JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ENHANCING CYBERSECURITY THROUGH EXPLAINABLE ARTIFICIAL INTELLIGENCE IN INTRUSION DETECTION SYSTEMS.

-Dr. Devendra Singh, Professor, Haryana Institute of Public Administration, Gurugram

ABSTRACT

Known as a cyber-physical system (CPS), a cyber-physical system is a network that consists of both cyber and physical components that communicate with one another in a feedback loop. Not only is a CPS required for day-to-day operations, but it is also required for the approval of vital infrastructure, which is the basis for cutting-edge smart devices. The most recent advancements in explainable artificial intelligence have been utilized to assist in the creation of robust intrusion detection systems for use in CPS environments. The objective of this study is to develop an Explainable Artificial Intelligence Enabled Secure Cyber-Physical System Intrusion Detection Technique, which will be referred to as XAIID-SCPS. The identification and classification of intrusions that occur within the CPS platform is the major emphasis of the XAIID-SCPS technique that has been recommended. For the purpose of selecting features, the XAIID-SCPS technique requires the utilization of a Hybrid Enhanced Glowworm Swarm Optimization (HEGSO) algorithm. For the purpose of intrusion detection, the improved Elman Neural Network (IENN) model was utilized, and the Enhanced Fruitfly Optimization (EFFO) approach was utilized to optimize the parameters of the model. In addition, the XAIID-SCPS technique combines the XAI approach LIME in order to enhance the level of comprehension and explainability of the black-box method for intrusion classification that is exact. The results of the simulation demonstrate that the XAIID-SCPS methodology works very well in contrast to other approaches, with a maximum accuracy of 98.87% overall.

KEYWORDS : *security; intrusion detection, cybersecurity , artificial intelligence.*

INTRODUCTION

Because of a cyber-physical system, also known as a CPS, physical devices that are equipped with sensing capabilities are able to connect with controllers or the internet whenever it is necessary. It is possible to use communication technologies such as wireless technology or short-distance communication in order to

continually update the condition of the physical environment or the status of the physical device to a distant server or controller. Because of recent developments in wireless communications and sensor technology, the CPS has been utilized in a variety of application domains, such as the chemical and aviation industries, the electronics industry, the fabrication of materials with automated supply chains, and smart industries, which includes transportation. These applications have been made possible by the CPS. In addition, the implementation of CPS applications in a number of different sectors opens the door to new security issues and concerns, making it more difficult to prevent infrastructure or sensitive data from being attacked by cybercriminals. The attacks include both physical assaults that might result in system failures or interruptions in the supply chain, as well as cyberattacks that make use of equipment that is linked to the internet. Consequently, as compared to the conventional information technology and network substructure, the security architecture of CPS is highly challenging.

Intrusion detection systems (IDS) are extremely successful in identifying and preventing data breaches because they are able to stop and identify intrusions. Identification may be broken down into two categories: misuse detection and anomaly detection (AD). Antidepressant medication is dependent on behavior, whereas abuse detection is dependent on patterns or facts. A number of false alerts are generated as a result of the enhanced detection rate of current IDS. In an intrusion detection system (IDS), it is essential to reduce the number of false positives. As a result of the fact that machine learning approaches have the potential to extract meaningful information from databases, a variety of intrusion detection systems were created with their aid. It's possible that these types of technologies will cut down on false positives. Machine learning (ML) strategies that included IDS, rule association, and GA were implemented with the help of artificial neural networks (ANNs), which stood for artificial neural networks. Several different machine learning strategies are merged in ensemble learning. It was observed by the authors that a collaborative strategy that makes use of machine learning techniques is more effective in reducing false positives.

The objective of this study is to develop an Explainable Artificial Intelligence Enabled Secure Cyber-Physical System Intrusion Detection Technique, which will be referred to as XAIID-SCPS. For the purpose of selecting features, the XAIID-SCPS technique requires the utilization of a Hybrid Enhanced Glowworm Swarm Optimization (HEGSO) algorithm. For the purpose of intrusion detection, the Improved Elman Neural Network (IENN) approach was applied, and an Enhanced Fruitfly Optimization (EFFO) algorithm was utilized for the purpose of parameter optimization. In addition, the XAIID-SCPS technique combines the XAI methodology LIME in order to enhance the explainability and comprehension of the black-box method for the purpose of exact intrusion classification. In order to evaluate the simulation values of the XAIID-SCPS approach, benchmark intrusion datasets may be utilized through testing.

Intrusion Detection Systems, often known as IDS, are crucial security measures that are designed to safeguard network infrastructures from harmful activities and unauthorized access by identifying them. Since their

inception in the middle of the 1980s, they have made major advancements in order to keep up with the ever-increasing complexity of crimes that include the use of computers. It is possible to divide intrusion detection systems into two distinct categories: network intrusion detection and prevention (NIDS) systems and signature and statistical anomaly detection (SAA) systems. In order to sift through network data for evidence of malicious activity, SAA systems employ heuristic behavioral analysis and signature analysis. Traditional security mechanisms, such as firewalls, may be unable to identify and maybe prevent potentially harmful actions and attacks. These systems have the ability to recognize and possibly prevent such attacks. In light of the fact that cyberattacks are becoming increasingly complex and attempt to compromise the availability, integrity, and confidentiality of network systems, it is very necessary for the detection and response to cyberattacks to become more precise. In light of the fact that there are only little distinctions between benign and malicious behavior, Pietraszek believes that around 99 percent of the warnings that are generated by intrusion detection systems are not connected to any cybersecurity concerns. Support vector machines (SVMs), neural networks (NNs), and fuzzy logic are some of the approaches that have been proposed by researchers as potential ways to enhance the capabilities of intrusion detection systems (IDS).

It has been established that these strategies have the ability to reduce the number of false positives and increase the detection rates for a variety of attack types, including denial-of-service (DDoS) attacks. In light of the fact that ChatGPT and other AI models may employ machine learning and natural language processing strategies in order to interpret detailed patterns and behaviors in network traffic, there is a growing interest in the possibility that these models might enhance the capabilities of intrusion detection. It is possible that the incorporation of AI models into intrusion detection systems (IDS) might make it possible to improve the identification of complex attacks, reduce the number of false positives, and provide more efficient reaction mechanisms. By implementing ChatGPT or other equivalent AI models, intrusion detection systems (IDS) have the potential to improve the detection of complex attacks, reduce the number of false positives, and enable more effective response mechanisms. The purpose of this study is to evaluate the potential for ChatGPT to improve the accuracy of IDS and boost its capabilities in terms of cybersecurity.

Systems for Detecting Intrusions

Intrusion detection systems (IDS) have been rapidly created in both research and business as a response to the increasing amount of cyberattacks that are being launched against commercial organizations and organizations that are run by governments throughout the world. The costs associated with combating cybercrime continue to rise year after year. Web-based assaults, denial of service attacks, and hostile insider attacks are among the categories of cybercrime that have the most damaging impact. As a consequence of these malicious incursions into computer systems, businesses or industries run the risk of losing their intellectual property. Intrusion detection systems, firewalls, and antivirus software are some of the tools that organizations employ to counteract actions of this nature.

Intrusion detection, sometimes known as ID, is known to be a crucial component of cyber security. It gives us the capacity to identify harmful network behavior before it puts the availability, integrity, or security of information at risk. By analyzing the events that occur within an information system, it is a technique that may be used to locate security breaches. In today's increasingly digital environment, it is unfathomable to conceive of a scenario in which one would be able to have a professional or personal life without having access to a network connection. The proliferation of Internet of Things (IoT) devices has provided attackers with the largest potential to execute an intrusion attack at this point in time. Therefore, it is of the utmost importance to fulfill the task of preventing the network devices from being penetrated. It is because of this that the question arises as to how one should effectively fight against both known and unanticipated threats. There is no straightforward solution to this problem because the number of dangers continues to grow with each passing year.

Artificial Intelligence and Trust Administration

There is still a problem that has not been overcome, and that is the lack of information and confidence that surrounds artificial intelligence in compared to traditional model-based optimization. To give just one example, deep reinforcement learning is unable of taking into consideration the basic factors that are responsible for behavior. The most serious aspect of this problem is that it has an effect on the management of trust in cyber security (for instance, malicious vehicle identification). In addition, recent research has shown that the Bayesian inference is extremely sensitive to the presence of inadequate data. As a consequence of this, there is an increasing demand for statistical artificial intelligence (AI) algorithms that can quantify uncertainty. This is especially true when it comes to the process of mapping algorithm design and huge data inputs to predicted wireless key performance indicators (KPIs). Recent years have seen a spike in research on "explainable AI," which refers to the process of developing statistical AI algorithms with the purpose of explaining the black-box model of artificial intelligence. This is in contrast to the traditional approach of striving to construct statistical AI algorithms that are naturally interpretable. For this reason, a trustworthy artificial intelligence system need to be able to offer an explanation for the decisions it makes, so that a human expert may grasp the underlying facts and the rationale behind them.

Reasonable Artificial Intelligence

During the course of the last several years, artificial intelligence has demonstrated that it is a tremendous success, surpassing expectations in a wide variety of applications. The empirical success of deep learning (DL) and machine learning (ML) models may be attributed to the combination of efficient learning algorithms and the huge parametric space that these methods provide. Machine learning and deep learning models are considered to be advanced black-box models due to the fact that the parametric learning space is comprised of millions of parameters and hundreds of layers on top of that. As a result of the "black-box" nature of the models, AI professionals (such as engineers and developers) are required to seek out a firsthand understanding of the process by which a model operates. In order to avoid the possibility of making illogical judgments and to

get in-depth explanations of their behavior, transparency, which is the opposite of a black-box model, is becoming an increasingly necessary need. To provide one example, binary predictions are not appropriate for use in precision medicine due to the fact that patient prescriptions are extremely sensitive. In addition, for the purpose of cyber security, the erroneous projections might leave the system vulnerable to breaches and lead to a situation in which important systems are protected by zero-trust security.

Because of this, explainable artificial intelligence is the focus of a significant amount of research in the field of artificial intelligence today, and it is vital to the improvement of the practical application of AI-based solutions. The incorporation of interpretability into the design process of AI-based solutions has the potential to improve their implementability in a number of ways. Firstly, it can promote impartiality in decision-making by identifying and correcting bias in the training dataset (also known as an imbalanced dataset). Secondly, it can enhance the robustness of AI-based solutions by drawing attention to potential adversarial perturbations that may alter the prediction. Lastly, it can strengthen the credibility of AI-based solutions by providing meaningful variable inference and causality of model reasoning. When it comes to the interpretability of AI and ML, explanation techniques and approaches may be grouped accordingly using the following categories: Some examples include model-specific and model-agnostic, intrinsic and post-hoc, premodel, in-model, and postmodel, feature engineering, and so on. Other examples include intrinsic and post-hoc. Over the course of this study, feature engineering and the rule-based model will be covered in order to investigate the decision tree technique for identifying the characteristics of malicious attacks and increasing confidence in intrusion detection systems.

OBJECTIVES

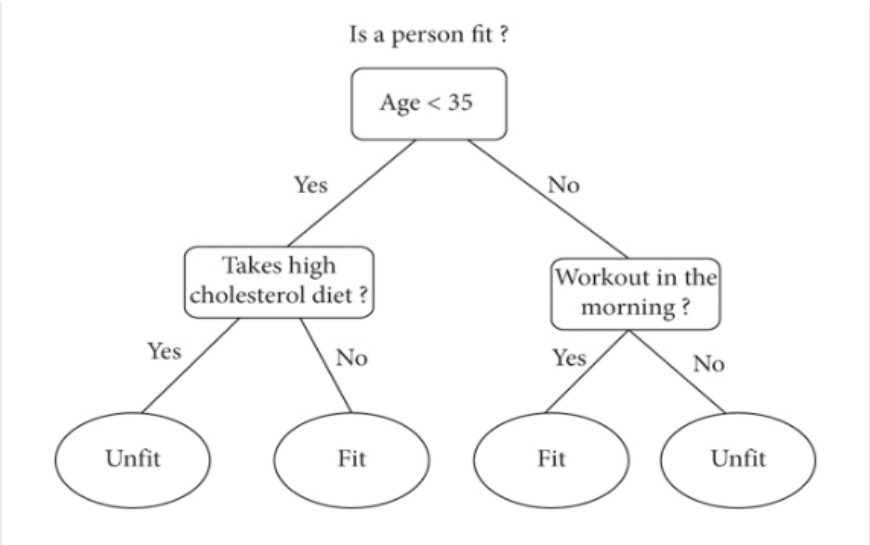
1. Carry out study on the subject of artificial intelligence.
2. Conducting research on cybersecurity.
3. Find out more information regarding intrusion detection systems, often known as IDSS.

RESESEARCH METHODOLOGY

Decision Tree

When it comes to methods of supervised learning, decision trees are considered to be nonparametric. Through the use of a tree-like structure, it provides an illustration of the likelihood of the outcomes of events and makes judgments. By taking this method, decision rules that are basic and easy to understand are produced, together with a conditional control statement. An additional supervised model, such as the support vector machine (SVM), does not allow for the observation and interpretation of the logic of the data. This method is also referred to as a supervised learning model in the field of machine learning. This is due to the fact that it permits the automatic development of prediction models via the utilization of algorithms that make use of a specified collection of observations (data) as a training dataset. When the decision model is constructed from the top

down, it takes the form of a tree structure, with leaf nodes and decision nodes following the root node at the top of the tree. As can be seen in Figure 1, the leaf nodes provide a conclusive classification, whereas the root nodes are significant predictors.



Rule formation is accomplished by the decision tree through the process of separating criteria. Although there are numerous approaches that may be utilized in the construction of decision trees, the Iterative Dichotomizer, often known as ID3, algorithm is widely regarded as being among the most effective. A greedy, top-down search of the training data sets that are supplied is carried out by ID3 in order to construct a decision tree. This search involves testing each attribute at each node. Based on a set of statistical measurements known as information gain, it decides which feature to test at each node in the tree. The degree to which a specific attribute splits the training samples into groups according to the desired classification is what this metric attempts to determine.

Entropy

In information theory, it is a metric that is used to represent the degree of impureness that a certain collection of samples possesses. The expression of the entropy S with regard to this c-wise categorization is as follows in the event that the target attribute takes c different values.

$$E(S) = \sum_{i=1}^{c} - P_i \log_2 p_i,$$

Pi represents the probability that S is a member of class I, and E represents the entropy of the situation. When written as the logarithm of base 2, the entropy is a measurement of the length of the encoding, and it is given in bits.

Acquire Knowledge

The expected minimization in entropy is computed by splitting the instances according to the qualities that they possess. Gain (S; A) of an attribute A S is a description of the information gain on the set of instances that is gained from the attribute.

$$\text{Gain}\,(S, A) = E\,(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} E\,(S_v)\,,$$

The subset of S that contains all of the possible values for attribute A, and values (A) is the subset of S that contains the attributes for which attribute A has a value. Using this metric, qualities are ranked, and a decision tree is formed. At each node along the route from the root, the attribute that has the largest information gain among those that have not yet been taken into consideration is the one that achieves the highest ranking.

RESULT

Here, we respond to the inquiries that follow:

(i) Q1-*Features*: are all elements that are crucial to the prediction process?

(ii) Q2-*Rules*: Which rules did the DT extract?

(iii) Q3-*Accuracy*: how precise DT is in comparison to other cutting-edge techniques?

Datasets: In order to apply the DT model, a network intrusion dataset that was collected during the 1999 KDD Cup is utilized. The data was given by the DARPA 98 Intrusion Detection Evaluation, which was conducted at the MIT Lincoln Laboratory. Using a large number of Internet-connected personal computers, these statistics were gathered in order to simulate a variety of intrusions and to represent a small United States Air Force installation that is staffed by knowledgeable individuals. There are 42 attributes that are utilized in this dataset. Among the five primary classifications, four are considered to be attacks, while the fifth is considered to be "normal," which means that it does not pose a threat. Information on the assaults is shown in Table 1.

Types of attack	Examples	Quantity	Proportion (%)
Denial of service (DoS)	smurf, apache2, pod, etc.	229,853	73.90
Remote-to-local (R2L)	imap, worm, phf, etc.	16,189	5.2
User to root (U2R)	perl, rootkit, and so on	228	0.07
PROBING	nmap, portsweep, etc.	4,166	1.34
Normal	Usual traffic patterns	60,593	19.48

Previous research has utilized this data to report on the efficiency of different supervised classification methods in multiclass settings. Additionally, the classification methodology has been utilized to foresee personal assaults. The process of detecting intrusions is, in general, the same as solving a classification problem, such as a binary classification problem, which includes identifying whether the behavior of network traffic is hostile or benign. In the course of this inquiry, we are not taking into consideration multiclass prediction. In a binary setting, our primary objective is to provide light on the approach that algorithms use in order to arrive at a conclusion on what constitutes normal or malignant behavior.

Assessment Measure

The model's accuracy is evaluated using the following assessment measures**.**

Accuracy

In the context of a classification problem, the definition of precision is the ratio of the number of true positives (tp) to the total number of true positives plus the number of false positives (fp). To be official,

$$\text{Precision} = \frac{t_p}{t_p + f_p},$$

where fp stands for false positives (recognized wrongly) and tp stands for true positives (identified properly).

Remember

In the context of a classification problem, recall is defined as the ratio of the number of true positives (tp) to the total number of true positives plus the number of false negatives (fn). To be official,

$$\text{Recall} = \frac{t_p}{t_p + f_n},$$

In which the term "fn" is used to refer to false negatives, which are incorrectly rejected.

F1-Results

The F1-score is a measure that attempts to represent the balance that exists between the accuracy and recall of a classifier model. It is determined by taking the harmonic mean of the accuracy measures and the recall measurements by combining them:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Q1: Features

The datasets include 42 characteristics. 41 characteristics out of the 42 are categorized into four distinct classes:

(i) Each TCP connection has unique qualities known as basic features (BFs).
(ii) The characteristics within a link that the domain knowledge suggests are called content features (CFs).
(iii) The properties that are calculated with a two-second time window are known as traffic features (TFs).
(iv) Host features (HFs) are characteristics intended to evaluate

It is important to note that the performance of the model is significantly impacted by the core concepts of machine learning. It gives us the ability to get rid of components of the model that are unnecessary and do not contribute to an improvement in the accuracy of the predictions. It is also possible to lessen the likelihood of drawing conclusions based on noise by reducing the amount of duplicate attributes. Therefore, in order to ensure that the model is trained in a more expedient manner, fewer attributes also indicate a simpler procedure. DT is able to choose essential features of the model in a straightforward manner when employing a natural approach. A predictor variable is responsible for producing splits, and the amount and quality of those splits are used to derive it.

Ascertaining the relevance of each characteristic may be accomplished through the construction of a prediction model by making use of the data. Reporting on varied degrees of importance is made possible via a mechanism that is incorporated into DT. The datasets that we had were randomly separated into three categories: twenty percent was used for validation, sixty percent was used for training, and twenty percent was used for the test set. With the use of training datasets, DT takes into account all of the features and decides which one best distinguishes class labels based on the measures of entropy (III-A1). The order in which the key traits are sorted is seen in Figure 1.

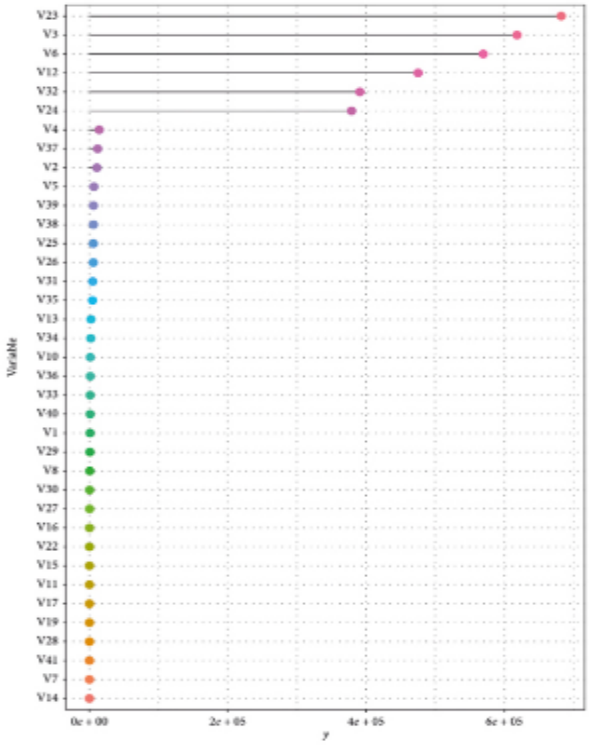


Figure 1 Ranking the features based on information gain.

It should be mentioned that the feature V23 has the highest ranking ever. Those blows that last for more than two seconds are the source of this phenomenon.

By relying exclusively on human evaluation, we are unable to determine with absolute certainty which quality is crucial to preserve during the modeling process. As a consequence of this, one of the categories of traffic characteristics (T) that we carried out was feature selection. The characteristic is referred to as "count," and it is a numerical attribute that is computed over a period of time that is equal to two seconds. In the same vein, we saw that feature V3, which is a part of the basic features (B) category and is referred to as "service," is positioned in the second position. This feature functions in accordance with the characteristics of each and every one of the distinct TCP connections, such as http and telnet. The feature in question is the one that possesses the discrete property. When it comes to the basic (B) category, "Flag," which is another feature, comes in third place. The normal or error status of the connection is referred to by this feature, which is a discrete property inside the connection. In Table 2, a description of the top 10 key qualities is presented to the reader. As can be seen in the table, the following feature categories are ranked according to their level of significance: one for traffic (T), two for basic (B), three for content (C), and four for hosts (H).

Table 2 Description of top ten features used in a decision tree.

Sr. No.	Features	Description	Type	Label
V23	Count	Number of connections to the same host as the current connection in the past two seconds	Continuous	T
V3	Service	Network service on the destination, e.g., telnet	Discrete	B

V6	Flag	Normal or error status of the connection	Discrete	B
V12	Logged_in	1 if successfully logged in; 0 otherwise	Discrete	C
V32	Dst_host_count	% of connections that have continuous H “SYN” errors	Continuous	H
V24	Serror_rate	% of connections that have continuous H “SYN” errors	Continuous	H
V37	dst_host_ srv_diff_host_rate	The percentage of connection that has different destination machine	Discrete	H
V2	Protocol_type	Type of the protocol, e.g., TCP and UDP	Discrete	C
V5	dst_bytes	Number of data bytes from destination to source	Continuous	C
V39	dst_host_srv_serror_rate	Server error rate	Continuous	H

The findings imply that integrating the attributes of BCTH might result in the development of a more comprehensive intrusion detection model.

4.3. Q2: Guidelines

The rules were developed from the data that was contained in our training set. In Figure 3, the DT that was built by utilizing our dataset is displayed. The leaf node is located at the bottom of the tree, whereas the root node is located at the top. As the rules are explored, they are done so from the highest nodes to the lowest nodes. There are a total of 19 phases involved in the process of constructing this tree. As the phases progress from node to node, they are shown in a depth-first fashion. Each traversal from the root node to the leaf node provides the governing rule for the decision. This rule is supplied by the chosen option. In this case, the leaf nodes are the decision nodes.

Figure 2_ Decision tree from the training dataset. Pink nodes are malicious nodes, and green nodes are normal nodes.

We utilized the R rattling package to extract the rules from our model. Table 3 displays the five principles for identifying malicious and regular nodes for demonstration purposes.

Table 3 Decision rule description from the training sets for malicious and normal nodes. The bold text is the clauses for the rules.

Leaf numbers	Rules	Interpretation
63	[V42 = normal. Cover = 681132 (47%)] V23 < 79.5 V37 < 0.495 V39 < 0.905 V5 ≥ 8.5	**IF** V23 < 79.5, V37 < 0.495, V39 < 0.905, V5 ≥ 8.5, **THEN** the class is normal, which covers 47% of terminal nodes, overall 681132 cases
125	[V42 = normal. Cover = 43147 (3%)] V23 < 79.5 V37 < 0.495 V39 < 0.905 V5 < 8.5 V3 = auth, finger, http, IRC, smtp, telnet, tftp_u, time, X11	**IF** V23 < 79.5, V37 < 0.495, V39 < 0.905, V5 < 8.5, V3 = auth or finger or http or IRC or smtp or telnet or tftp_u or time or X11, **THEN** the class is normal which covers 3% of terminal nodes, overall 43146 cases
14	[V42 = malicious. Cover = 3559 (0.01%)] V23 < 79.5 V37 < 0.495 V39 ≥ 0.905	**IF** V23 < 79.5, V37 < 0.495, V39 ≥ 0.905, **THEN** the class is malicious, which covers very small amount of terminal nodes, overall 3559 cases
60	[V42 = malicious. Cover = 1538 (0.005%)] V23 < 79.5 V37 < 0.495 V39 < 0.905 V5 ≥ 4.995*e* + 04 V3 = finger, http	**IF** V23 < 79.5, V37 < 0.495, V39 < 0.905, V5 ≥ 4.995*e* + 04, V3 = finger or http, **THEN** the class is malicious, which covers very small amount of terminal nodes, overall 1538 cases
4	[V42 = malicious. Cover = 709860 (49%)] V23 ≥ 79.5 V6 < 2	**IF** V23 ≥ 79.5, V6 < 2, **THEN** the class is malicious, which covers 49% of terminal nodes, overall 709860 cases

4.4. Q3: Precision

In this study, we compared the performance of the decision tree to that of two classification techniques that are extensively utilized: logistic regression and support vector machines. The results of the comparison are presented in Figure 4. According to the data, DT algorithms have a higher level of performance in terms of F1-score, accuracy, and recall. However, the improvement is not particularly significant. When it comes to successfully distinguishing between malicious and benign nodes, DT did not vary in terms of accuracy or recall. In contrast to DT, support vector machines (SVM) perform badly when it comes to predicting normal nodes, but they perform comparably well when it comes to predicting dangerous nodes. There was a significant difference between the performance of LR and that of SVM when it came to predicting normal nodes. However, when it

came to identifying dangerous nodes, SVM and DT performed better. When compared to other methods, the SVM has a poor recall in the data that we analysed.

Furthermore, when it comes to the F1-score, DT consistently demonstrates the best performance. In addition to the fact that it does not require linear/normal or additive features, DT also does not require the definition of possible interactions, which is one of the reasons why it performs better than comparable methods. Every aspect of the problem, including missing covariate values, multicollinearity, and outliers, is taken into consideration automatically. A further point to consider is that although LR is a linear model, it might not perform as well as DT in particular datasets due to the fact that some attributes might not be linear. In a manner analogous to this, we employ SVM with a linear kernel; despite this, it yielded better results than LR but was not as effective as DT.

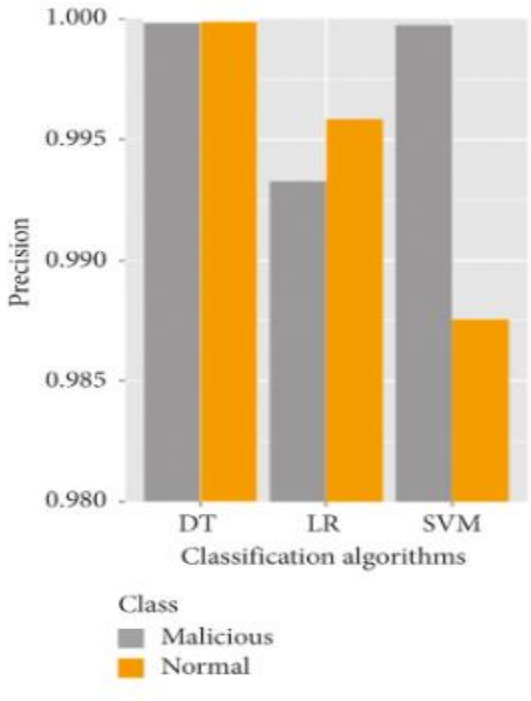


Figure 3 Performance of the state-of-the-art methods in predicting the classes between malicious and normal nodes. LR: logistic regression, DT: decision tree; SVM: support vector machines. (a) Precision

Providing a trustworthy artificial intelligence solution for intrusion detection systems is one of the advantages offered by the explainable analysis of the decision tree algorithm. To be more specific, the rule-based model (such as causal reasoning and underlying data evidence) and feature engineering have been utilized by the decision tree algorithm in order to analyze its conclusions. In the future, it is very necessary for cyber-security systems to have an understanding of the characteristics of harmful assaults and the ways in which they might be susceptible to intrusion detection systems.

It is interesting to note that a substantial body of research that utilized KDD intrusion detection data established a high level of prediction accuracy. The initial dataset from KDD served as the foundation for this study. By

comparing the various approaches to one another, the earlier study indicates which supervised classification algorithms perform better in the datasets. Our technique, on the other hand, is unique. Discovering the origin of the security breach and telling them about it is equally as important as identifying the breach itself. The algorithms that are used in machine learning are not ideal for application in a range of scenarios since they are difficult to understand and may include inference procedures that are difficult to perform. When it comes to circumstances in which it would be vital to grasp the option, such as when identifying rogue nodes in a communication network, this is especially true. Everyone who was involved in the attacks, including the security professionals and network engineers, has the "right to an explanation." As a result of the fact that the capacity of a decision tree to understand human idea processes commonly mirrors human thought processes, it is simple to comprehend the facts and arrive at conclusions that are insightful.

The following is the composition of the most common decision rules: The next step is to make a forecast, provided that the preconditions are met. It is most possible that choice rules are the most straightforward models for prediction to comprehend. Its IF-THEN form semantically reflects human language and the way people think if the condition is created from intelligible qualities, the length of the condition is brief (a limited number of feature = value pairs combined with an AND), and there aren't too many rules. In addition, the condition's duration is brief. It is the datasets that are responsible for teaching the decision tree these IF-THEN rules. With regard to the identification of rogue nodes, these principles have the potential to be of critical importance in stopping significant attacks and preventing additional damage to the communication infrastructure. Rules such as these, for example, may serve to notify persons responsible for security: If V23 is greater than 79.5 and V6 is less than 2, then there is a 97% probability that the class is malicious. If the number of connections to the host is greater than 79.5 and the error status of the connection is less than two, then there is a 97% possibility that malicious nodes are present. Therefore, it is required to exercise a higher level of vigilance. When it comes to working with network security workers who are always on the lookout for potentially harmful activities on the network, this advice is not only easy to adhere to but also quite beneficial.

A other categorization method, such as support vector machine (SVM) or even deep neural network, does provide us with a more accurate prognosis. Despite this, the decision tree is often considered to be among the most successful methods in terms of its capacity to be interpreted. The tree represents the human vision of the natural world. The questions and answers that are included in a decision tree that was developed by a human being are based on the person's comprehension and reasoning. In the field of data science, these rules are frequently generated by an algorithm that, through the process of analyzing the entire dataset, learns which questions to ask. With regard to the identification of harmful nodes, the algorithm will investigate whole datasets in order to establish a separate set of criteria that will be used to differentiate between "normal" and "malicious" nodes. A mathematical division of the entire network dataset is performed by the decision tree algorithms, which then decide its classification. For the purpose of this technique, decision trees are utilized to

first acquire knowledge regarding network intrusion data and then to classify newly acquired network data in a manner that is simple for humans to grasp and apply for decision-making purposes.

CONCLUSION

For the purpose of analyzing and comprehending the 1618 predictions that were generated by AI algorithms, a system known as XAI is utilized. 1619 Artificial intelligence has the capability to analyze datasets and keep track of a wide range of various security concerns and malicious behaviors in the realm of cybersecurity. Integration of artificial intelligence and human beings is the only method that can effectively deal with the increasing number of cyberattacks and the various cybersecurity problems. With the intention of utilizing explainability as a means of bridging the gap between people and machines, this article from 1623 provides an analysis of work that has been given over the past five years on the subject. After conducting a comprehensive analysis of the two ecosystems, namely XAI and Cyber1626 Security, an investigation into the components of Cyber1626 Security that are most significantly influenced by the utilization of AI was carried out. This study is distinguished from others by its assessment of how each technique provides explainability for a variety of application areas. This study also highlights the 1630 lack of formality and the importance of defining a standard. As of 1631 After everything was said and done, 1632 open challenges and the most important topics were investigated. For the purpose of ensuring that 1633 ad hoc frameworks and models are developed for safety rather than the 1634 application of generic models for post-hoc explanation, a large amount of effort will be required.

Traditional AI models, such as deep learning and machine learning, are able to make decisions that are more transparent and explainable thanks to XAI, which is a powerful framework. Nevertheless, cyber security is a domain in which openness and explainability are required in order to defend against the dangers associated with cyber security and to assess the security judgments that are generated. Because of this, we have included in this paper a comprehensive review of the most recent research on the use of XAI for applications related to cyber security. We completed the implementation of the essential concepts and taxonomies of the most recent XAI models by making use of important resources such as a generic framework and datasets that are easily accessible. In addition, we investigated the most cutting-edge XAI-based cyber security systems from a variety of application scenarios. These scenarios included the use of XAI to defend against a wide range of cyberattacks, the use of XAI for cyber security in a variety of industrial applications, and the detection of cyberthreats that target XAI models and related defensive strategies. The presentation covered a number of common types of cyberattacks, including malware, spam, fraud, distributed denial of service attacks, phishing, network intrusion, and botnet attacks. They were given the appropriate defensive mechanisms that employ XAI against them, which were taken into consideration. XAI was used in a variety of industrial areas, such as human-computer interaction, smart healthcare, smart financial systems, smart agriculture, smart cities, and smart transportation, and a comprehensive description of its use was provided. The presentation included a

variety of cyberattack techniques that were intended at XAI models, as well as responses that corresponded to those strategies. For the purpose of following up on these, we brought attention to and discussed a few challenges, significant findings, and potential future uses of XAI in the field of cyber security research. As a result of our study, we predict that researchers, developers, and security specialists who are interested in applying XAI models to challenging challenges in cyber security domains would find it to be a beneficial resource.

REFERENCES

1. CISA, "What is Cybersecurity? | CISA," What is Cybersecurity? https://www.cisa.gov/uscert/ncas/tips/ST04-001 (accessed Jul. 01, 2022).
2. D. S. Berman, A. L. Buczak, J. S. Chavis, and C. L. Corbett, "A Survey of Deep Learning Methods for Cyber Security," Information, vol. 10, no. 4, Art. no. 4, Apr. 2019, doi: 10.3390/info10040122.
3. "Number of internet users worldwide 2021," Statista. https://www.statista.com/statistics/273018/number-of-internet-usersworldwide/ (accessed Jul. 01, 2022).
4. "2021 Cyber Attack Trends Mid-Year Report | Check Point Software." https://pages.checkpoint.com/cyber-attack-2021- trends.html (accessed Jul. 01, 2022).
5. "Cyberattack disrupts unemployment benefits in some states," Washington Post. Accessed: Jul. 02, 2022.
6. "Threat Landscape," ENISA. https://www.enisa.europa.eu/topics/threat-risk-management/threatsand-trends (accessed Jul. 02, 2022).
7. D. Gümüşbaş, T. Yıldırım, A. Genovese, and F. Scotti, "A Comprehensive Survey of Databases and Deep Learning Methods for Cybersecurity and Intrusion Detection Systems," IEEE Systems Journal, vol. 15, no. 2, pp. 1717–1731, Jun. 2021, doi: 10.1109/JSYST.2020.2992966.
8. S. Zeadally, E. Adi, Z. Baig, and I. A. Khan, "Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity," IEEE Access, vol. 8, pp. 23817–23837, 2020, doi: 10.1109/ACCESS.2020.2968045.
9. S. M. Mathews, "Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review," in Intelligent Computing, Cham, 2019, pp. 1269–1292. doi: 10.1007/978-3-030-22868-2_90.
10. "Explainable Artificial Intelligence for Tabular Data: A Survey | IEEE Journals & Magazine | IEEE Xplore." https://ieeexplore.ieee.org/document/9551946 (accessed Jul. 02, 2022).
11. B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation,'" AI Magazine, vol. 38, no. 3, Art. no. 3, Oct. 2017, doi: 10.1609/aimag.v38i3.2741.
12. "A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques | IEEE Journals & Magazine | IEEE Xplore." https://ieeexplore.ieee.org/document/9614151 (accessed Jul. 02, 2022).

13. H. Jiang, J. Nagra, and P. Ahammad, "SoK: Applying Machine Learning in Security - A Survey," Nov. 2016.

14. "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection | IEEE Journals & Magazine | IEEE Xplore." https://ieeexplore.ieee.org/document/7307098 (accessed Jul. 05, 2022).

15. D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," Cluster Comput, vol. 22, no. 1, pp. 949–961, Jan. 2019, doi: 10.1007/s10586-017-1117-8.