Alkadhim Journal for Computer Science, Vol. 3, No. 4 (2024)

Open Access ISSN: 3007-1437 (Online), ISSN: 3007-1429 (Print)

IRAQI Academic Scientific Journals

Alkadhim Journal for Computer Science (KJCS)

Journal Homepage: https://alkadhum-col.edu.iq/JKCEAS

KJCS ALKADHIM JOURNAL FOR COMPUTER SCIENCE

Analyzing Big Data to Mitigate Cyber Attacks Using Machine Learning Classifiers: A Comparative Study of Efficient Classifiers

Ali Jawad Kadhim Abboodi

Imam AL-Kadhim university college- Baghdad-Iraq

Article information

Article history:
Received:June,1, 2025
Accepted:November,20, 2025
Available online: December, 25, 2025

Keywords:
Big data analysis, cyber security, machine learning.

**Corresponding Author:*
ali jawad kadhim abboodi
ali.jawad@iku.edu.iq

DOI:
https://doi.org/10.61710/kjcs.v3i4.113

Abstract

This article discusses a set of modern approaches and techniques in the field of cyber security and big data analysis using machine learning methods, with the aim of improving the ability to detect attacks and mitigate their impact. With the advancement of technology and the world entering a new phase of digital transformation, data and information are increasing in large, uncontrolled quantities, and are vulnerable to corruption and loss. It has become impossible to rely on old methods to preserve the enormous volume of data. The methodology relies on data analysis through sequential stages based on segmentation, classification, and monitoring of network anomalies. It also reduces the size of redundant features to speed up the process of analyzing big data, focusing on attempts to hack operating systems and security systems. Principal Component Analysis (PCA) was applied to reduce the data dimensions and identify influential features, facilitating the training of classifiers. The experiment included the use of multiple binary classifiers, including k-Nearest Neighbours (k-NN), Support Vector Machines (SVM), Bayesian algorithms, and artificial neural networks (ANNs), due to their high efficiency in monitoring network disturbances. The classifiers were combined using two methods: soft voting and majority voting, to achieve higher performance and better detection accuracy. Two data processing methods were applied: the first was to divide the data into subsets, each with its own processing path, and the second was to use sensors that collect and analyse user data via parallel paths to detect anomalies. An experiment was conducted on Internet of Things (IoT) device data, where the combined classifiers (such Support Vector (SV), Weighted Voting (WV), and Majority Voting (MV)) demonstrated higher performance than the individual classifiers, with the static SV classifier achieving a 2.5% increase in classification accuracy (ACC) compared to the best baseline classifier. The results confirm that systematically combining classifiers enhances the effectiveness of cyber detection systems in large and complex data environments.

1. Introduction

The increasing volume of data and information within communications networks has increased the inevitability and likelihood of their exposure to numerous risks, necessitating protection against loss, damage, cyber-attacks, and other risks. This is especially true given that information and data have become the true wealth of all fields

and disciplines, particularly in the financial, economic, military, and healthcare sectors[1]. This has prompted the authorities responsible for protecting this data to provide advanced systems that keep pace with all new and emerging developments, providing the greatest opportunity to protect massive data from cyber-attacks. With the increase in risks and threats, it has become necessary to devise new methods and techniques to protect data and information from these attacks, to ensure the security of data and information from cyber-attacks[2]. This is primarily due to the fact that modern cyber systems include technologies of new and increasing types of elements, both physical (devices and equipment) and social (expertise and training courses for available personnel) and raising the level of awareness among individuals. Therefore, cyber security systems and the Internet of Things (IoT) have become increasingly[3] There is an increasing demand for developing information security and protection systems, and providing networks with all the means, both physical and software, to combat cyber-attacks. Each component has its own characteristics, whether hardware or software, and modern systems. Modern methods have developed numerous technologies and systems that rely on artificial intelligence. Big data analysis using classification, clustering, and descriptive segmentation techniques are among the most important modern methods. Encryption processes are also popular and effective solutions. Elliptic curve coherence (ECC) and linear curves all operate efficiently and reduce the likelihood of hacking[4]. They are highly efficient and fast in standard encryption operations. During this period, many methods and devices have emerged to protect cyber security from encryption and detection through fingerprints, voice, and faces. However, hackers have developed more sophisticated methods for intrusion, and existing methods are insufficient to mitigate these attacks. In recent years, several studies have emphasized the need for comprehensive risk assessment methodologies that not only identify potential vulnerabilities but also determine the potential impact of cyber incidents[5]. The proposed approach requires a highly sophisticated management process based on technical methods such as digital engineering, encryption, early detection systems, antivirus tools, highly efficient software, advanced hardware, and personnel with extensive experience in cyber security, etc. The other approach relies on collecting big data in central storage units, classifying the data according to formal classifications, and analyzing these data classifications using modern artificial intelligence methods [6]To build such a system, we utilize mathematical statistics, linear optimization, graphical tools, probability theory, and various AI-based techniques and machine learning classification. These methods are the most commonly used for anomaly detection in big data in the most stringent cyber security systems. Analyzing data according to different classifications and methods using machine learning techniques and inference from the application of comprehensive parallel processing of information on traditional computing tools is It is one of the best and easiest fields for processing databases with high fluidity[8]. As for other fields, it relies on the use of big data technology in databases., may be less accessible. Therefore, everyone is in urgent need of using big data analytical methods using machine learning, and we will work to achieve this approach through the following steps[9].

The methodology focused on several research papers that explored modern and innovative approaches to detecting cyber-attacks through big data analysis using machine learning techniques, which are highly effective in detection processes. The solutions derived include the following contributions:

- A comparative study was conducted on different methodologies that employed various techniques in big data analysis and machine learning, demonstrating their effectiveness in detecting cyber intrusions.

The article structure includes the following:

The structure is divided into several Sections: the introduction, previous work, and an analysis of relevant reviews. The next section also examines the basic approach of machine learning-based detection methods for cyber-attacks in communications networks to combat digital anomalies. The third section presents the data used in proposed sets and structures for network intrusion detection systems and their applications. The final section of the research describes the results, recommendations, and suggestions for researchers in future studies.

The key contributions of this study include the following:

1. Identify the level of risk to big data in government institutions and the potential for achieving cyber security behavior.

2. Explore the role of artificial intelligence technology in supporting cyber security behavior within institutions.

3. Demonstrate the importance of machine learning algorithms in detecting breaches of digital data systems.

4. Determine the level of participation and cooperation in the banks studied in Iraq.

5. Test the relationship and impact between big data and machine learning behavior to reduce cyber security breaches in corporate networks.

2. Related Work

This section highlights some of the literature on the term "big data." This section reviews related work related to these two goals: detecting intrusions within the Internet and predicting cyber-attack events. Recently, cyber security event prediction has gained increasing attention after several restrictions were imposed on some websites that were frequently breached [11]. The authors proposed machine learning classifiers that predict the likelihood of exploiting security vulnerabilities in the future. Smith et al. explored 422,775 reviews to identify seven attributes for assessing customer satisfaction, making them easy to understand and evaluate overall customer satisfaction [12]. Selecting and implementing cyber security countermeasures is typically a costly task. Therefore, most current studies on cyber security event prediction focus on producing accurate predictions using hybrid methods. Some networks are efficient. Branitsky et al [13]. By leveraging massive datasets that include information about past cyber-attacks, such as DDoS vectors and man-in-the-middle attacks, predictive models can predict the likelihood of specific threats and help organizations prioritize their security efforts accordingly. In order to spot intrusion, the traffic created in the network can be broadly categorized into following two categories-normal and anomalous[2]. The difficulty of avoiding cyber tacks in some situations, and the ability of some precautionary measures to mitigate cyber risks. are also important. The dominance of some approaches in abandoning old approaches and replacing them with artificial intelligence is costly for economies. On the other hand, unsupervised learning techniques are used to detect previously unknown or emerging threats. Unlike supervised learning, unsupervised learning does not rely on labeled data; rather, it identifies patterns and anomalies within the data without prior knowledge of what constitutes a threat Derbeko et al [15]. Data mining can play a massive role in the development of a system which can detect network intrusion his paper, Mohamed presents a comparison of five different machine learning models on two well-known datasets for detecting anomalies in IoT networks[13]. To build the safety model, Sarker et al, used ten common classification techniques in machine learning, such as naive Bayes, logistic regression, nearest neighbors algorithm, stochastic gradient, decision tree, stochastic forest, adaptive reinforcement, support vector machine, and extreme gradient reinforcement[14] . Alakito and Ugountimellien used a data-splitting approach on a massive dataset (CIC-Bell-IDS2017) to

independently train three machine learning classifiers before and after feature selection. Then, a big data analytics tool was used to scale up and select features to standardize the data and select the most relevant feature set[16].

3. Material and methodology

3.1 Cyber security

Cyber security refers to protecting electronic systems and networks from threats to critical data and information, and protecting them from infiltration and hacking within the infrastructure of important big data, such as hacking, cyber-attacks, and cyber fraud[16]. Cyber security is one of the most important areas of modern study, prompting countries to invest large sums of money to protect their digital systems, particularly in the financial, military, and economic fields. Governments have encouraged academic institutions to motivate students to study this important field of study[17].

3.2 Artificial Intelligence

It is a science that enables machines to think and perform human tasks with high efficiency, precision, and extreme speed, while reducing effort and financial costs. Artificial intelligence is the most prominent among modern sciences and technologies that use computers to perform impossible tasks and difficult tasks that humans are unable to accomplish. Artificial intelligence has many important fields that are relevant to business, including the military, medical, educational, and agricultural fields[18]. It analyses data, makes decisions, and develops systems and programs that mimic human intelligence. It is used in many fields, such as e-commerce, the military, industry, medicine, and education. To solve data extraction and digital information tasks, modern machine learning methods are used to protect various data. Classification algorithms are the most important for these tasks, and the k-nearest neighbour method is the most common. To solve regression problems, we use linear regression, clustering, and random forests with supervised learning. Naive Bayes (SVM) and k-means are among the methods used in this research. In this paper, we will analyse Principal Component Analysis (PCA)[19]. We use Gaussian Bayes, decision trees, and the two-layer perceptron algorithm, which we implemented in this intrusion detection system. Artificial intelligence provides effective solutions that enable early detection of threats and analysis of behavioural patterns within networks, as well as the ability to respond quickly and effectively to attacks. Graphical Analysis Programs: Such as Big Data Analytics, which enable a comprehensive study of the characteristics of systems and networks[20].

3.3 Advanced Detection Systems Based on Artificial Intelligence (AI) Techniques

Machine learning-based Intrusion Detection Systems (IDSs) (Artificial Neural Networks(ANN) are among the most important AI-powered IDSs. They are used to detect threats through pattern recognition. (SVMs) and partial data classification make the detection process simpler and less complex to classify activities as normal or malicious[21]. Machine learning-based IDSs, such as supervised and unsupervised learning, offer a number of advantages. Predictions based on anomalous network movements have a superior ability to detect anomalies in data. These patterns are useful for handling spatial data (such as network packets)[22]. LSTMs are effective at detecting time-sequenced

attacks such as denial-of-view and Distributed Denial-Of-Service (DDoS) attacks [23]. Autoencoder detect anomalies in data. Anomaly detection systems rely on algorithms such as isolation forest, one-class SVM, and K-Means/DBSCAN to cluster suspicious activity.[24] Principal component analysis (PCA) detects abnormal changes in data. Third, hybrid detection systems combine signature-based and anomaly-based detection and use techniques such as decision trees, neural networks, rule-based logic, and AI classifiers. Generative AI-based systems (such as GPT or GANs) are used to detect, simulate, and understand attacks. GANs are used to generate attack data and test defensive systems[14].

$$count\{z_i | R(z_i) \neq c_i\} \rightarrow min \quad (1)$$

The analysis of database and information subject to classification in cyberspace. We impose the following equation on:

$$F(z) = (v_{1,\dots,.}\, v_n)^T \ .(z - \bar{x}) \quad (2)$$

This problem involves describing the properties of the classified objects, their assigned ci, the vector zi, and the class, where M is the number of elements in the user's candidate data set[25] . This requires developing an algorithm in the R language that allows approximating The information based on the set P represents the vectors. {zi}:

$$f(z) = z^t x^2 . w \quad (3)$$

There are many techniques used to solve problems related to big data analysis, which have the ability to detect all suspicious and anomalous cases and patterns within the web. To protect this data, we use machine learning techniques. These techniques are developed and trained to detect many hidden patterns in the data and provide appropriate analysis[26].

Let's delve into the essence of advanced detection methods and systems based on artificial intelligence techniques, including machine learning[27]. Experimenting with advanced structures to analyse and process the main components of big data reduces the margin of error from the size of the processed data. This is to preserve the original data as much as possible, and this is the essence of the method for handling the z-vector into a new cyberspace:

where v1, ..., and vn are the orthogonal eigenvectors. Matrices are sets of training-ready data (ordered in descending order of eigenvalues). The values of the variables are mathematically expressed as $\bar{x}$, which is the predicted random vector from the training data, ∠T • z is the principal component of the vector z, and the dimension of the new space that can be chosen is n. This method studies the most important combinations of linear features and excludes useless information. To build good classifiers using a machine learning approach, a high-level method is adopted, based on the support vector machine, based on the

various classes and their distance from the closest objects. The mathematical formula for this level is as follows:

$$F^{\{(1)\}(z)} = \text{sign}\left(-b + \sum_{i=1}^{ms} w_i \; x_i^t z \right) (4)$$

The symbols we represent in the previous equations are wi, xi, (i = 1, ..., Ms), and b is the common displacement coefficient. Where wi equals the product of the non-zero Lagrange multipliers with the desired output values, and xi are the vectors responsible for the support (i = 1, ..., Ms). This mathematical model allows us to partition the two basic sets in a linear order. Therefore, we apply several custom transformations to the equation. The k-nearest neighbor method allows you to associate a parsed vector with a class label containing data from the k-nearest examples and perform supervised learning on this vector. The mathematical z formula for this approach is given below:

$$F(2)(z) argmax_{c \in c} \sum_{i=1}^{k} [x_i \, \epsilon c] \qquad (5)$$

Here, the values of x1,...,xk are the evaluation vectors, and C is the smallest value representing the classes among all the training vectors. This method requires no prior variables for initial training.. To run it, you must memorize all the rules of the pre-trained set .To find the linear coefficients for the desired output values, we will focus on the training vectors and linear regression of the variables within the planned system. This is done using the linear equations that follow :We have a symbol for the equations X and y, where X is the symbol for the trained elements in the matrix, y is the symbol for the outputs, and w is the symbol for the vectors and elements for the loadings and weights. This includes the number of sample elements in the matrix (X). This is usually greater than the matrix determinants and the attributes (the number of columns in the matrix X, and the number of required variables). (w1,....., wn). It can be difficult to find a solution to the equations within the system in the model. It is preferable to use the least squares method, and we obtain w = (XT • X)1 • XT • y. This is the value resulting from the equations to produce the following model.

4 - Big Data Analysis Methods in Machine Learning

In fact, the methods we mentioned previously are a collection of different classifiers for hybrid attack detection and protection systems. The method used is to use data classified on two different datasets to determine the effectiveness of the system on both datasets.

• An Internet of Things traffic dataset.

• A computer network traffic dataset that includes host scanning and distributed denial of service (DDoS) attacks [28]

4.1 Detecting Attacks on the Internet of Things Infrastructure

A dataset was selected to train the model on suspicious attacks, which contains massive amounts of data from the digital space of the Internet infrastructure and cloud data. This massive dataset was created based on four devices that coordinate the size of records in tables and provide codes that distinguish fields from each other for ease of processing. The number of records in the dataset is 7,009,270. The training dataset (botnet_attacksN_BaIoT1detection) in Figure (1) illustrates how the experiment was conducted on two botnets (Mirai and BASHLITE) infected with the breach. These botnets consist of a set of physical devices for early warning, such as alarm bells, cameras, thermal systems, and optical detectors. Software applications also rely on highly skilled operators with experience in identifying suspicious attacks on the digital network[29].

Figure 2, there are nine datasets. These categories illustrate the changes that occur to the data after attacks. One of the categories is considered benign, and the remaining eight were attacked [21]. This procedure reduced the size of the analyzed sample by 1.67% and trained the classifiers using CICIDS2017 dataset .

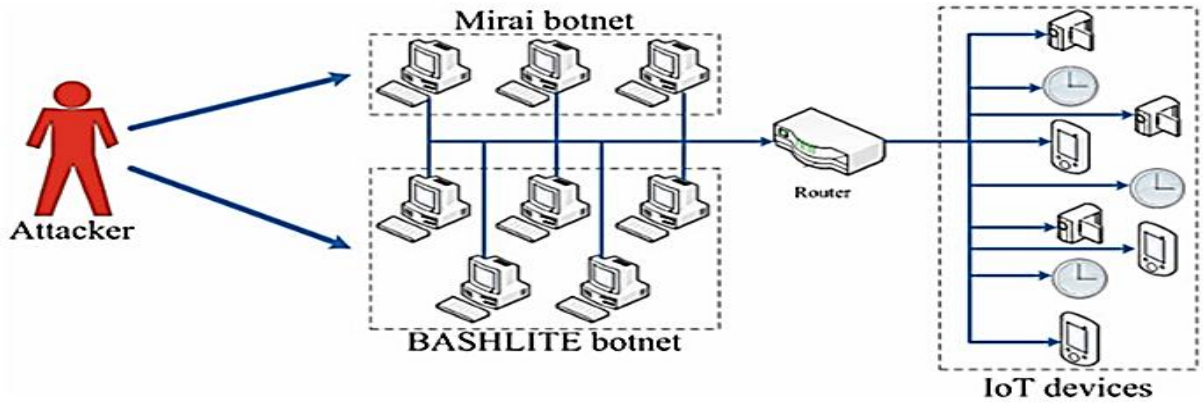


Figure 1. The figure shows a set of devices exposed to an attack from networks (Mirai BASHLITE), which collect all home appliances[12]

The results of applying principal component analysis (PCA) when processing the three components are shown in Figure 3. When using the principal component parts of the sample with 28,000 test elements, the elements are randomly selected. Through experimentation, there are elements that are close to each other from different classes. There are vectors of different dimensions, [18]. some of which are larger in size, which helps improve the model's performance and makes data classification effective over the same period. By removing some unnecessary features from the data, the training process increases and performance is significantly improved, exceeding 98% on the databases used. Conducting training operations systematically on large datasets transforms this curve and improves performance to a horizontal line, indicating excluded components that are not of significant use. The correlations between the 11 principal components and classes depend on the classification of the first rows from the bottom level of the first column[13]. The test results are randomly generated and all principal components are sorted, as shown in the following figure1 [16] . The last component is the most

important in terms of the strength of its correlation with the predicted pairwise class score. It is the highest among the last ten components, or elements, and is equal to 0.57, as illustrated in the detection system environment shown in Figure 3, designed to detect Internet-based attacks within the global network. The idea behind this architecture is to build three layers to enhance performance[24] .

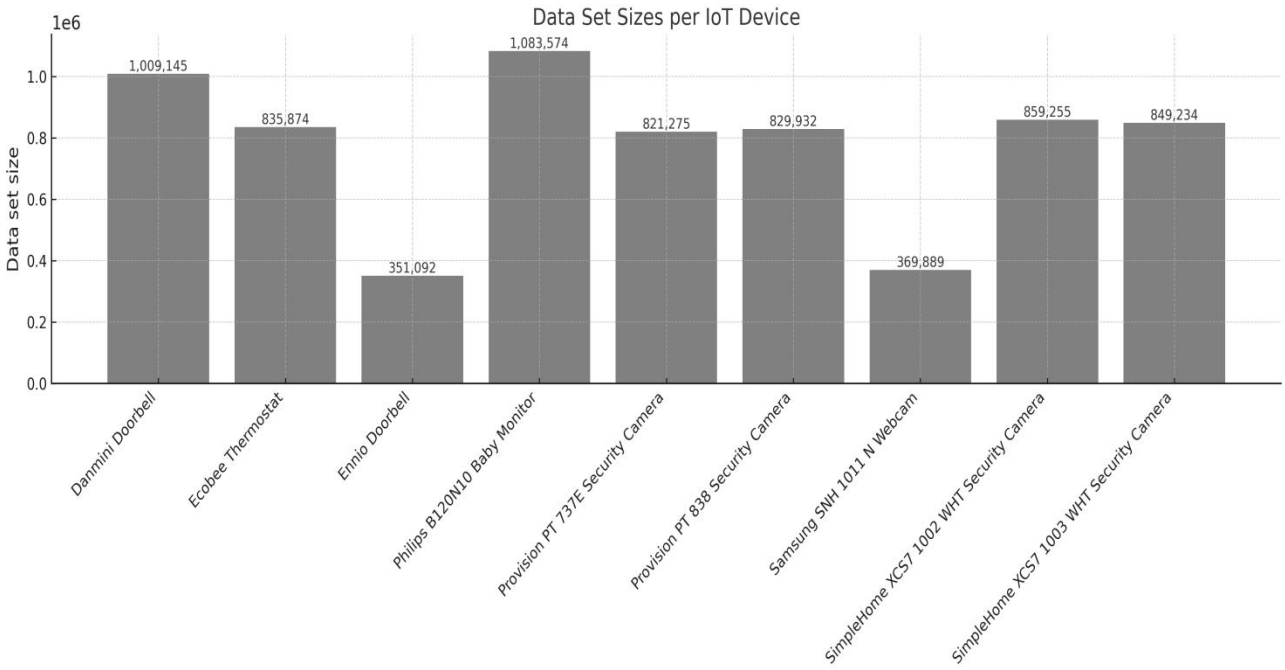


Figure 2 shows the dataset grouped by the number of IoT device records[13]

The model utilizes two approaches to developing classifiers: one for training and one for analysis and processing. Both approaches are compatible, facilitating smooth operation of the device..

- Each approach or method contains three stages:

- Data retrieval and analysis in a streamlined manner.
- Feature vector compression and identification.
- Classifier classification and training of each classifier individually.
- The training steps and the generation of the training sample are performed through several layers responsible for segmenting and analyzing the data. This is the function of the first layer of the model. The second layer prevents overlap between data and separates the classifiers into subsets.

• This allows the classifiers in the second layer to be processed. Also, the branches are processed in parallel using a backup copy of independent sets.

- As we mentioned in the first and second steps, the third step is complementary in that it extracts the classification output from the previous operations and provides reliable information from the analysis and training outputs of the trained indices. The architecture of the Intrusion Detection System (IDS) designed to detect network attacks in the mobile Internet of Things (IoT), shown in Figure 3, follows two approaches to intrusion detection. In the first, attacks are trained, and in the second, test cases are analyzed. These steps are performed in three layers to evaluate performance, sensitivity, and accuracy. Figure 3 illustrates these steps.

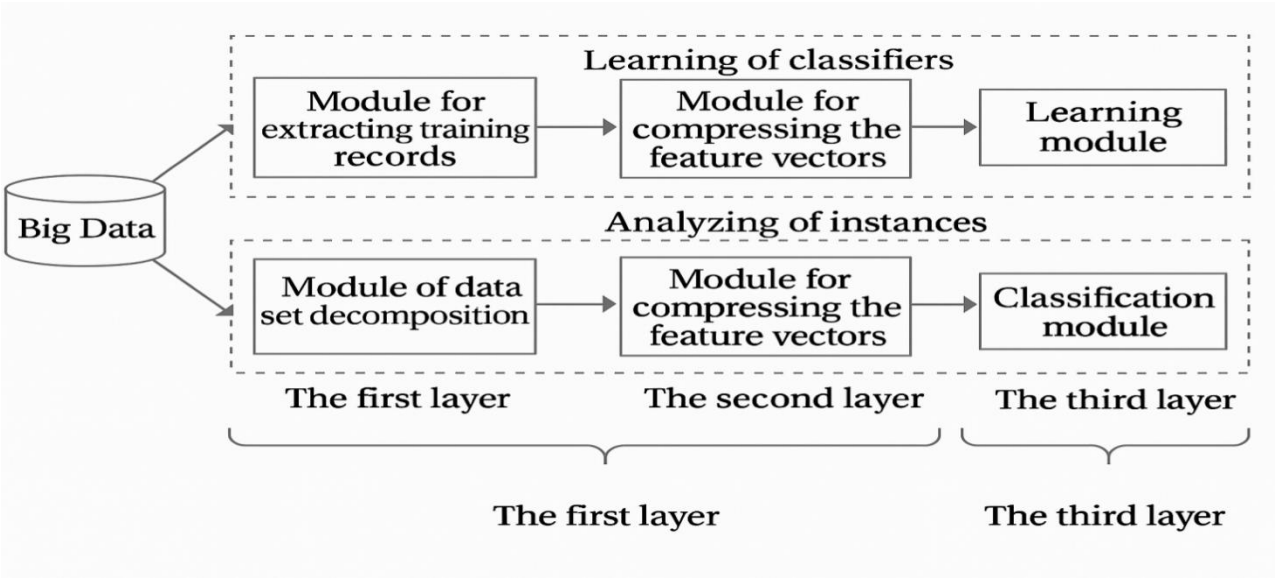


Figure: 3 Multi-layer model for big data classification using feature compression, feature reduction, and three-layer Analysis [7].

The main stages or steps in building a model that processes data:

1. The first step in building the model is to process the data and divide it into main classifiers. The model is built on two layers for detection.

2. In the first layer, a training sample is created and the data is analyzed. A copy of the good results is transferred to the second layer.

3. The second step is to combine the results of the first and second steps of the data processing operations within the layers.

4. In this stage, unnecessary data is removed from subsets that are not related to the system so as not to weaken the algorithm's performance (weight reduction), randomly selected from some subsets.

5. This approach facilitates training cases and increases the efficiency of algorithms designed to process big data.

First, the ensemble workflow map consists of several small parts ($1 \leqslant i < j \leqslant Dk$) to facilitate data processing. We denote each small part by its cluster number D, identify the elements of the matrix, extract the correlation coefficient for each pair of elements from the small part branches, and remove one element ej from Dk, where T is a specified threshold. When analyzing the data, the instance is iterated multiple times until the desired high-accuracy results are obtained, and the same iteration is applied to each element instance. Principal components are analyzed within the classifier. The sub data are compressed to reduce dimensionality and size. After determining the dimensions of the vectors and components, we reduce the sample size from 113 to 10 components. This is a significant percentage, allowing us to preserve 98 percent of the original data after model training. In the next layer, the prediction layer for the features analyzed from the trained elements, this layer fine-tunes the parameters before prediction. The performance of this process is evaluated using a defined classification criterion.

Second, to test the selected data, the classifier was exposed to two datasets from the network on which it was trained. Each dataset consists of thousands of data. The first dataset, on which the model was trained, contains 8,228,295 datasets. The second dataset, the most common on the internet, represents two types of attacks and contains 700,000 active elements. Each shard contains the same number of elements. The CICIDS2017 dataset is the most widely used dataset for detecting distributed denial-of-service attacks and is effective in detections that rely on modern automated methods.

5. Comparative Results Between Classifiers

By comparing the results of Tables 1 and 2, we discovered methods and techniques with high detection capabilities. This was demonstrated through an experiment using several classifiers characterized by accuracy and speed in the training process, as well as their efficiency in detecting and processing anomalies affecting the internet and data. The experiments involved comparing two types of big data related to the Internet of Things (IoT) using the CICIDS 2017 dataset. In fact, comparisons between the most common big datasets revealed the difference between the true positive rate and the false positive rate. A multi-layered scheme was used to integrate all classifiers and perform vector analysis after processing and analyzing all principal components in the formula denoted by Principal Component Analysis (PCA). The symbol "cb" represents the normal traffic class (benign class), and it also represents the Bayesian Gaussian Naive (GNB) algorithm. It also represents the supporting vector machine, the decision tree (DT), and the artificial neural network (ANN). The number of classifiers exceeded 60, all operating on the same scheme. Training time was reduced by training the classifiers on two equal sets of data distributed across two branches to minimize sample size, with parallel training. This approach is successful because splitting the training process into two main sets accelerates results and reduces training time due to the parallelism. The classifier design exhibits high sensitivity in identifying objects belonging to two binary categories. To construct the final model, the binary classifiers were combined into a multi-category model F(i), where i = 1…5. Three clustering methods were used: majority voting (MV), weighted voting (WV), and flexible voting (SV). The maximum

training sample size was 28,700 items, with 2,800 items per category and per IoT device. Testing was conducted using items not used during training (fresh data). The training and testing process was repeated 10 times with randomized data distribution to ensure reliability. Performance and accuracy were determined by the True Positive Rate (TPR) and False Positive Rate (FPR). Comparing the two positive rates showed a clear improvement in accuracy for the combined classifiers (MV, WV, SV) compared to the basic classifiers. The best performance was achieved with the SV classifier, with an increase in accuracy of 4.685% (ACC). The model outperformed the auto-encoding model in the TPR-FPR coefficient (99.8% vs. 99.3%). Using this index, we define two rates for true and false positives using the TPR-FPR metric, and their results are shown in Tables 1 and 2, as an indicator of the classifier's maximum achievement. The basic classifiers improved their ACC scores compared to the SVM, k-NN, GNB, ANN, DT, and SV basic classifiers, and their combinations. Using the combined MV, WV, and SV classifiers for seven IoT devices, we demonstrate the ability of these seven classifiers to reduce redundant features, thus reducing the classifier size. Using this classifier, we obtain highly accurate results, and the classifiers used show similarity in the compared results. As shown in Table 3, a simplified comparison between all the classifiers on the devices is presented.

As shown in Tables 1 and 2, through performance evaluation of Accuracy Index and Key Performance Indicators (KPIs), the Accuracy Index (ACC) actually improved after comparing the basic classifiers SVM, k-NN, GNB, ANN, and DT with the MV, WV, and SV composite classifiers for seven IoT devices. A 4.685% increase in Accuracy Index (ACC) compared to the maximum accuracy value was observed, a characteristic feature of the basic classifiers. This represents good and distinctive performance in the results presented in the device comparison in Tables 1 and 2.

Table (1) Combinations of Maximum Classifier Performance Indicators using the CICIDS 2017 dataset) .

Classifier		Device				
		Provision PT Ej737 Security (%) Camera	Ecobee Thearmostat (%)	Ennio Doorbell (%)	Philips B120N10 Beaby Monitor (%)	Danmini Doorbelle (%)
svm	ACC	98	97.98	99.28	89.8	98.83
	TRR-FPR	99	99.75	98.99	99.9	99.69
k-nn	ACC	98	99.69	98.45	96.8	99.71
	TPR-FPR	98	99.78	99.69	99.87	99.85
GNB	ACC	75	75.91	68.98	79.29	88.69
	TPR-FPR	98	94.91	99.66	99.3	96.87
ANN	ACC	88	99.58	71.84	91.2	98.47
	TPR-FPR	98	99.89	98.99	99.9	89.99

DT	ACC	89	99.76	98.89	98.5	98.99
	TPR-FPR	99	97.98	89.99	79.82	99.99
SV	ACC	99	96.99	87.99	95.57	97.77
	FPR-TPR	99	97.88	98.99	98.77	96.38
WV	ACC	98	88.99	98.99	97.88	99.43
	TPR-FPR	99	99.87	99.88	96.99	98.87

An increase in the ACC index of 4.685% compared to the maximum value of the ACC index, which is a characteristic of the basic classifiers. In fact, we processed the data in parallel by splitting the branches of multiple clusters. We also allocated a separate parallel thread to the problem parts in a partitioned manner. This approach outperforms the auto-supervisor proposed by Meidanetal (2018) in terms of the TPR–FPR ratio (99.8% (for DT) compared to 98.6%)[14]. The dependence of processing time on the number of threads was demonstrated by our dependence on the device when the number of threads varied in order. Data processing improved by 7369 and 7,065 times when moving from one thread to eight threads, E respectively, for the test and training sets.

Table (2) The Output Of The Classifications and Their Groups For The Maximum Values of The Indicators

Device					
Classifier		Samsung SNH 1011 N Webcam (%)	Provision PT 838 Security Camera (%)	Simple Home XCS7 1003 WHT Security Camera (%)	Simple Home XCS7 1002 WHT Security Camera (%)
Svm	ACC	9815	99.22	98.74	92.14
	TRR-FPR	98.87	99.8	99.82	98.82
k-nn	ACC	97.48	99.68	99.63	89.65
	TPRf-FPR	99.77	99.75	98.71	89.75
GNB	ACC	75.98	66.32	70.84	68.99
	TPR-FPR	99.71	99.78	96.88	99.17
ANN	ACC	88.7	99.85	99.69	97.63

	TPR-FPR	99.75	99.65	97.99	99.88
DT	ACC	98.05	98.58	99.88	99.11
	TPR-FPR	99.85	97.88	96.99	99.81
SV	ACC	98.80	99.87	98.89	99.05
	FPR-TPR	99.81	98.76	99.25	98.74
WV	ACC	98.88	99.99	99.36	98.99
	TPR-FPR	99.73	98.89	99.87	99.88

The GNB has the shortest training time among the classifiers due to its k-NN scope. The training process is limited to maintaining the match between the vectors, and the testing time depends on the type of classification. The neural network has the shortest data processing time among the base classifiers and between the clustering constructs. Multi-voting and weighted voting have the longest testing time for the vectors compared to other data clustering constructs.

Table (3) Results of the top classifier for each device for the detection system.

Classifier	ACC	TPR-FPR
SVM	96.38%	99.47%
k-NN	97.33%	99.64%
GNB	77.97%	97.35%
ANN	89.42%	97.35%
DT	97.63%	93.96%
SV	95.86%	98.20%
WV	96.46%	99.52%

We used the two-way interrogation method within data traffic for training. We studied various types of data, the characteristics of each type, and the attacks they are exposed to, which are the most dangerous. We identified training methods for the proposed models, which rely on machine learning

techniques and anomaly detection algorithms within the global network. These modern methods are used in early detection of malicious intrusions, enabled by the implementation of difficult-to-penetrate logistics. The combination of artificial intelligence, human expertise, and advanced technical devices has made it impossible to hack and infiltrate the network. The training and testing samples were formed by dividing the original large dataset roughly in half: 257 training items, 105 items, and 257 test items. Each of these samples contains approximately equal amounts of items from the three categories (port scanning, DDoS attacks, and benign traffic) in the databases from each sensor on multiple network analyzers and balancers.

The proposed system architecture is designed to simplify data processing and analysis to facilitate data flow within the network. The model has several features for detecting suspicious activity within the network. The detection process involves several steps and levels of analysis and early prediction of suspicious activity, which in turn relies on sequential and systematic algorithm optimization.The network packet intrusion detection system relies on machine learning methods and electronic devices that operate on thermal sensors and analyze anomalies within the data packets that are classified and categorized using machine learning algorithms. The anomaly detection system has superior ability to analyze all fragmented data, and the main task of the balancers is to distribute the network load among several analyzers. To process this, we used and studied several classifiers: decision trees, logistic regression, support vector machines, and clustering. Detailed performance indicators for each category are presented in [Figure 1]. The weighted voting method performs slightly better than other clustering formulas. Accuracy increased by 4.5-5% using weighted aggregation compared to the same benchmark shown by the baseline classifiers. Compared to the methods studied by Sharaf al-Din et al. our approach shows similar F-score values (97-98%). However, our approach trained this system in several stages, and the results within the curve were more accurate each time, with the percentage increasing by a constant amount until the percentage was similar to the F-score (99%).

Tables 4, 5, and 6 show the results after training on different samples. Detailed performance indicators are presented for each category. The decision tree exhibits slight deviations from the mean values of the indicators. Both logistic regression and the supporting vector machine rely on custom parameter initialization compared to the two previous classifiers, and the accuracy has increased by 5% compared to the same indicator shown by the base classifiers.

Table 4: Performance score for the secure data flow metric.

Classifiers and their weighted ensembles	Benign		
	f-Precision (%)	R-Precision (%)	F- Recall(%)
. Logistic regression	97	93	91
Decision tree.	89	88	93
Soft voting machine	98	72	91

Weighted voting	97	98	98
Support vector	98	99	93
machine	99	94	95

Table 5: Performance score for protection against distributed denial-of-service attacks.

Classifiers and their weighted ensembles	Benign		
	F-measure(%)	f-Precision (%)	F- Recall(%)
Soft -voting	75	98	90
- Support- vector machine	90	95	93
Logistic- regression	95	96	96
Weighted -voting	98	96	97
Decision- tree	96	94	98
Adeaboost	95	98	97

Table6 : Performance score for model validation.

Classifiers and their weighted ensembles	Benign		
	Precision (%)	Recalle(%)	F-measure(%)
Decision- tree	98	97	98
Logistic- regression	90	99	95
Support- vector machine	80	100	90
- voting Adiaboos-	96	99	98
Soft voting	95	98	98
- Weighted- voting	97	98	97

With this technique, can add multiple new classifiers to the detector without pre-training them.

This approach is ideal for detecting new attacks. It relies on testing the training and base classifiers at a specific time. It doesn't require parallel training of datasets, which significantly reduces the training and testing time. Increasing the number of threads consumes a lot of memory, and the memory consumption is increased due to the increased training phases. Reducing the number of training cycles increases detection speed.

.As shown in Figure 4, the high-performance representation diagram with the flowing data is displayed. In the results presentation program, Tables 4, 5, and 6 are presented as charts illustrating the best results by comparing them to the previously mentioned methods. Figure 5 shows the performance of the DDoS protection metric from the results in Table 5 using bar graphs and curved lines, as seen in Figure 6. This clearly demonstrates the superiority of the model's actual response rate compared to the other models discussed by the researcher. The researcher also included a comparison in Figure 7 to illustrate the degree of superiority among the best-performing methods shown in the bar graph. All these charts are table data after application using the SPSS statistical graphing editor.

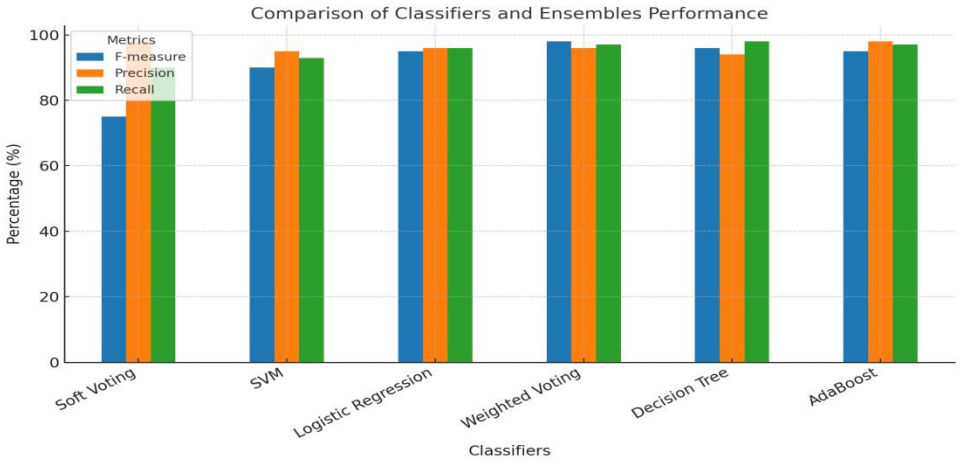


Figure 4: High-performance representation diagram with streaming data flow

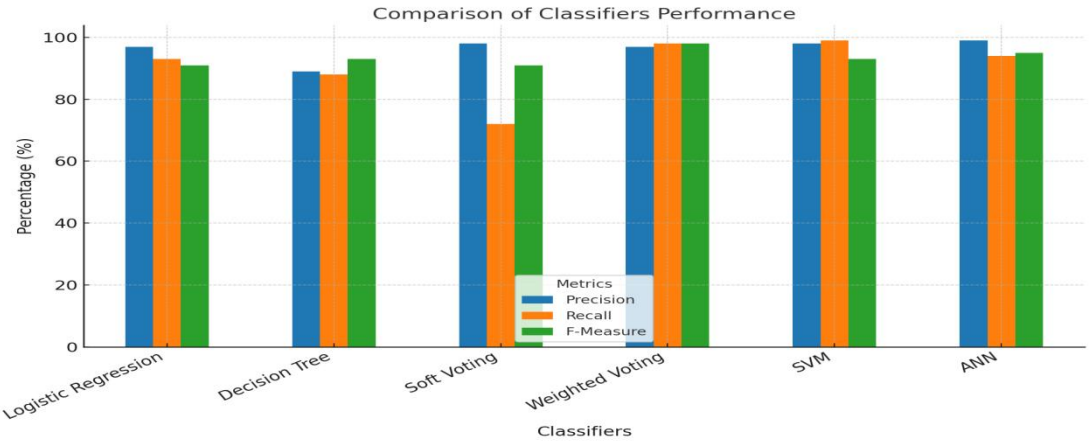
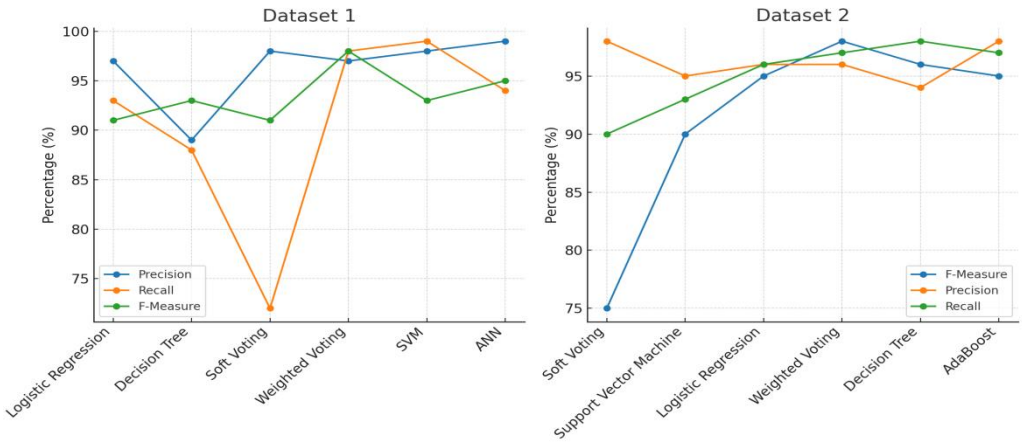


Figure 5: Performance of the metric for protection against distributed denial of service attacks.



. Figure 6: Model validation and performance results using bar charts.

Figure 7: Comparison of the highest performing methods.

6. Discussion:

Through the presentation of the methods and approaches we have discussed, as a comparative study of the main models in the process of detecting cyber-attacks, some methods and approaches have demonstrated success in detection processes. The results confirm that the systematic combination of classifiers enhances the effectiveness of cyber detection systems in large and complex data environments. They also highlight the effectiveness of modern artificial intelligence methods in preventing denial-of-service attacks. In fact, the detection of dual classifiers, which operate sequentially and process the data used for training and classification simultaneously to avoid wasting time and reduce irrelevant features to decrease the data size during processing, is consistent with previous machine learning methods that use the nearest neighbor (k-NN) technique. These results are consistent with previous studies in methodologies and research that have shown similar effectiveness. These methods include: Supporting Vector Machines (SVMs), k-NNs, GNBs, Artificial Neural Networks (ANNs), Decision Trees (DTs), Statistical Analysis (SV), and their combinations. Using the combined MV, WV, and SV classifiers for seven IoT devices, we demonstrate the ability of these seven classifiers to reduce redundant features, thus decreasing the classifier size. In fact, improving the model's accuracy and contextualization will enhance its practical application. Furthermore, research into transfer machine learning methods can enable the model to easily detect cyber-attacks using new and large datasets, while simplifying the retraining process. In conclusion, the study's analysis of the methods mentioned in this comparison reveals that each has its advantages and disadvantages in this field. However, it shows that the combined SV method for detecting seven IoT devices is the optimal method because it can reduce redundant features and yields better results in the detection process, reducing detection time and efficiency by focusing on the main data features and avoiding unnecessary redundancy in detection processes, as shown in Tables 1 and 2.

7. Conclusion:

This paper presents a comparative study of methods and techniques used in big data analysis through machine learning, which is considered crucial in mitigating cyber-attacks. These attacks are currently the cornerstone of security for digital networks and communications. The core of this comparative study, which examines multiple models, lies in simplifying the problem of data classification to a high degree of accuracy. This is achieved by applying principal component analysis, first processing the underlying dataset, and then employing advanced machine learning techniques (such as vector support machines, nearest neighbors, logistic regression, game theory, decision trees, and simple Gaussian Bayesian algorithms) to build and share reliable and intelligent classifiers. This approach was validated using two proposed test datasets. The first, selected from the internet, comprises 7,009,270 cases. The second, selected from the CICIDS 2017 dataset, illustrates two types of attacks: distributed denial-of-service (DDoS) attacks and window scan attacks. It contains over 500,000 cases, equally divided into two sections. Numerous datasets were analyzed, resulting in a comparison between two primary classifiers: majority voting and weighted voting. These are

two different models for distributed intrusion detection system architectures. This system is classified as a distributed network cyber-attack detection system. Furthermore, it is recommended to develop and train staff in artificial intelligence techniques to enhance their efficiency. It is also recommended to increase training programs on data models and design methodologies capable of faster attack retaliation. Technological advancements in computer hardware and processors capable of handling big data with high efficiency and accuracy must be kept pace. Strengthening research and development in the academic field and generating new and innovative ideas is also recommended.

References:

[1] Behiry, M. H., & Aly, M. (2024). Cyberattack detection in wireless sensor networks using a hybrid feature reduction technique with AI and machine learning methods. *Journal of Big Data*, *11*(1), 16.

[2] Choudhury, S., & Bhowal, A. (2015, May). Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. In 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM) (pp. 89-95). IEEE.

[3] Inuwa, M. M., & Das, R. (2024). A comparative analysis of various machine learning methods for anomaly detection in cyber-attacks on IoT networks. Internet of Things, 26, 101162.

[4] Sarker, I. H. (2021). Cyber Learning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks. Internet of Things, 14, 100393.

[5] Alaketu, M. A., Oguntimilehin, A., Olatunji, K. A., Abiola, O. B., Badeji-Ajisafe, B., Akinduyite, C. O., ... & Okebule, T. (2024). Comparative analysis of intrusion detection models using big data analytics and machine learning techniques. Int. Arab J. Inf. Technol., 21(2), 326-337.

[6] Alaketu, M. A., Oguntimilehin, A., Olatunji, K. A., Abiola, O. B., Badeji-Ajisafe, B., Akinduyite, C. O., ... & Okebule, T. (2024). Comparative analysis of intrusion detection models using big data analytics and machine learning techniques. *Int. Arab J. Inf. Technol.*, *21*(2), 326-337. Alaketu, M. A., Oguntimilehin, A., Olatunji, K. A., Abiola, O. B., Badeji-Ajisafe, B., Akinduyite, C. O., ... & Okebule, T. (2024). Comparative analysis of intrusion detection models using big data analytics and machine learning techniques. Int. Arab J. Inf. Technol., 21(2), 326-337.

[7] Nabi, F., & Zhou, X. (2024). Enhancing intrusion detection systems through dimensionality reduction: A comparative study of machine learning techniques for cyber security. Cyber Security and Applications, 2, 100033

[8] Azam, Z., Islam, M. M., & Huda, M. N. (2023). Comparative analysis of intrusion detection systems and machine learning-based model analysis through decision tree. Ieee Access, 11, 80348-80391.

[9] Nyame, L., Marfo-Ahenkorah, E., Abrahams, A., Ashley-Osuzoka, J., Ashong, G., & Aboagye, D. (2024). Rise in Cyber Threats in the United States and the Need for Advanced Cyber Risk Mitigation Tools and Adequate Skills to Combat Cyber Threats.

[10] Alauthman, M., Aldweesh, A., Al-Qerem, A., Daoud, I., Alkasassbeh, M., & Gawanmeh, A. (2025, April). Evaluating Reinforcement Learning Reward Functions for APT Detection in Industrial IoT Systems. In 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA) (pp. 1-6). IEEE.

[11] Huertas, L. M. (2025). Using Improvement Science and Participatory Action Research to Enhance Critical Thinking in First-Generation Hispanic Female STEM Students (Doctoral dissertation, Barry University).

[12] Erendor, M. E. (Ed.). (2024). Cyber Security in the Age of Artificial Intelligence and Autonomous Weapons. CRC Press.

[13] Alsodi, O., Zhou, X., Gururajan, R., Shrestha, A., & Btoush, E. (2025). From Tweets to Threats: A Survey of Cybersecurity Threat Detection Challenges, AI-Based Solutions and Potential Opportunities in X. Applied Sciences, 15(7), 3898.

[14] Sindiramutty, S. R. (2023). Autonomous threat hunting: A future paradigm for AI-driven threat intelligence. arXiv preprint arXiv:2401.00286.

[15] Wang, L., Chen, J., & Zhang, X. (2021). Real-time AI-driven cybersecurity threat detection and response. Journal of CyberIntelligence, 14(2), 78-95.

[16] Johnson, M., & Miller, R AI-powered (2022).defense mechanisms: utomated containment and response strategies. Cybersecurity Review, 10(4), 33-50.

[17] Priyadharshini, S. L., Al Mamun, M. A., Khandakar, S., Prince, N. N. U., Shnain, A. H., Abdelghafour, Z. A., & Brahim, S. M. (2024). Unlocking Cybersecurity Value through Advance Technology and Analytics from Data to Insight. Nanotechnology Perceptions, 202-210.

[18] Smith, J., & Johnson, (2022). RPredictive analytics in cybersecurity: Leveraging AI for proactive defense. Cyber Threat Intelligence Journal, 15(3), 45-62.

[19] Rajagopal, N. K., Qureshi, N. I., Durga, S., Ramirez Asis, E. H., Huerta Soto, R. M., Gupta, S. K., & Deepak, S Future of Business Culture: An Artificial Intelligence-Driven Digital Framework for Organization Decision-Making Process. Complexity, 54:2022: 7796507. https://doi.org/10.1155/2022/7796507

[20] Derbeko P, Dolev S, Gudes E, Sharma S (2016) Security and privacy aspects in MapReduce on clouds: a survey. Comp Sci Rev 20:1–28. https://doi.org/10.1016/j.cosrev.2016.05.001.

[21] Muthusubramanian, M., Mohamed, I. A., & Pakalapati, N. (2024). Machine learning for cybersecurity threat detection and prevention. Int. J. Innov. Sci. Res. Technol, 9(2), 1470-1476.

[22] Abimbola, O., & Idris, O. O. (2025). A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review. Path of Science, 11(3), 4009-4020.

[23] World Journal of Advanced Research and Reviews. GSC Online Press; 2024. p. 1778–90. Available from: https://dx.doi.org/10.30574/wjarr.2024.23.2.2550

[24] Inuwa, M. M., & Das, R. (2024). A comparative analysis of various machine learning methods for anomaly detection in cyber attacks on IoT networks. *Internet of Things*, *26*, 101162.

[25] Rodrigues GA, Serrano AL, Vergara GF, Albuquerque RD, Nze GD. Impact, Compliance, and Countermeasures in Relation to Data Breaches in Publicly Traded US Companies. Future Internet. 2024 Jun 5;16(6):201.

[26] Kryparos G. Information security in the realm of FinTech. InThe Rise and Development of FinTech 2018 Feb 15 (pp. 43-65).

[27] Joseph Nnaemeka Chukwunweike, Moshood Yussuf, Oluwatobiloba Okusi, Temitope Oluwatobi Bakare, Ayokunle J. Abisola. The role of deep learning in ensuring privacy integrity and security: Applications in AI-driven cybersecurity solutions [Internet]. Vol.

[28] Darem AA, Alhashmi AA, Alkhaldi TM, Alashjaee AM, Alanazi SM, Ebad SA. Cyber threats classifications and countermeasures in banking and financial sector. IEEE Access. 2023 Oct 23;11:125138-58.

[29] Abimbola, O., & Idris, O. O. (2025). A Critical Cyber security Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review. Path of Science, 11(3), 4009-4020.

[30] Abimbola, O., & Idris, O. O. (2025). A Critical Cyber security Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review. Path of Science, 11(3), 4009-4020.