Chapter 1
Black Box Models for eXplainable Artificial Intelligence

Krishna Keerthi Chennam, Swapna Mudrakola, V. Uma Maheswari, Rajanikanth Aluvalu, and K. Gangadhara Rao

Abstract Machine learning algorithms are becoming popular nowadays in cyber security applications like Intrusion Detection Systems (IDS). Most of these models are anticipated as a Black Box. Previously black box was a model where the user cannot see the internal logic. To reach the goal of overwhelming the crucial weakness, the cost may vary. This is related to both ethical and practical problems. Explainable Artificial Intelligence (XAI) is crucial to converting the machine learning algorithms to appreciate the management by accepting the human experts to understand the data evidence. Important role of trust management is to accept the impact of malicious data to identify the intrusions. This chapter addresses the XAI method to appreciate trust management using the decision tree models. Basic decision tree models are used to simulate a human contact to decision making by dividing the options into multiple small options for the IDS area. This chapter aims to implement the arrangement of issues labeled in the various black box methods. This survey helps the researcher to understand the classification of various black box models.

Keywords Black box · Cyber security · Decision trees · Intrusion detection system · Artificial intelligence

K. K. Chennam (✉) · S. Mudrakola
Vasavi College of Engineering, Hyderabad, India
e-mail: krishnakeerthich@gmail.com

S. Mudrakola
Matrusri Engineering College, Hyderabad, India

V. U. Maheswari
KG Reddy College of Engineering, Hyderabad, India

R. Aluvalu · K. G. Rao
CBIT, Hyderabad, India

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_1

1.1 Introduction to Machine Learning

There was a huge increase in artificial intelligence (AI) in a glimpse. Machine learning is a subset of AI. The main importance of machine learning is identifying the structure of data or format suitable data models used by the users. However, Machine learning is related to computer science and varies from former computational methods. Previously, Algorithms were written exclusively programmed instructions for computers to solve problems. Now machine learning (Othman et al. 2018) algorithms are used to educate the computers on data inputs and data statistics, analysis is used to produce output values within a range. Automatically decision is taken based on the sample data with the help of models and inputs. Many technologies are using machine learning (Gilpin et al. 2018) algorithms and get benefited. Facial recognition is one of the technologies which permit social media platforms like Facebook and Instagram's to help the users tag and share friends' photos (Logas et al. 2022). Movies or television shows using optical character recognition technology help to change images to text into movable (Jiang et al. 2022). Self-driving cars also depend on machine learning to map the routes (Saha and De 2022). Machine learning is consistently improving technology, which requires continuously improving methodologies for analyzing may affect the machine learning process (Pazzani et al. 2001). Supervised and unsupervised learning are two basic machine learning methods. Along with these two methods k-nearest neighbor algorithm, decision tree learning methods and deep learning are other important concepts in machine learning.

Firstly, supervised learning purpose is to learn by similar outputs by identifying errors and changing the models depending on the output (Cai et al. 2022). This model also uses the patterns to identify the labeled values and unlabeled data also. Supervised learning algorithms will make sure to identify the images and produce labels to the particular image by seeing the cat image, supervised learning will be able to identify and label it as an animal. Unsupervised learning is to identify the secret patterns in the data and automatically identify the classification of raw data. This is used for transactional data and complex data is more expansive and unrelated to organize properly (Kotenko et al. 2022). Example like unsupervised learning will be able to tag all cat images and group it.

Machine learning is based on statistics with basic knowledge by understanding and supporting machine learning algorithms. Correlation is used to identify the relation among two dependent or independent variables. Regression was used for identifying the relation among dependent and independent variable. When an independent variable is given and needs to identify the dependent variable, the regression statistics used to identify it is called regression enables prediction capabilities. To identify the pattern k-nearest neighbor algorithm is used for regression and classification. Small and positive integer is k value. Example of separating the square and circle shapes into two different classes, this classification is used.

Decision tree is a predictive algorithm based on the models, observations, analysis and gives target data values. This model is created to predict the target based input values. The data attributes identified based on the observation are branches

the conclusion of data target values is nothing but leaves. Deep learning is introduced based on neural networks with multiple layers in artificial neural networks based on hardware. The output is connected to an input to the next layer in the deep learning process. Computer vision and speech recognition have realized significant advances in deep learning approaches (Li et al. 2022). Humans can give biased decisions that lead to negative results, machine learning helps to overcome such issues and give unbiased decisions. Black box (Guo 2020; Perarasi et al. 2020a) systems exploit sophisticated machine learning models to identify separated secure data. Medical status, risk of insurances, eligibility score for credit cards acknowledge using machine learning algorithms construct predictive models and map the features into class in the learning phase (Svenmarck et al. 2018). The learning process is formed by the digital trances that are left after operating daily activities like social media activities, purchases, etc. Huge data may handle human biases and prejudices. Decision models are accomplished by inheriting biases, wrong decisions and illegal activities. Various scientific communities studied the issues of discussing machine learning decision models. Even though illustratable machine learning is the important case and accepted newly considering the situation, many ad-hoc distributed results.

The rest of the chapter is organized as follows. The First section discusses the importance of cyber security in XAI. Next section discusses Deep learning using XAI which follows the Intrusion Detection System (IDS). Section 1.5 is about applications of cyber security in XAI. Section 1.6 discusses the comparison of XAI using black box methods and finally about the conclusion.

1.1.1 Motivation

The unique aim of the chapter is to reach the novelty in research work using machine learning. AI understands different technologies under the same umbrella like machine learning to predict the results. Machine learning ultimately reaches the goal to reach for accurate results with training the model.

1.1.2 Scope of the Paper

Machine learning is one of the best options in career applications for smart systems to handle business attacks. Target is to calculate human intelligence and be able to make decisions more precisely under any situation. AI handles the different technologies that come under the same domain like pattern recognition, big data, machine learning, artificial intelligence and various other technologies. This is the reason AI is having much future scope in many applications.

1.2 Importance of Cyber Security in eXplainable Artificial Intelligence

Industries progressively improved with a better complex cyber security (Pienta et al. 2020) ecosystem depending on various types like users, technology and processes to functional roles. Cyber security is dependent on relations between users and groups, users, organizations and technology, technology and users. From the above trusting peers, cyber security prevents separately to defend against cyber attacks. AI models cite the knowledge from the gathered data. Actually, no human will believe the AI system for the possible and desirable quality of data, difficult methods and accountability, trained AI engineer. AI is trust related software that gives solutions to cyber-attacks. You may ask how to trust the AI models in cyber security, which are developed based on data analysis and predict the solutions from the data. The simple answer for this question is that XAI (Guo 2020; Arrieta et al. 2020) will justify reliability, ability, and trustworthiness. Main challenge for AI is the inability to understand and compare between transition models. A simple example is Autonomous vehicles (Perarasi et al. 2020b). Trustworthy AI should explain its decisions to allow the human expert to understand the underlying data evidence and causal reasoning.

Complex black box models study from machine learning and deep learning parameters. Based on the black boxes models, AI engineers identify direct models to make decisions and identify the behavior of models. Cyber security is liable for attacks and targets the trusted security in critical systems. Therefore XAI from AI plays an important role in developing the solution based AI with interpretability. Interpretability further assures uniformly in decision-making to detect the imbalanced dataset. Interpretability strengthens the powerful solution based AI using highlighting hidden could change the prediction. The decision tree model is developed based on the Intrusion detection system attacks (Svenmarck et al. 2018; Stampar and Fertalj 2015). The intrusion detection system developed fast in study and organization research in exchange for increasing cyber attacks on government and commercial enterprises internationally and action on cost is increased consistently (Lee et al. 2001). The main harmful cyber crimes are from vicious associates, denial of servers, web attacks, and organizations may lose the intellectual property related to vicious attacks in the system. Organizations install various firewalls, software like antivirus and intrusion detection systems against those attacks. Intrusion detection is a crucial role in cyber security, grants to determine vicious network activities previously compromises data connection, availability and opportunity. It is a method to identify security breaches by interrogating models in the data system.

Day-to-day, the digital system is adopted by the world. The network access leads to a lack of security issues that the Internet of Things devices (Lee et al. 2001; Chennam et al. 2022). Intrusion attacks with high possibilities on Internet of Things devices connected to the internet lead to network devices safely from intrusion. An IDS was developed to avoid important data from vicious acts. Important data with network access needs to be permanently protected from all pursuit to consume, expose, alter, disable, steal or gain unauthorized access. Traditional intrusion detection systems,

mainly signature-based, identify only popular attacks and may not identify new attacks. Machine learning is the best approach which is exclusively developed to maintain detection accuracy.

Artificial Intelligence (AI) has helped all the industries with effective results in deploying various applications to monitoring, Decision Making, Solving Complex problems, creative approaches, observation analysis, Language Recognition and Learning. Artificial intelligence has collaborated with additional technology like Machine Learning, Neural networks and Deep Learning. Artificial intelligence is used to compute the programs and prepare the system to behave like a human brain (Uma Maheswari et al. 2021; Deshpande et al. 2020). The AI has excelled in thinking, retrieving and taking decisions sometimes faster than the human brain. AI applications are used in medical Care, Teaching and Learning, Law, Commerce and public Departments etc. The above applications are intended to say that algorithms rule the world by AI, which is inevitable (Swapna et al. 2022).

XAI advantages are mainly concerned with ethics and continuous improvements. XAI required enough trust to handle the AI. For decades various AI models gave biased results or not perfect results which lead to ensuring the safety in AI decisions without any faults. To justify the final decisions taken by the AI required logical reasoning in decision making. AI helps to identify the malware weekly updated and all possibilities of pattern recognition, behavioral attacks of ransomware able to identify before entering into the system. Bots help to clear maximum chunks in internet networks. Stolen login details can create false account details, tampering data; bots can be the correct menace. Handling automatic threats is not possible alone. AI and machine learning will heal to construct the good bots to identify the engine crawlers, bad bots etc. AI starts to identify the data and accepts to provide cybersecurity to understand the strategy consistently.

1.2.1 Importance of Trustworthiness

The importance of Trustworthiness is an essential aspect to measure the safety, performance and reliability. The qualities requirements to say as trustworthy are the system must be accountable, fair, reliable behavior, reasonable and acceptable. The author Stephen Hawking says “AI can spread faster and can be violent if it is not controlled properly”. The AI systems need to be authorized and validated in each design and implementation phase. The AI systems need to be authorized and validated in each phase of the design and implementation. There are different algorithms used to predict the risk of the system, and it occur due to low-quality data training, narrow perception of the problem, technical issue management etc. can lead to unrecoverable loss of people, properties and loss the trust on AI practices. AI applications are used in important applications like facial recognition software, Tagging picture in television media, Health care practices and self-driving car are the high-risk applications, wrong decision may cause life. The author Davinder Kaur has raised some questions

Table 1.1 Questioner table states the importance of Trustworthy in AI

Research questionnaire	Proposed solution
Purpose of proving AI is trustworthy	Decisions taken by the AI system should be ethical practice, robust in nature, Lawful and acceptable
What protocols are used to work AI systems?	We can empower and help to maintain the AI system lawful practices
Why human control involved	AI systems need to collaborate with human intervention and machines in cognitive decision making
Reasons for AI acceptable	AI systems have proven to be trustworthy, fast and usable

to understand the requirements needed to conclude the AI system as worthy (Kaur et al. 2022) (Table 1.1).

The Black Box Model uses AI methods, the results are obtained, but its design will not help to justify the result. The explanations are required to extract the output function. We need to apply some techniques to find the reason to conclude (Zhang et al. 2022). The Post-Hoc Explainable is a reverse engineering process that starts to reach the initial state from the destination. Explainable algorithms like Support Vector Machine (SVM), Multi-Layer Neural Network, Convolution Neural Network and Recurrent Neural Network (Hermansa et al. 2022). XAI uses machine learning techniques to justify the results. The reasonable techniques are explained by simplifying the problem, Feature Connectivity, Local Reasoning, Visible Reasoning and Multi Classifier. The importance of AI is used to make better decisions, explain deep learning, Model Debugging, and build the latest model (Brito et al. 2022) (Fig. 1.1).

Machine Learning (ML) methods Contribution for XAI—The Machine Learning method works for limited data. The ML required defined features to the drive result. The complex problems will simplify and solve phase-wise, network designs are

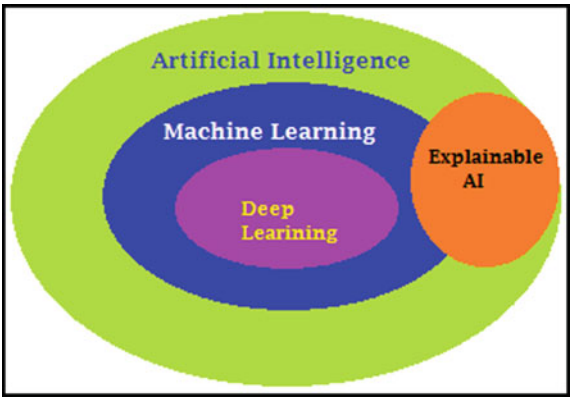

Fig. 1.1 Representation of AI, DL, ML, XAI Association

kept simpler, less trained data and good results are obtained on more and less data size (Lötsch et al. 2022). The ML concepts are classified into three classes: supervised, unsupervised, and Reinforcement Learning. The methods are Artificial Neural Network, SVM, Self Organizing Map, Model-based Reinforcement Learning, clustering, Dimension reduction, Regression, Classification, Transfer Learning, and NLP (Aliramezani et al. 2022).

1.3 Deep Learning (DL) Methods Contribute to XAI

The Deep Learning methods require a huge amount of training data. The feature extractions are undefined at the initial stage Based on the feature's importance, the feature is used for learning. The network training takes more time, based on hidden layers. As no of hidden layers increases, depth analysis is performed (Raza et al. 2022). The DL algorithms are CNN (Conventional Neural Network), MLP (Multi-Layer Perceptron), DNN (Deep Neural Network), and RNN (Recurrent Neural Network). All the methods extract the results using the above Neural Network. The above methods measure the performance using ROC, F1-Score, Accuracy, Recall and Precision metrics (Zhang et al. 2022).

Block Diagram of XAI is shown in Fig. 1.2, Data training is a machine learning process to teach the environment about the possible case studies and the latest and most possible cases, calculus applied to find the break-even point and threshold calculations, etc. Different Machine learning methods are used to extract the results. The Machine learning approaches discussed section I. The recommendation and conclusions are obtained based on the methods of ML Programs selected. The Explainable AI will strength up the decision by explaining the reason for obtained results. (Hoffman et al. 2018) Functionality difference between AI and XAI, ML functions will learn from training data and decisions are made based on learned function. The XAI will train the data, ML process will process the input to obtain the result. The Explainable AI will reason the result and explanation will be verified by the user to test the accuracy behavior. The conclusions and recommendations are assessed by the user (Hermansa et al. 2022) (Table 1.2).

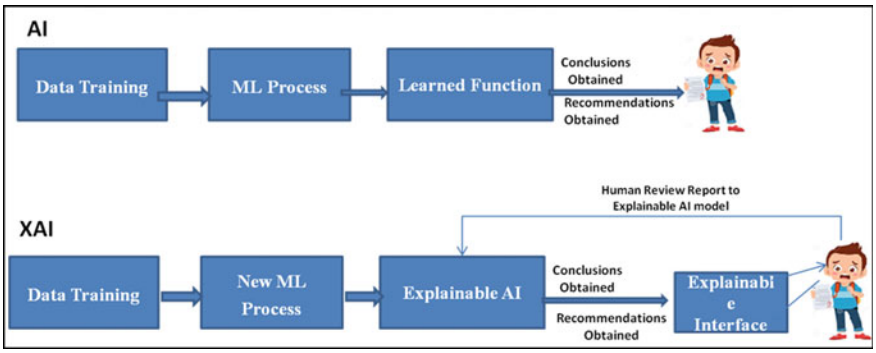

Fig. 1.2 Phases in AI and XAI model

Table 1.2 The details of the Explained AI and Non-XAI for User Perception

S. No.	User perception	AI/ML/DL algorithms	Explainable AI
1	Did the result obtained cause only this?	Unable to understand clearly	Understanding clarity
2	Alternate options for result opted	Unknown reason for not selecting another choice	Known reason for not selecting another choice
3	Success in result obtaining	Unknown for success	Known for success
4	Failure in result obtaining	Unknown for failure	Known for failure
5	Trust the system	Unpredictable	Predictable
6	Error in system	Correct based on the false positive	Reasoning for false positive

1.4 Intrusion Detection System

This section first explains the concept of IDS and then provides the details about the classification of IDS based on its deployment and the detection methodology. An IDS is the combination of two words, "intrusion" and "detection system." Intrusion refers to unauthorized access to the information within a computer or network system to compromise its integrity, confidentiality, or availability. The detection system is a security mechanism for detecting of such illegal activity. So, IDS is a security tool that constantly monitors the host and network traffic to detect any suspicious behavior that violates the security policy and compromises its confidentiality, integrity, and availability. The IDS will generate alerts about detected malicious behavior to the host or network administrators. Figure 1.3 depicts a passive deployment of NIDS, where it is connected to a network switch configured with the port mirroring technology. The task is to mirror all the incoming and outgoing network traffic to NIDS for performing traffic monitoring to detect intrusions. NIDS deploys in between the firewall and the network switch to allow all the traffic to pass through NIDS. IDS have grown

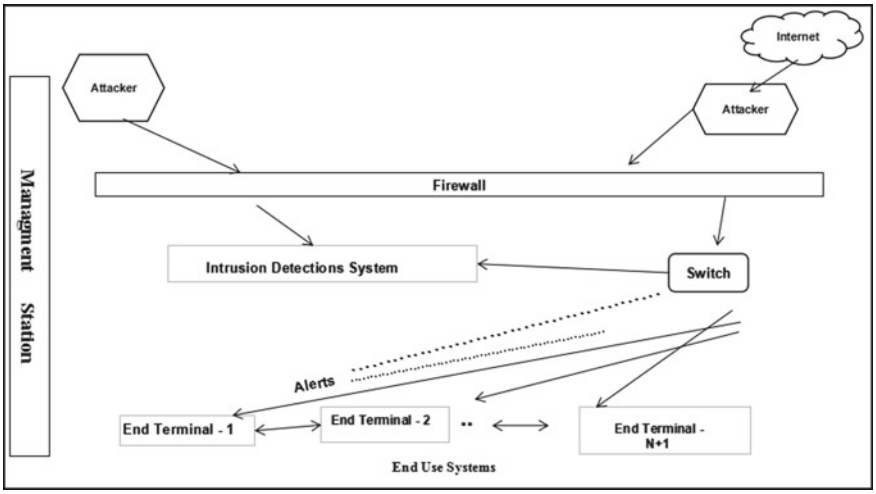

Fig. 1.3 Intrusion detections system network (Chebrolu et al. 2005)

quickly in exploration and industry in light of the expanding digital assaults against state run administrations and business ventures worldwide. The yearly expense of battling digital wrongdoing is constantly expanding (Stampar and Fertalj 2015). The most terrible digital wrongdoings are those brought about by noxious insiders, refusal of administrations, and online assaults. Businesses or organizations can lose their licensed innovation because of these pernicious assaults into the framework. To retaliate against such demonstrations, associations convey a firewall, antivirus programming, and an interruption discovery framework.

Recently detailed predictions, the deep neural network appropriately suitable for these various models are useful and difficult to identify. Autonomous vehicles need various parameters to deal with deep neural networks (Ye et al. 2004). Handling network administrators is difficult if deep neural network models from machine learning are implemented. Deep neural networks are also called a black box models. Decision making process issues are solved with black-box models using trial and error methods till they reaches for feasible solutions. Intrusion detection system implements machine learning methods to improve the accuracy of familiar attacks' analysis and identifies abnormal traffic issues and autonomous vehicle network issues. Machine learning algorithms can identify attacks to interpret the results. The main challenge is to combine the intrusion detection system with deep learning models to ensure security policies against attacks.

Various models proposed for intrusion detection systems like statistical methods (Lazarevic et al. 2003) proposed Markov model (Ye et al. 2004). Neural network (Novikov et al. 2006), fuzzy logic (Toosi and Kahani 2007) discussed. SVMs discussed the huge accuracy in implementing intrusion detection systems (Lahre et al. 2013; Zhang and Shen 2005; Ilgun et al. 1995). Training techniques designed to implement intrusion detection systems, experts considered the rules. Like, discussed

in Rudin (2019), need corporation rules against expertise in analytical models. The disadvantage of such analytical models is the extraction of huge rules, which leads to maximum difficulty in models. One of the critical aspects of a supervised classification model is feature selection. Identifying required features lessens the algorithm computation time. Intrusion detection systems expanse with selections (2017) with various feature. Author (Chebrolu et al. 2005) restricted fundamental features in composing an intrusion detection system is essential for recent detections. (Zaman and Karray 2009) Implements the selection models to construct a lightweight intrusion detection system.

As discussed in Vimalkumar and Radhika (2017), intrusion detection system models with a ratio profits as various techniques selections and two types of techniques like SVMs and endorsed maximum accurate levels denial of service attacks. One of the disadvantages of the model is that it is highly computational with a separate ratio and classification algorithm. Balakrishnan et al. (2014) examined different algorithms like k-means clustering algorithm, Naive Bayes, OneR algorithms for common traffic with accurate results and Denial of Service attacks. A genetic algorithm is implemented to ensure the identification of various models of intrusion with maximum efficiency as discussed by Farrahi and Ahmadzadeh (2015). Appliance different machine learning algorithms identify Denial of Service intrusions and establish the maximum efficiency with multilevel perception (Castelvecchi 2016). Peng et al. (2018) recommended an intrusion detection system planned on a decision tree to boost the accuracy detection. Naive Bayesian and KNN models are performing research to find an accuracy to identify the intrusions with better performance, maximum speed and less false alarm rates like in (Othman et al. 2018; Gilpin et al. 2018).

1.4.1 Classification of Intrusion Detection System

Intrusion detection systems are classified into various types. Firstly the IDS is classified into a network intrusion detection system that helps to identify the packet search, anonymous packets, inbound and outbound packers in firewalls and controls the network traffic. Next classifier is a host intrusion detection system that will monitor the host activities, clients can be ignored in IDS, able to identify the attacks and detected attacks can stop within the network. Another classifier is a protocol-based intrusion detection system used to analyse and monitor communication and monitor HTTP protocols and SQL protocols. Along with all the classifiers the hybrid intrusion detection systems works more efficiently by combining two or more IDS classifiers like host intrusion detection system and network-based intrusion detection system helps to monitor the system, application events and file systems. This hybrid IDS uses for critical server situations (Lin et al. 2022) (Fig. 1.4).

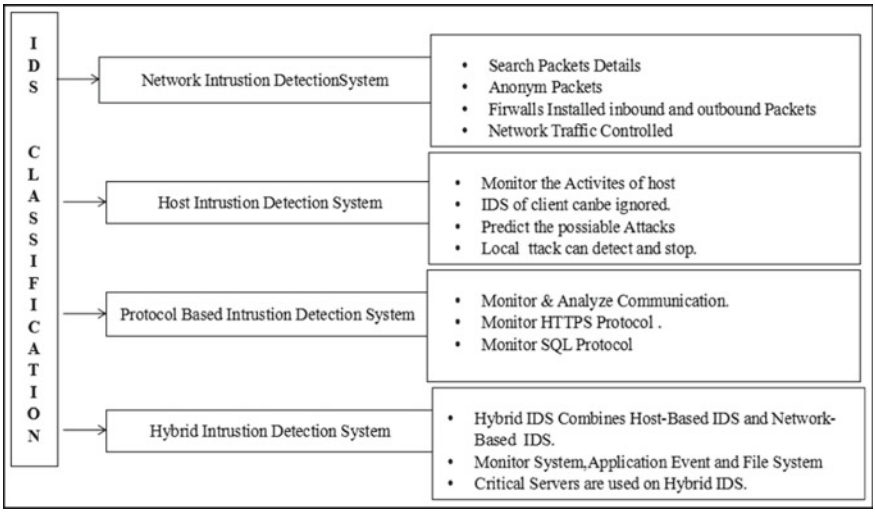

Fig. 1.4 Classification of Intrusion detections system

1.5 Applications of Cyber Security and XAI

It may be described as the method to relax the safety if you want to shield reputation casually, less business or monetary lack of a cluster Cyber security manifestly needs betters for safety with the notion to the business enterprise that regular customers use the network over the internet. Many intelligent tactics and strategies can be cast-off to install in it. The finest huge reality around safeguarding data is no longer a non-prevent process. The business enterprise owner must hold stuff modernized in mandate to keep the danger low (Bonfanti 2022). A business wants to appear a massive harm when they are no longer sincere about the protection in their online display. Nowadays, everyone connects from innovative cybercrime issues to single user issues like attacks, thefts, blackmails, illegal photographs. All are predicted at risk according to the financial supplier of businesses. Giving security to those various fields is vital to understanding general operations like handling information corresponding to credit cards and authorization information. Handling such sensitive data fails is one of the possible cyber attacks. Handling such sensitive data fails is one of the possible cyber attacks. To Gain knowledge on vicious emails, you need to study cyber security. Ransomware is another sort of vicious software program, considered to extract forex with the aid of gadgets or desktops and demand for money. Though after paying money to hacks, it's not guaranteed that the malware is removed from the gadgets or not (Urooj et al. 2022).

Social engineering is a tactic that fighters use to faux you into illuminating sensitive facts. They can importune a monetarist charge or development gets right of entry to your reserved information. Social engineering may be collective with a number of the pressures registered above to fashion you extra in all likelihood to attach on

links, switch malware, or perceive a malicious purpose. Most commercial enterprise operations goals run on the net, exposing their information and assets to diverse cyber threats. Since the information and gadget assets are the pillars upon which the agency operates, it drives the missing maxim that a threat to those people is surely a hazard to the institution itself. A hazard may be everywhere among a minor worm in a code to a complicated cloud hijacking liability. Risk evaluation and estimation of the reconstruction fee assist the agency in living organized and to appear beforehand for capacity losses (Rajanikanth 2021). Thus, understanding and formulating cyber security goals specific to each agency is important in protecting precious information. Cyber security is an exercise formulated for the project of complicated information on the net and on gadgets safeguarding them from attack, destruction, or unauthorized get right of entry to it. Cyber security aims to ensure threat-loose and stable surroundings for retaining the information, community, and gadgets guarded in opposition to cyber terrorizations. Cyber security has become a main challenge over the past 10–12 months within the IT world. All people are dealing with quite a few cybercrime issues in the existing world. As hackers are hacking principal touchy facts from authorities and a few corporation agencies, the people are very much involved as cyber security attacks can result in the whole thing from wholesale fraud to blackmail massive companies. There are many sorts of cyber-crimes rising wherein everybody wishes to be aware of the scams, and they're one of a kind measures and gear which may use for averting the cyber-crimes. Every agency desires to stabilize its private information from getting hacked. Getting hacked isn't always dropping the private information, but dropping the connection with clients within the market. The Internet is the modern day's fastest developing infrastructure. In modern-day technical surroundings, many new technologies are converting humanity. But because of technology, we're not able to guard our non-public facts in a green way, so cyber-crimes are significantly growing on day by day basis. The majority of the transactions in each business and private sector are achieved through an online transaction, so it's crucial to have knowledge that requires an excessive best of safety, retaining higher transparency to everybody and having more secure transactions. So cyber security is the present-day issue. Advanced technology like cloud services, mobiles, E-commerce, net banking and plenty of extra require excessive requirements and a more secure system of safety. All the gear and technology concerned for those transactions keep the maximum touchy and crucial consumer facts. So presenting vital safety to them may be very crucial. Cyber security and safeguarding touchy information and infrastructures are crucial to each nation's pinnacle precedence safety (Bendovschi and Ionescu 2015).

Organizations may see the loss if they are not transparent in security while stepping online. Nowadays, all know the progressive cyber defense agendas. Cyber security may lead to other results from natural theft, loss of photographs, and blackmail attempts at different levels. It all depends on the business levels, monetary services, infirmaries etc. Trusting and providing security in operations is compulsory. Cyber threat investigators train the users on position and recent susceptibilities also. Various kinds of cyber security phishing scam emails from different sources. The main aim of the data is to maintain the security and privacy of credit card details for user

logins which is the highest attack on the organization. Ransomware is one malicious software to blackmail the organization by leaking data or blocking the systems until the organization pays the demanded price. Malware is another kind of software to receive prohibited policy after using it to the system (Gazet 2010).

Enlightening sensitive data is another gimmick in social websites where malicious users can miss-use. The main aim of the huge business operations is to work on the online data and resources, leading to different kinds of cyber threats. The data and resources are an important support for any organization. If any organization lacks security for these two data and resources, it's a big threat from a small bug to difficult data seizing. The reconstruction cost of the organization is very high and may lose the customers due to a lack of trust in the organization. The objective of cyber security for any organization is securing the data and valuable customer information. Cyber security is one of the best practices for organizations to secure difficult online data and provide security from various attacks, destroying the data or checking the authorization or authentication for genuine users to access the data. The main achievement of cyber security is to make sure the surroundings are harmless surroundings for storing data, networks, and resources in contrast to cyber attacks. Among IT industry cyberattacks is one of the main issues from the past 10 years. Presently many common people are facing the cyber attacks problem with cybercrime. Hackers are able to get the sensitive data from the common people and from small organizations to huge organizations also. Different types of cyber crimes are increasing day to day life which gives alert to everyone to be understanding about the cybercrime now to avoid cyber crimes.

Every customer and organization tries to secure the data from the cyber-attacks discussed by Bendovschi in 2015. As infrastructure is emerging rapidly nowadays, the technologies are increasing and changing today. Due to this, we are losing hold on personal data and secure data, which regularly leads to increased cybercrimes (Hussain et al. 2021).

Non-commercial and commercial transactions mainly are happening through the internet. It's mandatory to ensure the experts provide high security and maintain transparency for more safety. New methods like cloud service providers, smartphones, E-commerce, net banking or mobile banking, telecom services need huge security in the implementation. Various tools and techniques tangled the critical sensitive client data.

XAI object detection software applications are designed to identify the objects on the image, video or online live streaming. The computer vision techniques are used to find objects, count similar objects, identify the object, identify the location, and read an image. The algorithms used are R-CNN, HOG, R-FCN, Single Short Detector, and YOLO (you only look once) (Kose et al. 2019). The Deep Neural Network is a 'black box' in behavior. The CNN algorithm will train the neural network; the reasoning for the output is evaluated using techniques Predicate Logic for self-monitoring methods (Floreano and Wood 2015).

Explaining Autonomous Drones application is specially designed to provide carrier service on mountains or hills. The uneven heights of the land and rocks have an issue in flight traffic plan, drone root plan, and physical location features. The

AI Drones are designed with Common Ground Learning and Explanation System (CGLES) with Explainable intelligence to what is comparable, why it is a drone, where explanation. Simultaneously performance prediction with explanations, Pre-Detection and Post-Detection Explanation used to determine the performance of the CGLE system. Pre-Detection is used to define the plan and execution performance, and Post-Detection is used to determine and find the betterment (Tseremoglou et al. 2022).

Explaining forecasting and packing for Air Cargo loading application will predict and decide to air cargo service to help accept or reject the booking based on the estimation of passenger aircraft belly. The cargos services are unpredictable due to the last hrs cargo service details being released. The training data and historical studies will help to predict the certainty. The author proposed a novel framework to provide process consideration based on the balance of aircraft capacity and dimensions of the ship (Han and Liu 2022).

Explaining Structure Health Monitoring through AI and ML applications are used to detect the extract the features and predict the patterns. Machine Learning is used for transparent processing, and some limiting features will undergo black-box execution. The XAI preprocessing problem uses applications in ML algorithms. An explanation is retrieved, interpretations and finally, build XAI model. The first ML algorithms are SHM systems. In this phase, it is required to remove the noise in the data and improper information. Supervised, unsupervised and reinforcement are the approaches. Four algorithms and ML algorithms are used to understand the problem, confirm the method, apply the method, explain the output, and interpret (Chou et al. 2022; Swapna et al. 2022; Kanaparthi and Swapna 2022).

Explainable AI for Deep Learning Models applications is specially used in applications for big data as it is difficult to handle for a complex task (Anders et al. 2021). Large scales of application are Speech Recognition, Text Analysis, Problem Solving and Image Classification. The above applications run very successfully for ML and AI concepts. But the decision extracted is not transparent. The author took an image classification problem in identifying the object and Interpreted the reason for classification. The process has been decomposed into classification steps. The input(x) was applied to a black box AI system and was predicted as a Rooster(x). The prediction has an AI explanation (we will verify the predictions, Identify the issues and differences, understand the problem, ensure the problem) (Han and Liu 2022).

Explaining AI for the Breast Cancer Detection Case based on reasoning applications designed in old patients or new patients, the general classification is the black-box approach. The measures are considered based on the terms of values and quality-wise. It has an automatic interpretation. The Databases are designed and retrieved using queries. The automatic case retrieval system will extract similar cases and retrieve preprocess queries stored in temporary memory. Automatic classification is used to classify the database, the reasoning based on Quantitative and Qualitative. The visual reasoning with query's and class classifications. The three algorithms are tested KNN, WKNN and RBIA to find the clinical validation to improve the study of CBR, and the third is to size the tumor. Finally, the knowledge discovered from the

medical dataset analyzer not only from CBR (Zhang et al. 2022; Swapna and Hegde 2021).

Various methods in XAI to predict the outputs are:

i. SHAP (Fidel et al. 2020)
ii. LIME (Visani et al. 2020)
iii. SHAPASH (Ghosh and Sanyal 2021)
iv. EXPLAINER DASHBOARD
v. DALEX (Baniecki et al. 2020)
vi. EXPLAINABLE BOOSTING MACHINES (Naser 2021).

SHAP (Shapley additive explanations) is a python tool used to visualize the model's output using Machine Learning implementation and helps visualize the output with more explanation. This algorithm will help explain the reason for the output of the prediction model. In the SHAP feature, importance is the first step to finding the important attribute from another attribute, and they are evaluated using slandered deviation and mean. It will help to remove the impurity in the decision. The SHAP will plot different plots like the SHAP Summary Plot used to combine the features and plot data points. SHAP Dependence plot will help plot the marginal effect of features, SHAP Force used for error analysis, explanation for findings for prediction (Kuzlu et al. 2020).

LIME (Local Interpretable Model-Agnostic Explanations) is another python package used to predict the classification and regression of the data. All the sample data will be extracted from the feature, and observe the results. The explainer will help to predict the result for each output. The linear regression is regularized. The difference between output and actual will use r2. The difference between actual and predicted output gave by linear regression function. It will explain the black-box model machine learning model. Local Interpretation will help to calculate trustiness and it also prove the untruth the model and visual explanation (Främling et al. 2021).

SHAPASH is a python library to Interpretable the Machine Learning results. It uses to build the Web Application to interpret the decision of the data scientist, users or customers, Stake holders of the business and Evaluators of the system. It provides the visualization explanation, to understand by the common users. The shapash has five step processes.

Step 1: Regression model is Build.
Step 2: SmartExplainer of Shapash is compiled and displayed on the webapp.
Step 3: Smart Predictor is predicted from the SmartExplainer.
Step 4: SmartPredictor can be saved in pickle File.
Step 5: Finally predictions can be made (Ghosh and Sanyal 2021).

EXPLAINER DASHBOARD is an interactive dashboard works for the machine learning models. It helps to analyze the reasons for the predictors and explanation on the working of the model. The Explainable Dashboard will help to build the unclouded machine learning model and Explainable. Classifier Dashboard has features like Classification Stats, Individual Predictions, and Own Condition prediction, Feature Dependence, Feature Interaction and Decision Tree. This can be

supported by Colab Programming. The dashboards are different types like single tab dashboard, Multi-TabDashboard, Documentation Dashboard etc. are the sample types of dashboards.

DALEX is a model Agnostic Language for Exploration and eXplanation is a Machine Learning analysis model helps to learn the behavior of the Model Predictor in Classification process and Applying Regression methods. This approach will also help to build the relation between the dependent variable to predict the outputs. It is also an interactive model for exploration of predictions. Explanatory Model Analysis. This model works with different levels of explanations like predict Level Explanation, Model Level Explanation, Save and Loading Explanation. Plot the graphs for user visualizations (Baniecki et al. 2020).

EXPLAINABLE BOOSTING MACHINES (EBM) are the cyclic Gradient boosting Adaptive model, tree-based classification and interactive model. This method for predication is said to be black box model, which said to be more accurate. Limitation is EBM will take long time to train the model but very fast at prediction. The process for EBM is train the model for classifiers, Visualization is explained in terms of local and Global. The Specific attribute analysis need to perform. All the Attribute Mean Absolute Score need to calculate (Naser 2021) (Table 1.3).

Table 1.3 Comparison of XAI techniques

References	Methods	Intrinsic	Post-hoc	Global logic	Local logic	Specific model	Agnostic model
Roth et al. (2021)	Decision trees	✓		✓		✓	
Das and Rad (2020)	Rule lists	✓		✓		✓	
Dieber and Kirrane (2020)	Lime		✓		✓		✓
Dhanorkar et al. (2021)	Sharply explanations		✓		✓		✓
Schlegel et al. (2019)	Saliency maps		✓		✓		✓
Fouladgar and Främling (2020)	Activation maximization		✓	✓			✓
Kłosok and Chlebus (2020)	Surrogate models		✓	✓			✓

(continued)

Table 1.3 (continued)

References	Methods	Intrinsic	Post-hoc	Global logic	Local logic	Specific model	Agnostic model
Ryo et al. (2021)	Partial dependence plot		✓	✓	✓		✓
Barbado et al. (2022)	Rule extraction		✓	✓	✓		✓
Adadi and Berrada (2018)	Model distillation		✓	✓			✓
Baur (2018)	Sensitive analysis		✓	✓	✓		✓
Keane et al. (2021)	Counterfactual explanations		✓		✓		✓
Heide et al. (2021)	Prototype and criticism		✓	✓	✓		✓
Anders et al. (2021)	Layer wise relevance program		✓	✓	✓		✓

1.6 Comparison of XAI Using Black Box Methods

The depth analysis of classifications of black-box models discusses in this section using XAI. The reverse engineering approach is used in the black box also familiar with black box predictors observing the input and output of the black box. Assigning decisions to black boxes is difficult to interpret may have differences and trust problems. Former datasets and training models handling human decisions could depend on (Pedreshi et al. 2008). These methods are acutely covered up inside the classified trainer. Improving the black-box model is a high risk, as discussed in Pasquale (2015), and carried by secret algorithms, legal protections, and differences consciously or unconsciously may lead to invisible or impossible. Automated differences are not new and not compulsorily due to the black box (Kuppa and Le-Khac 2020). Comparison methods for black-box models using XAI are discussed in the below table. The data types are used to analyze the black box models using XAI. General is the explanatory method to reach every black-box method. Random is the type indicated to randomly select any random perturbation of a data set. Code indicates the source code is available. The tabular method data set analyses the comparisons for black-box models using XAI (Table 1.4).

Table 1.4 Comparison methods for black box models using XAI

Reference	Explanator model	Black box types	Dataset technique
Fidel et al. (2020)	Decision tree	Neural network	General, dataset
Krishnan et al. (1999)	Decision tree	Neural network	General, dataset
Kuppa and Le-Khac (2020)	Decision tree	Neural network	General, random
Zhang and Shen (2005)	Decision tree	Neural network	General, random
Chipman et al. (1998)	Decision tree	Tree ensemble	Dataset
Peng et al. (2018)	Decision tree	Tree ensemble	General, random, dataset
Roth et al. (2021)	Decision tree	Tree ensemble	General, random
Hara and Hayashi (2016)	Decision tree	Tree ensemble	Random
Farrahi and Ahmadzadeh (2015)	Decision tree	Tree ensemble	Random
Främling et al. (2021)	Decision tree	Tree ensemble	General, random
Pasquale (2015)	Decision tree	Tree ensemble	Dataset
Adadi and Berrada (2018)	Decision rules	Neural network	Random
Adadi and Berrada (2018)	Decision rules	Neural network	General, random
Farrahi and Ahmadzadeh (2015)	Decision rules	Neural network	General, random, code, dataset
Ryo et al. (2021)	Decision rules	Neural network	Random, dataset
Chebrolu et al. (2005)	Decision rules	SVMs	Dataset
Balakrishnan et al. (2014)	Decision rules	SVMs	Dataset
Vimalkumar and Radhika (2017)	Features importance	AGNostic black box	General, code, dataset
Urooj et al. (2022)	Features importance	AGNostic black box	General, random, code, dataset
Rudin (2019)	Decision tree	AGNostic black box	General, dataset
Lahre et al. (2013)	Features importance	AGNostic black box	General, random, code, dataset

(continued)

Table 1.4 (continued)

Reference	Explanator model	Black box types	Dataset technique
Heide et al. (2021)	Features importance	AGNostic black box	General, random, code, dataset
Brito et al. (2022)	Features importance	Tree ensemble	Dataset
Barbado et al. (2022)	Features importance	SVMs	Code, dataset

1.7 Conclusion

AI takes final decisions with the help of difficult model analysis to calculate potentially secret patterns and low signals using huge data sets. Approaching the real-time application for trusting the AI-related solution. Analyzing and considering the AI-related solutions is required for trusting the applications in real-time. Cyber security systems are important applications sensitive to systems that are at risk in vicious attacks. Accordingly, decision tree algorithms for vicious nodes describe by handling the available datasets. Performance is calculated based on major tasks in the dataset like identifying ranks, decision tree extraction, and state of the art algorithms correlation. All the methods do not have the same level of improvement as the vicious node predictions. The various methods in XAI and the network traffic model are calculated with a double-time window as the important predictors in the decision tree and related to deep root node algorithms. The next second-highest rank is the feature-based network service for the personal TCP connection. This book chapter addresses problems like black-box model types, mapping the cyber security in XAI, and the role and importance of IDS in XAI. Black box working process and explaining decisions even without understanding the depth of opaque decision systems work regularly. Various approaches introduced in black boxes and a few scientific queries are still unable to answer. Research exercises neglected regular reading formalism by defining, describing, and identifying various types. XAI research is dependent on the application domains and affects the various huge applications. The evidence is the main drawback of formulation and unambiguous definitions. The work is affected by challenges with open problems in XAI. Finally, the further intriguing point is that clarifications are significant alone and indicators may be advanced straightforwardly from clarifications. AI is a powerful device that can be utilized in numerous areas of data security. There exist some vigorous enemies of phishing calculations and organization interruption identification frameworks. AI can be effectively utilized to create validation frameworks, assess the convention execution, survey the security of human collaboration verifications, brilliant meter information profiling, etc. Even though AI protects different frameworks, the AI classifiers are defenseless against vindictive assaults. Some work has been coordinated to work on the adequacy of XAI calculations and safeguarding them from different assaults.

References

Abduljabbar, R., Dia, H., Liyanage, S., Bagloee, S.A.: Applications of artificial intelligence in transport: an overview. Sustainability **11**(1), 189 (2019)

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Aliramezani, M., Koch, C.R., Shahbakhti, M.: Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: a review and future directions. Prog. Energy Combust. Sci. **88**, 100967 (2022)

Anders, C.J., Neumann, D., Samek, W., Müller, K.R., Lapuschkin, S.: Software for dataset-wide XAI: from local explanations to global insights with Zennit, CoRelAy, and ViRelAy. arXiv preprint arXiv:2106.13200 (2021)

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)

Aseen, I.S., Kumar, C.A.: Intrusion detection model using fusion of chi-square feature selection and multi class SVM. J. King Saud Univ.-Comput. Inf. Sci. **29**(4), 462–472 (2017)

Balakrishnan, S., Venkatalakshmi, K., Arputharaj, K.: Intrusion detection system using feature selection and classification technique. Int. J. Comput. Sci. Appl. **3**(4), 145–151 (2014)

Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: dalex: responsible machine learning with interactive explainability and fairness in Python. arXiv preprint arXiv:2012.14406 (2020)

Barbado, A., Corcho, Ó., Benjamins, R.: Rule extraction in unsupervised anomaly detection for model explainability: application to OneClass SVM. Expert Syst. Appl. **189**, 116100 (2022)

Baur, T.: Cooperative and transparent machine learning for the context-sensitive analysis of social interactions (2018)

Bendovschi, A.C., Ionescu, B.Ş.: The gap between cloud computing technology and the audit and information security. Audit Financ. **13**(125) (2015)

Bonfanti, M.E.: Artificial intelligence and the offence-defence balance in cyber security. In: Cyber Security: Socio-Technological Uncertainty and Political Fragmentation, pp. 64–79. Routledge, London (2022)

Brito, L.C., Susto, G.A., Brito, J.N., Duarte, M.A.: An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. Mech. Syst. Signal Process. **163**, 108105 (2022)

Cai, D., Wang, W., Li, M.: Incorporating visual information in audio based self-supervised speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2022)

Castelvecchi, D.: Can we open the black box of AI? Nature **538**(7623), 20 (2016)

Chebrolu, S., Abraham, A., Omas, J.P.: Feature deduction and ensemble design of intrusion detection systems. Comput. Secur. **24**(4), 295–307 (2005)

Chennam, K.K., Uma Maheshwari, V., Aluvalu, R.: Maintaining IoT healthcare records using cloud storage. In: IoT and IoE Driven Smart Cities, pp. 215–233. Springer, Cham (2022)

Chipman, H.A., George, E.I., McCulloh, R.E.: Making sense of a forest of trees. In: Weisberg, S. (ed.) Proceedings of the 30th Symposium on the Interface, pp. 84–92. Interface Foundation of North America, Fairfax Station, VA (1998)

Chou, Y.L., Moreira, C., Bruza, P., Ouyang, C., Jorge, J.: Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inf. Fusion **81**, 59–83 (2022)

Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv preprint arXiv:2006.11371 (2020)

Deshpande, N.M., Gite, S.S., Aluvalu, R.: A brief bibliometric survey of leukemia detection by machine learning and deep learning approaches. Lib. Philo. Pract. 4569 (2020)

Dhanorkar, S., Wolf, C.T., Qian, K., Xu, A., Popa, L., Li, Y.: Who needs to know what, when?: broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle. In: Designing Interactive Systems Conference 2021, pp. 1591–1602 (2021)
Dieber, J., Kirrane, S.: Why model why? Assessing the strengths and limitations of LIME. arXiv preprint arXiv:2012.00093 (2020)
Farrahi, S.V., Ahmadzadeh, M.: KCMC: a hybrid learning approach for network intrusion detection using k-means clustering and multiple classifiers. Int. J. Comput. Appl. **124**(9) (2015)
Fidel, G., Bitton, R., Shabtai, A.: When explainability meets adversarial learning: detecting adversarial examples using SHAP signatures. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
Floreano, D., Wood, R.J.: Science, technology and the future of small autonomous drones. Nature **521**(7553), 460–466 (2015)
Fouladgar, N., Främling, K.: XAI-PT: a brief review of explainable artificial intelligence from practice to theory. arXiv preprint arXiv:2012.09636 (2020)
Främling, K., Westberg, M., Jullum, M., Madhikermi, M., Malhi, A.: Comparison of contextual importance and utility with LIME and Shapley values. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, pp. 39–54. Springer, Cham (2021)
Gazet, A.: Comparative analysis of various ransomware virii. J. Comput. Virol. **6**(1), 77–90 (2010)
Ghosh, I., Sanyal, M.K.: Introspecting predictability of market fear in Indian context during COVID-19 pandemic: an integrated approach of applied predictive modelling and explainable AI. Int. J. Inf. Manag. Data Insights **1**(2), 100039 (2021)
Gilpin, H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the 2018 IEEE 5th International Conference on Data Science and advanced Analytics (DSAA), pp. 80–89. IEEE, Turin, Italy (2018)
Guo, W.: Explainable artificial intelligence for 6G: improving trust between human and machine. IEEE Commun. Mag. **58**(6), 39–45 (2020)
Han, H., Liu, X.: The challenges of explainable AI in biomedical data science. BMC Bioinform. **22**(12), 1–3 (2022)
Hara, S., Hayashi, K.: Making tree ensembles interpretable. arXiv preprint arXiv:1606.05390 (2016)
Heide, N.F., Müller, E., Petereit, J., Heizmann, M.: X 3 SEG: model-agnostic explanations for the semantic segmentation of 3D point clouds with prototypes and criticism. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 3687–3691. IEEE (2021)
Hermansa, M., Kozielski, M., Michalak, M., Szczyrba, K., Wróbel, Ł, Sikora, M.: Sensor based predictive maintenance with reduction of false alarms—a case study in heavy industry. Sensors **22**(1), 226 (2022)
Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
Hussain, F., Hussain, R., Hossain, E.: Explainable artificial intelligence (XAI): an engineering perspective. arXiv preprint arXiv:2101.03613 (2021)
Ilgun, K., Kemmerer, R.A., Porras, P.A.: State transition analysis: a rule-based intrusion detection approach. IEEE Trans. Softw. Eng. **21**(3), 181–199 (1995). In: Proceedings of the IEEE Symposium on Security and Privacy (1999)
Jiang, R., Wang, L., Tsai, S.B.: An empirical study on digital media technology in film and television animation design. Math. Probl. Eng. **2022** (2022)
Kanaparthi, S.H., Swapna, M.: A statistical review on Covid-19 pandemic and outbreak. Lecture Notes in Networks and Systems vol. 301, pp. 124–135 (2022)
Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A.: Trustworthy artificial intelligence: a review. ACM Comput. Surv. (CSUR) **55**(2), 1–38 (2022)
Keane, M.T., Kenny, E.M., Delaney, E., Smyth, B.: If only we had better counterfactual explanations: five key deficits to rectify in the evaluation of counterfactual XAI techniques. arXiv preprint arXiv:2103.01035 (2021)

Klesel, P.H.M., Wittmann, H.F.: Explain it to me and I will use it: a proposal on the impact of explainable AI

Kłosok, M., Chlebus, M.: Towards Better Understanding of Complex Machine Learning Models Using Explainable Artificial Intelligence (XAI): Case of Credit Scoring Modelling. University of Warsaw, Faculty of Economic Sciences, Warsaw (2020)

Kose, N., Kopuklu, O., Unnervik, A., Rigoll, G.: Real-time driver state monitoring using a CNN based spatio-temporal approach. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 3236–3242. IEEE (2019)

Kotenko, I., Izrailov, K., Buinevich, M.: Static analysis of information systems for IoT cyber security: a survey of machine learning approaches. Sensors **22**(4), 1335 (2022)

Krishnan, R., Sivakumar, G., Bhattacharya, P.: Extracting decision trees from trained neural networks. Pattern Recogn. **32**, 12 (1999)

Kuppa, A., Le-Khac, N.A.: Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)

Kuzlu, M., Cali, U., Sharma, V., Güler, Ö.: Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access **8**, 187814–187823 (2020)

Lahre, M.K., Dhar, M.T., Suresh, D., Kashyap, K., Agrawal, P.: Analyze different approaches for ids using KDD 99 data set. Int. J. Recent Innov. Trends Comput. Commun. **1**(8), 645–651 (2013)

Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., Srivastava, J.: A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of the SIAM International Conference on Data Mining, pp. 25–36. SIAM, San Francisco, CA, USA (2003)

Lee, W., Stolfo, S.J., Chan, P.K., et al.: Real time data mining based intrusion detection. In: Proceedings of the DARPA Information Survivability Conference and Exposition II. DISCEX'01, pp. 89–100. IEEE, Anaheim, CA, USA (2001)

Li, J., Chen, J., Bai, H., Wang, H., Hao, S., Ding, Y., et al.: An overview of organs-on-chips based on deep learning. Research **2022** (2022)

Lin, I.C., Chang, C.C., Peng, C.H.: An anomaly-based IDS framework using centroid-based classification. Symmetry **14**(1), 105 (2022)

Logas, J., Schlesinger, A., Li, Z., Das, S.: Image DePO: towards gradual decentralization of online social networks using decentralized privacy overlays. In: Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1), pp. 1–28 (2022)

Lötsch, J., Kringel, D., Ultsch, A.: Explainable artificial intelligence (XAI) in biomedicine: making AI decisions trustworthy for physicians and patients. BioMedInformatics **2**(1), 1–17 (2022)

Naser, M.Z.: An engineer's guide to explainable artificial intelligence and interpretable machine learning: navigating causality, forced goodness, and the false perception of inference. Autom. Constr. **129**, 103821 (2021)

Novikov, D., Yampolskiy, R.V., Reznik, L.: Anomaly detection based intrusion detection. In: Proceedings of the International Conference on Information Technology: New Generations (ITNG'06), pp. 420–425. IEEE, Las Vegas, NV, USA (2006)

Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T., Al-Hashida, A.Y.: Intrusion detection model using machine learning algorithm on big data environment. J. Big Data **5**(1), 34 (2018)

Pasquale, F.: The Black Box Society: The Secret Algorithms that Control Money and Information. Harvard University Press (2015)

Pazzani, M.J., Mani, S., Shankle, W.R., et al.: Acceptance of rules generated by machine learning among medical experts. Methods Inf. Med. **40**(5), 380–385 (2001)

Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–568. ACM (2008)

Peng, K., Leung, V., Zheng, L., Wang, S., Huang, C., Lin, T.: Intrusion detection system based on decision tree over big data in fog environment. Wirel. Commun. Mob. Comput. **2018**, Article ID 4680867, 10 pages (2018)

Perarasi, T., Vidhya, S., Leeban Moses, M., Ramya, P.: Malicious vehicles identifying and trust management algorithm for enhance the security in 5G-VANET. In: Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore (2020a)

Perarasi, T., Vidhya, S., Leeban Moses, M., Ramya, P.: Malicious vehicles identifying and trust management algorithm for enhance the security in 5G-VANET. In: Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India (2020b)

Pienta, D., Tams, S., Atcher, J.: Can trust be trusted in cybersecurity? In: Proceedings of the 53rd Hawaii International Conference on System Sciences, Maui, HI, USA (2020)

Rajanikanth, A., et al.: Data security in cloud computing using ABE-based access control. In: Architectural Wireless Networks Solutions and Security Issues, pp. 47–61. Springer, Singapore (2021)

Raza, A., Tran, K.P., Koehl, L., Li, S.: Designing ECG monitoring healthcare system with federated transfer learning and explainable AI. Knowl.-Based Syst. **236**, 107763 (2022)

Roth, A.M., Liang, J., Manocha, D.: XAI-N: sensor-based robot navigation using expert policies and decision trees. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2053–2060. IEEE (2021)

Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)

Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M., Hartig, F.: Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography **44**(2), 199–205 (2021)

Saha, D., De, S.: Practical self-driving cars: survey of the state-of-the-art (2022)

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.A.: Towards a rigorous evaluation of XAI methods on time series. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 4197–4201. IEEE (2019)

Stampar, M., Fertalj, K.: Artificial intelligence in network intrusion detection. In: Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1318–1323. IEEE, Opatija, Croatia (2015)

Svenmarck, P., Luotsinen, L., Nilsson, M., Schubert, J.: Possibilities and challenges for artificial intelligence in military applications. In: Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting, Bordeaux, France (2018)

Swapna, M., Viswanadhula, U.M., Aluvalu, R., Vardharajan, V., Kotecha, K.: Bio-signals in medical applications and challenges using artificial intelligence. J. Sens. Actuator Netw. **11**(1), 17 (2022)

Swapna, M., Hegde, N.: A multifarious diagnosis of breast cancer using mammogram images—systematic review. In: IOP Conference Series: Materials Science and Engineering, vol. 1042, no. 1, p. 012012. IOP Publishing (2021)

Swapna, M., Uma Maheswari, V., Aluvalu, R., Vardharajan, V., Kotecha, K.: Bio-signals in medical applications and challenges using artificial intelligence. J. Sens. Actuator Netw. **11**(1), 17 (2022)

Toosi, A.N., Kahani, M.: A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers. Comput. Commun. **30**(10), 2201–2212 (2007)

Tseremoglou, I., Bombelli, A., Santos, B.F.: A combined forecasting and packing model for air cargo loading: a risk-averse framework. Transp. Res. Part E: Logist. Transp. Rev **158**, 102579 (2022)

Uma Maheswari, V., Aluvalu, R., Chennam, K.K.: Application of machine learning algorithms for facial expression analysis. Mach. Learn. Sustain. Dev. **9**, 77 (2021)

Urooj, U., Al-rimy, B.A.S., Zainal, A., Ghaleb, F.A., Rassam, M.A.: Ransomware detection using the dynamic analysis and machine learning: a survey and research directions. Appl. Sci. **12**(1), 172 (2022)

Vimalkumar, K., Radhika, N.: A big data framework for intrusion detection in smart grids using Apache spark. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 198–204. IEEE, Udupi, India (2017)

Visani, G., Bagli, E., Chesani, F.:. OptiLIME: optimized LIME explanations for diagnostic computer algorithms. arXiv preprint arXiv:2006.05714 (2020)
Ye, N., Zhang, Y., Borror, C.M.: Robustness of the Markov-chain model for cyber-attack detection. IEEE Trans. Reliab. **53**(1), 116–123 (2004)
Zaman, S., Karray, F.: Lightweight ids based on features selection and ids classification scheme. In: Proceedings of the International Conference on Computational Science and Engineering, pp. 365–370. IEEE, Vancouver, BC, Canada (2009)
Zhang, Z., Shen, H.: Application of online-training SVMS for real-time intrusion detection with different considerations. Comput. Commun. **28**(12), 1428–1442 (2005)
Zhang, Y., Weng, Y., Lund, J.: Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics **12**(2), 237 (2022)