Machine learning -based intrusion detection with encrypted data (IOT).

Sarat Sai Bammidi, Masters in Information Systems Security Concordia University, Montreal, Canada.

E-mail: sai.sarat2@gmail.com

***Abstract*— The Internet of Things (IoT) idea refers to "things" as physical devices that are capable of communication. It introduces a range of readily available, dependable, and necessary activities and services. The IoT demands comprehensive security measures that put communication first and prioritise its protection by confidentiality, integrity, and authentication services. Data inside sensor nodes must be encrypted, and the network must be protected from disruptions and assaults. It is necessary to find a solution for the problem of communication security in an IoT network. Cyberattacks are still a possibility even though the IoT network is secured by encryption and authentication. Therefore, having an intrusion detection system (IDS) technology is essential. The common and prospective security risks to the IoT environment are examined in this article. Then, a review of IoT IDSs is provided in relation to the methodologies, datasets, and machine learning (ML) algorithms, based on comparing and evaluating recent studies in the field of IoT intrusion detection. The merits and weaknesses of contemporary IoT intrusion detection strategies are evaluated in this study. Additionally, contemporary datasets gathered from actual or simulated IoT environments are studied, high-performing ML methods are found, and the gap in contemporary studies is noted.**

***Keywords*-Packet headers, anomaly intrusion detection, supervised machine learning algorithms, Internet of thing, Intrusion detection system techniques, Intrusion detection system datasets.**

1 INTRODUCTION

With the development of the Internet, a lot of sensitive information, such as financial transactions, are conveyed through this medium. Network security is now a very significant area of informatics. This necessitated the deployment of encryption technologies and protocols in order to protect this content. The most popular encryption-based communication protocols include SSL/TSL, IPsec, SSH, and others [1].

The Internet of Things (IoT) is a smart network that uses well-established protocols to connect physical objects to the Internet. Smart, small sensors connect devices wirelessly in an IoT network. Without human intervention [2], IoT devices can communicate with one another. It communicates with them, adds more objects, and works together with them to create new applications and services using special addressing mechanisms.

Organizations typically employ a variety of mechanisms to encrypt their data and uphold privacy. The complexity of protocols, volume, and encryption have all created significant difficulties for the IDS system in identifying hostile activities offers security remedies for nefarious assaults or security lapses. It is a piece of hardware or software that keeps a computer system secure by spotting malicious activities. Two well-liked IDS categories that have been extensively employed to give security solutions are the signature-based intrusion detection system (SIDS) and the anomaly-based intrusion detection system (AIDS). A typical NIDS examines the headers and payloads of packets to detect malicious or unusual traffic behavior and report it. Although the data conveyed in packet payloads is encrypted, the only information that makes sense in encrypted packets1 are (i) TLS handshake packets and (ii) TCP/IP packet headers. Therefore, it appears that even well-known intrusion detection technologies do not sufficiently analyze encrypted communications.[3][4] or instance, Snort's SSL Readme page alerts users about this while investigating port 443.

1.1 BACKGROUND

A form of security system called machine learning-based intrusion detection with encrypted data employs machine learning algorithms to find attempts at unauthorized system access while encrypting sensitive data to safeguard user privacy. Data is analyzed in plain text form in typical intrusion detection systems, leaving it open to interception by attackers. Data is encrypted before analysis in machine learning-based intrusion detection, increasing data security.

These systems machine learning algorithms can be trained to spot trends in encrypted data that point to an intrusion attempt. These algorithms can also be utilized to differentiate between typical and anomalous system behavior, enabling the early identification and mitigation of security concerns.

As the volume of sensitive data being communicated online keeps increasing, the usage of encrypted data in machine learning-based intrusion detection systems has grown in significance in recent years. This method allows for good intrusion detection while adding an extra degree of protection to ensure that data is kept private.

1.2 PROBLEM STATEMENT

Attacks with no signatures in databases, such as novel types of attacks or zero-day attacks, cannot be detected by signature-based detection. The size of the signature database continuously expanding is another important aspect. Machine learning (ML) techniques, which are an enhancement in artificial intelligence

(AI), have been employed in recent years to enhance IoT intrusion detection systems (IDS). To support the creation of IoT IDSs, numerous publications, including [5][6], analyzed and contrasted various applicable ML algorithms and approaches using various datasets. However, it's still unclear which ML model was more effective for developing an effective IoT IDS based on a recent dataset gathered from an IoT environment. Therefore, it is currently necessary to conduct an updated evaluation in order to pinpoint these crucial issues. The SIDS can detect and stop similar assaults from happening in the future, even though they cannot stop every intrusion based on previously discovered symptoms of penetration. SIDS is almost impossible to use to identify intruders given the exponential rise in cyberattacks and attackers' use of sophisticated methods to mask attack patterns. However, there are so many problems in intrusion detection systems, we mainly concentrate on review of IOT intrusion detection systems using the supervised Machine learning.

1.3 PROJECT OBJECTIVE

- This paper provides a survey of IoT IDSs. The purpose of this research is to advance our understanding of the characteristics (motive and capabilities) of IoT cyber assaults.
- Additionally, the report provides an overview of recent studies on IoT intrusion detection utilizing machine learning algorithms for IoT networks based on datasets. Techniques and evaluation metrics to find the most accurate ML algorithm for IoT intrusion detection and the latest IoT dataset.
- The advantages and disadvantages of the hybrid, anomaly-based, signature-based, and specification-based types of IDSs approaches are then contrasted.

This paper is set up like follows: The most frequent cyberattacks in the IoT environment are described in section 2. The benefits and drawbacks of IoT intrusion detection methods are described in section 3. The performance measures, datasets, and supervised ML algorithms used in current IoT intrusion detection research was addressed, analyzed, and compared in Section 4. Section 5 presents the paper discussion, further work and contribution. Section 6 presents the conclusion.

2 IOT CYBER ATTACKS

Cyberattacks on IoT (Internet of Things) targets include those on wearables, industrial control systems, smart home devices, and other items that are a part of the IoT ecosystem. These assaults take advantage of flaws in the hardware and software of these devices, giving attackers access without authorization, the ability to steal data, or the ability to modify the targets.

The below are some common IOT cyber-attacks:

1)Physical Attacks: These attacks alter hardware components. The majority of IoT devices frequently operate outside, where they are very vulnerable to physical attacks [7].
2)The unlawful identification of systems, services, or vulnerabilities is a component of attacks referred to as reconnaissance. An example of a reconnaissance attack is the scanning of network ports [8].
3) Denial-of-service (DoS): This kind of attack seeks to deny access to a computer or network resource to the users who are being targeted. Due to their constrained memory and compute resources, the majority of IoT devices are vulnerable to resource enervation attacks.
4) When unauthorized users get access to networks or devices that they are not authorized to use, access assaults occur. There are two different kinds of access attacks: the first is a hacker gaining physical access to an actual thing. The second method is remote access through IP-connected hardware [9].
5) Attacks on privacy: The amount of information that is easily available through remote access techniques has made IoT privacy security increasingly challenging.
6)Cyber-crimes: With the use of the internet and smart products, users and data are utilized for hedonistic acts like fraud, brand theft, identity theft, and theft of intellectual property.
7) Destructive attacks: Exploiting space results in extensive disruption, damage to property, and loss of human life. Among destructive assaults, terrorism and retribution are two instances.
8) Supervisory Control and Data Acquisition (SCADA) Attacks: As active devices in real-time industrial networks, SCADA systems are connected to industrial IoT networks, enabling the remote monitoring and control of processes even when the devices are situated in faraway locations. Eavesdropping, man-in-the-middle, masquerade, and malware are the most particular and frequent types of SCADA assaults [10].

3. IoT INTRUSION DETECTION SYSTEM

A security tool called an IoT intrusion detection system (IDS) watches and examines network traffic in order to find and stop unauthorised access to IoT networks and devices. It is a software-based system that detects possible security threats and takes appropriate action using a variety of technologies, including machine learning, anomaly detection, and signature-based detection [11].

3.1 IoT Intrusion Detection Types

There are categorised into three types

- **Host-Based IDS(HIDS):** It is a kind of intrusion detection system that is deployed on certain hosts or endpoints, like servers or workstations, in order to monitor their actions and find potential security concerns.

- **Network-based IDS (NIDS):** It is a kind of security system that keeps an eye on network traffic for any

unusual activity and tries to identify and stop any security risks.

- **Distributed IDS (DIDS)**: It is a kind of intrusion detection system made to track network activity and find potential security risks on various devices or places. In contrast to conventional IDS systems, which are normally installed on a single host or network segment, DIDS is made up of a network of sensors that are dispersed throughout the network and collaborate to detect and analyse network traffic.

3.2 IOT Intrusion Detection Techniques

There are basically four types of methodologies for deploying the IOT intrusion Detection.

- Anomaly based IDS in IoT: An IDS (Intrusion Detection System) for the Internet of Things (IoT) that bases its detection of potential security threats on anomalous or unexpected behaviour patterns in IoT network traffic is known as an anomaly-based IDS [12].
- Signature based IDS in IoT: An IoT (Internet of Things) security solution known as a signature-based IDS (Intrusion Detection solution) uses pre-defined signatures or patterns of known attacks to identify potential security vulnerabilities in IoT network traffic.
- Specification based IDS in IoT: An IoT (Internet of Things) security solution called a specification-based IDS (Intrusion Detection solution) uses pre-established standards or rules to find potential security threats in IoT network traffic.
- Hybrid IDS in IoT: In order to provide comprehensive and efficient security for IoT networks, hybrid IDS (Intrusion Detection System) in IoT (Internet of Things) networks integrates different detection methodologies, such as signature-based, anomaly-based, and specification-based IDS.

4. SUPERVISED ML BASED IOT INTRUSION DETECTION

Computer systems can now predict events more accurately without explicit training thanks to machine learning (ML). It belongs to the category of artificial intelligence (AI). ML algorithms predict new output values using historical data as input. Reinforcement learning, unsupervised learning, and supervised learning are the three basic categories used to classify machine learning algorithms.

In this work, recent studies using supervised machine learning methods for IoT intrusion detection were examined, contrasted, and explored. Utilising labelled datasets, supervised learning places an emphasis on pattern recognition. For supervised learning to take place, the machine must be given sample data with various properties (expressed as "X") and the correct value output of the data (represented as "y"). The output and feature values of the dataset are known; hence it is referred to as being "labelled".
In order to create a model that can recreate the same underlying principles with new data, the algorithm next analyses data patterns [12].

TABLE 1: Comparison of different anomaly-based IDS techniques.

Technique	Strength	Limitations
Utilizing a fusion based technique to decrease the damage caused by strikes.	• Low communication overhead	• High energy consumption
Detecting Wormhole attacks using node position and neighbor information.	• Low resource consumption • Real time • Energy efficient	• Only One type of attack can be detected
Detecting sinkhole attacks by analyzing the behavior of devices	• Detection accuracy is high	• Detect limited number of attacks
A lightweight technique for identifying normal and deviant behavior	• Lightweight implementation • Detection accuracy is high	• High computational overhead
A request-response method's correlation functions are used to look for unusual network server activity	• Consuming modest resources • Lightweight detection system	• High computational overhead

TABLE 2: Comparison of different signature-based IDS techniques.

Technique	Strength	Limitations
Detecting network attacks by signature code in IP based ubiquitous sensor networks	• High detection accuracy • Low energy and resource consumption	• Can detect limited number of intrusions
The pattern-matching engine is used to detect malicious nodes using auxiliary shifting and early decision techniques	• Low memory and computational complexity • Maximum speed up	• Not real-time • Can detect limited number of intrusions
Detection of malware signature detection using reversible sketch structure based on cloud.	• Fast • Low communication consumption • High detection accuracy	• High memory requirement • Has a limited ability to identify assaults

TABLE 3: Comparison of different specification-based IDS techniques.

Technique	Strength	Limitations
Mitigation of black hole attacks Using an effective strategy in routing protocol for low-power and lossy (RPL) Networks	• Low delay • High detection accuracy of the infected node	• Only black hole attacks can be detected
Detecting internal attacks by designing a secure routing protocol based on reputation mechanism	• Detection accuracy is acceptable • Low delay	• Needs skilled administration
Topology assaults detection on RPL using semi-automated profiling tool.	• Detection accuracy is high • Low energy consumption • Low computation overhead	• High overhead
Sinkhole attacks are detected using a constraint based specification intrusion detection approach.	• Low overhead • Minimal energy usage	• Not real-time
Using a game-theoretic method to identify deceptive attacks in IoT network with honeypots.	• High detection accuracy • Real-time	• Needs additional resources. • High converge time

4.1 Datasets Used for IoT Intrusion Detection

Datasets are used to train and assess supervised machine learning models. The effectiveness of any IDS ultimately depends on the dataset's quality, particularly its ability to accurately identify assaults [13]. Researchers here take into

consideration and use six datasets, including NSL-KDD, UNSWNB15, CICIDS 2017, Bot-IoT, DS2OS, and IoTID20, to train and test IoT intrusion detection models.

- NSL-KDD: The KDD99 dataset has been upgraded with the NSL-KDD dataset. Since redundant records are not included in the train set, the classifiers won't be biased in favor of more frequent records. The proportion of records in the original KDD data set that are selected from each group of difficulty levels is inversely related to the number of records in each group [14]. The 41 features in the NSL-KDD dataset are divided into three groups: fundamental characteristics, content characteristics, and traffic characteristics.
- UNSW-NB15: In 2015, the UNSW-NB15 dataset was released. The UNSW cyber security lab's synthetic environment was established to build it. UNSW-NB15 uses the IXIA Perfect Storm tool to depict nine significant attack families. The IXIA tool has made it possible to create current representations of both normal and pathological network traffic in a synthetic environment. There are nine different attack categories with 49 features each, including analysis, fuzzers, backdoors, denial-of-service, exploits, reconnaissance, generic shellcode, and worms [15].
- CICIDS 2017: The CICIDS 2017 dataset was produced in that year. It comprises seven popular families of attacks that are safe and meet real-world requirements, including botnet, brute force, XSS, SQL injection, infiltration, and port scanning. The dataset has been fully annotated with 83 network traffic features that were calculated and extracted for all attack and benign network flows [16].
- BoT-IoT: The BoT-IoT dataset was developed by creating a testbed network environment in UNSW Canberra's Research Cyber Range Lab. This dataset includes real and simulated IoT network traffic as well as many sorts of attacks, including data collecting (probing attacks), denial of service attacks, and information theft. For potential multiclass classification purposes, it has been labelled with the label characteristics suggesting an attack flow, the assaults category, and the attacks subcategory [17].
- DS2OS: The DS2OS IoT platform was used to record the traces in this dataset. Two types of labelled and unlabelled datasets exist. The unique qualities specify the data items for unsupervised ML models in an unlabelled dataset. A labelled dataset is also used for supervised ML models and contains information about the class of each data occurrence [18].
- IoTID20: IoT networks employ the IoTID20 information to identify unusual behaviour. IoT gadgets and networked structures make up the testbed for the IoTID20 dataset. The dataset includes a wide range of IoT threats as well as other flow-based features. A flow-based IDS can be analysed and judged using the flow-based features. The IoTID20 dataset's final version has three label features and 83 network features [19].

Year	Dataset link (URL)	No. of Instances	No. of Features	Dataset collection performed on IoT environment	Type of dataset
2009	https://www.unb.ca/cic/datasets/nsl.html	148,519	41	No	Imbalanced
2015	https://research.unsw.edu.au/projects/unsw-nb15-dataset	2,540,044	49	No	Imbalanced
2017	https://www.unb.ca/cic/datasets/ids-2017.html	2,830,743	83	No	Imbalanced
2019	https://ieee-dataport.org/documents/bot-iot-dataset	73,370,443	29	Yes	Imbalanced
2018	https://www.kaggle.com/datasets/francoisxa/ds2ostraffictraces	409,972	13	Yes	Imbalanced
2020	https://sites.google.com/view/iot-network-intrusion-dataset/home	625,783	83	Yes	Imbalanced

Table 4: Data Set Characteristics

4.2 Supervised ML Algorithms Used for IoT Intrusion Detection.

4.3 Methodology.

There are so many supervised ML algorithms for IoT intrusion detection, some of them are described as below:

- Logistic regression (LR): A probability-based tool for predictive analysis is logistic regression (LR). It uses a more powerful approach for binary and linear classification problems because it using the sigmoid function, expected values are converted to probabilities between 0 and 1. It is a classification model that is very effective with linearly separable data classes and is reasonably easy to apply.
- Naive base (NB): Is a collection of categorization techniques based on the Bayes theorem. Instead of being a single method, it is a family of algorithms that all follow the same guiding principle, which is that each pair of attributes should be categorised separately [20].
- Artificial neural networks (ANN): The widely used ML technology known as (ANN) was developed using the biological neural network seen in the human brain as a basis. The weight values of each artificial neuron are output to the layer below. the inputs from neurons in the previous layer are processed by a feed-forward neural network, a type of ANN. A key subset of feed-forward neural networks (MLP) is multilayer perception. The back propagation algorithm is the most well-known MLP training technique that adjusts the weights between neurons to lower error. Although the system can exhibit slow convergence and be in

danger of a local optimum, it can also quickly adjust to new data values [21].

- Support Vector Machine (SVM): To minimise the distance between two classes, this algorithm seeks out a hyperplane. The categorization offers a learning base for future data processing. The algorithm divides the groups into several configurations using hyperplanes (lines). SVM generates a learning model that divides fresh examples into various groups. SVMs are classified as non-probabilistic or binary linear classifiers based on these functions. SVMs may employ techniques like Platt Scaling [20] in circumstances requiring probabilistic classification.
- Decision Tree (DT): Each internal node of a decision tree (DT) indicates an evaluation of an attribute. Each leaf node indicates the categorization results, and each branch shows the outcome of an evaluation. To create decision trees, a variety of algorithms are often employed, including ID3, CART, C4.5, and C5.0. A decision tree is created by examining the samples, and it is then applied to accurately classify new data [22].
- Random Forest (RF): Random Forest (RF) is a technique used to create a forest of decision trees. This algorithm is frequently used due to its fast operation. Countless decision trees can be used to create a random forest. By averaging the outcomes of each component tree's forecast, this method generates predictions. Random forests exhibit compelling accuracy results and are less likely to overfit the data than a traditional decision tree technique. This method works well while examining plenty of data.
- Ensemble Learning (bagging and boosting): A well-known ensemble learning technique for increasing the effectiveness and precision of ML systems is the boosting method. The continual addition of models to the ensemble is the basic tenet of the boosting technique. Strong learners are effectively promoted from weak learners, or "base learners." As a result, it helps to lessen variation and bias while increasing forecast accuracy. Boosting is an iterative process that modifies an observation's weight findings in accordance with the most recent categorisation.
  Examples of boosting methods include adaboost (AB), gradient boosting machines (GBM), and extreme gradient boosting (XGBoost). Bootstrap aggregating, sometimes known as bagging. It is one of the earliest and most fundamental ensemble ML methods, and it is effective for problems that just need a little amount of training data. This strategy involves training several original models with replacement using random subsets of data obtained through the bootstrap sampling technique. By using majority voting, the various output models produced from bootstrap samples are integrated [23].

4.3 Evaluation Parameters

Accuracy, precision, recall, and F1-score are a few examples of measures that can be used to gauge how effective ML algorithms were. The parameters True positive (TP), False positive (FP), True negative (TN), and False negative (FN) are used to calculate performance measures. These parameters are described below for IDSs.

a) Precision: It is the proportion of relevant occurrences among those that were retrieved. Higher precision is regarded as better for model performance.

Precision= True Positive/True Positive + False Positive.

b) Recall: It is the proportion of events that were determined to be pertinent. It is also known as True Positive Rate (TPR) and is computed using.

Recall= True Positive/True Positive + False Negative.

c) Accuracy: The most straightforward performance metric, it is only the proportion of accurately predicted observations to all observations. Model precision is determined using.

Accuracy=True Positive + True Negative/Total.

d)F1-Score: The weighted average of recall and precision is defined by the harmonic mean of recall and accuracy, which was obtained using.

F1 score= 2 * Precision*Recall/Precision+Recall.

e) ROC curve: It is a receiver operating characteristic curve that displays how well a classifier performs at different threshold levels.

f) Area Under Curve (AUC): It has a tight relationship to the ROC idea. It stands for the ROC curve's area under the curve. As a performance metric for classification models in ML, it has been widely employed. It has a value range of 0 to 1. The model performs better the higher the value [24].

4.4 Supervised ML Algorithms for IoT Intrusion Detection: Analysis and Comparison

Several recent studies are examined and compared based on the ML algorithms (classifier), datasets, type of classification, and performance of the classifier in order to review research on intrusion detection using ML in the IoT environment. These algorithms' effectiveness is determined by several metrics. In this study, the algorithms are compared with an emphasis on accuracy. In this part, a thorough analysis of 21 studies (published between 2019 and 2022) was conducted, and Table 5 compares the results. According to Ibrahim et al. [23], RF has comparable greater performance, and the system has outstanding accuracy. a knowledgeable Anomaly Detection IoT

(AD-IoT), a proposed anomaly detection system, utilising the UNSW-NB15 dataset and RF to find binary labelled classification. The outcomes showed that the AD-IoT was capable of achieving the highest classification accuracy while lowering the false positive rate. The datasets KDD99, NSL-KDD, and UNSW-NB15 were used by Samir et al. in [24] to evaluate the number of ML models.

The abundance of input traits that are susceptible to overfitting was the main factor in the decision to use the aforementioned classification techniques. In order to find the ideal input parameters for RF, AB, XGB, and GBM, random search was employed. RF outperforms other classifiers in terms of precision. However, out of all the classifiers, AB performs the worst. The results indicated that there is a wide range in the performances of the classifiers using Friedman test statistics and 10-fold validation. Next, compared to other classifiers' average classification time for a single case, CART classifies instances of the CIDDS-001, UNSW-NB15, KDDTrain+, and KDDTest+ faster. Using the UNSWNB15 dataset, Vikash et al. [25] proposed (UIDS) an IDS. Analysis revealed that only 13 parameters were necessary to lower the false alarm rate (FAR) of the UNSW-NB15 dataset. Seven ML methods were tested by Jadel and Khalid [26]. Except for the Naive Bayes (NB) and Quadratic algorithms (QDA), all of the algorithms had the best success rates in identifying practically all attack types. It can be shown that Adaboost, KNN, and ID3 were the algorithms with the best performance. Compared to KNN, ID3 is substantially faster. The full dataset with the top seven features determined in the feature selection process is what determines how accurate the algorithms are. The role of a selection of selected machine learning (ML) approaches for IoT intrusion detection was examined by Aritro et al. [27] using datasets and flows at the application layer (host-based) and network layer (network-based). The findings for both datasets showed that RF was the most accurate algorithm while LR was the fastest algorithm in terms of speed.The classifiers random forest (RF) and extra trees (ET), which are the best of the two, fared better than the others. The performance results were strikingly close to those obtained with all characteristics, even though RF's features selection method only selected 14 features. In addition, the LR classifier had the lowest accuracy in comparison to the others. Although Andrew et al. used various techniques, the results demonstrate that RF performed better with the non-weighted dataset in terms of precision and accuracy. However, using a weighted dataset allowed ANN to perform more accurately in binary classification. In multi-classification, KNN and ANN performed remarkably well for weighted and non-weighted datasets, respectively. The results demonstrated that ANN correctly identified the type of assault. In order to identify between legitimate traffic and attack traffic, K. V. V. N. L et al. tested four ML algorithms on IoT traffic. All of the analysed data can be carefully categorised into the appropriate classifications using decision trees. Comparing decision trees to other classifiers, they were likewise the most accurate. Using the IoTID20 dataset, Pascal et al. proposed a novel hybrid feature selection-based anomaly-based detection method for Internet of Things networks. The most accurate binary classification was achieved when RF was applied using mean imputation. Overall, there weren't many differences in each imputation strategy's accuracy. The highest level of multi-class classification accuracy was likewise reached using RF on a regression-imputed dataset. Additionally, RF's cluster-based classification accuracy was higher while requiring less training time than other cutting-edge supervised ML-based techniques. Khalid et al. evaluated the effectiveness of four ML techniques for classification goals. Both the Bot-IoT dataset and the IoTID20 dataset were used in the study; however, only 5% of the Bot-IoT dataset was entirely selected for the experiment whereas the second dataset was fully picked. The outcomes demonstrated that other classification algorithms were surpassed by the SLFN classification approach. Maryam et al. suggested employing binary classification to apply the three ML algorithms RF, GDBT, and SVM to the NSL-KDD dataset. According to the findings, SVM had the lowest accuracy on the fog layer whereas RF had the maximum accuracy. In order to identify cyber-attacks and anomalies in IoT networks, Souradipst et al. [76] suggested the B-Stacking approach as an intrusion detection model. Boosting and stacking are two ensemble methods that are combined to form B-Stacking. KNN, RF, and XGBoost were selected as the level-0 weak learners. Also employed is XGboost as the level 1 learner. The model exhibited a high detection rate and a low false alarm rate, according to experimental findings on two well-known datasets. The suggested model's lightweight design makes it possible to use it on Internet of Things nodes with meagre power and storage resources. A dataset called as IoT2020 dataset that was created from the IoTID20 dataset was used by Jingyi et al. utilising DT, RF, and GBM ML algorithms. The DT method outperformed the other algorithms in terms of accuracy, while RF had a higher AUC score, per the data. An anomaly intrusion detection in an IoT system was proposed by Abdulaziz et al.

In order to evaluate the effectiveness of five supervised machine learning models in identifying and categorising network activity, feature engineering and a data pre-treatment framework were used. Based on experimental evaluation, the detection phase that separates normal from anomalous network activity has an accuracy of 100%. The implemented models obtained 99.4-99.9% accuracy in classifying network traffic into five categories of attacks. An IoT anomaly-based IDS based on a unique feature selection and extraction approach was proposed and implemented by Khalid et al. Four ML techniques were used to train and test the model framework on the IoTID20 and NSLKDD datasets. The suggested ML ensemble-based hybrid feature selection strategy was tested using the system, which achieved a maximum detection accuracy of 99.98%.

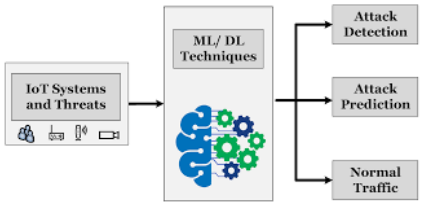

Figure 1: A review of IDS using ML and Deep learning in IOT.

Year	ML algorithm (classifier)	Dataset	Classification type	Classifier accuracy
2019	LR, SVM, DT, RF, ANN	DS2OS	Multiclass	LR=0.983, SVM=0.982, DT=0.994, RF=0.994, ANN=0.994.
2019	RF	UNSW-NB15	Binary	RF=99.34
2019	LR, NB, DT, RF, KNN, SVM	KDD99, NSL-KDD, UNSW-NB15	Binary	Accuracy of the algorithms depend on the used dataset
2019	For the level-1 model, DT For level 2 model, RF	CICIDS2017, UNSW-15	2 level classification (binary then multiclass)	Both datasets' specificity was 100% for the model, while its precision, recall, and F score were all 100% for the CICIDS2017 dataset and 97% for the UNSW-NB15 dataset
2019	RF, AB, GBM, XGB, DT (Cart), MLP, extremely randomized trees (ETC)	CIDDS-001, UNSW-NB15, NSL-KDD	Binary	Average accuracy value for 4 datasets using holdout are: RF=94.94, GBM=92.98, XGB=93.15%, AB=90.37, CART=91.98, MLP=82.76, ETC=82.99
2019	DT, NN, SVM	UNSW-NB15	Multiclass	DT=89.76, NN=86.7 , SVM=78.77, Proposed model: 88.92
2019	NB, QDA, RF, ID3, AB, MLP, KNN	BoT-IoT	Binary.	NB=0.78, QDA=0.88, RF=0.98, ID3=0.99, Adaboost=1.0, MLP=0.84, KNN=0.99
2019	SVM, LR, D T, KNN, RF	UNSW-NB15, their own dataset	Binary	The accuracy depends on the dataset and the algorithm
2020	RF, XGB, DT, MLP, GB, ET, LR	UNSW-NB15	Binary	Results with all features: RF=0.9516, XGB=0.9481, DT=0.9387, MLP=0.9371, GB=0.9331, ET=0.9501, LR=0.8984
2020	KNN, SVM, DT, NB, RF, ANN, LR	Bot-IoT	Binary, multiclass	On binary classification: KNN=0.99, SVM=0.99, DT=1.0, NB=0.99, RF=1.0 ANN=0.99, LR=0.99
2020	SVM, NB, DT, adaboost	Their own synthetic called (Sensor480)	Binary	SVM=0.9895, NB=0.9789, DT=1.0000, Adaboost=0.9895
2020	RF	IoTID20 dataset	Binary based on the attack type	The accuracy result depends on the attack type
2021	SVM	NSL-KDD, UNSW-NB15.	Binary, multiclass	The accuracy depends on the dataset, the type of classification and number of features
2021	RF, SVM, ANN	UNSW-NB15.	Binary, multiclass	All features: RF with Binary=98.67, Multi-class=97.37, SVM in Binary=97.69, Multiclass=95.67, ANN in Binary=94.78, multiclass=91.67
2021	LR, SVM, DT, ANN	IoTID20, BoT-IoT	Multiclass	The results are based on the dataset and the categories of attacks
2021	SLFN	IoTID20	Binary	The proposed model=0.9351
2021	SVM, GBDT, RF	NSL KDD	Binary	SVM=32.38, GBDT=78.01, RF=85.34
2021	B-stacking	CICIDS2017, NSL-KDD	Multiclass	Accuracy for CICIDS2017 is 99.11% Accuracy for NSL-KDD approximately is 98.5%
2022	DT, RF, GBM	IoT2020	Binary	DT=0.978305, RF=0.978443, GBM=0.9636
2022	Shallow neural networks (SNN), bagging trees (BT), DT, SVM, KNN	IoTID20	Binary, multiclass	For binary classification all models achieved 100% For multiclass: SNN=100%, DT=99.9%, BT=99.9%, SVM=99,8%, KNN=99.4%
2022	ANN, DT (C4.5), Bagging, KNN, Ensemble	IoTID20, NSL-KDD	Binary, multiclass	Accuracy depends on feature selection approaches, datasets, and attack type for multiclass classification

Table 5: Comparison of the selected supervised ML based IoT IDS.

5 Related works

IoT intrusion detection systems employing supervised machine learning methods have been the subject of substantial research. The following are some examples of related works. Arsalan Tariq et al. (2018)'s "A Machine Learning Approach for IoT Intrusion Detection System": Using machine learning algorithms like Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbour (KNN), the authors of this work proposed an IoT intrusion detection system. On the CICIDS2017 dataset, they tested the suggested system, and they got an accuracy of 98.5%. Ahmed Abdelwahab et al.'s "IoT Intrusion Detection System Using Machine Learning Algorithms" (2019): With the help of machine learning algorithms like SVM, Decision Tree, and Naive Bayes, the authors of this paper proposed an IoT intrusion detection system. On the IoT-23 dataset, they tested the suggested system, and they got a 99.7% accuracy rate.Sanaa Ghouzali et al.'s "An Intelligent IoT Intrusion Detection System Using Machine Learning Techniques" (2020): In this paper, the authors developed a machine learning-based IoT intrusion detection system that makes use of SVM, KNN, and Logistic Regression. On the UNSW-NB15 dataset, they tested the suggested system, and they got a 99.2% accuracy rate. Karim Bouzidi et al.'s "IoT Intrusion Detection System Based on Machine Learning Techniques" (2021): In this paper, the authors developed a machine learning-based IoT intrusion detection system that makes use of Random Forest, SVM, and Multilayer Perceptron (MLP) algorithms. On the IoT-23 dataset, they tested the suggested system, and they got a 99.8% accuracy rate.

5.1 Discussion

According to existing writing, there have been significant efforts put towards creating IDSs for the IoT. Utilizing popular datasets like NSL-KDD, UNSW-NB15, and CICIDS2017, several researchers have evaluated the performance of their systems. These datasets were not used to analyze IoT environment traffic. Therefore, extensive research should be done using current datasets like IoTID20, which contains features of IoT network traffic. The state of the art also demonstrates that some models perform well, especially tree-based algorithms like decision trees, random forests, and boosting. The results of ML algorithms' performance rely on the dataset used, the features, and the classification category.

5.2 Contribution

- **Team Contribution:** The workload is shared equally among the team members, who all read through all the papers before outlining their findings, issues, and solutions for machine learning-based intrusion detection systems in encrypted data. Everyone focused on the introduction, background, methodology, problem assessment, future work, and conclusion sections. Despite having some shared headings and topics, everyone nonetheless chose to write their own report in their own words on their interested topic.
- **My Contribution:** After extensive study and research, I concentrated more on the background, issue description, methodology, machine learning techniques utilised (SVM, KNN, etc.), related works, and conclusion in IOT. The following are some of the papers I looked at:

  A) "IoT Security: Review, Blockchain Solutions, and Open Challenges" by Mohsen Ali, Saman Taghavi Zargar, and Mohammad Mehdi Azizi Abarghouei.

B) "A Machine Learning Based Intrusion Detection System for IoT Networks" by Arif Sari and Murat Kantarcioglu.
C) "Anomaly-based Intrusion Detection System for IoT Networks using Machine Learning Techniques" by Bhavya Rani, Ravi Kant Sahu, and S. K. Srivastava.
D)"A Comprehensive Survey of IoT Intrusion Detection Systems: Recent Advances and Future Challenges" by Mehdi Bahrami, Reza Ebrahimi Atani, and Majid Naderi.
E) "Deep Learning Based Intrusion Detection System for IoT Networks" by Shuai Zhang, Yuxin Zhang, and Yue Wang.
F)"An IoT-based Wireless Sensor Network for Smart Agriculture" by Li Li, Li Li, Li Wang, and Min Li.
G)"A Survey of Intrusion Detection Techniques in IoT Networks" by Rasha Ibrahim, Ahmed Al-Dubai, and Imed Romdhani.
H) "Machine Learning Techniques for Intrusion Detection: A Survey" by Md. Rakibul Islam and Md. Rafiul Islam.
I)"A Survey of IoT Security: Threats, Vulnerabilities, and Countermeasures" by Nadeem Javaid, Ashfaq Ahmad, Muhammad Yousaf, Syed Ali Hassan, and Zahid Anwar.
J)"A Review of Intrusion Detection Systems in IoT Environments" by Maged M. Abdelsamea, Tarek R. Sheltami, and Abdulrahman A. Mirza.
K) A Review on IoT Intrusion Detection Systems Using Supervised Machine Learning: Techniques, Datasets, and Algorithms Azeez Rahman Abdulla, Noor Ghazi M. Jameel.

- Yes, I confirm that work load is equally distributed, however I focussed more on Machine Learning (IOT)intrusion detection systems in encrypted data.

6 Conclusion

The broad adoption of IoT devices across sectors and communities over the past ten years has been one of the most significant technological advancements. There are a number of challenges that have emerged with the development of IoT. IoT security is one of these challenges and cannot be ignored. IoT networks are susceptible to a range of dangers. Cyber-attacks are still a possibility even though the IoT network is protected by encryption and authentication. Utilizing IoT IDS is crucial and required as a result. The performance of many recent studies that used diverse approaches, datasets, ML algorithms, and their performance for detecting IoT intrusions was thoroughly analyzed and compared in this work. The analysis identifies the IoTID20 dataset as the most recent IoT dataset for intrusion detection. Additionally, tree-based ML algorithms like DT, RF, and boosting algorithms outperformed in the majority of studies. Numerous findings were made that required further investigation, such as the need for real-world IoT intrusion detection datasets for building and testing machine learning models, as well as for real-time and lightweight IDSs that require little in the way of resources and detection time. When creating new IoT IDSs, all these factors should be taken into consideration. Further research should be done to address current IoT threats as well as the need to find the best IDS placement techniques that increase IoT security while reducing the risk of cyberattacks.

REFERENCES:

[1] RFC 4252 – The Secure Shell (SHH) Authentication Protocol, http://tools.ietf.org/html/rfc4252, 03-03-2010
[2] S. Li, L. D. Xu and S. Zhao. “The internet of things: A survey”. Information Systems Frontiers, vol. 17, no. 2, pp. 243-259, 2015.
[3] “Snort: README.ssl,” Available: https://www.snort.org/faq/readme-ssl, 2020, Accessed on Oct. 28, 2020.
[4] P. Kotzias, L. Bilge, P.-A. Vervier, and J. Caballero, “Mind your own business: A longitudinal study of threats and vulnerabilities in enterprises.” in NDSS, 2019
[5] A. Khraisat and A. Alazab. “A critical review of intrusion detection systems in the internet of things: Techniques, deployment strategy, validation strategy, attacks, public datasets and challenges”. Cybersecurity, vol. 4, no. 1, pp. 1-27, 2021.
[6] N. Mishra and S. Pandya. “Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review”. IEEE Access, vol. 9, pp. 59353-59377, 2021.
[7] M. M. Hossain, M. Fotouhi and R. Hasan. “Towards an Analysis of Security Issues, Challenges, and Open Problems in the Internet of Things”. In: 2015 IEEE World Congress on Services. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 21-28, 2015.
[8] S. Ansari, S. Rajeev and H. S. Chandrashekar. “Packet sniffing: A brief introduction”. IEEE Potentials, vol. 21, no. 5, pp. 17-19, 2003.
[9]]L. Liang, K. Zheng, Q. Sheng and X. Huang. “A Denial of Service Attack Method for an IoT System”. In: 2016 8th International Conference on Information Technology in Medicine and Education (ITME). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 360-364, 2016.
[10] K. Tsiknas, D. Taketzis, K. Demertzis, and C. Skianis. “Cyber threats to industrial IoT: Asurvey on attacks and countermeasures”. IoT, vol. 2, no. 1, pp. 163-186, 2021.
[11] N. Chakraborty and B. Research. “Intrusion detection system and intrusion prevention system: A comparative study”. International Journal of Computing and Business Research, vol. 4, no. 2, pp. 1-8, 2013.
[12] S. Raza, L. Wallgren and T. Voigt. “SVELTE: Real-time intrusion detection in the internet of things”. Ad Hoc Networks, vol. 11, no. 8, pp. 2661-2674, 2013.
[13] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani. “A Detailed Analysis of the KDD CUP 99 Data Set”. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-6, 2009.
[14] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani. “A Detailed Analysis of the KDD CUP 99 Data Set”. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-6, 2009.
[15] N. Moustafa and J. Slay. “UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set)”. In: 2015 Military Communications and Information Systems Conference (MilCIS). IEEE. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 1-6, 2015

[16] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani. “Toward generating a new intrusion detection dataset and intrusion traffic characterization”.In: The International Conference on Information Systems Security and Privacy. vol. 1, pp. 108-116, 2018.
[17] N. Koroniotis, N. Moustafa, E. Sitnikova and B. Turnbull. “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset”. Future Generation Computer Systems, vol. 100, pp. 779-796, 2019.
[18] F. X. Aubet. “Machine Learning-Based Adaptive Anomaly Detection in Smart Spaces”. B.Sc. Thesis, Department of Informatics, Technische Universität München, Germany, 2018.
[19] I. Ullah and Q. H. Mahmoud. “A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks”. In: Canadian Conference on Artificial Intelligence. Springer, Berlin, Germany, pp. 508-520, 2020.
[20] A. Churcher, R. Ullah, J. Ahmad, S. U. Rehman, F. Masood, M. Gogate, F. Alqahtani, B. Nour and W. J. Buchanan. “An experimental analysis of attack classification using machine learning in IoT networks”. Sensors, vol. 21, no. 2, p. 446, 2021.
[21] R. Olivas. “Decision Trees,” Rafael Olivas, San Francisco, 2007.
[22] M. Ahmad, Q. Riaz, M. Zeeshan, H. Tahir, S. A. Haider, M. S. Khan. “Intrusion detection in internet of things using supervised machine learning based on application and transport layer features using UNSW-NB15 data-set”. Journal on Wireless Communications and Networking, vol. 2021, no. 1, pp. 1-23, 2021.
[23] I. Alrashdi, A. Alqazzaz, E. Aloufi, R. Alharthi, M. Zohdy and H. Ming. “Ad-iot: Anomaly Detection of IOT Cyberattacks in Smart City Using Machine Learning”. In: 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, pp. 0305-0310, 2019.
[24]S. Fenanir, F. Semchedine and A. Baadache. “A machine learningbased lightweight intrusion detection system for the internet of things”. Revue D Intelligence Artificielle, vol. 33, no. 3, pp. 203-211, 2019.
[25] V. Kumar, A. K. Das, and D. Sinha. “UIDS: A unified intrusion detection system for IoT environment”. Evolutionary Intelligence, vol. 14, no. 1, pp. 47-59, 2021.
[26] J. Alsamiri and K. Alsubhi. “Internet of things cyber attacks detection using machine learning”. International Journal of Advanced Computer Science and Applications, vol. 10, no. 12, pp. 628-634, 2019.
[27] A. R. Arko, S. H. Khan, A. Preety and M. H. Biswas. “Anomaly Detection In IoT using Machine Learning Algorithms”. Brac University, Bangladesh, 2019.
[28] A Review on IoT Intrusion Detection Systems Using Supervised Machine Learning: Techniques, Datasets, and Algorithms Azeez Rahman Abdulla, Noor Ghazi M. Jameel Technical college of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Kurdistan Region, Iraq.
[29] "IoT Security: Review, Blockchain Solutions, and Open Challenges" by Mohsen Ali, Saman Taghavi Zargar, and Mohammad Mehdi Azizi Abarghouei.
[30]"A Machine Learning Based Intrusion Detection System for IoT Networks" by Arif Sari and Murat Kantarcioglu.
[31]"Anomaly-based Intrusion Detection System for IoT Networks using Machine Learning Techniques" by Bhavya Rani, Ravi Kant Sahu, and S. K. Srivastava.
[32]A Comprehensive Survey of IoT Intrusion Detection Systems: Recent Advances and Future Challenges" by Mehdi Bahrami, Reza Ebrahimi Atani, and Majid Naderi.
[33] "Deep Learning Based Intrusion Detection System for IoT Networks" by Shuai Zhang, Yuxin Zhang, and Yue Wang.
F)"An IoT-based Wireless Sensor Network for Smart Agriculture" by Li Li, Li Li, Li Wang, and Min Li.
[34]"A Survey of Intrusion Detection Techniques in IoT Networks" by Rasha Ibrahim, Ahmed Al-Dubai, and Imed Romdhani.
[35]"Machine Learning Techniques for Intrusion Detection: A Survey" by Md. Rakibul Islam and Md. Rafiul Islam.
[36]"A Survey of IoT Security: Threats, Vulnerabilities, and Countermeasures" by Nadeem Javaid, Ashfaq Ahmad, Muhammad Yousaf, Syed Ali Hassan, and Zahid Anwar.
[37]"A Review of Intrusion Detection Systems in IoT Environments" by Maged M. Abdelsamea, Tarek R. Sheltami, and Abdulrahman A. Mirza.