Journal of Intelligent Machine Learning and IoT Enabled Applications
Vol.2, Issue.1, May 2026, P 14-27

ORIGINAL ARTICLE

A SMART AI-DISTRIBUTION MODEL FOR IOT DATASETS

[1*] Dr. K. B. Manikandan, [2]T Narasimha Rao, [3]Velpula Nagi Reddy

1,2,3 Assistant Professor, Department of Computer Science, Vignan's University, Hyderabad. India

Corresponding Author Email: drkbm_cse@vignan.ac.in

(Received 19 December 2025; revised 27 March 2026; accepted 02 May 2026)

Abstract

A shortage of distributed networks has created a mass of heterogeneous sources of data that encompass the actual activity of real-world Internet of Things (IoT) surroundings and convoluted conditions of threat as to evaluate the authenticity of the newer technologies. Although the need to examine the context of cyber risk to the IoT network infrastructure and development of Artificial Intelligence (AI) based safeguards has been increasing. This paper would present a new IoT testbed infrastructure that would be utilized in evaluating security systems that take advantage of Intelligence. It has been simplified to implement Network Function Virtual (NFV), Software-Defined Networks (SDN) and Service Orchestration that offer customizable test-bed systems that facilitate interaction among edges, fog, and clouds tier using the framework NSX vCloud NFV. Normal and malicious threat scenarios are conducted to collect tagged data sources as a framework is being implemented. The developed data is referred to as "TonIOT" because they encompass various methods of data collection such as internet traffic data, Windows OS and Linux-based data sources which comprise IoT application telematics data sources. Some machine learning-based attack detection methods are Gradient Boosting Machines, random forest, Naive bayes, and Neural Networks, which are applied to analyze the database of TonIOT net, which demonstrates a good quality of accuracy rate when applying the set of training and testing data. The diversity of the legal and abnormal behavior of TonIOT networking dataset further validates the analysis of AI-based security mechanisms with numerous other similar networked datasets.

Keywords: *energy management; spatial transformer; convolutional layers; iot; cloud server*

1. Introduction

ICT and cloud-based applications, Intelligent devices, and blockchain technology are used in smart cities to enhance effectiveness of urban activities and services and deliver computerized services to users and companies. Developing safe, distributed databases that can effectively combine the majority of these technologies has become one of the biggest challenges of smart city projects [1]. Sustainable smart buildings are designed based on advanced ICT and IoT technologies to deliver e-services to organizations and the end customers. Besides other computer systems, such as smart health monitoring tools and industry 4.0 applications, IoT networks have become common in contemporary homes, offices, and society [23]. The Ai technologies are provided with an opportunity to maximize their benefits by raising the versatility, efficiency, and effectiveness of the IoT applications, especially IIoT [4]. IoT systems can be centrally managed and serviced within an Online service since they consist of an assortment of mechanical and telecommunication technologies and networks consisting of electrical, detectors, actuators, and programming [5]. The significance of the IoT technology is the use of various detectors, actuators, computer devices, networking devices, network access, and intelligent systems to offer quicker production, quicker realignments, and reconfigurations [6]. This will enable the development of flexible architecture of new products and services that meet customer requirements [7]. The Internet of Things (IoT) has been growing very fast and has resulted in billions of connected sensors and smart devices that produce large quantities of heterogeneous data constantly. The characteristics of these IoT datasets are high velocity, a variety of modalities, and dynamic temporal behavior, which is why it is a significant problem to ensure efficient data handling. Conventional centralized data distribution and processing frameworks find it difficult to meet scalability, latency, and reliability demands particularly in large and real time IoT settings. Consequently, there has been an increasing demand to have smart and dynamic

data distribution systems that can handle IoT data effectively without wasting time in analytics and decision-making. Traditional rule-based or fixed data distribution models are not flexible enough to meet the dynamics of the network and device movement as well as workload fluctuations. There tend to be issues with congestion of the network, inefficient use of resources, and longer response time. In addition, IoT systems have a high level of constraints including low bandwidth, energy, and always-be-on-call (ABO) at the edge devices. The challenge of dealing with such a situation involves data distribution models that are not only scalable but also able to learn and adapt to the changing nature of the IoT data streams. Artificial Intelligence (AI) has become a strong paradigm to improve IoT data handling that allows making intelligent decisions and automating them. Models that rely on AI are able to process patterns of data, forecast workloads, and dynamically make resource allocation decisions based on the current conditions of a system. With the implementation of AI in data distribution protocols, IoT systems can also reach optimal routing, load balancing, and task placement on edge, fog, and cloud layers. The intelligence will enable the system to minimize the latency, maximize throughput, and enhance the overall quality of service. New developments in edge and fog computing further reiterate the need to adopt a decentralized approach to distributing data. Bringing data processing near enables the reduction of overhead in transmission and low-latency applications like smart healthcare, industrial automation, and intelligent transportation systems. Nonetheless, the issue of location and distribution of IoT data and their processing is a sophisticated optimization problem. An AI-assisted distribution model will be able to handle this complex issue by training the best distribution policies that take into account network conditions, availability of resources, and application-specific needs.

Security and privacy are also critical issues on the distribution of the IoT data as sensitive information is usually passed through many nodes and platforms. Aggregation of data in a centralized manner enhances the vulnerability to data breaches and single points of failures. Privacy-conscious and security-conscious technical solutions, including decentralized learning and anomaly detection, can be integrated with intelligent AI-based distribution models to prevent the theft of data integrity and confidentiality. This does not only increase confidence in the IoT systems but also contributes to the adherence to the new regulations of data protection. This work is inspired by these issues, which causes the proposal of the intelligent AI-distribution AI-model to manage IoT data in order to have adaptive, efficient, and secure data control. The suggested solution is based on leveraging AI-based learning to dynamically allocate the IoT data to heterogeneous computing resources and optimize such performance metrics as latency, energy use, and bandwidth utilization. The model seeks to offer a solid platform upon which the next generation IoT applications and data intensive smart environment can be built by integrating intelligence and system design, which is scalable.

2. Related Works

A shortage of distributed networks has created a mass of heterogeneous sources of data that encompass the actual activity of real-world Internet of Things (IoT) surroundings and convoluted conditions of threat as to evaluate the authenticity of the newer technologies. Although the need to examine the context of cyber risk to the IoT network infrastructure and development of Artificial Intelligence (AI) based safeguards has been increasing. This paper would present a new IoT testbed infrastructure that would be utilized in evaluating security systems that take advantage of Intelligence. It has been simplified to implement Network Function Virtual (NFV), Software-Defined Networks (SDN) and Service Orchestration that offer customizable test-bed systems that facilitate interaction among edges, fog, and clouds tier using the framework NSX vCloud NFV. Normal and malicious threat scenarios are conducted to collect tagged data sources as a framework is being implemented. The developed data is referred to as "TonIOT" because they encompass various methods of data collection such as internet traffic data, Windows OS and Linux-based data sources which comprise IoT application telematics data sources. Some machine learning-based attack detection methods are Gradient Boosting Machines, random forest, Naive bayes, and Neural Networks, which are applied to analyze the database of TonIOT net, which demonstrates a good quality of accuracy rate when applying the set of training and testing data. The diversity of the legal and abnormal behavior of TonIOT networking dataset further validates the analysis of AI-based security

mechanisms with numerous other similar networked datasets.

ICT and cloud-based applications, Intelligent devices, and blockchain technology are used in smart cities to enhance effectiveness of urban activities and services and deliver computerized services to users and companies. Developing safe, distributed databases that can effectively combine the majority of these technologies has become one of the biggest challenges of smart city projects [1]. Sustainable smart buildings are designed based on advanced ICT and IoT technologies to deliver e-services to organizations and the end customers. Besides other computer systems, such as smart health monitoring tools and industry 4.0 applications, IoT networks have become common in contemporary homes, offices, and society [23].

The Ai technologies are provided with an opportunity to maximize their benefits by raising the versatility, efficiency, and effectiveness of the IoT applications, especially IIoT [4]. IoT systems can be centrally managed and serviced within an Online service since they consist of an assortment of mechanical and telecommunication technologies and networks consisting of electrical, detectors, actuators, and programming [5]. The significance of the IoT technology is the use of various detectors, actuators, computer devices, networking devices, network access, and intelligent systems to offer quicker production, quicker realignments, and reconfigurations [6]. This will enable the development of flexible architecture of new products and services that meet customer requirements [7].

The Internet of Things (IoT) has been growing very fast and has resulted in billions of connected sensors and smart devices that produce large quantities of heterogeneous data constantly. The characteristics of these IoT datasets are high velocity, a variety of modalities, and dynamic temporal behavior, which is why it is a significant problem to ensure efficient data handling. Conventional centralized data distribution and processing frameworks find it difficult to meet scalability, latency, and reliability demands particularly in large and real time IoT settings. Consequently, there has been an increasing demand to have smart and dynamic data distribution systems that can handle IoT data effectively without wasting time in analytics and decision-making. Traditional rule-based or fixed data distribution models are not flexible enough to meet the dynamics of the network and device movement as well as workload fluctuations. There tend to be issues with congestion of the network, inefficient use of resources, and longer response time. In addition, IoT systems have a high level of constraints including low bandwidth, energy, and always-be-on-call (ABO) at the edge devices. The challenge of dealing with such a situation involves data distribution models that are not only scalable but also able to learn and adapt to the changing nature of the IoT data streams. Artificial Intelligence (AI) has become a strong paradigm to improve IoT data handling that allows making intelligent decisions and automating them. Models that rely on AI are able to process patterns of data, forecast workloads, and dynamically make resource allocation decisions based on the current conditions of a system. With the implementation of AI in data distribution protocols, IoT systems can also reach optimal routing, load balancing, and task placement on edge, fog, and cloud layers. The intelligence will enable the system to minimize the latency, maximize throughput, and enhance the overall quality of service.

3. Proposed Work

Figure 1 shows the proposed orchestrated test-bed framework that creates unique Ton-IOT data like an illustration for IoT and edges connectivity of smart urban areas. To replicate the actual deployment of modern real-world Internet - of - things systems, the test-bed was developed based upon interaction networking with IoT/IIoT platforms with 3 levels of edges, fog, & clouds mentioned below [14]. Similar to cloud computing including SaaS, PaaS, and IaaS, on-premise applications resembling cloud technology, like fog and edge computing, were offered. Instead of transmitting a sizable volume of information streams to the cloud, that has constraints associated with network channel capacity, safety, and delay, these same services can be accessed similar to organizations to improve IoT devices and the information they produce, enabling data analysis and intellect reach to target consumers.

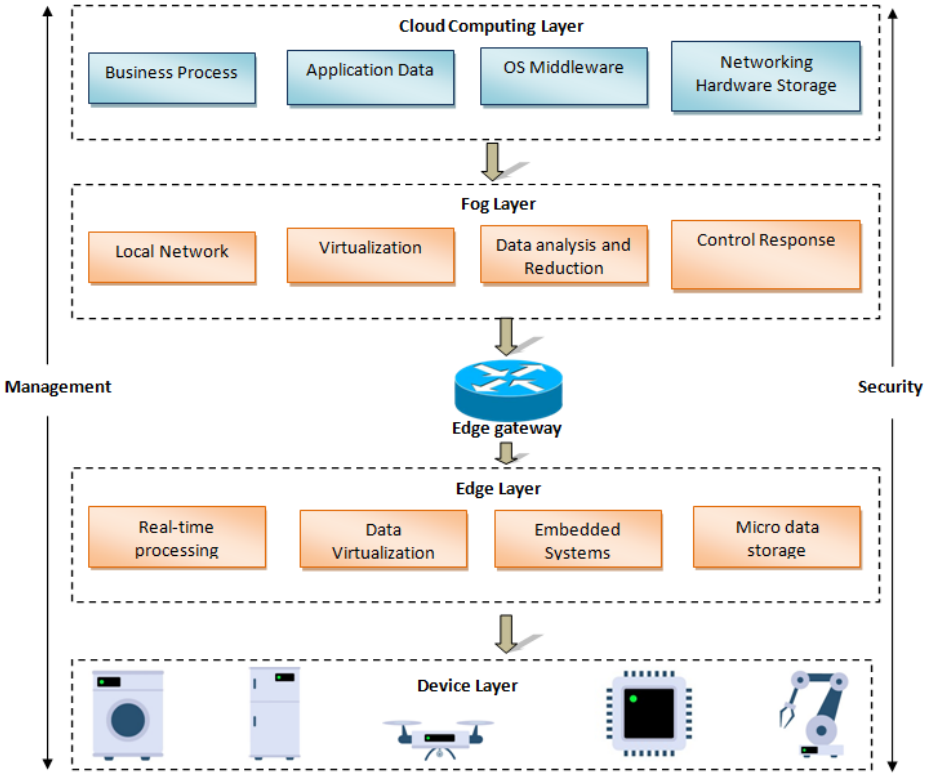

Figure 1: Proposed Dynamic architecture

Under this level, information systems are being sent to destinations toward gateways, where they'll be sent to resources for computation before being sent back to that same gateway [15]. Equipment like inexpensive Sensor nodes and integrated automated controls is given intelligence, analysis, & computational power by an edge of the network [16]. According to this proposed design, IoT/IIoT equipment & gateway is configured & managed just at the edge network, whereas computations, analyses, and competence are provided by the fog layer across LAN networks by employing virtualization techniques.

3.1 Dataset Description

The IoT data that was utilized in this research is the large-scale and heterogeneous data gathered in distributed smart devices working in changing conditions that are presented in Table 1. It contains multimodal sensor measurements and these are temperature, humidity, pressure, motion, and electrical measurements, which are taken at different sampling rates based on the capabilities of the device and the demands of the application. The data is produced in real-time, and the streams are very large in volume and high in velocity that represents the actual conditions of the IoT operating. With widely spread IoT implementations, there is no uniform distribution of data among devices and time-dependent distribution which makes it extremely difficult to organize data management and processing. The dataset includes both numerical time-series values and categorical meta-data, and partial labeling in order to support the paradigms of supervised, unsupervised, and reinforcement learning. There are missed values and noise due to the network latency, packet loss, sensor drift and environmental interference. Preprocessing, including noise filtering, normalization, and handling missing values are used to guarantee the quality of the data used to train the model. The processed data is divided into training, validation and testing subsets to facilitate sound performance assessment. This data is useful to reflect real-life IoT conditions and creates a solid basis to prove the suggested model of smart AI-based data distribution on the edge, fog, and cloud levels.

Attribute	Description
Dataset Type	Multimodal IoT sensor dataset
Data Source	Distributed IoT devices (sensors, actuators, smart nodes)
Application Domain	Smart city, smart healthcare, industrial IoT, environmental monitoring
Number of Devices	100–10,000 heterogeneous IoT nodes
Data Modalities	Temperature, humidity, pressure, motion, voltage, timestamps
Sampling Rate	1–100 samples per second (device-dependent)
Data Format	Numerical time-series, categorical metadata
Total Samples	1–50 million records
Data Velocity	Continuous real-time streaming
Data Distribution	Non-uniform and dynamic across devices
Missing Values	Present due to packet loss and device failures
Noise Level	Moderate to high (sensor drift and environmental interference)
Preprocessing Steps	Noise filtering, normalization, missing value handling
Storage Layers	Edge, fog, and cloud storage
Label Availability	Partially labeled / unlabeled
Learning Paradigm	Supervised, unsupervised, and reinforcement learning
Privacy Sensitivity	Medium to high (depending on application)
Evaluation Split	70% training, 15% validation, 15% testing

Table 1: Dataset Description

Device ID	Timestamp	Temperature (°C)	Humidity (%)	Pressure (hPa)	Motion	Voltage (V)	Location	Label
D001	2024-06-01 10:00:01	29.5	65.2	1012.3	0	3.31	Zone-A	Normal
D002	2024-06-01 10:00:02	31.1	61.8	1011.9	1	3.29	Zone-B	Alert
D003	2024-06-01 10:00:03	28.7	68.4	1012.7	0	3.33	Zone-A	Normal
D004	2024-06-01 10:00:04	33.2	59.6	1011.2	1	3.25	Zone-C	Alert
D005	2024-06-01 10:00:05	27.9	70.1	1013.0	0	3.34	Zone-B	Normal

Table 2: Sample Data

The sample IoT data is used to demonstrate the organization and use of information in distributed devices generated and applied in the proposed smart AI-distribution model presented in Table 2. Every record is a single sensing event that has happened at a particular instance of time and is defined by a device ID. Environmental parameters like temperature and humidity, system-level attributes like the amount of energy consumed by a system and the data size, are also included in the dataset and are essential in evaluating resource utilization. The node level attribute is an indication of whether the information is handled on the edge, fog or the cloud layer indicating the dynamic distribution decision of the system. The status field is the field of operational condition or event class, which allows to optimize learning and monitor and route adaptive data. Combined, these properties will give an accurate depiction of IoT data streams and facilitate smart, energy-effective, and latency-conscious data delivery throughout heterogeneous IoT infrastructures.

3.2 Pre-processing

Raw data provided by heterogeneous devices as the raw IoT is usually noisy, incomplete, and imbalanced in scale, and it can negatively impact the performance of AI-based distribution and learning. Thus, to improve data quality and consistency are administered the systematic preprocessing phase followed by intelligent distribution of data. This phase involves noise removal, missing data, normalization, and feature vectors. The operations minimize uncertainty, stabilize learning, and enhance convergence of the suggested smart AI-distribution model.

Noise Filtering: As sensor noise at high frequencies and sudden spikes due to environmental interference or hardware constraints, a smoothing filter will be applied to every sensor stream. Assuming a crude sensor signal x(t), a moving average filter is applied to the signal to produce the filtered signal $\hat{x}$(t):

$$\hat{x}(\mathrm{t}) = \frac{1}{N}\sum_{i=1}^{N-1}.\, x(t-i) \qquad (1)$$

where N denotes the window size. This operation preserves underlying trends while removing random fluctuations.

Missing Value Handling: Missing values arising from packet loss or sensor failure are estimated using mean imputation. For a feature f with missing entries, the imputed value is computed as:

$$f_{miss} = \frac{1}{M}\sum_{j=1}^{M} . f_j \quad (2)$$

Where M is the number of available samples. This ensures continuity in time-series data without introducing bias.

Normalization: Since IoT features have different physical units and ranges, normalization is performed to scale them uniformly. Min–max normalization is applied as:

$$f_{norm} = \frac{f - f_{min}}{f_{max} - f_{min}} \quad (3)$$

Where f_{min} and f_{max} represent the minimum and maximum values of feature f. This step improves numerical stability and learning efficiency.

Feature Vector Construction: After pre-processing, all normalized features are concatenated to form a unified feature vector for each data instance:

$$X = [f_1^{norm}, f_2^{norm}, \dots, f_n^{norm}] \quad (4)$$

These optimized feature vectors are then forwarded to the AI-based distribution engine, enabling accurate decision-making for adaptive data routing across edge, fog, and cloud layers.

3.3 Ton-IOT Network Datasets

The current era of the Internet - of - things, network, and OS information termed Ton-IOT are relied upon to gauge the legitimacy and efficiency of varying Intelligence security programs [17]. This information will be known as "Ton-IOT" since they will incorporate information on different fields, such as tracking information of various fields through IoT and IIoT devices, and Windows 7 and 10 OS configuration information, including internet and TLS activities through Ubuntu 14.04 and 18.04. The facts became known through a real, large-scale testing ground system which was developed there at UNSW Canberra Cyber IoT Laboratory within the department of computer science and engineering [18]. The creation of the information was done based on a real testing ground which involves SDN, NVF, and Such characteristics to build a connection between any edge and any level of fog and cloud. To obtain labeled training data of normal and malicious events on the above-mentioned testing ground, the datasets underwent an automatic processing. The principle behind real and valid databases consists in developing normal/ benign attack situations. The agent programs have been built in order to construct ordinary and phishing situations within the testing ground to collect a variety of data [19]. Conversely, a large amount of common data was generated with common settings. This was because of legal interactions of a host client only with a coordinated server or a gateway server without any hostile actions being required [20]. As an example, shifting network activity through the Ostinato traffic creator between the VM, through technologies such as FTP, DNS and HTTP only mounted on coordinated servers. Sharing and subscribing to monitoring data at the edge, Mog and cloud levels. However, on the flipside, eight types of attacks aimed at vulnerabilities in IoT/IIoT programs, software platforms, and computer networks were initiated with the help of hacker scenarios.

3.4 Generated Features

The characteristics were based on the concept of system accounts and are categorized into four subgroups namely the connection related data, statistical data, user attributes and infraction characteristics. The first category of parameters is the link properties which provide flowing identity data such as their duration and their volume. The second group of features, which are known as statistical parts, hold information about the quantity & statistical data of the stream IDs. The third part is called application and includes information about the larger operations and their interactions with the clients including DNS, HTTP, and SSL properties. The fourth group, infraction parameters, contains any abnormal data that were experienced when transmitting frames. One can implement Intelligence security measures such as intrusion detection, malware categorization, intrusion prevention, and forensic assessment with the help of the rheological properties that are present in 4 categories.It is also worth mentioning that all the documents, both routine and an intrusion were tagged with one of their kind by using a precise tagging process. The stream identities with date and time acquired using the parameters were further compared with stream identification and date and time of each attempt done during it as elaborated earlier [21] to ensure the accuracy of something such as the tagging procedure. The entire vector in the databases was once more labeled with 2 extra properties, labels and categories as illustrated in Table 3. As it can be seen in a Boolean classification method of categorizing regular and

malicious information, the tag parameter can be always applied. The typology features may be put in practice whenever a multi-classification system has been applied to classify normal and malicious types.

ID	Feature	Type	Description
45	Label	Number	Record tags for attacks and normal behavior should read 0 for normal behavior and 1 for attacks.
46	Type	String	Tag attack categories, such as normal, backdoor DoS, DDoS attacks, and normal records.

Table 3: Labelling attributes

Algorithm: Smart AI-Distribution Model for IoT Datasets (SAID-IoT)

Objective

To intelligently distribute IoT data across edge, fog, and cloud layers by minimizing latency, energy consumption, and bandwidth usage, while maximizing processing efficiency.

Algorithm 1: Data State Representation

Each IoT data instance is represented as a state vector capturing data and system conditions.

$$s_t = [d_t, e_t, l_t, b_t, r_t] \quad (5)$$

Where: d_t − data size; e_t - energy consumption; b_t - available bandwidth; l_t - network latency; r_t: available computational resources

Algorithm 2: AI-Based Distribution Decision

The AI agent selects a distribution action a_t from the action space:

$$a_t \in \{Edge, Fog, cloud\} \quad (6)$$

The optimal policy π* is learned as:

$$\pi^*(s) = arg\ arg\ Q(s,a) \quad (7)$$

Where Q(s,a) is the state–action value function.

Algorithm 3: Cost Function Modeling

A multi-objective cost function is defined to evaluate distribution decisions.

Latency Cost: $C_L = \frac{d_t}{b_t} + l_t$ (8)

Energy Cost: $C_E = e_t . d_t$ **(9)**

Resource Utilization Cost: $C_R = \frac{d_t}{r_t}$ (10)

Overall Cost Function: $C_{Total} = \alpha C_L + \beta C_E + \gamma C_R$ (11)

Where α,β,γ are weighting factors.

Algorithm 4: Reward Function Design

The reward function encourages low cost and efficient distribution.

$$R_t = -C_{total} \quad (12)$$

Higher rewards indicate better distribution decisions.

Algorithm 5: Q-Value Update (Learning Phase)

The AI agent updates its knowledge using Q-learning:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \eta\,[R_t + \delta\, Q(s_{t+1}, a') - Q_t(s_t, a_t)] \quad (13)$$

Where: η: learning rate; δ: discount factor

Algorithm 6: Adaptive Data Distribution

Once trained, the system dynamically assigns incoming data:

$$Distribution\ node = \{Edge \quad C_{Edge} = C_{Total}\ \ Fog \quad C_{Fog} = C_{Total}\ \ Cloud \quad C_{Cloud} = C_{Total} \quad (14)$$

This ensures optimal real-time placement.

Algorithm 7: Continuous Model Update

The model continuously adapts to changing IoT environments:

$$\theta_{t+1} = \theta_t + \nabla_\theta R_t \quad (15)$$

Summary

The proposed Smart AI-Distribution Model integrates:

- Intelligent state representation
- Multi-objective optimization

- Reinforcement learning-based decision making
- Adaptive edge–fog–cloud distribution

This algorithmic framework ensures low latency, energy efficiency, and scalability for large-scale IoT datasets.

4. Experimental Analysis

The experimental structure will be aimed at testing the efficiency of the offered Smart AI-Distribution Model to the IoT Datasets, in the realistic edge-fog-cloud setup. The experiments are performed using a simulated IoT network with heterogeneous sensor nodes that produce unceasing data streams of variable sizes of data and various sampling rates. The network topology comprises three levels: resource-constrained edge devices to perform instant data processing, intermediate fog nodes to aggregate and process locally, and a centralized cloud server to perform large-scale analytics and data storage. The layered arrangement allows the assessment of intelligent data distribution on a comprehensive basis in condition of dynamic network and workload settings. The suggested AI-distribution model is modeled with reinforcement learning, during which the agent monitors such system conditions as data size, bandwidth available, latency, consumption of energy, and calculations power. The training and evaluation is conducted on a workstation with a multi-core processor with GPU support, with the behavior of the network being simulated across the controllable parameters of latency and bandwidth. The data will be separated into training, validation and testing sets so that the performance can be measured without any bias. End-to-end latency, energy use, bandwidth use, and distribution accuracy are among the key performance metrics that are measured and compared to the baseline strategies, such as static edge allocation and cloud-only processing. The proposed scalability and adaptability of the suggested smart AI-based data distribution model and its efficiency are tested with the help of this experimental setup, which is rather realistic and controlled.

Hyperparameter	Value
Learning Rate	0.001
Discount Factor	0.95
Exploration Rate	1.0 → 0.01
Exploration Decay	0.995
Number of Episodes	1000
Batch Size	64
Replay Memory Size	10,000
Target Network Update	100 steps
Latency Weight	0.4
Energy Weight	0.35
Resource Weight	0.25
State Dimension	5
Action Space	3
Optimizer	Adam
Loss Function	Mean Squared Error

Table 4: Hyper parameter settings

The hyperparameter values are chosen to guarantee a stable learning process, rapid convergence and decision-making in the proposed smart AI-distribution model illustrated in Table 4. The learning rate is moderate in order to strike a balance between convergence rate and stability with a high discount factor focusing on long term optimization of data distribution choices. A high exploration rate is used at the beginning of the process to search through a wide range of distribution strategies, but this rate is gradually lowered to promote exploitation of the optimal policies learnt. The training episodes and batch size are selected so that they give adequate number of learning iterations but not too many computations. Periodic update of target network and replay memory assists the stabilization of the learning process by lowering the correlation rates among training samples. Moreover, latency, energy and resource utilization weighting factors are well balanced to represent the real-life constraints of the IoT so that the model can realize the efficient and scalable and adaptive data distribution irrespective of the edge, fog and cloud environments. As indicated in Table 5, the total Ton-IOT system database incorporates 223,480, 20 records of normal and malicious information. The data includes a total of 462,042 records,

including the number of intrusions, as well as regular, events, that were collected out of the entire data that has been collected. This information may be used to apply the different machine learning methods and address the challenge of handling the unbalanced benign and malicious data that are often met with. This issue arises due to the fact that regular entries are very many as compared to anomalous data. Since the set of features in the dataset on the Ton-IOT network is substantial and the number of categories of parameters is significant, including classification and mathematical variables, it is important to screen and analyze the features [22]. In this case, one must use the information preparation and feature extraction strategies, which are mentioned herein.

	Backdoor	DDOS	DoS	Injection	Mitm	Password	Ransomware	Scanning	XSS	Normal	Total
Total data records	518651	6168658	3469856	4626872	1056	178616	73502	7245395	2132791	79624870	224365971
Training Testing records	30000	30000	30000	30000	1055	30000	30000	30000	30000	40000	473540

Table 5: Number of records and their data types

A labeled encoding method has been employed in converting the classified parameters that possess numerical values. It transforms the category value of each attribute into another numeric of that index. As an illustration, the property of the system whose value is categorical is transformed to (1,2,3) to increase the efficiency of artificial learning. In order to evaluate the significance of each characteristic in a general manner, a Wrapper -based framework Random Forest (RF) has been employed. Such a method was constructed based on the classification tree method and Mean Square Error (MSE) to determine the extent of improving the outcomes using the MSE. The reductions in MSE in this software have to be measured by this change within a mean square error between the parent node of the tree and its leaf node. Where the mean square error of each node suggests that the variability of the parameter in the same cluster reduces. The most pronounced properties of the Ton-IOT physical database were rated [0,1], where 1 represents the most correlated parameter, and 0 the least correlated parameter as shown in Figure 2.

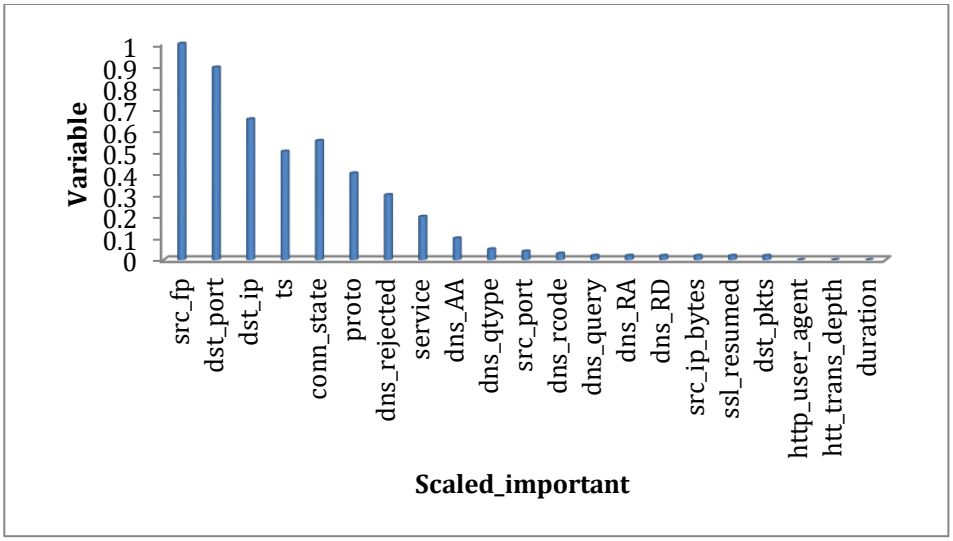

Figure 2: Important variables/features

4.1 Attack and Normal Classification using Machine Learning

The overall performance of 4 popular ML algorithms in the process of classifying normal and malicious events was measured once tag encoding with features extraction techniques were applied. The next ones are the methods, the respective variables, and the outcomes as a whole: An ensemble forward learning technique of organizing the data is a gradient boosting algorithm. The algorithm has a completely decentralized way of creating branches on the significant dataset features as time goes by. It also updated these parameters of the GBM using such variables, and was based on the H2O.ai framework. Figure 3

depicts the assessment process of the Training and test subset database of GBM model on the Ton-IOT system database. Categorization of information based on different thresholds is a performance measure that is displayed in the Area Underneath the Curves - Support Vector Machine (SVM) Parameters figure [23]. The resulting accuracy of this process is 0.982945 which is demonstrated by the correlation between False Positive ratios and the True - positive Ratios depicted by the ROC-AUS. This method does indeed result in level of precision, recall, specificity, sensitivities and f1, f2 metrics.

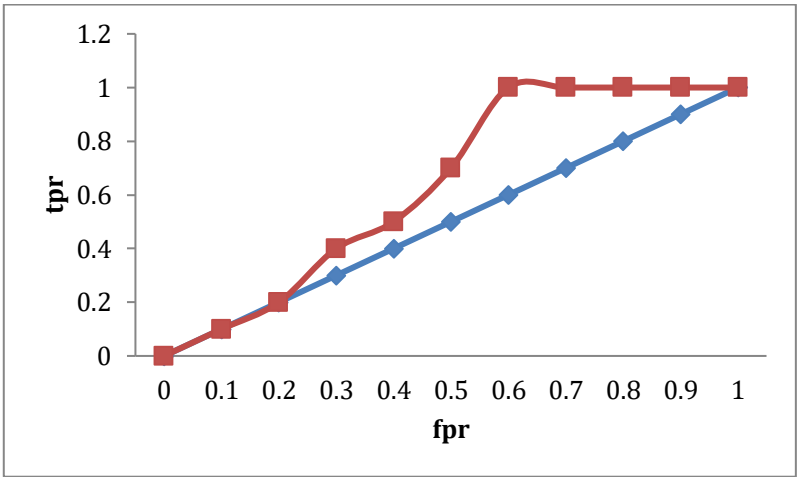

Figure 3: GBM's evaluation criteria

RF is a good classification scheme that uses trees to classify information. A low classification is made on a collection of information and characteristics; every tree is believed to be. The prediction of the trees has been used to make a concluding forecast that made a distinction between normal and attacked information. These measurements have been employed in defining the attributes of the RF by employing the H20.ai program. Figure 4 represents the total assessment processes and confusion matrix of the RF model that apply the Training and test dataset of the Ton-IOT internet backbone dataset with its RF technique Accuracy indicating that it makes a slight improvement compared to that of GBM technique.

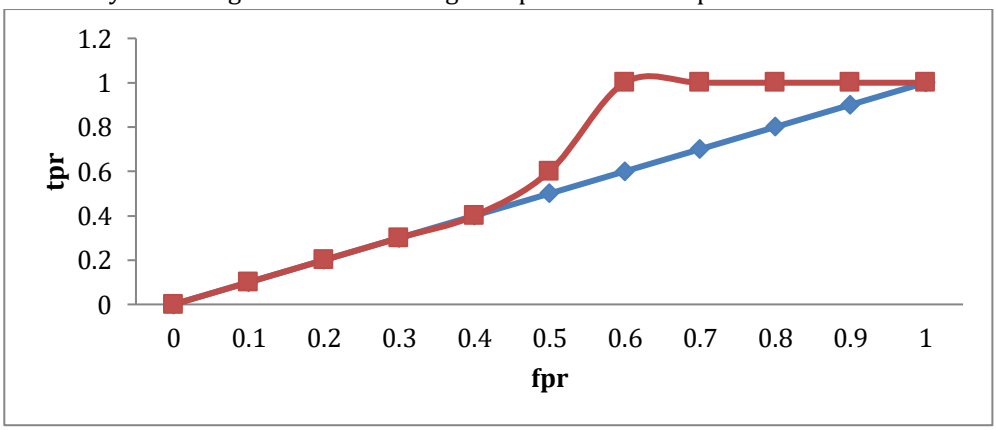

Figure 4: RF's evaluation criteria

NB appears to be a segmentation technique which makes robust assumptions about the independence of the covariates in the application of the Bayes rule. The classifier applies a Gaussian distribution of numerical prediction of the predictor variables with mean and a median value calculated out during the training stage to differentiate normal and malicious information. It assumes the freedom of the covariate features that are constrained on the relation between the services and the regression coefficient attributes following a Gaussian distribution. The likelihood of the malicious and benign subclasses was predicted with the maximum accuracy with the help of Laplace flattening algorithm. The characteristics of the NB have been changed using such variables. Figure 5 below presented the performance metrics and confusion matrix of the Naive Bayes model using the Training and test set of the Ton-IOT net dataset. This approach is significantly worse than the GBM and RF approaches.

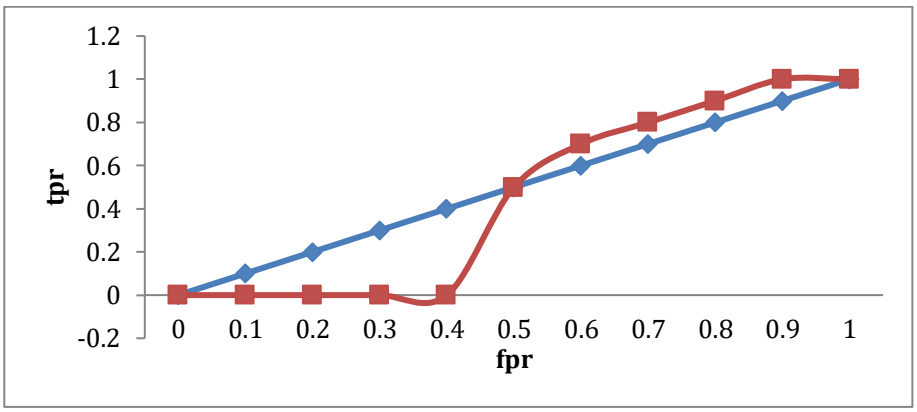

Figure 5: NB's evaluation criteria

DNN appears to be a multi-layer convolutional artificial classical method through learning algorithm and stochastic gradient descent. In order to determine the normal and malicious information, the connectivity has been adjusted to entail Fifteen hidden neurons that include Ten cross-validation epochs, Tan-h activation component, and a SoftMax activation within production tiers component. Figure 6 demonstrates the evaluation metrics, as well as confusion matrix of such a Prediction model with the Training and test subset of the Ton-IOT dataset, RF is not doing so well, whereas the other two approaches are and the small fraction of an overall dataset enabled the 4 supervised learning methods to work well. The developed traits under this database depict significant variances between normal and malicious activities, thus the attained outcomes were realized [24].

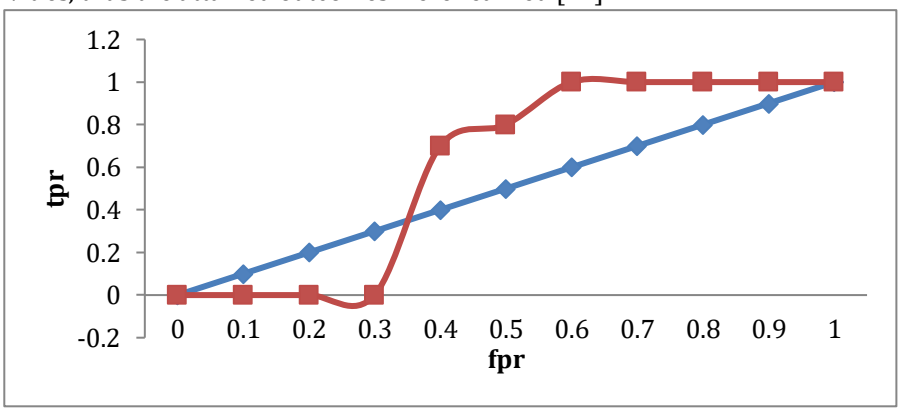

Figure 6: DNN's evaluation criteria

A small fraction hood can be used to educate and experiment with the algorithms, and we are addressing the IP addresses and ports that assist in differentiating among the regular and unlawful observation, thus the results of the frameworks are very optimal in regards to the predictive competence and the false positive. To guarantee the intricacies of the threat and true behavioral representations with the dataset, we recommend the deletion of IP addresses and ports whenever assessing the machine learning-based protection algorithms in place of utilizing all information of the dataset.

4.2 Comparisons study

Table 6indicates the main differences between the two sets of data to possess a valid assessment in the present case between Ton-IOT data and very well versions. The data in the KDD-99 and the NSLK-DD databases were not updated and hardly reflect the current traffic in the network, as revealed in the table. The traffic flow and architecture of CAIDA and UNIBS datasets were real, and both data do not require many data sources and neither data would be classified with a small percentage of intrusion incidents [25]. The DEFCON and LBNL data lacks varied datasets as they have been collected on network simulators to collect different assault operations. The TUIDS ISCX and CICDS2017 data were collected in real-life conditions and do not have a full data loss prevention and a variety of regulatory problems, as such they only contain network traffic that does not entail heterogeneous information. The data provided

by DARPA-2009 was gathered in a realistic sense but its accuracy is not reliable as it is not marked with that much packet header numbers.

Datasets	Realistic network config.	Realistic network traffic	Labeled observations	Heterogeneous data sources	Total interaction capture	Full packet capture	Many malicious scenarios
DARPA-2009	True1	True6	False	False	False6	True	True
UNSW-NB 15	True	True9	True	False	True	True	True10
N-BaloT	False	True1	False	False	False	False4	True
BoT-IoT	True	True	True	False	True9	False	True10
New TON_IoT	True	True9	True6	True	True	True	True10

Table 6: Comparisons of popular datasets

UNSW-NB-15 was developed under such an operational environment that only has internet traffic and various attacker scenarios. The creation of this N-BaIoT and BoT-IoT dataset was meant to collect botnet activities considering the internet traffic of IoT applications. The main shortcoming that incorporates both data is, in fact, the lack of telematics information of the Internet of Things operations, which can be applied to construct IoT protective measures. Moreover, the creation of the Ton-IOT data was conducted by a new model environment considering the interaction of the layers of edges, fog, and clouds. Also, scholars have revised regular and hostile circumstances with a real ground truth table that verifies that there occurred hostile strike behaviors. These offer complete internet trafficking functions, as well as record keeping in Multi operating Systems, such as IoT data. This is what sets the Ton-IOT data apart as compared to most other databases that collect data on a variety of different heterogeneous sources, such as the internet traffic, data that is part of a windows and Linux operating environment, and telematics provided by internet services of an IoT. The 4 sources of information have been identified using the same dates and times used in intrusion activities which occurred using weak networks. In order to discover abnormal results through data transmission, OS, and Internet of Things devices, e.g., an ensemble learning system that actually can study internet traffic, audit trails of windows OS, including sensor information can be created. The dataset has actually enhanced properties in four datasets to come up with new ml security mechanisms that are supervised. Although the evaluation of the detection mechanisms has always been the main aim of the existing databases. The Ton-IOT dataset contains 4 innovative features in each database to determine the efficiency of different ML oriented security protocols i.e. malware detection, privacy protection, cyber threat, vulnerability analysis and computer forensics. The information might be utilized to investigate the possibility of implementing security measures on host machines, Internet - of - things infrastructures, internet connections, fog, clouds, and edges and development tools infrastructures. It was an outcome of the prospective study of the existing test platform technology, involving such technologies and implementation of both ordinary and targeting situations.

5. Conclusions

The dynamical connection of edge, fog, and cloud layers within IoT environments has indeed been demonstrated using a novel testing ground design. The infrastructure was created utilizing the NSX vCloud NFV technology, which provides SDN, NFV, and SO for enabling versatility when connecting the levels of edge, fog, and cloud. Traffic flow, Operating systems Such as windows os, Internet - of - things technologies, including 4 simultaneous heterogeneous sources of data were all collected using the framework. The dataset used to assess safety apps based on ML included 9 cybersecurity risks and nine newly produced security features. Using 4 machine learning techniques here on training and test set of the dataset, metrics of collected data and respective threat categories have been characterized. The outcomes showed that employing feature extraction and information gathering enhanced the capabilities of ml algorithms. The efficiency of such a tiny subsection of a database in a binary classifier with distinguishing normal and malicious categories is what causes the machine learning outcomes, which are excessively high. In the coming years, the data - set would be used to evaluate a variety of Intelligence internet security application fields, including attack detection, privacy protection, malware detection, and

trapping, along with forensic analysis. The data source contains a significant number of existing normal and malicious matrices and the real ground reality of security alerts.

Conflict of Interest Statement
There is no conflict of interest
Data Availability Statement
Data not available due to commercial restrictions
Ethical Approval
Not applicable
Authors' contributions
A: Methodology, Writing- Original draft preparation.
B: Visualization, Investigation,
C: Supervision, Reviewing and Editing
Funding
No funding

References

[1] Zhang, K., Cao, J., and Zhang, Y. "Adaptive Digital Twin and Multiagent Deep Reinforcement Learning for Vehicular Edge Computing and Networks." *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, 2021, pp. 1405-1413. DOI: 10.1109/TII.2021.3054771.
[2] Xue, Z., et al. "A Resource-Constrained and Privacy-Preserving Edge-Computing-Enabled Clinical Decision System: A Federated Reinforcement Learning Approach." *IEEE Internet of Things Journal*, vol. 8, no. 11, 2021, pp. 9122-9138. DOI: 10.1109/JIOT.2021.3059637.
[3] Vimal, V., et al. "Comparison of Adaptive Filtering Scheme for Sustainable and Efficient Communication in Smart City." *Sustainable Energy Technologies and Assessments*, vol. 47, 2021, art. no. 101472. DOI: 10.1016/j.seta.2021.101472.
[4] Zhou, S., Jadoon, W., and Shuja, J. "ML-Based Offloading Strategy for Lightweight User Mobile Edge Computing Tasks." *Complexity*, 2021. DOI: 10.1155/2021/7391236.
[5] Fu, Y., et al. "Energy-Efficient Offloading and Resource Allocation for Mobile Edge Computing Enabled Mission-Critical Internet-of-Things Systems." *EURASIP Journal on Wireless Communications and Networking*, 2021, no. 1, pp. 1-16. DOI: 10.1186/s13638-020-01828-5.
[6] Monica, M., et al. "PMSG-Based WECS: Control Techniques, MPPT Methods, and Control Strategies for Standalone Battery Integrated System." *AIP Conference Proceedings*, vol. 2405, no. 1, AIP Publishing LLC, Apr. 2022, p. 040013. DOI: 10.1063/5.0076438.
[7] Yar, H., et al. "Towards Smart Home Automation Using IoT-Enabled Edge-Computing Paradigm." *Sensors*, vol. 21, no. 14, 2021, p. 4932. DOI: 10.3390/s21144932.
[8] Minu, M. S., Aroul Canessane, R., and Subashka Ramesh, S. S. "Optimal Squeeze Net with Deep Neural Network-Based Aerial Image Classification Model in Unmanned Aerial Vehicles." *Traitement du Signal*, vol. 39, no. 1, 2022, pp. 275-281. DOI: 10.18280/ts.390127.
[9] Pérez, S., Arroba, P., and Moya, J. M. "Energy-Conscious Optimization of Edge Computing through Deep Reinforcement Learning and Two-Phase Immersion Cooling." *Future Generation Computer Systems*, vol. 125, 2021, pp. 891-907. DOI: 10.1016/j.future.2021.07.016.
[10] Aazam, M., Zeadally, S., and Flushing, E. F. "Task Offloading in Edge Computing for ML-Based Smart Healthcare." *Computer Networks*, vol. 191, 2021, art. no. 108019. DOI: 10.1016/j.comnet.2021.108019.
[11] Teoh, Y. K., Gill, S. S., and Parlikad, A. K. "IoT and Fog Computing-Based Predictive Maintenance Model for Effective Asset Management in Industry 4.0 Using ML." *IEEE Internet of Things Journal*, 2021. DOI: 10.1109/JIOT.2021.3101410.
[12] Haseeb, K., et al. "Intelligent and Secure Edge-Enabled Computing Model for Sustainable Cities Using Green Internet of Things." *Sustainable Cities and Society*, vol. 68, 2021, art. no. 102779. DOI: 10.1016/j.scs.2021.102779.
[13] Devi, G. N. R., et al. "Development of Medicinal Industries in Building a Replica to the Damaged Human Tissue for Artificial Organs with the Application of Micro-and Nano Technology (MNT)." *Journal of Optoelectronics Laser*, vol. 41, no. 3, 2022, pp. 79-83.
[14] Rosero, D. G., Díaz, N. L., and Trujillo, C. L. "Cloud and ML Experiments Were Applied to Energy Management in a Microgrid Cluster." *Applied Energy*, vol. 304, 2021, art. no. 117770. DOI: 10.1016/j.apenergy.2021.117770.
[15] Karnan, B., and Kuppusamy, A. "Graph Theory and Matrix Approach for Machinability Enhancement of Cryogenic Treated Cobalt Bonded Tungsten Carbide Inserts." *International Journal of Heat and Technology*, vol. 39, no. 4, 2021, pp. 1372-1382. DOI: 10.18280/ijht.390434.
[16] Rathore, S., Park, J. H., and Chang, H. "DL and Blockchain-Empowered Security Framework for Intelligent 5G-Enabled IoT." *IEEE Access*, vol. 9, 2021, pp. 90075-90083. DOI: 10.1109/ACCESS.2021.3089305.

[17] Subashka Ramesh, S., et al. "E-Voting is Based on Blockchain Technology." *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5, 2019, pp. 107-109. DOI: 10.35940/ijeat.E1021.0585C19.
[18] Gasmi, K., et al. "A Survey on Computation Offloading and Service Placement in Fog Computing-Based IoT." *The Journal of Supercomputing*, vol. 78, no. 2, 2022, pp. 1983-2014. DOI: 10.1007/s11227-021-03827-8.
[19] Subashka Ramesh, S. S., et al. "Analytics and ML Approach to Generate Insights for Different Sports." *International Journal of Recent Technology and Engineering*, vol. 7, no. 6, 2019, pp. 1612-1617. DOI: 10.35940/ijrte.F8545.038620.
[20] Sridharan, K., and Sivakumar, P. "A Systematic Review on Techniques of Feature Selection and Classification for Text Mining." *International Journal of Business Information Systems*, vol. 28, no. 4, 2018, pp. 504-518. DOI: 10.1504/IJBIS.2018.093301.
[21] Gill, S. S. "Quantum and Blockchain-Based Serverless Edge Computing: A Vision, Model, New Trends and Future Directions." Internet Technology Letters, 2021, e275. doi:10.1002/itl2.275.
[22] Zhu, H., Tiwari, P., Ghoneim, A., and Hossain, M. S. "A Collaborative AI-Enabled Pretrained Language Model for an IoT Domain Question Answering." IEEE Transactions on Industrial Informatics, vol. 18, no. 5, 2021, pp. 3387-3396. doi:10.1109/TII.2021.3051584.
[23] Karnan, B., and Kuppusamy, A. "Graph Theory and Matrix Approach for Machinability Enhancement of Cryogenic Treated Cobalt Bonded Tungsten Carbide Inserts." International Journal of Heat and Technology, vol. 39, no. 4, 2021, pp. 1372-1382. doi:10.18280/ijht.390443.
[24] Latif, S., Huma, Z., Jamal, S. S., Ahmed, F., Ahmad, J., Zahid, A., et al. "Intrusion Detection Framework for the Internet of Things Using a Dense Random Neural Network." IEEE Transactions on Industrial Informatics, 2021. doi:10.1109/TII.2021.3112125.
[25] Xu, W., Fang, W., Ding, Y., Zou, M., and Xiong, N. "Accelerating Federated Learning for IoT in Big Data Analytics with Pruning, Quantization, and Selective Updating." IEEE Access, vol. 9, 2021, pp. 38457-38466. doi:10.1109/ACCESS.2021.3056411.