Development a data mining techniques to detect and prevention cyber attack for cybersecurity

Ahmed Shihab Ahmed[1,*], Hussein Ali Salah[2]

[1]*Department of Basic Sciences, College of Nursing, University of Baghdad, IRAQ*

[2]*Department of Computer Systems, Technical Institute- Suwaira, Middle Technical University, IRAQ*

[*]*Corresponding author: ahmedshihabinfo@conursing.uobaghdad.edu.iq*

Received 12 Aug, 2023, Accepted 27 Nov. 2023, published 1 Jun 2024.

DOI: 10.52113/2/11.01.2024/1-27

ABSTRAC: The identification and forecasting of cyber-attacks is crucial process. In this article, we describe a paradigm for cyber security that makes use of data mining to forecast cyberattacks and identify appropriate countermeasures. The framework's two primary elements are the surveillance and prevention of cyberattacks. The system constructs a predictive model to forecast future cyberattacks after first extracting appropriate timing with cyberattacks from previous data that used a decision tree based on the J48 algorithm. A variety of cyber-attacks, involving DDoS, port scans, and SQL Injection, are described in the datasets. The suggested framework effectively recognizes cyberattacks and gives patterns associated with them. The suggested predictive algorithm for identifying cyberattacks has a 99% average prediction performance. The predictions model's test outcomes demonstrate how effective it is at spotting potential cyberattacks in the future. Moreover, solutions like malware detection and monitoring were provided using data mining. Given the state of computer networks today Users of computer networks ought to take security very seriously. In this article, implications of data mining for risk evaluation and identification were highlighted, along with a unique method for quickly and accurately detecting malware.

Keywords: Cyber Attack Prediction, Data Mining, Decision Tree, Accuracy, Malware, Security Prevention.

1. Introduction

A group of policies and tools collectively referred to as "cybersecurity" are meant to safeguard our information systems, communications, and data from intrusions, attacks, and interruptions [1]. They employ a range of cyber protective actions in an attempt to protect the privacy, accuracy, and accessibility of information as well as data monitoring systems. In an expanding cooperative effort, investigators and experts in cyber security from institutions, businesses, educational establishments, and government organizations have been developing a variety of cyber defense systems to guard against potentially unfriendly assaults on cyber infrastructure. Responsive security measures like detection systems for intrusions form the second layer of defence in networked computers (IDSs). IDSs employ data gathered from log data and network connections to identify and discover intrusion occurrences, assess the level of damage produced by intrusions, find hackers, and stop future assaults [2].

Data mining is the practice of analyzing data from a certain source and combining that input into valuable information. KDD, which stands for "knowledge discovery," is another name for this process. Regarding online data security. Data is being mined using various techniques to find potentially dangerous circumstances [3].

To identify cyberattacks, there are a few simple methods that can be utilized, such as the following: With encryption, data can be secured by converting it data into an unreadable format (cypher text). The only people who can understand this communication are those who have the secret key. Intrusion detection is the method of analyzing data gathered from computer networks and systems to ascertain whether a network intrusion or safety infraction has taken place. A kind of evaluation in which assessors probe a network or security schemes in search of weaknesses [4] is known as penetration testing. Web users must become knowledgeable on how to protect themselves from online fraud as well as identity theft. The right behaviour and effective system security can lead to a reduction in risks and an improvement in online safety. Due to scarce resources and inadequate cyber security understanding, small and medium-sized businesses face a variety of security issues. There are numerous categories into which these problems can be divided [5].

Cybersecurity is different from data security in terms of scope and desired use. Although the phrases are frequently used interchangeably, cybersecurity is a subtype of information security while information security is a larger category. Endpoint protection, encryption keys, and cybersecurity are a few instances of data protection. Another illustration is physical security. It is also closely related to information security, which safeguards data against risks like natural catastrophes and lapses in server reliability. Cybersecurity primarily focuses on the threats that could be created by technological advances in addition to the methods and apparatuses that could be employed to prevent or lessen such risks. Data protection is a related topic that focuses on preventing an organisation’s information from being accidentally or maliciously leaked to third party companies who are not authorized to see it [6].

In recent years, the scientific community has paid a lot of attention to the topic of cyber security. Data protection, systems engineering, and other electronic assets of a company from cyberattacks include datasets, equipment, software, and associated infrastructures. Both purposeful and unintended cyberattacks are possible. Deliberate cyber-attacks occur when specific computer software or other technology infrastructure is used to disrupt an institution's system. For instance, maliciously inserted software defects in computer applications. Ignorance, attention problems, or a lack of knowledge about cyber security are the causes of accidental cyberattacks. For instance, programmer, operator, and server manager errors. Organizations frequently experience cyberattacks as a result of improper adherence to security protocols. However, as cybercriminals get more skilled and exchange security flaws with one another via the deep web, cyberattacks are still rising [7]. In order to avoid cyberattacks, detection and avoidance of intrusion technologies are being developed [8].

The contribution of the article is as follows:

(1) use previous data to find patterns connected to cyberattacks; (2) use the trends to forecast upcoming cyberattacks on a real system; and (3) manage processes in making the right interventions to lessen cyberattacks. Employing the six available to the public intrusion prevention networks (IDS) datasets supplied by the Canadian Institute of Cyber security [26], we assess the efficacy of our suggested method. Every dataset's predictive algorithm has a 99% average accuracy, demonstrating the excellence of the conceptual model. As an appropriate security measure, it is also vital to adopt effective as well as intelligent strategies for early diagnosis of cyber risks in order to reduce the danger of catastrophic cyberattacks including data breaches, computer hacking, and DDoS attacks. One of the challenges for security professionals is malware identification. Malware detection is aided by the integration of data mining techniques such as categorization, SVM, extrapolation, decision trees, network mining, and KNN methods with anti-threat systems.

As a result, in this Article, we describe a system that uses a data mining technique to identify patterns linked to cyber-attacks from previous network data and then uses these trends to forecast potential attacks. The suggested cyber security architecture can be used to foresee cyberattacks on a real-time network.

The remaining portions of the essay are structured as follows: The related work is discussed in Section 2. The classifications for cyber security are explained in Section 3. The architecture that utilizes data mining to anticipate cyberattacks for cyber security is covered in Section 4. Introduction to Data Mining (Dl) and Machine Learning (Ml) for Cyber Security is covered in Section 5. A technique for combining network data security protection with web data mining technologies 6. The experimental findings and discussion are presented in Section 7. Lastly, Section 8 presents the paper's conclusion.

2. Related Work

The involvement of computer security has changed to offer the best prevention for data over the network as a consequence of the growth of the web and the enormous amount of information being sent each second, as well as the techniques for safeguarding and maintaining it and differentiating those who are approved to view it. In this work, the researchers investigate the function of data mining techniques in computer security. Cybersecurity and public safety are just two applications of data mining for safety. National security issues include assaults on institutions and the damage of crucial infrastructure, including electricity grids and copper telephone wiring. Protecting networked and personal computer infrastructure from hazardous malware, including Trojan horses and malware, is the focus of cybersecurity. Moreover, products like vulnerability scanning and monitoring are being delivered through data mining [9].

This lesson examines cybersecurity procedures that safeguard computer networks from intrusions, hacker attacks, and information theft, as well as the place of machine learning in this field. Also, the most important research on the use of deep computing and machine learning in cybersecurity is summarized in this section. Findings demonstrate the importance of machine learning and deep learning methods in preventing unauthorized access to computer systems and in managing system infiltration by foreseeing and comprehending the behaviour and traffic of harmful software [10].

Due to the increased use of the internet, detection and prevention of intrusions are becoming essential issues. Many methods to avoid or find intrusion in a network have been suggested in the past. Yet the majority of methods used nowadays in IDS detection are unable to resolve this issue in an efficient manner. In addition to this, machine learning (ML) has been included in a number of applications due to its success in producing accurate results for each application. So, the focus of this work is How deep learning and data mining can be utilized to identify IDS in a network" in the near future. Detection accuracy rates, fewer instances of false alarms, and reduced communication costs are all benefits of using effective methods like categorization, analysis, etc. in machine learning (ML) [11].

This research examines the typical configurations and operating principles of Dos assaults, as well as the available defence detection techniques. A computerized data mining model based on the AKN method is suggested after investigating many widely used computer data mining techniques, including clustering, categorization, neural network, extrapolation, connection, and webpage data mining algorithms. Moreover, preventive detection trials against Dos attacks are carried out using computerized data mining techniques relying on the AKN algorithm, and it is compared to conventional algorithms. A detection accuracy rate of more than 97% and a recognition efficiency increase of more than 20% are demonstrated by the test findings, which demonstrate that studies based on the AKN method have stronger defensive detection effects than conventional techniques [12].

From the point of thought, they should prioritize data security personalities. Consumers of computer systems should be very protective, deepen their understanding of network data security, and appreciate the relevance and value of network security for computers. Web usage mining technologies is a solid option between them [13]. The suggested framework effectively recognizes cyberattacks and provides patterns associated with them. The suggested predictive algorithm for identifying cyberattacks has a 99% overall accuracy rate. Future cyberattacks can be predicted using the patterns extracted from the forecasting model based on past data.. The forecasting model's results from experiments demonstrate how effective it is at spotting potential cyberattacks in the future [14].

A comprehensive intrusion prevention architecture with a detector for handwriting attack predictions and a database to detect outlier was proposed in the article [15]. The database contains all of the anomalous behaviours that have been independently or jointly detected by the computer or user. It's frequently utilized as an internet alternative due to how quickly it can be implemented [16]. Wang and Jones have presented a comparison analysis using probability and future ML techniques and procedures, including Nave Bayes and Gaussian in addition to those of decision tree and random forests, for the identification of attackers and their hostile behaviors. Today, a large number of retraining data sets built from KDD99 have been utilized for effective operation, and each approach has been applied to attack types like DoS, Probe, R2L, and U2R with an appropriate analysis of the outcomes. The data transformation used to create these datasets supports the claim that the KDD characteristics are distinct compared to the others and utilize different feature scaling. Both decision trees (DT) and random forests (RF) exhibit reliable procedures and outcomes in the detection of denial-of-service attacks. In contrast, the outcomes of Gaussian and Naive Bayes are significantly better in a handful of the different attack categories, such as Probe, R2L, and U2R. In light of the results, the article concludes that probabilistic strategies for malware detection are much more durable than response-based strategies.

3. Cyber Security’s Classifications

Several strategies can be used to duplicate an attack. The assault may take the form of a sequence or signature that is utilized to spot the deviance. They will undoubtedly be able to identify the majority, or most, of the common assault methods. Therefore, when dealing with small or unknown assault patterns, they end up being of little use. These systems make an effort to identify and distinguish "bad" activity. The main challenge is coming up with a trademark that incorporates all the different types of a persistent attacks. Many different machine learning approaches have been used to find abuse of these systems. By connecting the usual operations with the acts that would be anticipated of an attacker, such detection methods are shown to be helpful in identifying epidemics on the network [17].

A system to recognize and categorize network events based on artificial neural networks was presented by the researcher in [18]. (ANN). The sources of data are constructed using an array of formats, including constrained, imperfect, and nonlinear ones. They developed a data identification system that makes use of neural networks' superior analytical capabilities. By creating an architecture with four completely connected levels, a multi-layer categorization system employing MLP is employed to identify misuse. There are 2 output nodes and 9 input terminals in the neural network architecture. Three stages of data pre-processing were used, including: Procedure Identity (PID) - the set of guidelines that govern an event (TCP = 0, UDP = 1, ICMP = 2, and Unknown = 3). Aecopd, Source Port, Target, and Source (IP address corresponding to a source) Address of the final destination (IP address of a destination) g) Protocol ICMP (like echo requests, null, etc.) h) Direct Exposure Size g) ICMP Code (length of the data packets) 1. Source Data.

For 12,000 iterations of the chosen training examples, a back-propagation technique was used to train the neural network design. 11.532 data points were examined; 980 were picked at random for assessment, and the remainder of the 980 were utilized to educate the system. It took the model of neural networks 28.12 hours to finish. The findings show that the average root-mean-square error on the learning algorithm is 0.076186 and that it is 0.058918 on the testing data. Ultimately, depending on RMS, each datagram was assigned to either a regular or an assault set, yielding a reliability of 95%.

To identify the intriguing characteristics and categorize each link as a load bearing capacity or an attack, they used mining association rules. To determine the relationship between TCP/IP characteristics and the different types of attacks on the CICDoS2017 data set, they suggested classification models. Less restriction is kept because of the rules they created. For additional network records, a C4.5 predictor is used after the rules have been constructed. The CICDDoS2017 dataset served as the basis for the studies. In the first seven and following two weeks, respectively, a machine learning model and testing dataset are created. According to the findings, the average recognition rates were 98%, 96%, 87%, and 78%, correspondingly. Associations rule mining's key issue is that the resulting rules may represent correlation, although the method has promise for creating attack signatures. Further in [19], the authors proposed an algorithm to use the existing signature data and find the signature of the related attack in less time. They observed that the Trademark Apriori (SA) technique, which is founded on Apriori, requires less computation than their method. These techniques can be utilized in abuse detection methods like Snort to create new signatures. In light of the current signature, the suggested method discovers a fresh attack signature. The scan minimization technique is also used to cut down on the time needed to scan databases. Comparing this approach to the Signature Apriori procedure, it entails efficiently determining a new attacking signature. The data mining strategy has been used by the investigators to supplement IDS's network-based signature generation. By doing this, fingerprints are generated for both misuse and abuse according to traffic composition and misappropriations based on data transmissions. The common association rules method, known as the Apriori algorithm, is the foundation of the Signature Apriori (SA).

42 among of the 48 malware IRCs were successfully categorized by the naive Bayes classifiers (which produce a false negative rate (FNR) of 9.76%). [20]. It was harder to accurately classify IRC traffic as botnet or non-botnet in Stage (ii) by applying categorization. A methodology for vulnerability scanning using a naive Bayesian network was proposed by the author in [21] and is known as an adaptable detection system for intrusions. The new invasion signatures like DoS, r21, u2r, and probe are discovered using the CICDDoS2017 dataset, which contains 48 assaults. Eight characteristics in the Bayesian network make up the dataset, including the following: Connection type, Service, Land, Incorrect fragmentation, many unsuccessful login, signed in, root terminal, and is visitor logins. A connection tree reasoning technique is employed in the initial stage to determine whether the data is normal or malicious, with average detection rates of 92.74% for regular and 93.73% for incursion. The dataset is divided into six different categories in the second phase: DoS, poking and prodding, R2L, and U2R. According to the effectiveness, the diagnostic accuracy for DoS is 93.62%, for probe it is 98.36%, for R2L it is

23.48%, for U2R it is 8.48%, and for other categories it is 79.46%.

4. A Data Mining Structure for Cyber Security to Forecasting Cyber – Attacks

We provide a methodology for the cyber-attack forecasting for data security in this part. The framework consists of six fundamental steps, which we list below.

Step 1: Prepare the data

Step 2: Create a J48 decision tree using the prepared data

Step 3: The decision tree's feature retrieval

Step 4: Acquire future data from the connection step 4

Step 5: anticipate attacks.

Step 6: Intervention to lessen cyberattacks [22].

We now go into great depth on the application's steps. In predicting assaults, we first gathered previous system datasets from the source. For the sake of this article, we assume that a datasets D has a collection of records (R1, R2, R3,..., Rn) and a variety of characteristics (A1, A2, A3,..., Am). A dataset's attribute may be either quantitative or categorical. Keep in mind that the previous network collection must include both regular network events (that is, records devoid of any cyberattacks) and information affected by cyberattacks. To create an appropriate dataset for the forecasting model, we pre-process the information.

```
Algorithm: Predict Cyber Attack
Input: A dataset D= {$R_1$, $R_2$, $R_3$,…, $R_n$} /*n records in D*/
Output: A set of cyber-attacks C={$C_1$,$C_2$,$C_3$,….,$C_k$}
…………………………………………………………
Set Continue ← Yes, Check ← Yes, C ← NULL
WHILE Continue == Yes DO
    /*Step 1: Data pre-processing for dataset D*/
    D ← Pre-processing (D)
    /*Step 2: Build a J48 decision tree for dataset D*/
    T ← Build J48 DT (D)
    /*Step 3: Pattern extractions from the decision tree */
    P ← Pattern Extraction (T)
    WHILE Check == Yes DO
        /*Step 4: Obtain future record from the network */
        D ← Obtain Future Record ()
        D ← D U Ď)
        /*Step 5: Prediction of cyber-attacks */
        Ć ← Prediction of Cyber Attack (P, Ď)
        C ← C U Ć
        /*Step 6: Intervention to reduce cyber-attacks */
        Intervention to Cyber Attack (C)
        Check ← do You Want To Continue () /*Input will be either Yes or No */
    END WHILE
    Continue ← do You Want To Continue () /* Input will be either Yes or No */
END WHILE
Output C
```

Fig (1): Algorithm for prediction of cyber-attacks

Any entries with blank values are deleted. With publically available intrusion detection networks (IDS) datasets given by the Canadian Institute of Cybersecurity (CICDDoS2019), numerical data can be imputed rather than data being removed [23]. We retrieved six datasets from CIC that are connected to various computer security, such as DDoS, Port Scan, Bot, SQL Injection, and Heartbleed. On the information that was generated in Step 1, we constructed a classification method in Step 2 using the J48 decision tree technique [24]. J48 is a popular classification technique for forecasting. The J48 algorithm's ability to provide logic constraints that make it simple to comprehend why a specific incident happened is another crucial feature. The article employs the J48 decision tree algorithm from Weka. We applied the J48 default settings from Weka. Step 3 involved extracting the model's characteristics. We created a program for patterns retrieval from the J48 tree that accepts the output of the Weka J48 tree as input and outputs a set of logic rules for the decision tree [25]. Remember that the phrase "root to leaf path" refers to a logic rule. The number of patterns of a decision tree is the same as the number of leaves in that decision tree. A pattern indicates why a particular outcome happens. The application also generates information regarding the amount of recordings with the largest class value, the amount of records that were incorrectly classified, and the likelihood that leaf has the largest class value. The project's output is provided as a csv file, which enables trend analysis simple for the reader. Note the tables (Tables 5 to Table 7) in section 7 where we tabulated the structures for the reader's convenience. To make it simple to use the criteria to match with containing the new, they can be stored as records in a SQL database.

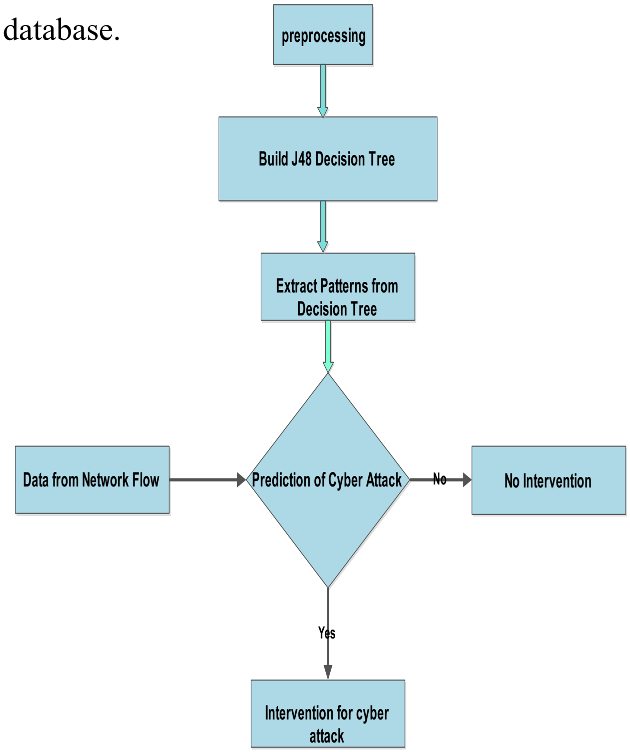
preprocessing
Build J48 Decision Tree
Extract Patterns from Decision Tree
Data from Network Flow
Prediction of Cyber Attack
No
No Intervention
Yes
Intervention for cyber attack

Fig (2): A block diagram of the proposed framework.

In Step 4, we obtained the records from the network for which we want to investigate if any cyber-attacks are happening. Note that, we assume the number of attributes, except the class attribute for both future records and historical records to be the same. The class attribute value for a future record would be assigned based on the prediction outcome of Step 5. In this step, we also updated the dataset D by adding new records Ď into D. The updated dataset D will be used for the purpose of future model building. In Step 5, we classified every record that we obtained in Step 4 based on the patterns that we extracted in Step 3. A record that we acquired in Step 4 ought to coincide with one of the Step 3 pattern. Future communication records may be recorded in a database in a manner that makes it simple to compare them to the patterns database. Any potential cyber-attacks on the infrastructure should be shown by a future record and its associated pattern. For further analysis, the anticipated attacks will be attached in C. At Step 6, if a cyberattack is discovered, a professional can intervene to address the attack. Our new framework ability to be deployed to a live network and forecast cyberattacks on the fly makes it feasible to effectively intervene and halt the attacks, which is one of its key strengths. The structure of a cyberattack will allow a specialist quickly comprehend the problem and execute the appropriate countermeasures if one is discovered by the forecasting model. If the prediction is used at the start of a network, when the first data with the objective from users is received, the data with the objective can then be approved or rejected based on the predictions model's results in order to protect the network. A specialist in connectivity can create defenses against attacks that the modeling can activate autonomously, protecting the network from damage. We do, however, think that some assaults will call for direct action from network specialists. We exhibit the suggested framework's method in Figure 1 and its block diagram in Figure 2, respectively. Incremental steps are those from Step 4 to Step 6, which are seen in Figure 1. The repeating processes can, however, be interrupted at any time by the user. A post - classification change detection method can be used on the dataset to determine whether the proportion of recordings for the class values has changed if the framework is unable to forecast cyberattacks [26]. In this case, the J48 decision tree on the data set must be rebuilt, and the user must then carry out steps 3 through 6 of the architecture. Moreover, some datasets' class imbalance problems can be resolved and prediction results can be improved by using a cost-sensitive J48 classification from Weka [27].

5. Introduction to Data Mining (DM) and Machine Learning (ML) for Cyber Security

The phrases databases for knowledge discovery (KDD), Data Mining (DM), and Machine Learning (ML) are sometimes used indiscriminately. According to studies (table 1), the KDD process is portrayed in its entirety and works with obtaining important, previously undiscovered knowledge/information from data. The procedure of DM as a particular stage in KDD that deals with the construction of algorithms for retrieving sequences from data is explicitly described and defined in [28]. This leads to the conclusion that they share traits with both ML and DM. Data collection, data filtering and which was before, feature extraction, deployment of DM techniques, and result explanation are the steps that make up the KDD procedure. DM is one of the steps used to apply techniques to identify data patterns [29].

These two phrases are frequently used interchangeably and in discussions simultaneously. Machine learning (ML) is a branch of article that enables machines to acquire knowledge on their own without even being pattern recognition, the inventor of ML. The categorization and prediction methods are the key areas of attention for algorithms that use machine learning. The machine learning algorithms discover insights into possible future situations by learning from strength and conditioning data. This is a discussion of the numerous classification methods used in cyber security generally [30].

• Decision Trees

The significant and well-liked strategies for categorization are decision trees. A decision tree is merely a straightforward flowchart with a tree-like structure, where each internal node represents a test with regard to an attribute, every branching representing the result of the test, and each leaf node acquiring a class label. Ross Quinlan created the algorithm for decision trees known as ID3 (Incremental Dichotomiser). Then he provided the replacement for ID3-C4.5, which has evolved into a standard for understanding methods.

• C4.5 Algorithm

This model draws inspiration from the ID3 algorithm and incorporates new traits to take into account the problems that ID3 has. It is supposed to use a top-down recurrent divided and conquest strategy and is thought to have a greedy algorithm. The approach is as continues to follow:

a) If S is tiny or all the data samples in S correspond to the identical class, the leaf node is tagged as the most frequently class in S. C4.5 employs the split and conquer technique for tree creation.

b) In the alternative, the dividing process's criteria is controlled by the method for selecting attributes. By determining the most effective method to divide the tuples into different classes, the dividing criteria determines which characteristic is to be examined at node S. A decision tree is created by repeating the technique [31].

• Naive Bayes Algorithm:

A streamlined form of the Bayesian learning process is used by the Naive Bayes algorithm (NB). There are statistical analyzers involved. These classifiers, which are based on the Bayes theorem, can be used to calculate the probability of participation. Conditional independent refers to the presumption that the impact of a selected attribute from a particular class does not dependent on the value from several other characteristics. Use of Naive-Bayes classifiers is one of the most effective, reliable, and best ways to avoid noisy data. The key benefit is that it just needs a minimal quantity of training data to roughly approach the classification constrictions.

• K-Nearest-Neighbor

The K-Nearest-Neighbor (k-NN) categorization is one of the easiest and most essential ones. It is founded on the training by equivalency method and performs well even when there is little or no real warning of the data distribution. The training data are described by'm' structural equation model properties, each instance duplicating a specific point in the m-dimensional space. As a result, it is clear that an m-dimensional pattern space is where all of the elements are kept. A k-nearest neighbor predictor examines the patterns field for the k training perceived that are somewhat similar to that of an original substance when dealing with unlabeled data samples [32].

• Support Vector Machine

It primarily aids in the creation of a hyper plane by plotting the input sequence onto a space with extremely high proportions. The data points can be divided into various classes by the hyper plane. Hyper planes that have the largest functional margin—the distance from any classes nearest training data point—achieve a high level of distinctiveness. It has been found that when margin increases, the classifier's training error decreases. For the two classes, the hyperplane serves as a decision boundary. Indeed, the identification of a misinterpretation caused by a certain approach is ensured by the permanence of a decision function. SVM is used for implementing categorization, prediction, and other tasks [33].

Table 1: References that use Data mining techniques to detect attacks and malicious software.

Ref.	Motivation	Methods	appropriate result
[34]	Manage and defend the computerized vehicle's components from assault.	CNN and CNN-LSTM	These techniques successfully recognize threats with a precision of over 97% and prevent their presentation on vehicle monitors.
[35]	Investigate network issues and find threats	K-NN and DNN	The DNN detects intrusions with a precision of greater than 92%.
[36]	Create a hacking detecting strategy for cyberattack protection and threat	CNN	According to CNN, the effectiveness was more than 99%.

	categorization.		
[37]	Create a hacking detecting strategy for cyberattack protection and threat categorization.	RNN	RNN achieves the highest precision at 98.27%.
[38]	Identify aberrant attack through the Internet by improving the computerized learning technique.	RBM and DBM	These techniques have achieved a remarkable 97.9% efficiency in aberrant detection of intrusions.
[39]	Create a program that recognizes threats and cyberattacks and shields networks of computers.	MLP and PID	These techniques identify intrusions with a 98.96% precision and help identify the types of intrusions.
[40]	Create a framework that secures data in cloud-IoT networks that analyzes and detects assaults utilizing URLs on edges machines.	a number of parallel complex models	These algorithms successfully identify typical queries with a high precision of exceeding 99%.
Our proposed	Identify aberrant attack through the Internet by improving the computerized learning technique.	C4.5, k-NN,DT,SVM	Every dataset's predictive algorithm has a 99.99% average accuracy

6. A Method of Integration of Network Information Security Prevention and Web Data Mining Technology

6.1. Full application of web data mining resources

Only Web application data, sign up information, proxies service data, and transaction data are present in Web data mining assets in terms of networking data protection, as follows: Web server data comes first. To play the significance and value of applications, telecommunications field must rely on computer Internet technology, and Web service is the most plentiful source that internet Connections technology has to offer. For instance, when a user accesses a page, the Web server of the page displays the pertinent data. The data can be classified into two groups, access information and log file, depending on the subject matter and service mode. The primary data source for Web data digger technologies is log records. The evidence in the log can be used to identify the hints in the attack network information, which can then be solved using the appropriate methods and processes. to guarantee the data security [41] of the system. Sign up data is the second type. The user must send the relevant information data to the server while logging into the networking data source. The sign up data must be provided, and it will be combined with the accessibility log to increase the data mining's precision. It can also monitor user access activity simultaneously. The third is that telecommunications field resembles commercial transactions, increasing the amount of data on as well as between sites. The information from these business interactions can be thoroughly examined using Online data mining techniques. To ensure the security of network information, certain poor information and problematic information must be eliminated [42].

6.2. Web data mining methods

The fundamental task to achieve networking data security protection is to precisely and quickly mine aberrant reality through the use of Web data mining technologies. This technique can not only make clear what constitutes cybersecurity danger behavior and illegal behavior, but it can also advance network data security preventive science, practicality, and applicability. There are primarily the following techniques to accomplish Web data mining, which are based on the advancement of network cybersecurity protection and Web data mining technologies in China: the first, association rules. Each piece of data in the networking data base has a strong association with the others. The connection between both the knowledge can be mine out when there are excessive security procedures in the internet data system and they build up. By using frequent patterns, we can, for instance, differentiate between users' typical and atypical access to a webpage and promptly take the necessary action in the event of unusual access. Analyzing classifications comes next. Many divisions are first established in accordance with the network information system's actual circumstances, and they are then classed and maintained in accordance with the pertinent data. Advanced neural network technology, simulations future technologies, statistical future technologies, and other techniques are frequently used in categorization data mining. The associated dataset can be mapped into a specific category using the categorization and processing requirements, protecting the privacy [43] of network data. Cluster analysis comes in third. The term "cluster analysis" describes how various groups handle and examine connected data in accordance with the data's actual circumstances. By using this technique, it is possible to differentiate one group from those other parties as well as efficiently assure that the data within a group have certain features. Sparse and 04205843 accurately differentiated by clustering analysis. Network of support data security protection with the appropriate data.

Fourth, an investigation of outliers. This type of data mining is also known as classifier analysis, because there is a very clear difference with other data and exceptions, or data that does not fit the normal pattern or behavior. The outlier analytical method can be separated into three steps [44] during the specific application procedure: the first step is to identify the heterogeneity, the 2nd stage involves evaluating the variability, and the third phase is to resolve the variability. Genuine but unexpected information can be mine and a lot of important data can be retrieved.

7. Result and Discussion

Using publicly accessible intrusion detection systems (IDS) statistics provided by the CICDDoS2019 [45], we assess the efficacy of the suggested methodology. Be aware that we employed 12-fold cross-validation, a least of 4 recordings per leaf, and a level of confidence of 0.35 for the J48 decision tree algorithm. Six datasets in the CICIDS are associated with different cyber-attacks, such as DDoS, Port Scan, Bot, SQL Injection, and Heartbleed. The datasets are summarized in Table 2, where we observe that the least and highest record counts are 160264 and 468923, respectively. Each dataset includes 80 characteristics, except the classifier. The final column of Table 2 displays the total number of base classifiers for each dataset. We list the total amount of records for each category feature values in a dataset in Table 3. Regular network behavior is indicated by the class value "Benign," while cyberattacks on the network are indicated by the base classifiers DDoS, Portscan, Bot, Brute strength, SQL Injection, and Heartbleed. Thursday-WorkingHours- PortScan.pcap ISCX dataset features 142334 BENIGN and 169832 Port scan events, whereas Thursday-WorkingHours-DDos.pcap ISCX dataset contains 89615 BENIGN and 137016 DDoS records. Working Hours Morning. Cap ISCX dataset on Thursday contains 178054 BENIGN and 2867 Bot entries.

- ❖ Collect the dataset (CICDDoS2019) 7 files in one file by instruction Group called new file (master.csv) (Example=1138472, special Attribute=1, frequent attribute=80)
- ❖ Binary Class (DDoS, BENIGN)
- ❖ Feature (Wheight correlation)

Table 2: A Summary On Cic ddos 2019 Datasets

Dataset	No. Record	No. Attribute	No. Class
Thursday-WorkingHours-DDos.pcap_ISCX	137016	80	2
Thursday-WorkingHours-.pcap_ISCX	169832	80	2
Monday-WorkingHours-WebAttacks.pcap_ISCX	2867	80	2
Monday-WorkingHours-Infilteration.pcap_ISCX	468923	80	2
TuesdayworkingHours.pcap_ISCX	160264	80	2

Table 3: A Summary of Class Distribution of The Cicddos2019 Datasets

Dataset	Class Value	No.Records	Class Value %
Thursday-WorkingHours-DDos.pcap_ISCX	DDoS	106722	47.38%
	BENIGN	128027	56.71%
Thursaday-WorkingHours-pcap_ISCX	FTP-Patator	8736	1.780%
	BENIGN	532163	98.927%
Monday-WorkingHours-WebAttacks.pcap_ISCX	Web Attack SQL Injection	31	0.012%
	Web Attack Brute Force	1608	0.924%
Monday-WorkingHours-Infilteration.pca	Infiltration	42	0.02%
	BENIGN	468923	99.99%

p_ISCX			
Tuesdayworkin gHours.pcap_ISCX	DoS GoldenEye	11284	1.4%
	BENIGN	160264	73.4%

Monday-working Hours.pcap ISCX has 160264 BENIGN, 11284 DoS GoldenEye, Thursday-Working Hours- WebAttacks.pcap ISCX dataset has 176145 BENIGN, 1608 Web Attack Brute Force, 31 Web Attack SQL Injection, and 1608 Web Attack XSS records, and Thursday-WorkingHours- Infilteration.pcap ISCX dataset has 468923 BENIGN and 42 Intrusion documents. This is a description of each assault on the datasets: DoS assault: In a denial-of-service (DoS) intrusion, an inordinate amount of network activity renders a computer system or network temporarily unreachable. The attacker stops authorized users from using the system. DDoS Attack: DDoS is an attack when many systems are used to take down a website, service, or networking that has been selected as the target. Often, the hacker does this using a network.

DDoS happens when numerous systems crash as a result of heavy network traffic. Only using or more network services, the hacker overloads the traffic or capabilities of a target device. Brute Force Attack: This form of system assault involves a lot of passwords being tried and tested by the attacker. Instead than attempting every possible combination, the attacker in this attack utilizes a login from a dictionaries file. Web strike: This is a further typical cyberattack in which the attacker produces a SQL command that compels the target systems to supply data relevant to the inquiry. Cross-Site Scripting (XSS) occurs when programmers fail to thoroughly test their code to detect the risk of script insertion. Assault by Intrusion: Typically, this assault originates within the network. Typical software (like Adobe Acrobat Reader) has flaws that an attacker identifies and uses to get access. The victim's machine has a backdoor that the attacker can use to launch various network attacks, including IP sweeps, thorough port scans, and Nmap service enumerations. Note that Table 3 shows that the proportion of Bot, Internet Assault, SQL Injection, and Intrusion assaults is quite low when compared to benign attacks for specific datasets. For instance, less than 1% of the overall records are affected by infiltration. Yet, as we can see from the reflection of the improvement in predictive performance and patterns, the predictive algorithm properly predicted the majority of these assaults.

Table 4: Accuracy of The Prediction Model On the Cicddos2019 Datasets

Dataset	Accuracy
Thursday -WorkingHours- DDos.pcap_ISCX	98.99 %
Thursday -WorkingHours-.pcap_ISCX	99.99 %
Monday-WorkingHours-WebAttacks.pcap_ISCX	97.96 %
Monday-WorkingHours-Infilteration.pcap_ISCX	99.72 %
TuesdayworkingHours.pcap_ISCX	99.99 %

For each dataset we used in our investigation, we showed the forecast performance of the models in Table 4. As shown in Table 4, the predictions model's efficiency for each data set is greater than 99%. Furthermore, we saw that the model properly detected the majority of the attacks in each dataset. From Table 5 to Table 7, the predictions model's characteristics for the information Thursday-WorkingHours-DDos.pcap ISCX, Thursday-WorkingHours-PortScan.pcap ISCX, and Monday-WorkingHours- Infiltertion.pcap ISCX are shown.

Table 5: Cyber-Attacks Patterns from Thursday-workinghours- Ddos.Pcap_Iscx Dataset

Majority Class	Class Information	Pattern
DDoS	records=1326744, DDoS= 1326744 other=0	Fwd Packet Length Max <= 23 and Total Length of Fwd Packets > 19 and Destination Port <= 128 and Packet Length Mean > 5.25 and Fwd IAT Max > 2389
DDoS	records=41, DDoS=40 other=1	Fwd Packet Length Max <= 23 and Total Length of Fwd Packets <= 19 and Bwd Packet Length Min <= 874 and Destination Port <= 80 and FIN Flag Count <= 0 and Total Length of Fwd Packets > 13 and Total Backward Packets <= 1 and Average Packet Size <= 8.64
DDoS	records=29, DDoS=29 other=0	Fwd Packet Length Max <= 20 and Total Length of Fwd Packets <= 18 and Bwd Packet Length Min <= 985 and Destination Port <= 80 and FIN Flag Count > 0 and ACK Flag Count > 0
DDoS	records=63, DDoS=61 other=1	Fwd Packet Length Max <= 23 and Total Length of Fwd Packets <= 19 and Bwd Packet Length Min <= 874 and Destination Port <= 85 and FIN Flag Count <= 0 and Total Length of Fwd Packets <= 13 and Bwd Packets/s > 7161.738505 and Destination Port > 53 and Init_Win_bytes_backward <= 322 and Init_Win_bytes_backward > 186
DDoS	records=13, DDoS=13 other=0	Fwd Packet Length Max <= 23 and Total Length of Fwd Packets <= 19 and Bwd Packet Length Min <= 874 and Destination Port <= 85 and FIN Flag Count <= 0 and Total Length of Fwd Packets <= 13 and Bwd Packets/s <= 5161.738405 and Flow IAT Min <= 1253264 and Fwd Packet Length Std > 1.432373 and Flow IAT Min > 164 and Total Backward Packets <= 0
DDoS	records=43, DDoS=43 other=0	Fwd Packet Length Max <= 23 and Total Length of Fwd Packets <= 19 and Bwd Packet Length Min <= 8745 and Destination Port <= 85 and FIN Flag Count <= 0 and Total Length of Fwd Packets <= 13 and Bwd Packets/s <= 5161.738405 and Flow IAT Min > 1253264 and Init_Win_bytes_forward <= 533 and Flow Bytes/s <= 5.3753167

We provided 6 DDoS attack type types in Table 5. There are 128027 records with PortScan assaults, and the predictive algorithm, which we can see in the characteristics, properly identified 128018 records. Six patterns relating to port scan attacks are shown in Table 6. We can observe from the patterns of the table that there are 147831 recordings involving port scan assaults, and 147816 data were successfully detected by the forecasting model.

Table 6: Cyber-Attacks Patterns from Thursday- working hours- Port scan. Pcap_Iscx Dataset

Majority Class	Class Information	Pattern

PortScan	records=373,
	PortScan= 373
	other=0
		Fwd Packet Length Max > 3 and Packet Length Mean <= 4 and Min Packet Length > 0

PortScan	records=93,
	PortScan=93
	other=0
		Fwd Packet Length Max > 3 and Packet Length Mean > 4 and Packet Length Std > 1531.555147 and Fwd Packet Length Max <= 164)

PortScan	records=21,
	PortScan=21
	other=0
		Fwd Packet Length Max > 3 and Packet Length Mean <= 4 and Min Packet Length <= 0 and Idle Min > 5991254 and Destination Port <= 342

PortScan	records=20,
	PortScan=20
	other=0
		Fwd Packet Length Max > 3 and Packet Length Mean > 4 and Packet Length Std <= 1532.554147 and Bwd IAT Min > 670 and Bwd Packets/s > 171.066403 and Total Backward Packets > 3

PortScan	records=7,
	PortScan=7
	other=0
		Fwd Packet Length Max <= 3 and PSH Flag Count <= 0 and Fwd Header Length > 142 and Init_Win_bytes_forward <= 322 and Total Fwd Packets > 6 and Total Fwd Packets <= 6

PortScan	records=62,
	PortScan=62
	other=0
		Fwd Packet Length Max <= 3 and PSH Flag Count <= 0 and Fwd Header Length > 143 and Init_Win_bytes_forward <= 342 and Total Fwd Packets <= 6

Table 7 shows two instances that show how our suggested predictive algorithm can recognize an infiltration cyber-attack. There are 21 entries in one pattern, of which 18 are infiltration. There are 18 records for other patterns, all of which are infiltration. It should be noted that the dataset contains 42 records with infiltration, or 0.02% of all records; 34 of these records were accurately identified by the forecasting model, proving that the predictive algorithm is also effective for category datasets.

There are 1816 entries with Bot assaults in the Thursday-WorkingHours- pcap ISCX dataset, and our predictive algorithm successfully identified each one of them. There are 6826 entries with FTP-Patator in the Tuesday-WorkingHours.pcap ISCX dataset, and the model properly detects 6826 of them. Furthermore, the predictive algorithm successfully identified every one of the 4786 entries with Patator in the same data. There are 31 entries with "WebAttackSQL Injection" in the Monday- WorkingHours- WebAttacks.pcap ISCX dataset, or 0.023% of all the entries in the dataset, and the predictions model accurately detects 17 of them. There are records of four assaults in the Tuesday-workingHours.pcap ISCX dataset: The DoS roots contain, the DoS slowhttptest, the DoS Hulk, and the DoS GoldenEye. There are 13 records using PortScan, and the model properly identified each one of them. Similarly, DoS-GoldenEye detected 11152 out of 11182 records. Yet we discovered that some datasets have a problem with minority class. Cost-sensitive classification methods can be utilized for good predictive outputs if the J48 decision tree is unable to accurately detect cyber-attacks [46]. To instantly forecast cyberattacks, the rules of the prediction models can be applied to a network flow. To ensure that the predictive algorithm did not produce any false detections, a domain expert might analyze the threats that the prediction algorithm identified. It is also considered that a domain expert can manage a brand-new cyberattack that does not fit with any known patterns. Moreover, the framework for addressing intrusions can be expanded to include remedies for specific assaults.

Table 7: Cyber-Attacks Patterns from Thursday-working hours-Infilteration.Pcap_Iscx Dataset.

Majority Class	Class Information	Pattern

Infiltration	records=21,
	Infiltration= 18
		Infiltration: Total Fwd Packets = 1123, Active

	other=1	Minimum > 4636862, ACK Flag Count > 0, and Bwd Packet Size Max = 26

Infiltration	records=18,
	Infiltration= 18
	other=0
		Total Fwd Packets > 1123 and Bwd Packet Length Mean <= 7.2

Table-5, Table-6 and Table 7 represent the results after applying some machine learning algorithms on data respectively. From Table 8, we can see that SVM, Decision Tree, K-Nearest-Neighbor and Naive Bayes performed equally well comparatively better than C4.5 Algorithm if we consider the accuracy.

Table 8: Results after applying machine learning algorithms on data.

Model	Accuracy
K-Nearest-Neighbor	91 %
Support Vector Machine	91 %
Naive Bayes	91 %
Decision Trees	91 %
C4.5 Algorithm	89 %

7.1. Malware detection using Data Mining

Malware is a harmful software program that, via viruses, Trojan horses, and warms, causes unusual behavior in computer programs. Data mining classification techniques can be used to find malware and notify it to the network administrator. Attacks by malware on the system are caused by visiting malicious website, downloading infected videogames or free apps, downloading infected music files, installing infected plugins or toolbars, and more. When installing any software, it's crucial to read the warning warnings, especiall;2y the ones regarding authorization for access emails or personal information.

7.2. Malware Statistics

According to studies, malware attacks are to blame for 85% of the system's damage [47]. Malware is discovered to be 96% distributed via email attachments. By 2021, there will be a 62% rise in mobile malware attacks. Android devices were the target of 99% of malware. 99% of malware was downloaded through unofficial apps. Nine of the ten payloads are ransomware. In a single week, malware infects 19 million websites worldwide. By 2021, 93% of banking firms will be the target of malware. 50% of the ransomware victims paid the demand. These days Maintaining the integrity, confidentiality, identification, and non-repudiation of data exchanged over the internet is a significant difficulty. According to their activity and characteristics kept in databases, data mining techniques aid in the earlier identification of malware.

7.3. Malware Detection

Both passive and active analysis methods are used to categorize a program as malware in behavior modification malware detection. Digital signal, which is difficult to examine and detect attacks, provides the basis for static analysis used for malware detection. Real-time code executions are a component of dynamic analysis that is used to examine infected files using a virtual machine. Malware is the term for harmful computer code that infiltrates a system via spam emails, email messages, and weak internet services. As a result, servers are brought down, network and machine configuration information is stolen, personal data is improperly accessed, key infrastructure is compromised, and so on. Application of Upcoming data mining approaches for extraction, categorization, and clustering are important tools for detecting malware [48]. The technique of malware detection through data mining is shown in the diagram below.

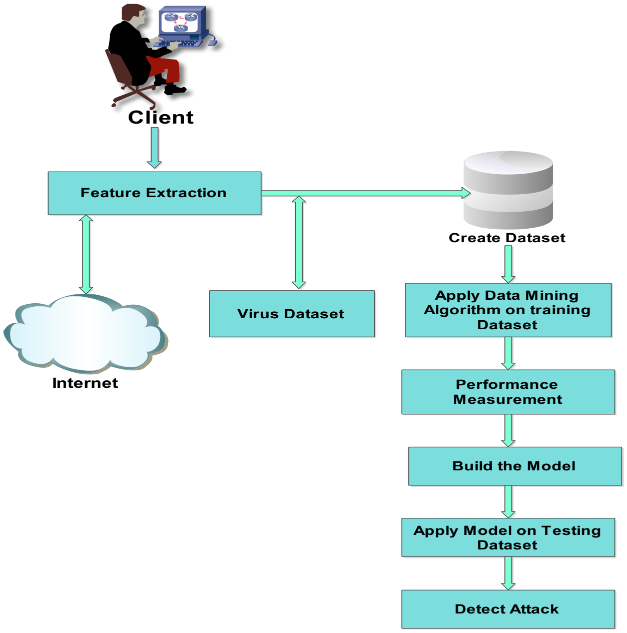

Fig. (3): Malware Detection Using Data Mining

When the client computer is online while being scanned, the machine data is retrieved by the antimalware program. This application creates a dataset by extracting characteristics from several files in order to begin the subsequent extraction procedure. Virus definitions are kept in viral files from the corpus dataset. Techniques such as static and dynamic analysis, nonlinear dynamic, and hybrid analysis are used to extract characteristics or patterns from data. To create assembly files, use the IDA pro technique in terms. To achieve better outcomes, arithmetic operations from assembly language are removed to create abstract assembly files. From the training dataset, extract frequent instruction associations. On the training dataset, data mining techniques like classification and association rule mining are applied based on their activity or signature to produce numerous instructions from assembling. Using statistical techniques, the algorithm's effectiveness in detecting malware is evaluated. The technique is taught until it performs as desired before being used to create the prototype. For the validation data, this training model is used to identify and indicate the type and state of malware [49].

In the method of extracting features for static and dynamic analysis PE packages are examined but not actually run. Data analysis patterns identification using approaches like the Windows API, N-grams, strings, Opcodes, or Control Flow Graph (CFG). Investigating or exploring all potential execution methods routes in malware samples is one of the useful techniques. The N-gram method is used to detect auto run viruses utilizing artificial neural network algorithms. Using the API call approach, it is possible to find malware that has hidden connections between its code segments.

Debugging or characterizing the code through actual runtime performance is done through nonlinear dynamic. Variable value, code input, and system setup all play a role in this process. New malware classifications are found using this analytic mechanism. detection of data analysis patterns using a debugger, simulators, emulation, and environment based on a virtual server. Hybrid methodological approaches combine the advantages of static and dynamic analysis. Packaged malware is first analyzed dynamically, and the concealed code is then recovered by contrasting the malware's runtime performance with an assessment of its example using a modelling approach. Dynamic analysis is used to find hidden files, whereas static modeling is used to keep track of unpacked files [50].

7.4. Techniques of Malware Detections:

7.4.1. Signature based malware detection:

Malware traces from earlier attacks. When vulnerable code is discovered, it is examined by extracting a malware signature made up of a unique byte sequence. When a sign is compared to an existing identity, an anti-malware application will identify the file as dangerous and pack malicious software. Anti-malware software in this case needs to wait until a device is attacked before looking for a signature. When categorizing a risk as malware, data mining approaches like classification algorithms are used since it saves time and increases forecast accuracy compared to the old approach. This approach is simple to use, has extensive malware information, is searchable, and is generally accepted. Bypassing the threat using some obfuscation and cryptography techniques is possible for signature databases. It is unable to recognize the polymorphic virus that replicates data in the sizable database [51].

7.4.2. Behavior Based Malware Detection:

Program behavior, computational efficiency, fast response, browsing patterns, cookie information, the types of attachments, and features are extracted all aid in the identification of suspicious activity or malicious programs. Using a data mining algorithm, assembly features and API calls are used in behavior-based identification. You can use unsupervised methods like clustering, SVM, and nearest neighbor technologies to analyze activity and find undetected malware. This technique aids in the detection of data flow dependencies in malicious software programs as well as polymorphic malwares. Complex behavioral pattern detection takes more time and storage capacity. The table 9 that follows outlines data mining methods for detecting malware:

Table 9: Data Mining Techniques for Malware Detection

Type of Malware	Data Mining Techniques	Data Analysis Method

Polymorphic Malware Detection[52]	K-means	Dynamic
Android Malware Detection[53]	SVM, J48, Naïve Bayes	Dynamic
API Malware Detection[54]	Naïve Bays, SVM, Decision Tree, Random Forest	Dynamic
Sequential Pattern Malware Detection[55]	All-Nearest-Neighbor, KNN, SVM	Hybrid
Frequent Pattern Malware Detection[56]	Graph Mining	Static
Behavioral Malware Detection [57]	Regression, SVM, J48	Dynamic

The above table shows various data mining methods for malware detection system based on their behavioral and signature characteristics. Static, dynamic, and hybrid data interpretation are utilized to extract hidden patterns in the information to increase the precision of malware detection. Selecting the right technique and analysis instruments to discover hidden risks and send alerts to protect data from additional attacks is a problem for cyber security specialists.

8. Conclusion

In order to decrease cyberattacks, we proposed in this article a framework for cyber security employing data mining. We first created a prediction system to anticipate future cyberattacks by extracting patterns from previous data of cyberattacks using the J48 decision tree method. We utilized the publicly accessible cyber security datasets given by the Canadian Institute of Cybersecurity to test the efficacy of the estimation method. When compared to trial findings, our proposed method accurately identified cyber-attacks like DDoS, PortScan, Bot, and SQL Injection. Our predictions model's overall accuracy in identifying cyber-attacks across all datasets is close to 99%. Future cyberattacks can be predicted using the effectiveness and derived characteristics for cyberattacks. The predictions model's empirical findings demonstrate the model's efficiency and ability to foresee potential cyberattacks. As an appropriate security measure, it is also vital to adopt effective as well as intelligent strategies for timely identification of cyber threats in order to reduce the danger of catastrophic cyberattacks including data breaches, ransomware attacks, and DDoS attacks. One of the challenges for security professionals is malware identification. Data mining methods such as classification, SVM, regression, decision trees, graph mining, and KNN algorithms can be combined with anti-threat systems to help identify malware before it enters the network, thereby defending your IT architecture against additional attacks.

Acknowledgment. Authors would like to thank Baghdad University, College of Nursing, Department of Basic Sciences and Middle Technical University, Technical Institute-Suwaira, Department of Computer Systems for the support given during this work.

References

[1] Rahman, M. A., Al-Saggaf, Y., & Zia, T. 2020, A data mining framework to predict cyber-attack for cyber security. In 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), IEEE, 207-212.

[2] Alloghani, M., Al-Jumeily, D., Hussain, A., Mustafina, J., Baker, T., & Aljaaf, A. J., 2020, Implementation of machine learning and data mining to improve cybersecurity and limit vulnerabilities to cyber attacks. Nature-inspired computation in data mining and machine learning, 47-76.

[3] Koroniotis, N., Moustafa, N., Schiliro, F., Gauravaram, P., & Janicke, H., 2020, A holistic review of cybersecurity and reliability perspectives in smart airports. IEEE Access, 8, 209802-209834.

[4] Gupta, B. B., & Sheng, Q. Z. (Eds.), 2019, Machine learning for computer and cyber security: principle, algorithms, and practices. CRC Press.

[5] Hasan, Z., & Jishkariani, M., 2022, Machine Learning and Data Mining Methods for Cyber Security: A Survey. Mesopotamian journal of cybersecurity, 2022, 47-56.

[6] Pan, B., Zhang, L., Li, C., & Chen, L., 2022, The relationship between cyber security and machine learning. International Journal of Basis Applied Science and Article, 8 (2022), 996-1003.

[7] Salloum, S. A., Alshurideh, M., Elnagar, A., & Shaalan, K., 2020, Machine learning and deep learning techniques for cybersecurity: a review. In The International Conference on Artificial Intelligence and Computer Vision Cham: Springer International Publishing (pp. 50-57)..

[8] Alazab, M., & Tang, M. (Eds.), 2019, Deep learning applications for cyber security. Springer.

[9] Afzaliseresht, N., Miao, Y., Michalska, S., Liu, Q., & Wang, H., 2020, From logs to stories: human-centred data mining for cyber threat intelligence. IEEE Access, 8, 19089-19099.

[10] Mohammed, I. A., 2020, Artificial intelligence for cybersecurity: A systematic mapping of literature. Artif. Intell, 7(9), 1-5.

[11] Le, D. N., Kumar, R., Mishra, B. K., Chatterjee, J. M., & Khari, M. (Eds.), 2019, Cyber Security in Parallel and Distributed Computing: Concepts, Techniques, Applications and Case Studies. John Wiley & Sons.

[12] Samtani, S., Abate, M., Benjamin, V., & Li, W., 2020, Cybersecurity as an industry: A cyber threat intelligence perspective. The Palgrave Handbook of International Cybercrime and Cyberdeviance, 135-154.

[13] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. 2020,

Cybersecurity data science: an overview from machine learning perspective. Journal of Big data, 7, 1-29.

[14] Mijwil, M., Salem, I. E., & Ismaeel, M. M., 2023, The Significance of Machine Learning and Deep Learning Techniques in Cybersecurity: A Comprehensive Review. Iraqi Journal For Computer Science and Mathematics, 4(1), 87-101.

[15] Kalinin, M., Krundyshev, V., & Zegzhda, P., 2021, Cybersecurity risk assessment in smart city infrastructures. Machines, 9(4), 78.

[16] Wang, L., & Jones, R., 2021, Big data analytics in cyber security: network traffic and attacks. Journal of Computer Information Systems, 61(5), 410-417.

[17] Saura, J. R., Palacios-Marqués, D., & Ribeiro-Soriano, D., 2021, Using data mining techniques to explore security issues in smart living environments in Twitter. Computer Communications, 179, 285-295.

[18] Sarker, I. H., Abushark, Y. B., Alsolami, F., & Khan, A. I., 2020, Intrudtree: a machine learning based cyber security intrusion detection model. Symmetry, 12(5), 754.

[19] Aiyanyo, I. D., Samuel, H., & Lim, H., 2020, A systematic review of defensive and offensive cybersecurity with machine learning. Applied Sciences, 10(17), 5811.

[20] Thach, N. N., Hanh, H. T., Huy, D. T. N., & Vu, Q. N., 2021, technology quality management of the industry 4.0 and cybersecurity risk management on current banking activities in emerging markets-the case in Vietnam. International Journal for Quality Research, 15(3), 845.

[21] Andrade, R. O., & Yoo, S. G., 2019, Cognitive security: A comprehensive article of cognitive science in cybersecurity. Journal of Information Security and Applications, 48, 102352.

[22] Ferrag, M. A., Maglaras, L., Janicke, H., & Smith, R., 2019, Deep learning techniques for cyber security intrusion detection: A detailed analysis. In 6th International Symposium for ICS & SCADA Cyber Security Research, 6, 126-136.

[23] Mohamed, N., Al-Jaroodi, J., & Jawhar, I., 2020, Opportunities and challenges of data-driven cybersecurity for smart cities. In 2020 IEEE systems security symposium (SSS) 1-7.

[24] Sreedevi, A. G., Harshitha, T. N., Sugumaran, V., & Shankar, P., 2022, Application of cognitive computing in healthcare, cybersecurity, big data and IoT: A literature review. Information Processing & Management, 59(2), 102888.

[25] Ferrag, M. A., Babaghayou, M., & Yazici, M. A., 2020), Cyber security for fog-based smart grid SCADA systems: Solutions and

challenges. Journal of Information Security and Applications, 52, 102500.

[26] Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H., 2020, Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative article. Journal of Information Security and Applications, 50, 102419.

[27] Zhao, S., Li, S., Qi, L., & Da Xu, L., 2020, Computational intelligence enabled cybersecurity for the internet of things. IEEE Transactions on Emerging Topics in Computational Intelligence, 4(5), 666-674.

[28] Alhayani, B., Mohammed, H. J., Chaloob, I. Z., & Ahmed, J. S., 2021, Effectiveness of artificial intelligence techniques against cyber security risks apply of IT industry. Materials Today: Proceedings, 531.

[29] Ullah, F., Naeem, H., Jabbar, S., Khalid, S., Latif, M. A., Al-Turjman, F., & Mostarda, L., 2019, Cyber security threats detection in internet of things using deep learning approach. IEEE access, 7, 124379-124389.

[30] Li, Y., & Hu, X., 2022, Social network analysis of law information privacy protection of cybersecurity based on rough set theory. Library Hi Tech, 40(1), 133-151.

[31] Chan, L., Morgan, I., Simon, H., Alshabanat, F., Ober, D., Gentry, J., ... & Cao, R., 2019, Survey of AI in cybersecurity for information technology management. In 2019 IEEE technology & engineering management conference (TEMSCON), 1-8.

[32] Ali, A., Septyanto, A. W., Chaudhary, I., Al Hamadi, H., Alzoubi, H. M., & Khan, Z. F., 2022, Applied Artificial Intelligence as Event Horizon Of Cyber Security. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS), 1-7.

[33] Altalhi, S., & Gutub, A. 2021, A survey on predictions of cyber-attacks utilizing real-time twitter tracing recognition. Journal of Ambient Intelligence and Humanized Computing, 1-13.

[34] Abie, H., 2019, Cognitive cybersecurity for CPS-IoT enabled healthcare ecosystems. In 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), 1-6.

[35] Sarker, I. H., Furhad, M. H., & Nowrozy, R., 2021, Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Computer Science, 2, 1-18.

[36] Ali, R., Ali, A., Iqbal, F., Khattak, A. M., & Aleem, S., 2020, A systematic review of artificial intelligence and machine learning techniques for cyber security. In Big Data and Security: First International Conference, ICBDS 2019, Nanjing, China, December 20–22.

[37] Al-Omari, M., Rawashdeh, M., Qutaishat, F., Alshira'H, M., & Ababneh, N., 2021, An intelligent tree-based intrusion detection model for cyber security. Journal of Network and Systems Management, 29, 1-18.

[38] Christen, M., Gordijn, B., & Loi, M., 2020, The ethics of cybersecurity (p. 384). Springer Nature.

[39] Zhang, F., Kodituwakku, H. A. D. E., Hines, J. W., & Coble, J., 2019, Multilayer data-driven cyber-attack detection system for industrial control systems based on network, system, and process data. IEEE Transactions on Industrial Informatics, 15(7), 4362-4369.

[40] Prasad, R., & Rohokale, V. 2020, Cyber security: the lifeline of information and communication technology. Cham, Switzerland: Springer International Publishing.

[41] Shetu, S. F., Saifuzzaman, M., Moon, N. N., & Nur, F. N., 2019, A survey of botnet in cyber security. In 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), 174-177.

[42] Thomas, T., Vijayaraghavan, A. P., & Emmanuel, S., 2020, Machine learning approaches in cyber security analytics, Singapore: Springer, 37-200).

[43] Alagheband, M. R., Mashatan, A., & Zihayat, M., 2020, Time-based gap analysis of cybersecurity trends in academic and digital media. ACM Transactions on Management Information Systems (TMIS), 11(4), 1-20.

[44] Loan, F. A., Bisma, B., & Nahida, N, 2022, Global research productivity in cybersecurity: a scientometric article. Global Knowledge, Memory and Communication, 71(4/5), 342-354.

[45] Kilincer, I. F., Ertam, F., & Sengur, A., 2021, Machine learning methods for cyber security intrusion detection: Datasets and comparative article. Computer Networks, 188, 107840.

[46] He, Q., Meng, X., Qu, R., & Xi, R., 2020, Machine learning-based detection for cyber security attacks on connected and autonomous vehicles. Mathematics, 8(8), 1311.

[47] Carley, K. M., 2020, Social cybersecurity: an emerging science. Computational and mathematical organization theory, 26(4), 365-381.

[48] Caramancion, K. M., 2020, An exploration of disinformation as a cybersecurity threat. In 2020 3rd International Conference on Information and Computer Technologies (ICICT), IEEE, 440-444).

[49] Sakthivel, R. K., Nagasubramanian, G., Al-Turjman, F., & Sankayya, M., 2022, Core-level cybersecurity assurance using cloud-

based adaptive machine learning techniques for manufacturing industry. Transactions on Emerging Telecommunications Technologies, 33(4), e3947.

[50] Pawlicki, M., Choraś, M., Kozik, R., & Hołubowicz, W., 2020, On the impact of network data balancing in cybersecurity applications. In Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, Proceedings, Part IV 20, Springer International Publishing, 196-210.

[51] Dasgupta, D., Akhtar, Z., & Sen, S., 2022, Machine learning in cybersecurity: a comprehensive survey. The Journal of Defense Modeling and Simulation, 19(1), 57-106.

[52] Chow, Y. W., Susilo, W., & Thorncharoensri, P., 2019, CAPTCHA design and security issues. Advances in Cyber Security: Principles, Techniques, and Applications, 69-92.

[53] Rathore, S., & Park, J. H., 2020, A blockchain-based deep learning approach for cyber security in next generation industrial cyber-physical systems. IEEE Transactions on Industrial Informatics, 17(8), 5522-5532.

[54] Maennel, K., 2020, Learning analytics perspective: Evidencing learning from digital datasets in cybersecurity exercises. In 2020 IEEE European symposium on security and privacy workshops (EuroS&PW), IEEE., 27-36.

[55] Lee, L. 2019, Cybercrime has evolved: it's time cyber security did too. Computer Fraud & Security, 2019(6), 8-11.

[56] Samtani, S., Kantarcioglu, M., & Chen, H., 2020, Trailblazing the artificial intelligence for cybersecurity discipline: a multi-disciplinary research roadmap. ACM Transactions on Management Information Systems (TMIS), 11(4), 1-19.

[57] Zahra, S. R., Chishti, M. A., Baba, A. I., & Wu, F., 2022, Detecting Covid-19 chaos driven phishing/malicious URL attacks by a fuzzy logic and data mining based intelligence system. Egyptian Informatics Journal, 23(2), 197-214.