independent Gaussian random variables. By this they define acceptable range for each variable. Multivariate models [33] have been defined followed by Univariate model .In this approach it defines the relationship between two or more metrics which were projected. Unauthorized users specified time series models for detecting metrics from various sources. This model includes interval timer, event counter which are used to note the time and values observed.

The major gain of Statistical-Based AIDS was it does not necessitate any previous knowledge for regular activity. When suspicious behaviours are detected, however, it sends out a warning with detailed information.
Although, this Statistical-Based AIDS have some disadvantages. Vulnerable to be trained by an attacker and it is very difficult to set parameters and metrics based on systems.

3.2.2 Knowledge-based techniques:
The Knowledge-based approaches have knowledge of different sources which assists to solve complex problems. [34] It has different techniques for various issues, among them expert system approach (Knowledge-based**)** is the mainly used loom which analyses and make decisions for specific domain problems. This expert system involves three steps to categorize audit data. They are:

a) Identification of various attributes and modules from training records.
b) Presumption of cataloguing regulations ,parameters or procedures
c) Accordingly assessment statistics were classified

It has different subtypes such as Finite state machine to specify states and transitions, depiction languages to specify N-grams, UML etc., professional systems to specify rule-based classification.

- The major advantage of this approach to detect anomaly are robustness and flexibility.
- It is complicated and lingering to construct a premium knowledge system/data by probing training data.

3.2.3 Machine Learning-based A-NIDS schemes:
Machine learning techniques [35] were allowed to develop the applications to improve accuracy and to predict the results without coding. They consider previous data as input to expect new outcomes. Machine learning schemes [2] also focused on constructing models. But these models increases accuracy and performance, based on executed outcomes. To update themselves, these techniques have different strategy which implements in many situations but the resources are very expensive and this consider as major drawback. Mainly Machine learning-based schemes were divided into Supervised and Unsupervised learning methods.

i) Supervised learning-based IDS techniques:
Data labelling [23] is the method of recognizing various facts (images, text files, videos, etc.) and adding more significant and enlightening labels to offer circumstance so as to machine learning sculpt can be trained from it. With the aid of this labelled training data, supervised learning-based IDS detects intrusions. Generally these techniques have two stages, namely training and testing.

In training stage, to define an algorithm different classes and data features are identified. By analysing these algorithms, the labelled output values which consist of data sources and various output data are named as intrusions or normal class. Later, by applying feature selection unwanted features will be

deleted. In this training phase, the classifier will be trained to understand the relationships between input and labelled output values.

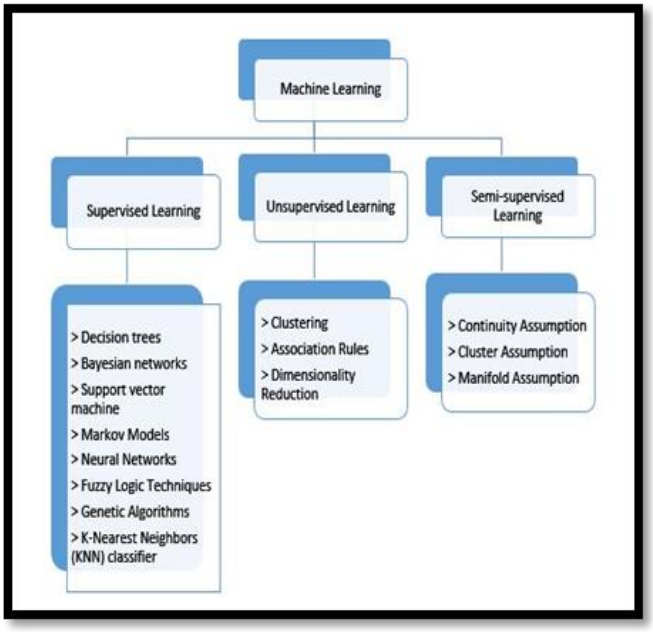

Figure 2 Machine Learning Techniques

In the testing phase, intrusions or normal class are classified by using trained model. The classifier will create a model with feature values to predict the source of input data. Figure 2 highlights the different Machine Learning techniques.

Different classification models are categorized using these algorithms. These classification models are created, but not only to manage the training records, but also defined to detect a new class of records with more accuracy. Some of the classification models are:

Decision trees: A Decision tree [3] can predict the class or value of the target system from prior knowledge. Based on some specific criterion, the data can be split continuously. The data and its properties are analysed by three elements in a decision tree. Decision node is the first and basic element which identifies the testing feature. Branch is the second one, which specifies the predicted decision from the test feature value. Leaf is the third component that consists of instance class. Dynamic datasets are handled more efficiently with decision trees.

Bayesian networks: Bayesian networks [17] are generally used to detect intrusions with the help of statistical schemes. These techniques can be used to provide probabilistic relations between groups of selected variables, to predict the further actions and integrate both data with prior knowledge. The advantage of this Bayesian networks is, they can work effectively to predict the results by analyse the behaviour of target systems in many situations and also used to detect errors, but the foremost drawback of this network is, the outcomes may be related to threshold based systems. So, they need further computational endeavour.

Support vector machine (SVM): SVM is a supervised learning algorithm that applies mainly for categorization problems in machine learning. [17] N-dimensional space can be categorized into several classes by creating a good line (or) decision boundary. In future, from these classes, a new data point will set in the classification. The decision boundary will be considered as best and also called as hyper plane. Support vector machines were defined by splitting hyper plane. Kernel function used to classify intrusions linearly. To classify intrusions, Kernel function maps training data. During the training phase of a support vector machine, feature selection is used to segregate various input points into specified classes.

Markov Models: Markov models are commonly used to model systems, their transitions, to recognize patterns. There are different types of Markov models. According to [17] related survey we have to discuss about two main models of Markov models. They are Markov chains and hidden Markov models.

The Markov chains model [25] used by autonomous systems which have fully observed states. This group of states are inter-related with transition probabilities to derive their model capabilities. These probabilities are predicted from general behaviour of the scheme. Anomaly exposure will be probably derived by comparing the incongruity score which are obtained from threshold value. The hidden Markov models used by autonomous systems which have partially observed states which mean the states and transitions will be hiding. Only productions will be observed.
These models are widely used for IDS (host& Network) which provides a good approach to predict the outcomes in the target systems.

Neural Networks: Neural Networks or Artificial Neural networks [9] are a series of methods which specifies different algorithms to simulate the operations of human brain. Due to its flexibility and adaptability it can be adopted in the area of anomaly intrusion detections. This loom used to create profiles, to predict next instructions, to detect intrusions based on models defined by patterns. However, there is no explicit explanation of the specific detection decisions.

Fuzzy Logic Techniques: Fuzzy logic is extracted as of Fuzzy [22] set presumption which handles imprecise data, estimated relatively than accurately inferred from conventional predicate reason. These strategies were used as Fuzzy variables in anomaly detection. The processing scheme is based on the existence of behaviour of the statistics i.e., if the data lies in the given interval, the observation will be normal. Even though Fuzzy logic techniques are more efficient to inquire network ports, they consume more resources and also become more controversial due to its logical implementations.

Genetic Algorithms: Genetic algorithms [18] are search based algorithms, a classification of evolutionary algorithms. These genetic algorithms are specified as machine learning based algorithms which are used to generate good quality

solutions for optimization and searching problems by replicating the procedure of natural selection, reproduction and mutation. It can anticipate the answer without assuming prior information or behaviour of the system, but it demands a lot of resources.

K-Nearest Neighbours (KNN) classifier: K-Nearest Neighbours [18] classifier is the one of the classification methods of Supervised Learning based techniques which applied as a representative quantile classifier in machine learning. The basic concept of this classifier identifies an unlabeled data taster to the class and named that sample of its k nearest neighbours. This algorithm presumes the correlated properties among the novel case/data and existed cases and put the original case into the class that is chiefly analogous to the offered categories.

ii) Unsupervised learning:

To create a group of unlabelled datasets as a cluster, to analyse the concept and behaviour of those unlabelled datasets [36] Unsupervised Learning algorithms are used. Without human intervention unsupervised learning algorithms identify the unknown patterns. Mainly these models are employed for three different tasks such as Clustering, Association & dimensionality.

A) Clustering:

Clustering and outlier detection: Clustering [24] is a simple strategy that observes data, which has similar qualities and grouped them as a cluster. It's a method of sorting unlabelled data into groups based on similarities and differences. By analysing the information consists of patterns, the new and unclassified data objects are processed as groups. Each cluster has respective points selected by using clustering methods.

There will be some suspicious points which are not related to any cluster. These types of suspicious points are called as outliers and considered as anomalies. By using raw audit data, Clustering detect intrusions with less effort. These clustering algorithms classified into exclusive and overlapping clustering.

Exclusive clustering is a type of grouping which is also known as “hard clustering”. It specifies only one cluster metrics. The K-means clustering technique was the best example for exclusive clustering.

K-means Clustering: In this process K number of clusters [37] is created that depends on each centroid group distance. To these K clusters different data points assigned and are clustered under the specified centroid category, if they are closest to it. A larger ‘K’ value with high granularity is indicatory of smaller groupings, however smaller ‘K’ value with low granularity is of larger groupings. K-means clustering mostly used in various areas like document clustering, Market segmentation, image segmentation, image compression etc.,

Overlapping Clustering: The type of clustering [24] varied from exclusive clustering. It categorizes data points which have discrete degree of memberships and allows them to multiple clusters.

Hierarchical Clustering: Hierarchical clustering [38] is an unsupervised clustering algorithm, which is also called as Hierarchical cluster analysis. It is secluded into two approaches. They are Agglomerative and Divisive.

Agglomerative Clustering is a process that follows bottom-up approach. At first, the data points are secluded as distinct groupings. Later based on common patterns and similar properties, these data point groupings are combined repeatedly until to achieve one cluster. To calculate and observe these common properties, four different methods are used.

- Wards Linkage defines distance among clusters by amplify in the addition of squared behind the clusters were amalgamated.
- Typical linkage defined by mean distance connecting two points in every group.
- The utmost distance involving any two points in any cluster is defined as entire linkage, moreover recognized as maximum linkage.
- A single linkage, also known as a minimum linkage, is the shortest distance amid two points in a cluster.

For all above distances, the most commonly used metric is Euclidean distance using the below formula

$$D_e = \left(\sum_{i=1}^{n} (p_i - q_i)^2 \right)^{1/2} \quad (1)$$

Divisive Clustering is a top-down approach. A based on [38] similarity between data points, as well as with their difference, the single data cluster is divided. Generally divisive clustering is not used, but considered in the specified context of hierarchical clustering. Dendogram, it is a tree representation, used to visualize the clustering process.

Probabilistic Clustering: It's an unsupervised learning method for solving soft clustering issues, such as cluster density estimation. The process of probabilistic clustering is based on the similar properties of data points that belongs to particular categorization are clustered. The Gaussian Mixture Model (GMM) is the one of the widely used probabilistic clustering methods. These GMM were categorized into mixture models that developed by probability distribution functions which are not specified. In GMMs, we assume appropriate existence of variables to cluster data points, because in general the variables are unknown.

B) Association Rules:

Association Rules were used to detect variables and their associations in specific datasets. These are rule-based [38] methods which are commonly used to analyse market strategies among various products and called as Market Basket Analysis. One of the most well-known algorithms for dealing with association rules is the Apriori algorithm.

C) Dimensionality Reduction:
In general, [38] to get best results we need more data and this more data makes visualization of datasets as more complex. In some cases we called it as over fitting. To reduce the size of data inputs and preserves the dataset integrity there is a technique called Dimensionality Reduction. This technique can be used for large number of dimensions (or) features for given datasets. Principal component analysis, singular value decomposition, Auto encoders are different dimensionality reduction methods.

(iii) Semi-supervised Learning:
It is one of classifications in machine learning algorithms [4] which considered as median of Supervised and Unsupervised learning algorithms. In training phase, combo of labelled (few labels) and unlabelled datasets were used. When dealing with unlabelled datasets, this is an effective strategy for developing relationships between items. When things are closer together, they have a common group or label, according to continuity assumption. In semi-supervised learning, boundaries are thin(less density) and plain which are added by decision boundaries. As per cluster assumption, data is classified into various clusters and the output label shared by data points within the cluster. Data occurs on multiples of particular dimensions as compared to the input, according to manifold assumption. To make use of distances and densities, Manifold assumption helps.

The working process of semi-supervised learning [4] is training the data model, to attain accurate results. The labelled and unlabeled data, where both are with pseudo labels were linked. After linking, if the results are not accurate, then these labelled and unlabeled data are trained. This process continues to decrease error rate and increases the accuracy rate of the model. Some of the uses of Semi-supervised learning algorithms were text document classifier, speech analysis, protein sequence classification and web content classification.

4. Performance Metrics

Performance metrics [1] are classified as many classes for IDS. Positive class and negative class are the two-class classifier which evaluates IDS performance. Positive class denotes attack and Negative class denotes no attack.

From this the general performance metrics are defined as:
- **True Positive (TP)** defined as the result exists where the system exactly estimates the attack.
- **True Negative (TN)** is the result where the system exactly estimates the no attack.
- **False Positive (FP)**is the result where the system wrongly estimates the attack.
- **False Negative (FN)** is the result where the system wrongly estimates the no attack.

From the above statements, we have some standard performance measures. They are:
- True positive rate = TP/ (TP+FN)
- False positive rate = FP/ (FP+TN)
- False negative rate = FN/ (FN +TP)
- Accuracy = (TP + TN) / (TP + TN + FP +FN)

Here Total no. of attacks = TP + FN
Total no. of normal instances = FP + TN

5. MLIDS in Encrypted Traffic

A unique technique to detect anomaly for encrypted online connections is presented in the publication [21]. This method of intrusion detection uses encrypted traffic analysis, which examines the stuffing of encrypted communication via simply data range and time exclusive of decryption. The classification begins by extracting data from encrypted traffic, i.e., collection of data extent and time for every consumer application. Second, accesses were differentiated in sequence consistency, and access probabilities are determined. Eventually, harmful activity is recognized using criteria derived as of the regularity of visits and HTTP traffic patterns. Pre-mechanisms were not required, and isolation was not breached, despite the fact that the technique uses a network analysis attack over security mechanisms. The technique relies on anomaly detection was based on the occurrence of comparable accesses and the features of typical HTTP accesses, although the situation is considerably more difficult to identify than in traditional anomaly detection. They tested the proposed system's accuracy using a real dataset collected at an access point as well as a DARPA IDS assessment dataset. The stated system, according to [21], identifies a variety of threats, however if the information is encrypted.

Provide a seven-step process for deep-learning-based traffic categorization that includes Problem Definition, Data Collection, Dataset Pre-Processing, Feature selection, Model Selection, Training and Validation, and Periodic Evaluation/Update. Stronger Encryption Protocols, Multi-Label Categorization, Middle Flow Classification, Zero-Day Applications, Domain Adaptation, and Multi-Task Learning are among the outstanding difficulties in traffic classification [42]. Figure 3 conveys network classifier and their steps for categorizing the network traffic.

Based on machine learning, a system was proposed to identify BT (Bit Torrent) traffic among the mainly prominent and troublesome P2P applications [38]. To avoid the difficulties stated, this method [38] is autonomous of the both IP addresses and payload content. This specified system relies solely on quantitative metrics derived from traffic patterns and LS-SVMs. Thirty flow-based statistical characteristics are derived from the flows, and using a feature selection method, the most efficient feature set is determined. With this method, an overall accuracy of 93.6 percent was reached. Forward session lengths, backward volumes and packet dimensions, forward inter-arrival period, normal and active time, and their modifications are the most efficient characteristics for recognizing BT traffic, according to the testing results [38]. Where the DPI approach fails to recognize encrypted communication, our technique performs better.

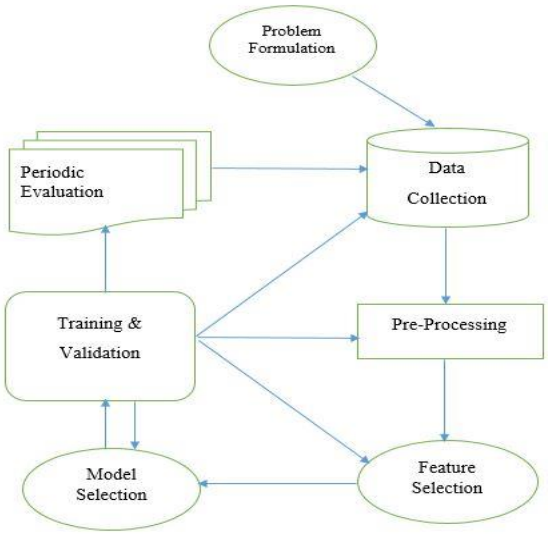

Figure 3 A frameworks for creating a network classifier

Unlike other systems, the design of novel structure does not require a learning period, a sophisticated setup, or knowledge of the service to be protected [39]. The architecture makes it possible to identify assaults regardless of connectivity, IP addresses, encryption scheme, server setup, accuracy, or the type of attack used. As a result, the design [39] may be utilized in a variety of ways without requiring any special configuration.

According to T. Kovanen et al. [25], encrypted traffic IDS employs a method that necessitates decryption of traffic prior to IDS analysis. Reverse engineering apps or the target's protocol stack can be used to access decrypted communications. Their approach replicated traffic to a CIDS that was capable to decrypt it and did deep packet inspection on it. They employed VPN and Shamir's secret contribution technique. The vulnerable congregation issue has been resolved. But, because this approach necessitated decryption, it is only applicable to a restricted number of network topologies.

For encrypted traffic IDS systems that need decryption, comparable detection approaches as normal IDS may be utilized. As a result, the correctness of their solutions isn't the primary concern here, as the IDS's accuracy is unrelated to encryption. Different sorts of behaviours and characteristics may be discerned from encrypted communication without decryption, according to the [25] analysis. The alternatives which do not need decryption are much more appealing since they might make use of the techniques utilized in passage scrutiny and encrypted traffic class. The feasibility of traffic analysis approaches for detecting activities on speedy connections have been investigated.

Demonstrates a machine learning method depending on 3 randomness tests (Entropy, Chi-square, and Arithmetic mean) [40]. Utilized a total of 193,647 packets to evaluate performance (163,462 plain and 30,185 encrypted). To discover the majority appropriate cataloguing approach to spot encrypted traffic, the packets were tested with four machine learning-based categorization methods—Nave Bayesian, Support Vector Machine, CART, and AdaBoost—the essential entropy-based recognition technique. The entropy-based detection approach was greatly surpassed by machine learning-based detection techniques. Recommends using CART, not only generate the maximum correctness grades (0.997 in F-measure) except the most effective in testing (about 0.560 s). However, when the amount of the dataset grows, throughput will alter.

Declares a generic model for encrypted traffic classification in DL-based mobile services and reviews classification job formulation, data preparation, pre-processing, model input design, pre-training design, and model architecture [41]. In addition, certain significant challenges such as class imbalance and semi-supervised learning-based traffic categorization are shown and explored.

The technique presented is very well-organized and effectual for intrusion detection on encrypted system data, as mutually the recognition model and the connected data being kept secret [43]. As an intrusion regulation, we used a decision tree that is analysed secretly across network data utilizing homomorphic encryption methods. This analysis shows that by making adjustments several specifications such as the list of regulations, features, and the no. of classifications includes partitioning of features, security and privacy can be improved, and the proposed discretization method was useful barely for escalating effectiveness of completion time excluding for improving solitude. Because [43] the protocol is quick, scalable, and parallelizable, the execution speeds imply that it can be employed in real-world applications. However, just a few detecting models and classification techniques were used in this strategy.

Deep packet scrutiny strategies frequently rely on evaluating the packet payloads content towards report particular actions, are used in traditional traffic monitoring procedures, including intrusion detection and prevention systems, ensuing in underprivileged adaptability for encrypted data.[44] offer an accurate signature-based intrusion detection system intended for encrypted network traffic . The work was broken down into two sections: (i) signature creation and (ii) edifice of the intrusion detection system. The network packet data, rigorously collected from packet headers, is used to construct the signatures, which is a strategy designed for encrypted networks. They [44] choose OpenCL for the engine implementation since it allows for consistent execution between GPU processors and an elevated processing unit. However, additional network packet metadata is not included in the signature generation step of this method.

6. Intrusion Detection Datasets

A Dataset improves the capability of detecting abnormal behaviour of the system, which plays key role for any IDS methodology. Abnormal behaviour can be referred as Intrusive behaviour. The availability of datasets is based on their usage i.e., few datasets are restricted due to some privacy issues when they are used in commercial products to

analyse network packets and some datasets are available publicly and considered as standard datasets. They are DARPA, KDD, and NSL-KDD, ADFA-LD etc.

6.1 DARPA/KDD Cup99:
It is the earlier dataset created by DARPA (Defence Advanced Research project Agency) in 1998. The created dataset named as KDD 98 (Knowledge Discovery & Data mining) dataset, which contributes IDS research. It is the most commonly panned dataset. The US Air Force LAN was modelled using these datasets. In Third International Knowledge Discovery & Data mining Tools competition KDD cup 99 used [7], which is derived from 1998 DARPA. The PCAP encoded records that make up the DARPA IDS [22] assessment dataset comprise several weeks of traffic.

6.2 CAIDA:
The dataset collection [23] was obtained in 2007 and comprises connected traffic patterns from widespread Distributed Denial of Service (DDoS) attacks. This sort of DoS assault tries to disrupt regular traffic on a processor or network by flooding it with large number of network packets, preventing valid users from accessing intended target. A downside of CAIDA is the lack of variety in the assaults. Another disadvantage is the data does not include predictions from the entire network, it's impossible to tell the difference between aberrant and usual traffic flows.

6.3 NSL-KDD:
It's a broad dataset derived from a prior collection of KDD cup99 datasets. The systematic swot of the cup99 dataset exposed noteworthy flaws to facilitate has a major impact on intrusion detection accuracy and contributes to a skewed evaluation of AIDS. [30] NSL-KDD was created using the KDD Cup99 dataset to remedy a problem with a large number of identical packets by removing duplication. There are 1, 25,973 files in the training dataset and 22,544 records in the testing dataset. The volume of the dataset was large enough that it was feasible to utilize the entire dataset devoid of case at haphazard. As a consequence, the outputs of diverse study projects have been uniform and comparable.

6.4 ISCR 2012:
Authentic network traffic [29] traces were studied in ISCX 2012 dataset to detect standard computer behaviour from regular traffic of various protocols. It is built on labelled real-world network traffic and includes a variety of types of attack.

6.5 ADFA-LD and ADFA-WD:
Two datasets namely ADFA-LD and ADFA-WD [30] representing structure with implementation of modern threats were published as standard datasets with scholars at the Australian Defence Force Academy. These datasets were developed via the analysis of system call oriented HIDS and comprise entries both from Linux and Windows operating systems. Because a few of the intrusion cases in ADFA-LD were originated commencing novel zero-day attacks, the specified dataset was well suited to demonstrate the distinctions between SIDS and AIDS techniques to detect the intrusions. System call patterns of many forms of intrusions are also included in ADFA-LD. The ADFA-WD is a current windows dataset that may be used to evaluate HIDS.

6.6 CICIDS 2017:

CICIDS 2017 [26] includes information on both innocuous activity and novel intrusions. The data records, destination address, IPs, protocols, and various intrusions are all used to label CICIDS dataset. The precise dataset was serene via a comprehensive connected architecture that includes modems, firewalls, switches, routers, and nodes with various operating systems. From the collected data traffic, the precise dataset comprises 80 net flow features.

6.7 BoT-IoT:
The BoT-IoT dataset [27] be urbanized in the UNSW Canberra Cyber Range Lab with creation of a replicated communication network. It has a mixture of regular and botnet traffic in the internet architecture. The root files for the dataset were accessible in a assortment of forms, along with the actual Pcap files, produced argus files, and CSV files. The records were split by attack scenario and Sub division to aid in the tagging process. Table 1 shows the comparison of different datasets.

Table 1. Comparison of Dataset

Dataset	Realistic Traffic	Label data	IoT traces	Zero-day attacks	Full packet captured	Year
DARPA 98	Yes	Yes	No	No	Yes	1998
KDDCUP 99	Yes	Yes	No	No	Yes	1999
CAIDA	Yes	No	No	No	No	2007
NSL-KDD	Yes	Yes	No	No	Yes	2009
ISCX 2012	Yes	Yes	No	No	Yes	2012
ADFA-WD	Yes	Yes	No	Yes	Yes	2014
ADFA-LD	Yes	Yes	No	Yes	Yes	2014
CICIDS2017	Yes	Yes	No	Yes	Yes	2017
Bot-IoT	Yes	Yes	Yes	Yes	Yes	2018

7. Conclusion and Future Scope

An Effective IDS is capable of detecting contemporary malware are becoming increasingly important for the protection of computer systems. To develop and implement such IDS systems, an exhaustive understanding of the advantages and weaknesses of current IDS research is required. In this study, we've detailed an overview of IDS and Machine learning-based IDS techniques, kinds, and technologies, as well as their pros and cons when it comes to encrypted traffic.

The most often used public datasets for IDS research were also explored. Successful detection systems must be capable to properly make out a diversity of threats, counting assaults that utilize avoidance strategies. Designing IDSs capable of defeating evasion strategies in encrypted networks is still a big issue in this field.

****No Funding Agents are involved in the preparation of Manuscript.**

**** The authors have no conflict of interest with the journal editorial board members and the suggested reviewers.**

References

[1]. Khraisat, A., Gondal, I., Vamplew, P., &Kamruzzaman, J. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, Vol.**2**, Issue.**1**, pp.**1-22**, **2019**.

[2]. Gulla, K. K., Viswanath, P., Veluru, S. B., & Kumar, R. R. (2020). Machine learning based intrusion detection techniques. In *Handbook of computer networks and cyber security*. Springer, Cham, pp.**873-888, 2020**.

[3]. Kumar, G. K., Viswanath, P., & Rao, A. A. (2011). Intrusion Detection Using an Ensemble of Decision Trees. In *IICAI*, pp.**382-392, 2011.**

[4]. Van Engelen, J. E., &Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, Vol.**109**, Issue.**2**, pp.**373-440, 2020.**

[5]. Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., &Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, Vol.**28**, Issue.**1-2**, pp.**18-28, 2009**.

[6]. Chiba, Z., Abghour, N., Moussaid, K., &Rida, M. (2019). Intelligent approach to build a Deep Neural Network based IDS for cloud environment using combination of machine learning algorithms. *Computers & Security*, *86*, pp.**291-317, 2019.**

[7]. Siddique, K., Akhtar, Z., Khan, F. A., & Kim, Y. (2019). KDD Cup 99 data sets: a perspective on the role of data sets in network intrusion detection research. *Computer*, Vol.**52**, Issue.**2**, pp.**41-51, 2019.**

[8]. Kumar, V., Srivastava, J., &Lazarevic, A. (Eds.). (2006). *Managing cyber threats: issues, approaches, and challenges*. Springer Science & Business Media, Vol.**5, 2006**.

[9]. Kim, H., Kim, J., Kim, Y., Kim, I., & Kim, K. J. (2019). Design of network threat detection and classification based on machine learning on cloud computing. *Cluster Computing*, Vol.**22**, Issue.**1**, pp.**2341-2350, 2019.**

[10]. Hatef, M. A., Shaker, V., Jabbarpour, M. R., Jung, J., &Zarrabi, H. (2018). HIDCC: A hybrid intrusion detection approach in cloud computing. *Concurrency and Computation: Practice and Experience*, Vol.**30**, Issue.**3**, e4171, **2018.**

[11]. Sultana, N., Chilamkurti, N., Peng, W., &Alhadad, R. (2019). Survey on SDN based network intrusion detection system using machine learning approaches. *Peer-to-Peer Networking and Applications*, Vol.**12**, Issue.**2**, pp.**493-501, 2019.**

[12]. Hu, J. (2010). Host-based anomaly intrusion detection. In *Handbook of information and communication security*. Springer, Berlin, Heidelberg, pp.**235-255, 2010.**

[13]. Liu, M., Xue, Z., Xu, X., Zhong, C., & Chen, J. (2018). Host-based intrusion detection system with system calls: Review and future trends. *ACM Computing Surveys (CSUR)*, Vol.**51**, Issue.**5**, **1-36**, **2018.**

[14]. Ahmed, M., Pal, R., Hossain, M. M., Bikas, M. A. N., &Hasan, M. K. (2009, April). NIDS: A network based approach to intrusion detection and prevention. In *2009 International Association of Computer Science and Information Technology-Spring Conference*. IEEE, pp.**141-144, 2009.**

[15]. Singh, R., Kalra, M., & Solanki, S. (2020). A hybrid approach for intrusion detection based on machine learning. *International Journal of Security and Networks*, Vol.**15**, Issue.**4**, **233-242, 2020.**

[16]. Çavuşoğlu, Ü. (2019). A new hybrid approach for intrusion detection using machine learning methods. *Applied Intelligence*, Vol.**49**, Issue.**7**, pp.**2735-2761, 2019.**

[17]. Ray, S. (2019, February). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, pp.**35-39, 2019.**

[18]. Chaudhari, R. R., &Patil, S. P. (2017). Intrusion detection system: classification, techniques and datasets to implement. *Int. Res. J. Eng. Technol.(IRJET)*, Vol.**4**, Issue.**2**, pp.**1860-1866, 2017.**

[19]. Aljamal, I., Tekeoğlu, A., Bekiroglu, K., &Sengupta, S. (2019, May). Hybrid intrusion detection system using machine learning techniques in cloud computing environments. In *2019 IEEE 17th international conference on software engineering research, management and applications (SERA)*. IEEE, pp.**84-89, 2019.**

[20]. Wahyudi, B., Ramli, K., &Murfi, H. (2018). Implementation and analysis of combined machine learning method for intrusion detection system. *International Journal of Communication Networks and Information Security*, Vol.**10**, Issue.**2**, pp.**295-304, 2018.**

[21]. Yamada, A., Miyake, Y., Takemori, K., Studer, A., &Perrig, A. (2007, May). Intrusion detection for encrypted web accesses. In *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*. IEEE, Vol.**1**, pp.**569-576, 2007.**

[22]. Li, J., Qu, Y., Chao, F., Shum, H. P., Ho, E. S., & Yang, L. (2019). Machine learning algorithms for network intrusion detection. *AI in Cybersecurity*, pp.**151-179, 2019.**

[23]. Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, pp.**381-386, 2020.**

[24]. Thaseen, I. S., Poorva, B., &Ushasree, P. S. (2020, February). Network intrusion detection using machine learning techniques. In *2020 International conference on emerging trends in information technology and engineering (IC-ETITE)*. IEEE, pp.**1-7, 2020.**

[25]. Kovanen, T., David, G., &Hämäläinen, T. (2016). Survey: Intrusion detection systems in encrypted traffic. In *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Springer, Cham, pp.**281-293, 2016.**

[26]. Koch, R., &Rodosek, G. D. (2010, September). Command evaluation in encrypted remote sessions. In *2010 Fourth International Conference on Network and System Security*. IEEE, pp.**299-305, 2010.**

[27]. Sharafaldin, I., Lashkari, A. H., &Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, *1*, pp.**108-116, 2018.**

[28]. Creech, G. (2014). *Developing a high-accuracy cross platform Host-Based Intrusion Detection System capable of reliably detecting zero-day attacks* (Doctoral dissertation, University of New South Wales, Canberra, Australia). **2014.**

[29]. Shiravi, A., Shiravi, H., Tavallaee, M., &Ghorbani, A. A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *computers& security*, Vol.**31**, Issue.**3**, pp.**357-374, 2012.**

[30]. Tavallaee, M., Bagheri, E., Lu, W., &Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE symposium on computational intelligence for security and defense applications*. Ieee, pp.**1-6, 2009.**

[31]. Hendry, G. R., & Yang, S. J. (2008, March). Intrusion signature creation via clustering anomalies. In *Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2008*. International Society for Optics and Photonics. Vol.6973, p.69730C, **2008.**

[32]. Butun, I., Morgera, S. D., &Sankar, R. (2013). A survey of intrusion detection systems in wireless sensor networks. *IEEE communications surveys & tutorials*, Vol.**16**, Issue.**1**, pp.**266-282, 2013.**

[33]. Ye, N., Emran, S. M., Chen, Q., &Vilbert, S. (2002). Multivariate statistical analysis of audit trails for host-based intrusion

detection. *IEEE Transactions on computers*, Vol.**51**, Issue.**7**, pp.**810-820, 2002.**
[34]. Walkinshaw, N., Taylor, R., & Derrick, J. (2016). Inferring extended finite state machine models from software executions. *Empirical Software Engineering*, Vol.**21**, Issue.**3**, pp.**811-853, 2016.**
[35]. Dua, S., & Du, X. (2016). *Data mining and machine learning in cybersecurity*. CRC press. **2016.**
[36]. Celebi, M. E., & Aydin, K. (Eds.). (2016). *Unsupervised learning algorithms*. Berlin: Springer International Publishing. **2016.**
[37]. Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, pp.**80716-80727, 2020.**
[38]. SeyedTabatabaei, T., Adel, M., Karray, F., &Kamel, M. (2012, July). Machine learning-based classification of encrypted internet traffic. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, Berlin, Heidelberg. pp.**578-592, 2012.**
[39]. Koch, R., Golling, M., &Rodosek, G. D. (2014). Behavior-based intrusion detection in encrypted environments. *IEEE Communications Magazine*, Vol.**52**, Issue.**7**, pp.**124-131, 2014.**
[40]. Cha, S., & Kim, H. (2016, August). Detecting encrypted traffic: a machine learning approach. In *International Workshop on Information Security Applications*. Springer, Cham. pp.**54-65, 2016.**
[41]. Wang, P., Chen, X., Ye, F., & Sun, Z. (2019). A survey of techniques for mobile service encrypted traffic classification using deep learning. *IEEE Access*, *7*, pp.**54024-54033, 2019.**
[42]. Rezaei, S., & Liu, X. (2019). Deep learning for encrypted traffic classification: An overview. *IEEE communications magazine*, Vol.**57**, Issue.**5**, pp.**76-81, 2019.**
[43]. Karaçay, L., Savaş, E., &Alptekin, H. (2020). Intrusion detection over encrypted network data. *The Computer Journal*. Papadogiannaki, E., & Ioannidis, S. (2021). Vol.**63**, Issue.**4**, pp.**604-619, 2020.**
[44]. Acceleration of intrusion detection in encrypted network traffic using heterogeneous hardware. *Sensors*, *21*(4), 1140.

AUTHORS PROFILE

Mr. K. R. Harinath is born in Nandyal, Andhra Pradesh,India in the year 1989. He is the Research Scholar of JNTUA, Ananthapuramu, University. He completed his Post Graduation M.Tech in CSE from SKD Engineering College,JNTUA and B.Tech in IT from MACET,JNTUA. He is currently working as Assistant Professor in the Department of CSE,RGM college of Engineeringand Technology,Nandyal,AP. He is a Lifetime member in IAENG. He has published 5 papers at international journal and conference. His area of interest includes Machine Learning. Email: harirooba007@gmail.com

Dr.G.Kishore Kumar is born in Nandyal,AndhraPradesh,India in the year 1980.He received Ph.D from JNTUA University.He Completed his Post Graduation M.Tech in CSE from JNTUA and B.Tech in CSE from RGMCET.He is currently working as Associate Professor in the Department of CSE,RGM College of Engineering and Technology,Nandyal,AP.He is a member in IEEE, IEAE. He has published 21 papers at international journal and conference.His area of interest includes Machine Learning and Data Mining. Email: kishorgulla@yahoo.co.in

