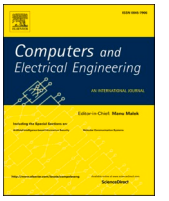
Computers and Electrical Engineering 119 (2024) 109643

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng



Security, privacy, and robustness for trustworthy AI systems: A review

Mozamel M. Saeed *, Mohammed Alsharidah

Department of computer science, Prince Sattam bin Abdulaziz University, Saudi Arabia

ARTICLE INFO

Keywords:
AI Systems
Privacy
Robustness
Security
Trustworthy

ABSTRACT

This review article provides a comprehensive exploration of the key pillars of trustworthy AI: security privacy and robustness. The article delved into security measures both traditional and cutting edge identifying emerging threats and challenges in ever ever-evolving landscape of artificial intelligence (AI) the discussion extends to advanced encryption techniques and imperative privacy preservation, emphasizing the ethical consideration inherent in safeguarding user data. The robustness and adversarial attack on AI, present techniques for the robustness model and ensure model interpretability and explainability through AI. The exploration of federated learning (FL) elucidates its conceptual foundations and intricate interplay between security, privacy, and collaborative model training. Differential privacy (DP) outlines insights into its application, and challenges. The ethical consideration section scrutinized bias and fairness in AI. The article concludes with an examination of emerging technologies in AI security and privacy anticipating challenges. This review article serves as a comprehensive guide to navigating the complex terrain of trustworthy AI.

1. Introduction

The increasing advancement in Artificial intelligence has allowed the formation of several systems depending on it. It gives crucial social and economic benefits to the community like application in entertainment, medicine, security, transportation, and finance [1]. However, several systems based on AI are found to be susceptible to undetectable attacks done by certain groups of people breaching individuals' privacy. Such violations can cause even loss of human life. These kinds of individual experiences can lessen trust in systems made of AI [2]. Practitioners who are working on AI such as developers, decision-makers, and researchers have reviewed the performance of the system i.e., accuracy to be the principal measure in the flow of work [1]. Developing an AI trustworthy system requires a comprehensive understanding if the model is unbiased or biased. For modern AI systems bias has been always a challenge. Several applications such as language translation to face recognition have depicted increased levels of system biases as exhibited in the performance of non-uniform among different test and group sets. This has a powerful implementation system's accountability and fairness, with highly noteworthy implications for society. Interpretability and explainability are needed for these systems in several conditions for instance: medical and enforcement of law, where decision-making of the black box is not supportable [3].

Despite the latest system of AI results in a remarkable rate of accuracy, they are not able to interpret their process of decision and failure causes and rate of success. Security and privacy are the way to the success of AI models with higher accuracy levels. Recent

* Corresponding autor.
E-mail addresses: m.musa@psau.edu.sa, mozamel8888@gmail.com (M.M. Saeed).

https://doi.org/10.1016/j.compeleceng.2024.109643
Received 8 February 2024; Received in revised form 12 July 2024; Accepted 29 August 2024
Available online 6 September 2024

studies have revealed that algorithms of AI can violate information taken out from social media to de-anonymize faces that are blurred and encourage spying which is unwanted with the help of CCTV cameras. These kinds of AI systems provide both opportunities and challenges. The monitoring system boosts society and individuals' safety, their vulnerability violations, and attacks also give chances for abuse. Specifically, adversarial attacks have had a negative impact on users' perception that those systems related to AI can be manipulated easily [3]. AI can disclose individuals' business secrets and privacy. The hackers can benefit from vector features developed by AI models to rebuild private input information like fingerprints, as a result, disclosing users' information which is sensitive. The attack on systems built on AI can cause serious security and economic outcomes [4]. In the current era, as the demand for developing trustworthy AI is increasing, it is essential to summarize the achievements of AI and examine its future direction for research [3]. The AI systems need to be robust to input perturbation and have the potential to make an accurate decision. Several research have observed that AI systems particularly those systems that opt for deep learning models might be sensitive to input perturbations that are unintentional and intentional, constituting a high risk to the security of applications that are critical in nature. For instance, models of spam detection can be tricked by well-designed emails [4]. Hence, the sender of spam emails can take benefits which can lead to a bad experience for the user.

Recently, a literature review on trustworthy AI has become apparent. As the demand for developing trustworthy AI is increasing, it is vitally important to summarize the latest achievements and explain the future direction of the research. In this review, we give a wide-ranging overview of trustworthy AI to assist individuals who are new to obtain the basic knowledge of what factors make AI systems Trustworthy and aid them in tracing recent programs in this field.

This review article initiates many possible interactions between various dimensions to explain other significant challenges around the AI system that have not drawn adequate attention. Previous review research was mainly based on trustworthy AI systems, but there is no paper that discusses the security, privacy, and robustness of AI with techniques by which we can make AI more trustworthy, further, we have added validation and verification methods and future challenges and applications of AI. It also focuses on particular solutions based on computation to gain knowledge of every dimension of AI. Other articles such as [5-8] focus on the requirement of trustworthy AI and why it is necessary. The present review study aims to elucidate the vital role of security, privacy, and robustness in establishing trustworthy AI. In Section 2 recent work by various researchers has been discussed, section 3 discusses how trustworthy AI is made with the help of security, privacy and robustness, section 4 discusses the security in AI, section 5 discusses Privacy in AI system, section 6 discusses robustness of AI systems, section 7 discusses techniques for enhancing AI trustworthiness, section 8 discusses validation and verification, section 9 discusses Applications of AI, section 10 discuses ethical considerations, section 11 discusses Future Challenges, section 12 discusses future applications. The Future Studies and conclusion are discussed in Sections 13 and 14, respectively.

1.1. Research questions

Q1. What are the key challenges and potential solutions in implementing and maintaining a balance between achieving optimal AI performance and preserving user privacy, particularly when employing advanced techniques for security and robustness in AI systems?

Q2. How does the integration of federated learning, differential privacy, and homomorphic contribute to enhancing the security, privacy, and robustness of AI systems, ensuring their trustworthiness in handling sensitive data?

2. Recent work

Intelligent Transportation Systems (ITS) have modernized the system of transportation by adding high-level technologies for safe and efficient mobility. Nonetheless, complex challenges are faced by these systems while assuring resilience and security against adversarial attacks. A recent study examines these challenges by initializing the Dampster–Shafer data fusion-based Adversarial Deep Learning (DS-ADL) Model. The model majorly focuses on decision level, feature level, and original image level. To improve the capabilities of systems, the study utilized Dempster–Shafer-based Multimodal Sensor Fusion allowing information fusion from various sensors for enhanced understanding of the scene. As a result, the system improved the abilities of decision-making and perception. The ability of the model to defend and detect adversarial while managing low latency fog-cloud operations underscores its significance for the real world in ITS [9].

Chinmay Chakraborty et al. [10], developed a detection system for cyber-attacks in the healthcare industry with the help of federated and centralized transfer learning mode. Further, an Edge of Things (EoT) structure was produced, linked with the health sector and cloud to transfer the data effectively, and the developed Centralized with Multi-Source Transfer Learning (CMTL) algorithm which was utilized for classification and detection of several threats like Malware attack, DoS attack, <am in the Middle attack and Injection attack. The developed framework performance was examined utilizing several datasets like Federated TON_IoT, X-IIoTID, and EMNIST. The study produced high accuracy and execution time as compared to other algorithms.

Another research proposes a detection system for anomaly by incorporating a deep learning technique known as Convolutional Neural Network (CNN) with Kalman Filter (KF) based Gaussian-Mixture Model (GMM). The developed model was utilized for detecting anomalous behavior in CPSs. The developed system consists of two major processes. The first step is the data pre-processing by filtering and transforming original data into a novel format and gaining data privacy preservation. In the second step, GMM-KF was added to the CNN model for estimation of accuracy and detection of anomaly [11].

A study proposed an IoT-based healthcare cyber-physical system that gives efficient utilization of resources at cloud and fog levels with less cost execution. Moreover, the study also extracts data from networking sites social media, and reviews of drugs for analysis. In this study, two different types of methods were utilized for the collection of data. The developed effective utilization of resources and

budget-friendly scheduling of tasks at the fog level and multi-objective heuristic approach Ant colony optimization task scheduling (MOHACO-TS) at cloud level. Both algorithms concentrated on accomplishing a maximum number of tasks with minimum time. The outcome of the study revealed that the proposed mechanism IoT-HCPS outperformed existing algorithms and techniques [12].

Senthil et al. [13], developed an evidence-based detection system for detecting selfish nodes. In this study, the security concern is managed by the trust authority (TA), which can find the node that is replicated, and executed by an adversary. An enhanced self-centered friendship (ISCF) tree was framed and a replica was assigned to the node. This model outperformed the existing model in regards to the detection of selfishness accuracy and time.

Ganesh et al. [14], represented a Public Key Encryption with an Equality Test based on Data Loss Prevention (DLP) with problems of double decomposition over near-ring. The Computation Diffie-Hellman was used in the structure of algebraic which requires DLP with a Double Decomposition problem for developing a Public Key Encryption with an Equality Test that gives the system a high level of security. The developed technique was secure and it resolves the quantum algorithm problem attacks in systems of IIoT. The search time of the developed technique was 150 ms.

Aravinthkumar et al. [15], present a novel scheme for VM selection optimization utilizing the SI method known as Analogous Particle swarm optimization (APSO). Moreover, optimal VM was utilized for detecting the disease of the kidney. In the study, a neural network (NN) was utilized as an automatic technique for kidney disease diagnosis. Different types of comparisons and experiments were executed to examine the developed system. The outcomes revealed that the APSO model worked well, the developed model enhanced the efficiency of the system by 5.6%. The predicting kidney disease precision employing the neural network was 95.7%.

Senthil et al. [16], employed algorithms of cryptographic as an effective control access mechanism for a system of Internet of Medical Things-based health care. The Rivest Cipher (RC6) algorithms were used to produce the key value, and the algorithm of elliptic curve digital signature encrypted the key value with the help of RC6, and the encrypted output was deployed to secure the hash algorithm (SHA256) for enhancing the integrity of data. The outcomes indicated that the developed system is highly robust against several known attacks, like DoS, sensor attacks, and router attacks.

Table 1 shows a recent paper based on security and privacy measures for AI. It discusses how different approaches were applied by the researcher for the preservation privacy and security of that data.

3. Building trustworthiness

The trustworthiness of a system is built on the foundation of security, privacy, and robustness, each plays a crucial role in ensuring users can depend on the system for its intended purpose. Here is how it is established through these key pillars.

3.1. Security

3.1.1. Protection against threats

A trustworthy system must be secure, guarding against various threats such as cyber-attacks, unauthorized access, and data breaches. Implementing strong authentication encryption and access control mechanisms builds a robust security posture [31].

3.1.2. Vulnerability mitigations

Regular assessments including penetration testing and vulnerability assessments, help identify and address potential weaknesses. Timely patching and updates are essential to stay ahead of evolving security threats [32].

Table 1
Recent Approaches.

Parameter	Approaches	Reference
Remote Diagnosis with Privacy-Preservation	Interoperable IoMT Approach	Subramaniam et al., [17]
AI-based healthcare system security	Quantum photonic convolutional neural network for artificial intelligence-based healthcare system security	Sita et al., [18]
Cyber-physical anomaly detection for inverter-based microgrid	Autoencoder neural network	Tabassum et al., [19]
Anomaly detection in cloud environments	Hybrid AI approach	Prakash et al., [20]
Malicious and selfish nodes in MANETs	Hybrid Bayesian and modified grey PROMETHEE	Suresh et al., [21]
Detection System for Secure MANETs	Blockchain-Enabled Lightweight Intrusion	Ilakkiya et al., [22]
Consensus algorithm in IoT systems	Lightweight scalable blockchain	Fathi et al., [23]
Data Security-Based Routing in MANETs	Protected route system	Hande et al., [24]
Internet of Medical Things	ECMQV-MAC authentication protocol and EKMC-SCP blockchain networking	Lin et al., [25]
Classification of malicious users in radio networks	Modified light GBM	Sekar et al., [26]
Malicious Attacks in Web 3.0	Federated Learning	Yuan et al., [27]
AI for VHetNets	VHetNets for AI and	Wang et al., [28]
Anomaly Detection in Cloud Computing	Knowledge Graph Embedding and Machine Learning Mechanisms	Mitropoulou et al., [29]
intrusion detection in IoT networks	XAI	Sharma et al., [30]

3.1.3. Incident response

A robust incident response plan ensures that, in the event of security incidents, the system can quickly and effectively contain and mitigate the impact, minimizing disruptions and protecting user data [32].

3.2. Privacy

3.2.1. Data protection measures

Building trust requires a commitment to protecting user data. Implementing privacy-preserving technologies such as encryption anonymization and access control ensures that sensitive information is kept confidential [33].

3.2.2. Transparent data practices

Clearly communicating how user data is collected, processed, and stored faster trust. Providing the user with control over their data including opt-in and opt-out mechanisms contributes to a transparent and respectful approach to privacy [33].

3.3. Robustness

3.3.1. Reliability and performance

A trustworthy system must be reliable and performant. This involves rigorous testing under various conditions to ensure the system can handle both expected and unexpected scenarios without compromising its functionality [34].

3.3.2. Resilience against adversarial attempts

Robustness includes resilience against adversarial attempts to manipulate or compromise the system. This involves measures such as adversarial training in machine learning models and ensuring that the system can recover gracefully from disruptions [34].

4. Security in AI

Data can be utilized according to the needs of the business; hence cyber-attacks can cause a greater challenge. Generally, cyber-attacks are malicious and try to violate an organization or an individual's data attempted by another organization or individual. SQL injection attack, phishing, denial of service (DoS), ransomware, Zero-day exploit and malware attack are frequent these days [35]. The incidences of these types of cybercrime can impact individuals or an organization, as this can lead to disruption and destructive bankruptcy. For example, an IBM report states that a violation of data can cost around 8 million dollars for the US [36], and the calculated yearly cost to the economy worldwide for cybercrime is about 400 billion dollars [37]. The number of cyber-crimes is increasing exponentially which is an alarming situation for researchers and professionals of cybersecurity [35].

4.1. Traditional security measures in AI

4.1.1. Access control

It is a procedure of security managing the utilization or access of resources such as data, files, and networks of computers. For instance, depending on the users' responsibilities, access based on a role control scheme can be employed to restrict access to the network, lowering the risk to the organization [38].

4.1.2. Firewall

It is a framework for a security network that controls and traces outgoing and incoming traffic of the network. It is a system based on a host or network that depends on security rules set to block or accept the traffic. It has the capability of traffic filtering from unauthorized sources to prevent attacks [39].

4.1.3. Anti-malware

It is frequently utilized to detect, prevent, and remove viruses from the computer. The new software for antivirus can secure users from several malware attacks like spyware, ransomware, trojan horses, etc [40].

4.1.4. Sandbox

It is used for preventing failures of software or systems. it is normally employed to run the coding of a program from untrusted suppliers, websites, users, or third parties [41].

4.1.5. Data encryption

To secure data and information from being changed and stolen, encryption is utilized by turning data into secret code that can be unlocked by a special digital key. Data which is encrypted can be secure and transferred between two computers [42].

4.2. Emerging threats and challenges

4.2.1. Adversarial attack

An adversarial attack has the potential to manipulate the model of machine learning (ML) at the testing or treating phase to mislead the AI model leading to incorrect output. Machine learning susceptibility to adversarial attacks is a major risk these days. The evolving of attack strategies and the potential for transferability across the different models make countering adversarial attacks an ongoing challenge [43].

4.2.2. Data poisoning attacks

Data poisoning attacks are one of the frequent attacks against models of machine learning. It can be done on the federated learning model by exposing data to adversary poisoning in a few devices that are participating in the test and learning process so the accuracy of the global model is made vulnerable. These can be done either by injecting poison from other devices or the data is injected with poison. These kinds of attacks may be non-targeted or targeted. As AI system are becoming more data-dependent the risk of data poisoning attack affecting models' integrity are increasing highly [44].

4.2.3. Evasion attack

It occurs in the test phase. The objective of the attacker is to design perturbations for sample testing to obtain inaccurate predictions from the model of the victim [45].

4.2.4. White-box attacks

In the white box attack, the hacker can use all data of the model such as information on gradient, parameters, and its architecture. The process of attack can be defined as a problem for optimization [46]. In recent years, the white box has been studied broadly as it can disclose the systems parameters and the architecture that assist people to know the machine learning model weakness more clearly and they can examine mathematically [4].

4.2.5. Black-box attacks

In these attacks, no knowledge about the machine learning model is exposed to the enemy. They can only add input data and the model's output query. The easiest way to perform a Blackbox attack is to keep inquiring about the user models and gradient approximation with the help of the numerical differentiation method [4].

5. Privacy in AI system

Decision-making systems based on AI examine a large number of datasets to make a conclusion, where performance and accuracy depend on the utilized data size for training. But the usage and availability of data at a considerate level can also have conflicting effects. Organizations such as private, government, or hackers can exploit the data of users. Datasets for AI training frequently include information about users that is sensitive, increasing the chances of a breach of privacy. It includes raw information access, inaccurate trained model inference, or access to the model that is unaccepted [33]. These demonstrate that it is essential to preserve the privacy of the user's data to prevent dangerous outputs and raise the trust of the user in the AI system. However, if an individual sees that AI-based systems are employing proper steps to preserve the privacy of individuals' data and identity, as a result, they will trust more in AI systems more. As the availability of data is increasing, the rate of data breaches has also increased. These types of incidents have exposed billions of data which contain sensitive information of users and caused significant abuse of the sensitive information. Moreover, these types of targeted and internal attackers can bring systems' efficiency down and hence lead to a decreased level of trust of the users in AI systems [47]. During the phase of collection of data, threats of privacy can occur as a result of collected data and its issue of shortage. The privacy of data can be at risk in the phase of pre-processing and modeling, only if the AI can identify again information that is sensitive from the data which is not sensitive [48]. To claim trust in the efficiency of machine learning (ML) systems for training, evidence is needed for access to the data, its protection, and its usage. AI have developed several mechanism and methods to tackle these issues under the of "privacy-preserving machine learning" (PPML) [49].

Various methods have been developed to establish the privacy of the user. One of the methods is de-identification which deletes personal identifier and their links from the data. There are 2 types of identifiers i.e., indirect and direct. Indirect identifiers can detect the identity of individuals when they are associated with other information and direct identifiers, associate directly with individuals' identity [50]. Khalil and Ebner designed a technique of masking that masks the sensitive characteristics and values. These kinds of methods are called techniques of suppression that raise the privacy of data but lessen the quality and utility of the data [51]. Few researchers developed federated learning to give privacy to the data of individuals utilizing collaborative learning where un-processed data is not sent to various devices. The model is passed after it is trained [52].

PPML's goal is to secure the information or model's privacy utilized in ML at evaluation and training during the execution. PPML has advantages for users' models, and for those individuals who trained or produced the data. PPML is highly motivated by privacy communities and cryptography and it is executed in practice utilizing a mixture of different techniques [49].

These types of techniques are strong for developing trust between the user of the model and the owners' data by making sure key information of private. They might be not very useful in every context; however, several techniques might be needed to protect individual data and information from malicious attacks [33].

6. Robustness of AI systems

AI system robustness has been largely researched and they are mainly concentrated on the Robustness of attacks particularly to interrupt systems which is widely known as adversarial machine learning (AML) [32].

On the research of AI system robustness to attacks and perturbations that might happen during the process and must thus be simple to deal with attacks, outcomes are more insufficient. Few researchers cope with the robustness of a perturbation that is natural and developed measures for raising it [53-55], however, they didn't perform a better systematic assessment of this.

The testing of the robustness of the AI system is done on a neural network based on datasets of test and specification of a few properties. A metric is utilized to carry scores of robustness. Aside from detecting frequent failure model modes, the outcomes can be utilized to understand how to improve those sets of data to comprehend further properties of custom, which can enhance both the process of testing and the neural network [32].

6.1. Techniques for adversarial robustness

6.1.1. Adversarial training

In the ongoing quest for adversarial robustness, a multifaced approach is adopted. Adversarial training stands out as a prominent strategy involving the integration of adversarial examples into the training. This process compels the model to adapt and learn to resist perturbation, thereby enhancing its resilience [56].

6.1.2. Regularization method

It penalizes complex models' input preprocessing techniques to cleanse data and the development of inherently robust architecture contributes significantly to fortifying all systems against adversarial manipulation [56].

6.1.3. Certified robustness

In the context of machine learning, it refers to the assurance that a model is resilient against adversarial attacks within the mathematically guaranteed margins. Certification involves establishing a mathematical proof that the model will maintain in a certain level of performance or accuracy even when subjected to carefully crafted adversarial inputs [56].

6.2. Explainable AI (XAI) and model interpretability

The attention to XAI has been given across various domains of application [57]. As a result, there is an increased number of tools of XAI and researchers are proposing new techniques to both academia and industry. The present system of XAI gives various sets of functions and dimensions for the analysis of data to comprehend difficult models of AI. Even though the pipeline of machine learning can give correct predictions, still it is insufficient in 2 crucial phases i.e., understanding and explaining. The phase of understanding includes assurance of quality and training of an AI model while the phase of explaining is essential when the model of machine learning is utilized in applications of the real world [58].

6.2.1. Understanding phase

6.2.1.1. Enhancing and debugging models of AI. The model of AI goes through iteration prior it is established fully [59], for instance, to enhance the model performance gradually. Throughout this price, the model source error is discovered and eliminated after carefully cross-checking and testing. However, explanation can decrease the steps of this process by assisting with sources of error in model recognition.

6.2.1.2. Bias detecting. The process of decision-making can be partially or fully initialized with the help of models of AI [59]. Nonetheless, if the models are trained on historical data that is biased, then this bias impacts the whole system negatively and thus makes false results of decisions. XAI is a significant tool that assists in detecting biases in models of AI.

6.2.1.3. Understanding scientifically. The Automated Statistician project [60] demonstrates its prediction by splitting down difficult datasets into sections of communicable and interpretable results to the individuals. This allows researchers to improve data features understating [59].

6.2.1.4. Robust model building. Models that are unlikely to be influenced by few alterations in the input are known as robust models and those that are able to explain are known as more robust models. This is instinctive which is having logical explanations for predictions which is unlikely to be affected by the noise [59,61].

6.2.1.5. AutoML. It has made explainability more essential, as the whole pipeline of the data is changed into a black box [62]. When utilizing AutoML, an individual does not have the capability for feature selection or a model's process for decision-making visibility.

6.2.2. Explaining phase

6.2.2.1. Better decision making. The model of AI can predict accurately whether a customer is possibly to leave in the future or not, which alerts the business, but it presents no solution for it. With the help of XAI, it can give more insights into the process of decision-making and also answer the "Why" part. The business with this knowledge can make improved goals and plans for the future. The potential of AI models for future prediction has attracted several users. Despite their utilization extensively, the result of the models is complex to defend. For instance, several of the parameters have been fine-tuned in the model of AI and machines that can initiate a self-driving car [63]. Several times it is a challenge for a user to understand the logic of ML/AI systems during security violations. Moreover, decision-making in the environment of Blackbox is difficult. The systems of AI few times give incorrect outcomes in the false positive form which misinterpret the expert of security jeopardizing the whole system's integrity [63]. Cybersecurity is one of the sectors in which the utilization of XAI can provide benefits because the result of the model will become more explainable. The traditional techniques of artificial intelligence and machine learning have no capability level as compared to the XAI for detecting automatic threats, evolving threats, and efficient and swift response-ability. It gives professionals for security high-quality level significant reasoning and identification of threats for the purpose of decreasing breaches of security and privacy of users and enhancing productivity overall. XAI also assists in the discovery of future risk, it represents an excellent solution when there is a need for accountability, interpretability, and explainability [63].

XAI serves as a critical dimension in the development and deployment of AI systems, aiming to make their decision-making process more transparent and interpretable. Interpretability techniques play a pivotal role in demystifying the intricate working of complex models. These techniques encompass various methodologies such as feature importance analysis, attention mechanism that highlights relevant portions of input data, and the generation of human-readable explanations for model decisions [64]. Couteaux et al. [65], developed XAI based DeepDream approach where the neuron activation was maximized by doing the image's gradient ascent. It has output curves that depict the feature's evolution during the maximization. It benefits the interpretability and visualization of neural networks and was utilized segmentation of tumors from CT scans of the liver. The system of criminal justice is another example of XAI. In Countries such as the US, algorithms based on automation are being utilized to track the location of the crime that is most likely to happen, who will commit a crime, and to appear in court will fail [66]. Soares et al. [67] developed an approach for the detection of COVID-19 with the help of a CT scan. It was observed that this approach crosses the other DL approaches like VGG-16, Google Net, and Res-Net with regard to precision, accuracy, and performance. CT scan was presented in which radiologists can identify COVID-19. Although systems based on intelligence have greater opportunities, the XAI popularity increases as assists in comprehending particular decisions made by the AI. It also motivates individuals to develop human-like solutions with a crucial understanding of the processing of the brain's natural information. Fig. 1 shows the factors that contribute to trustworthy AI.

7. Techniques for enhancing AI trustworthiness

Integrating federated learning, differential privacy and homomorphic encryption enhances AI security, privacy, and robustness.

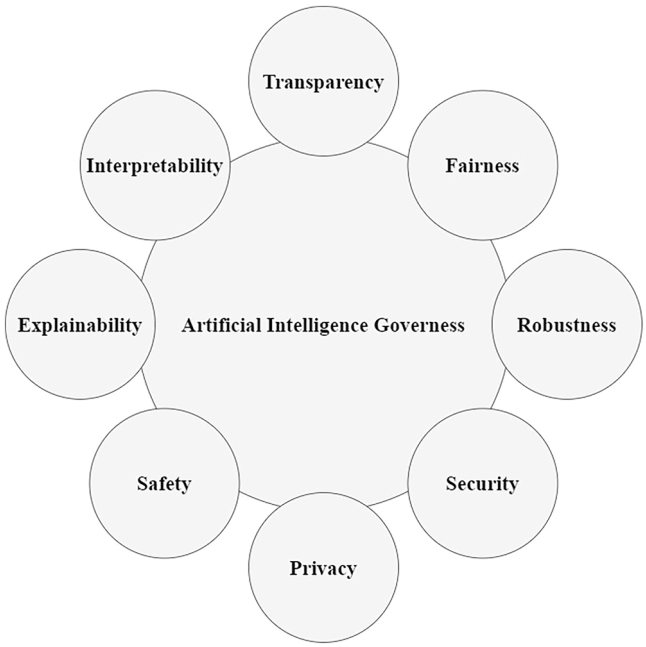


Fig. 1. Governance of AI [68].

Federated learning enables training without compromising privacy. Differential privacy safeguards individual data contribution and homomorphic encryption ensures the confidential processing of sensitive od sensitive data. Together these techniques fortify AI systems for responsible and privacy-preserving advancement.

7.1. Federated learning

Federated learning in AI enables model training across the decentralized devices preserving privacy to keep data localized. This collaborative approach is a more inclusive and secure deployment of the AI model without compromising it sensitive information. It is the method that allows machine learning training models on user's data with their devices, Fig. 2 illustrates the working of FL on various devices. It gives a solution to enhanced machine learning systems with trustworthiness, data of sovereignty, and finer alignment. It rapidly gained attention in industry and academics to do training model tasks by utilizing several parties. The objective was to address concerns about the privacy of data that obstacle algorithms of ML from properly utilizing several sources of data. The FL has applied to many domains like finance and healthcare [69].

7.1.1. Types of federated learning

7.1.1.1. Federated transfer learning (FTL). The Federated transfer learning is utilized when the included 2 data sets are not only dissimilar in size of sample but also in space feature. It might be employed to develop solutions for features and full dataset problems. Particularly, two features are understood by using restricted sets of samples and utilized for predicting results for samples for only one feature [70].

$$X_i \neq X_i, Y_i \neq Y_i, I_i \neq I_j \forall D_i, D_j, i \neq j$$

7.1.1.2. Vertical federated learning (VFL). The techniques of machine learning for data that is vertically partitioned have been recommended to preserve privacy, by adding cooperative analysis of statistics [71], rule mining association [72], secure linear regression [73-75], classification [76], and gradient descent [77]. Researchers [78,79] have represented VFL techniques for training models of logistic regression that secure the privacy of an individual.

7.1.1.3. Horizontal federated learning (HFL). Those datasets that differ in sample number but share a feature space are known as horizontal federated learning. A server that is honest-but-curious is generally considered in HFL systems [80-82]. Only the server has the capability to enter data participants' privacy where it hosts.

7.1.2. Privacy-preserving approaches

7.1.2.1. Perturbation. The perturbation is added in primary data, so the data that is perturbed are identical statically from the primary data. The schemes broadly opted for our global and local "Differential Privacy". After the local model is trained global approach adds noise to the framework, on the other hand, in every user's data, noise is added by the local one utilized for training [83].

7.1.2.2. Encryption. Before sharing the data of the user each parameter of the model is encrypted. The most popular opted scheme is Homomorphic Encryption, in which the server does not decrypt the model's parameters to collect them into the global model. Secure Multiparty Computation (SMC) is one more scheme that enables users to estimate the function of an objective without disclosing their

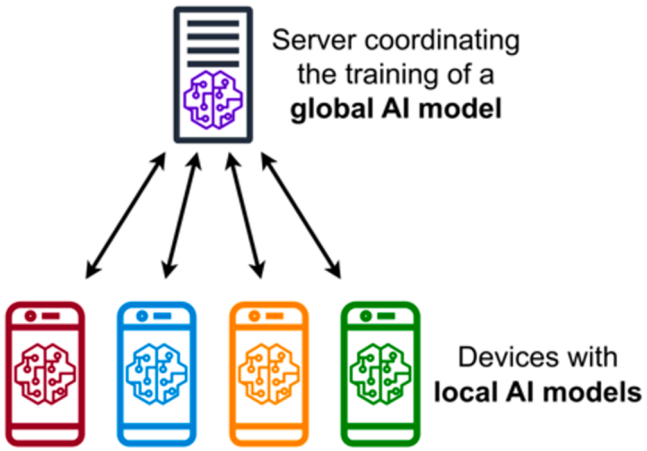


Fig. 2. Working of Federated Learning [69].

information [84].

7.1.2.3. Anonymization. It includes l-diversity and k-anonymity schemes. The L-diversity spread on k-anonymity so that attributes that are sensitive to the sample can be secure. If every sample from the sets of data cannot be redetected from the given data its condition is satisfied [85].

7.1.3. Application

These days, the power of computing of smartphones and storage capacity is growing rapidly. The following technology has been labeled mobile edge computing (MEC), however, leakage of personal data risk has also increased [85]. It can be solved by combining MEC and FL. Wang, X. et al., [86] examine the In-Edge AI' architecture which unites MEC and FL and it maximizes the allotment issues of resources. Moreover, Qian et al., [87] produced a service of privacy awareness placement to give service of high quality with the help of combining MEC and FL.

Furthermore, FL has a great opportunity in the healthcare sector. Pfohl et al., [88] investigate differential privacy for Electronic Medical Records (EMR) in federated settings. Huanga et al. [89] utilized EMRs in hospitals to detect heart disease patient mortality ratio. In the process of training, there was no transmission of parameters of data between the databases of hospitals. Lately, FL has been broadly utilized in the examination of biomedical images. Silva et al. [90] utilized Federated principal components analysis (fPCA) to take out MRI features from various hospitals. When FL is utilized in sales, finance, and various industrial sectors, where data cannot be compiled for training models of machine learning because these factors as protection of privacy and security of data, including the right of intellectual properties, it has several innovation mechanisms for modeling [91-93]. Gathering the records of medical from various institutions is a big challenge due to the requirement of maintaining the confidentiality of patients. Due to this, the data of medical is becoming very rare. The medical resources distribution and sickness diagnosis have undergone an exceptional transformation as a consequence of the development of artificial intelligence. Even though the collection and processing of data results in security problems such as private patient information exposure [94]. The researcher proposed methods to employ the data of patients without breaching their privacy. The data regarding medical have 2 challenges: a lack of inadequate labeling and data, these two can solved by employing federated transfer learning [95]. APOLLO network which was based on multi-federated learning was developed by researchers utilizing interlinked systems of healthcare to gather health outcome data and longitudinal real data and to assist physicians in future disease predictions of patients [96].

7.1.4. Challenges

7.1.4.1. Expensive communication. The first challenge faced in FL is expensive communication. Although networks of federated learning consist of a higher number of devices it can slower the network communication than the computation by the local magnitude of several orders. It is required to build an efficient communication method that repeatedly sends messages that are small as part of the operation of training. To lessen the communication setting: 2 key factors are: to lessening the rounds of communications in total, and lessen the transmitted message size [97].

7.1.4.2. Systems heterogeneity. Each device's communication capabilities, computation, and storage in federated learning may vary as a result of hardware variability (memory, CPU), power (level of battery), and connectivity of the network (Wi-Fi, 5 G, 4 G, 3 G). Moreover, the size of the network and constraints linked with the system on every device result in a fraction of very small of the devices being active immediately, for instance, 100 devices in active in millions of networks of devices. Every device might be unreliable and is not frequent for a device that is active to drop from the iteration as a result of constraints of energy or connectivity [97,98].

7.1.4.3. Statistical heterogeneity. Frequently devices collect and produce data in a diverse in distributed manner among the networks,

Table 2
Different AI methods.

Parameters	Adversarial learning	Federated learning	Explainable learning
Methods Overview	Models trained to resist adversarial attacks [102].	Decentralized model training across devices [69].	Models design for transparent explanation [64].
Potential Attackers	Adversaries inject subtle changes [102].	Threats to data privacy during collaborations [69].	Those manipulating model explanations [64].
Intrusion Detection	It utilizes adversarial training to detect and mitigate adversarial attacks and anomaly detection [102].	Collaborative learning improves detection [69].	Emphasizes clear explanations for anomalies [64].
Privacy Preservations	Addresses privacy concerns in model training.	Maintains data privacy in distributed settings.	Balances transparency with individual privacy.
Robustness	Focuses on robust model performance.	Improves robustness through collaborations.	Strives for robustness and interpretability.
Challenges	Vulnerable to sophisticated attackers.	Coordination challenges in decentralized setups.	Balancing transparency without sacrificing accuracy.
Research Focus Areas	Improving adversarial robustness.	Enhancing communication efficiency to collaboration.	Developing methods for more interpretable models.

for instance, users of smartphones have several utilizations of language regarding the prediction of the next word. Furthermore, data point numbers among devices might differ remarkably, and there might be primary structures present that take the association including devices and their linked distributions. The paradigm of development of data breaches commonly utilized independent and identically distributed (I.I.D.) presumption in optimized distribution, raised the chances of stragglers and might include complexity in evaluation, modeling, and analysis [97].

7.1.4.4. Privacy concerns. The concern for privacy is the most crucial in terms of applications of federated learning. The FL takes action toward data protection produced on every device by sharing updates of models. For instance, a part of the raw data and information of gradient. Hence, model update communication in the process of training can reveal information that is sensitive to the central server or the third party. Moreover, the latest method targets to improve FL privacy utilizing tools like differential privacy and secure military computation, these kinds of approaches sometimes give privacy at a reduced cost of the performance of the model or the efficiency of the system. Balancing and understanding both empirically and theoretically is a big challenge for learning federated learning [99-101].

Table 2 shows a difference in adversarial, federated, and explainable learning based on parameters such as method, attacker, detection, challenges, and future research.

7.2. Differential privacy

Differential privacy is a foundational concept in privacy-preserving data analysis, providing a rigorous framework for balancing the need for valuable insights from data with the imperative to protect from data with the imperative to protect individuals' privacy shown in Fig. 3. Development to address the challenges associated with sharing and analyzing sensitive information. The core principle is to enable the extraction of insights from datasets while simultaneously protecting the privacy of the individual's data points [103].

In the recent years, AI has attracted many researchers. Nonetheless, with all its advancement in every sector, issues have also been obtained regarding breaches of privacy, fairness of the model, and issues related to security. Differential privacy is a strong mathematical model that has numerous properties that can solve these issues, hence making it one of the most valuable tools. And this is the reason DP has been applied widely in AI systems.

The main purpose of DP is to provide privacy to the user with the help of hiding the data and information of individuals [105].

The differential privacy targets the differences among the results to query *f* between the datasets neighboring to privacy-preserving. The differential privacy gives a mechanism *M*, which is an algorithm of randomization that retrieves the database and executes a few functionalities [106].

An algorithm of randomization *M* gives ε-differential privacy for any neighboring pair datasets *D* and D`, and, for every set of result Ω, if *M* satisfies:

$$PrPr\ [M(D)\ \varepsilon\ '\Omega]\ \leq expexp\ (\varepsilon)\ \cdot PrPr\ [M(D')\varepsilon\ '\Omega]\ + \delta,$$

where $'\Omega$ indicates the output range of the algorithm *M*

The variable is explained as the budget of privacy, which manages the guarantee level of privacy mechanism *M*. A smaller ε shows stronger privacy. If $\delta = 0$, the randomized mechanism *M* provides ε-differential privacy [105].

7.2.1. Privacy

The mechanism of DP ensures that the outcome probability from the algorithm is consistent by updating users' records in the data training.

7.2.2. Security

With the help of the DP mechanism, it can lessen the influence of malicious attacks on tasks related to AI. These can warranty

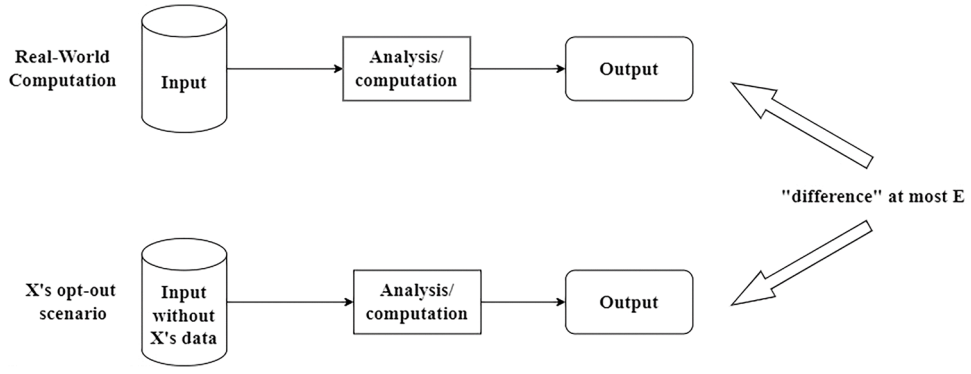


Fig. 3. Framework of differential privacy [104].

security in AI systems.

7.2.3. Fairness

Differential Privacy can assist in sustaining fairness in a machine-learning model by the repeated sampling of the training data. An algorithm is said to be fair when the outcomes are independent of attributes like gender and race.

The concerns for privacy are increasing nowadays around the globe for text, images, videos, and audio. Differential privacy has gained attention to secure and protect a broad range of datasets. However, still, it is still much unexplored in the field of AI, providing potential for privacy.

7.2.4. DP methods for image content

Data of images, particularly text, and faces, can show sensitive information about the individual. When the data is shared with third parties, the issues related to its primacy should be considered. Blurring, blacking, and pixelization are some of the most common technologies in obscuring the region of interest (ROIs) in pictures. For example, the algorithm known as boc blurring was developed in the privacy-protecting Google Street view to protect the license plates of vehicles and the faces of humans. The emergence of DP as modernized technology to solve this issue. It has accommodative adoptions in explaining the privacy of data of images [107,108].

7.2.5. DP methods for audio content

Amazon Echo and Apple Home Pod. Secondly, voice prints that are personal as a type of distinctive biometric detector carry in audio data, same as fingerprints are highlighted as strong indicators of privacy in the GDPR [109]. Moreover, voice play plays a major role in the system of voice authentication, and if it is disclosed various privacy issues will occur. To solve these concerns of privacy, a few solutions have been developed. Anonymization is frequently a method that is utilized. The other methods can although give privacy protection practically, however, they cannot make clear the amount of privacy that is protected from the attacks. The utilization of differential techniques can be successful in achieving this objective [110].

7.3. Homomorphic encryption

The homomorphic encryption development is one of the high levels of advance, enhancing the computations scope that can be applied to generate the encrypted data homomorphically. Homomorphic encryption serves a vital role in ensuring trustworthy AI by enabling computation on encrypted data. This privacy-preserving technique allows data to remain encrypted even during the process shown in Fig. 4. In the context of AI, it contributes to building a secure and transparent AI system by mitigating security and privacy risks. The main aim of homomorphic encryption is to enable the calculation of the data that is encrypted [111]. However, the information can remain private when it is processed, allowing the needful task to be completed with residing data in such an environment that is untrusted. In the globe distributed heterogeneous networking and computation is a highly important capacity. Computing data that is encrypted is an aim in cryptography to find its general method since it was developed in 1978, by Rivest et al. [112]. The high interest in this topic is a result of its several applications.

For many years the concern for security has been one of the crucial features of several research trying to give various methods and techniques [114]. Encryption has provided the most significant method to guarantee in privacy, security, and protection for AI existing systems. furthermore, homomorphic encryption has a high attention to its several benefits. It can be applied to the cloud that re-public, giving the overview of service providers that can take various operations on data that is encrypted with being decrypted. Hence this technique is utilized in several systems, with various aims, and gives high adaptability by being executed in several ways [115,116].

Fully homomorphic encryption is classified into two kinds of mathematical operation i.e., multiplication and addition. The cryptosystem is classified by its level of complexity particularly when discussing practical execution. To this date, there is no significant execution for these systems. Tawalbeh et al. [117] and Manish et al. [118], were involved in security with the help of the cloud to guarantee the secrecy of data, confidentiality, and privacy, however, encryption was one of the best methods and the more special it was homomorphic encryption.

This enables a cloud provider to execute any type of mathematical operation on encrypted data by decrypting it. For every user security is one of the highest concerns, even in an organization's productive scale or user-simple scale, depending on the storage of

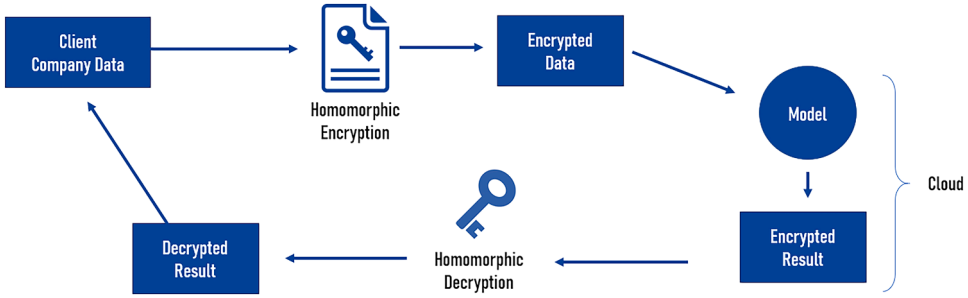


Fig. 4. Working mechanism of homomorphic encryption [113].

data, receiving and sending messages in the cloud are required to be secure especially when meeting with a third party. Several applications are based on homomorphic encryption providing excellent performance [114].

Homomorphic encryption is built on the principle of data processing without the need for decryption of it [119-121]. Confidentiality of data is one of the main concepts of executing the encryption system in suspicious server systems. Compared to other encrypted protocols, homomorphic encryption enables algebraic computations like multiplication and addition directly on data which is encrypted and acts as plain text. This is demonstrated as,

$$v = f(u) \Leftrightarrow Enc(v) = g(Enc(u))$$

Here u and v illustrate unencrypted vectors, while *Enc* acts as the operation of encryption executed on u and v vectors. a cryptosystem is represented as c with ciphertext (cj), plaintext (pi), and *Enc* (encryption function) where:

$$Enc(pi) = (cj)$$

We execute multiplicative and additive operations on pi such as:

$$\mathrm{E}\Delta : (x1)\Delta Enc(x2) = Enc(x1 + x2)$$
$$\mathrm{E}\Delta : (x1)\Delta Enc(x2) = Enc(x1 * x2)$$

A process of encryption satisfying both multiplicative and additive properties is regarded as homomorphic. A combination of AND and XOR Boolean functions are utilized in the encryption.

7.3.1. Challenges

Even though this method ensures the protection of data, confidentiality, and security there are still some challenges that need to be solved. The settings for security are important for wireless sensor networks (WSN) to make sure the sectary of transmitted data. The system is required to give a model which is more reliable for security in the WSN network. It will be important to provide systems that are efficient in the cost of energy [114].

Research presented the challenge of the polynomial ACD problem to estimate polynomial common divisors. Even if the system is simple, till now the robustness of the system is not guaranteed with several attacks on the systems. This problem needs to be solved to avoid attacks [122].

8. Post-quantum cryptography (PQC)

A significant amount of research has been done recently on quantum computers which are devices that use quantum mechanical processes to solve mathematical problems that are hard or unsolvable for traditional computers. Several of the current public key cryptosystems can be broken if large-scale quantum computers are ever constructed. This would put digital communications on the internet and elsewhere integrity and confidentially of serious danger. Developing cryptographic systems that are safe against both classical and quantum computers while still functioning with current networks and protocols of communication is the objective of post-quantum cryptography-resistant cryptography [123]. With the quantum computers looming, PQC safeguards data from being broken by these powerful machines. However, even PQC algorithms are vulnerable to side-channel attacks, where attackers exploit leaked information during encryption/decryption. For securing applications in the quantum era, such as Metaverse, Web 3.0, and deeply embedded systems, the merging threats and potential countermeasures should be considered [124].

8.1. SIKE on M4

SIKE, a PQC algorithm shows promise for securing resource-constrained devices like the ARM Cortex-M4 microcontrollers. While some implementations require optimization for speed, Sike Offers a path toward future-proof encryption on these low-power platforms. The post-quantum key encapsulation protocol SIKE has a smaller ciphertext and public sizes, but its execution time is slow. The

Table 3
AI Applications.

AI techniques	Domains	Task	References
Deep learning	Cybersecurity	Detection of malware	Wang et al. [156]
	Healthcare	Detection of Covid-19	Islam et al. [157]
	Business and Finance	Prediction of Stock trend	Anuradha et al. [158]
	Virtual Assistant	Chatbot	Dhyani et al. [159]
Machine learning	Cybersecurity	Attack and Anomaly Detection	Sarker et al. [160]
	Healthcare	Aid for COVID-19	Blumenstock et al. [161]
	Smart city	Smart parking pricing system	Saharan et al. [162]
Fuzzy logic-based approach	Cybersecurity	Network anomaly detection system	Hamamoto et al. [163]
	Healthcare	Heart disease detection	Reddy et al. [164]
	Business	Satisfaction of Customer	Kang et al. [165]
	Cybersecurity	Selection of optimum feature	Fatima et al. [166], Onah et al. [167]
Hybrid method	Applications for mobile	Personalized decision-making	Sarker et al. [168]
	Business	Satisfaction of customers	Kang et al. [165]

fastest SIKE implementation for the Cortex-M4 microcontroller is presented by a study that focuses on low-level operation optimization to achieve notable speedups (Around 22.97 percent) in earlier works, furthermore, optimization of energy consumption yields improvement of up to 11.9 percent. These optimizations make SIKE suitable for resource-constrained devices [125].

8.2. Curve448 and Ed448 on Cortex-M4

Curve448 and Ed448 offer high-security keys for key exchange and digital signature, respectively. However, their suitability for resource-constrained devices like the ARM cortex M4 microcontroller depends on implementation efficiency. The research was conducted on the first public key-cryptography algorithms, Curve448 optimized for a 32-bit ARM Cortex M$ microprocessor. Within 6,285904 clock cycles, the study accomplishes a scaler multiplication operation by concentrating on effective finite field and group operation. This paves the way for secure communication utilizing cortex M$ processor and is the fastest known implementation of Curve448 on this platform [126].

The effectiveness of cryptography for resource-constrained devices, such as those used in the Internet of Things (IoT) is a continuous area of research. For public key cryptography, which is utilized for secure key exchange and communication, this is particularly important. Due to its minimal resource needs, elliptic curve cryptography (ECC) is a preferred choice in this industry. A study optimizes two specific ECC algorithms i.e., Curve448 and Ed448 for ARM cortex M4 processes, which are frequently encountered in IoT devices, The study achieved a notable speedup over earlier work, and the study contributed to more secure communication via ECC [127].

8.3. SIKE round 3 on ARM Cortex-M4

SIKE Round 3 is a good fit for ARM cortex M4-microcontroller due to its tiny key sizes, still, it's not super speedy yet. A work of study presented the utilization of compressed SIKE for devices with limited resources, especially aimed at the STM32F407VG microcontroller. It achieves up to 25 percent speedups over earlier work by utilizing exciting optimization approaches and proposing new assembly designs for finite field operations. For embedded systems, this increases the practical value of compressed SIKE [128].

8.4. Kyber on 64-bit ARM Cortex-A

Kyber is a post-quantum cryptographic algorithm designed to be secure against quantum and classical attacks. It is used for public key encryption and key encapsulation. Implementing Kyber on a 64-bit ARM Cortex-A processor involves optimizing the algorithm to leverage the processor's performance and efficiency features. The research was conducted that presented optimized Kyber implementations for 64-bit ARM Cortex-A processors, a post-quantum cryptography technique. The study concentrated on enhancing the algorithms in a few areas, such as noise sampling, number theoretic transformations, and using integrated AES accelerators. On these processors, Kyber encryption operates more quickly on these processors because of these modifications, which produced notable speedups over earlier efforts. Notably, the utilization of AES acceleration achieved even greater improvements. Overall, the study established new performance benchmarks for Kyber encryption on a 64-bit arm cortex A processor [129].

8.5. Cryptographic accelerators on Ed25519

The cryptographic accelerators enhance the performance of ED25519 operations by using specialized hardware for faster modular arithmetic and parallel processing. They improve efficiency, reduce power consumption, and offer additional security features. This makes them ideal for high speech, and secure digital signatures in various applications. The Ed25519 digital signature method on FPGAs was utilized in a study. Compared to the methods, Ed25519 provides faster execution while keeping robust security. The study suggests two FPGA designs such as a high-performance one that prioritizes speed (21xfaster, processing over 6,200 signatures per second). Both approaches perform better than earlier implementations and are safe against side-channel attacks. It was concluded that Ed25519 is now more appealing for secure for secure hardware-based communication as a result [130].

8.6. Supersingular isogeny Diffie-Hellman (SIDH) key exchange on 64-bit ARM

The SIDH post-quantum key exchange protocol is designed to be secure against quantum attacks. Implementing SIDH on a 64-bit ARM processor involves optimizing the elliptic curve and isogeny computation to leverage the capabilities of the processor. A study optimized the SIDH exchange protocols for 64-bit ARMv8 processors, leading to notable speedups (up to 5x) over generic implementations. When affine and projective techniques were compared the study discovered that projective performs better overall even though affine may be superior in the last round. SIDH's significant computational cost persists even after optimization. But when compared to other post-quantum cryptography methods, it offers smaller key sizes [131].

9. Fault attacks as side-channel attack

Fault attacks are a type of side channel attack where an attacker induces errors in a cryptographic device such as by manipulating voltage, or temperature, or using laser beams. By analyzing the resulting incorrect outputs, the attacker can infer secret keys or internal

data. These attacks are often combined with other side-channel techniques like power analysis, to increase effectiveness. Countermeasures include error detection and correction, redundant computations, and other hardware and software protections to ensure the integrity of cryptographic operations.

9.1. Error detection in lightweight Welch-Gong (WG)-oriented stream cipher WAGE

In the lightweight stream cipher WAGE, error detection is crucial to ensure data integrity and security. WAGE uses a combination of simple error detection codes and redundancy techniques to detect and mitigate errors during the encryption and decryption process. For devices with limited resources, a study suggests error detection techniques for lightweight cryptographic algorithms ASCON. Although it can be susceptible to mistakes in hardware implementations ASCON provides hashing and encryption functions. The three error detection schemes- CRC, signature, and interleaved signature that can be used with ASCON in hardware were introduced. Two FPGA efficient (99.99 percent error coverage) with minimal to no performance impact. This makes ASCON more reliable for hardware-based security in devices with limited resources [132]. Data security on the devices becomes increasingly important as the IoT expands. Lightweight cryptography (LWC) specifically the ASCON algorithm recently standardized by the NIST, is significant for this purpose [133].

9.2. Error detection reliable architectures of Camellia block cipher

With the help of reliable architectures of Camellia block cipher, error detection is achieved through redundancy and parity checks. This includes duplicating critical operations and comparing results, as well as using parity bits to detect errors in data blocks. The measures ensure data integrity and enhance the security and robustness of camellia making it suitable for environments where reliability is essential. A study focuses on Camellia, a block cipher algorithm, and addresses the reliability and security issues in cryptographic implementations for resource-constrained devices such as medical implants. The study suggested error detection techniques that can be tailored to various S-box implementations according to the particular reliability and security requirements of the device. These techniques provide high error coverage with low performance and implementation overhead, making Camellia more reliable for these critical applications [134].

9.3. Fault diagnosis specifically for the low-energy midori block cipher

The fault diagnosis for the low-energy Midori block cipher involves techniques to detect and mitigate errors induced by faults. This includes implementing error-detection codes, redundancy in computations, and techniques like parity checks to identify discrepancies in data and operations. These measures are crucial to ensure the cipher's reliability and security in low low-energy environment where maintaining data integrity is paramount. Aghaie et al., [135] presented a work in which error detection technique for Midori, a novel energy-efficient block cipher intended for devices with limited resources such as medical implants. These techniques, which were initially created for Midori, aim to address both the S-box layer and the entire circular structure. Test results indicate that Midori has a high degree of error coverage with little effect on performance, which increases its reliability in practical applications.

9.4. Block cipher QARMA with error detection mechanisms

QARMA is a lightweight block cipher designed for efficient and secure encryption in constrained environments. It incorporates an error detection mechanism to ensure data integrity during the encryption and decryption process. A lightweight block cipher called QARMA, which is utilized in resource-limited devices such as medical implants, is introduced in a study along with error detection techniques. Both logic-gate and lookup table implementations are targeted by these initial error detection algorithms created especially for QARMA. This algorithm is known as reliable for practical uses because of the technique's ability to guard against both transient and permanent errors or negligible influence on performance [136].

10. Discussion

Traditional encryption methods like ECC and RSA are at risk of being scattered by future quantum computers. These powerful machines could crack the math behind them, leaving our sensitive data on smartphones, blockchains, emails, and web browsers. To address these threats, post-quantum cryptography (PQC) is an emerging algorithm. It is specifically designed to withstand attacks from both classical and quantum computers. Enhancing the security of algorithms of cryptography by securing them against active attacks of side-channel is vital in the engineering of cryptography. The BLAKE algorithm is an effective function that has been established based on Bernstein's ChaCha stream cipher. For security purposes the well-known company Google has replaced RC4 in TLS with ChaCha along with Bernstein's Poly1305 message authentication code, which shows that BLAKE's employment has significant importance. A study presented a high-performance fault scheme for the BLAKE. Particularly for round function for which approaches such as two faults diagnosis were developed and examined for the capability of error detection. The outcome of the study showed that with the help of simulations of injection-based error, it showed the coverage of error of nearly 100 percent by the proposed method, which will make BLAKE more reliable [137].

Binary extension finite fields GF (2^m) has been utilized in several modern error-correcting codes and public-key cryptosystems. It is vital for the application of current and post-quantum cryptography. Lately, schemes like the Itoh-Tsujii algorithm (ITA) and Fermat's

little theorem (FLT) have been researched to achieve better performance; still, the following operation is time-consuming, expensive, and difficult and might need thousands of numbers of gates for its operation [138]. Due to recent advancements in quantum computing, the development of public-key cryptosystems is needed to protect against attacks permitted by quantum computers. National Institute of Standards and Technology (NIST) in 2017 developed a project to regulate quantum computers based on algorithms of public-key cryptography. It was found that among major classes of post-quantum algorithms, lattice-based cryptography was the most quantum-resistant [139].

Recently study introduces error detection methods for WG-29, a stream cipher used in various applications. These are the first such methods designed specifically for WG-29 hardware implementations. The methods offer high error coverage (over 99.99 percent) with minimal impact on area, power, and delay (less than 40, 12, and 10 percent increase, respectively). This makes WG-29 hardware implementation more reliable. Additionally, these methods can be adapted with minor changes for other similar stream ciphers [140].

While encryption keeps our data safe, side-channel attacks exploit subtle leaks such as power usage, electromagnetic waves, or processing time, to steal the secret key. These attacks are particularly risky for lightweight cryptography, and encryption design for devices with limited resources. The need for secure, low-power, and low-energy cryptographic algorithms in the post-quantum age is discussed by Canto et al., [141]. It emphasizes code-based cryptography, emphasizes the mcEliece and Niederreiter public key cryptosystems in particular, which are strong contenders for NIST standardization. These systems are susceptible to errors brought on by external circumstances and deliberate attacks like differential fault analysis, even though they have strong error-correcting capabilities. To tackle these vulnerabilities, the study suggests effective fault detection algorithms such as normal parity, interleaved parity, and two forms of cyclic redundancy checks (CRC-2 and CRC-8). Implementing these techniques on the Kintex-7FPGA validates them, showing that they are feasible and have little overhead for usage in limited embedded systems.

The effectiveness of implementing the post-quantum key exchange protocol based on the Jao and DeFeo isogeny on embedded platforms powered by ARM processors was examined by Koziel et al., [142]. The authors use Montgomery multiplication and education to achieve a threefold speedup over the GNU Multi precision Library in their proposal of new primes to improve constant-time finite field arithmetic and speed up isogeny computations. They found that affine isogeny formulas perform noticeably quickly on ARM systems after analyzing projective isogeny formulas from Costello et al (Crypto 22016). Timings found from AARMv7A architecture-based platforms at 85-0.128-,170-bit quantum security levels are presented, together with optimized affine SIDH libraries for 512, 768, and 1024-bit primes.

Another study presented the security issues associated with newly emerging deeply embedded systems, such as wearable and implantable medical devices, which have a wider "attack surface" than more conventional embedded systems, like nano sensor networks and secure smart cards. Due to the severe limitations of these battery-powered systems, conventional solutions are frequently impracticable given the potentially fatal effects of security breaches in medical devices. Due to the security concerns are Multidisciplinary, university teaching, both graduate and undergraduate has not kept up with somewhat faster advancement in cryptography engineering research for these vital systems. The author offers a research and education integration plan tailored to medical device security in order to close this gap. They employed this tactic for two years at the graduate level, concerning side-channel attacks known as fault analysis attacks. In contrast to conventional embedded systems security research and instruction, the results show both the methodology's success and shortcomings. The study points out that their methods for integration are universal and adaptable to various vital infrastructures [143].

As wearable and implantable medical devices become more integrated into daily life, modern embedded computing systems are increasingly vulnerable to serious and even fatal security threats. These sophisticated biomedical systems data mining and processing for medical purposes, pose special security and privacy problems because of their delicate operating environment and strict size and limitation of energy. These challenges compose areas such as analytics of big data, secure machine learning, privacy-preserving data mining, and bioinformatics, all with substantial constraints [144].

Niasar et al., [145] point multiplication on curve448_ which NIST recommended for 224-bit security in elliptic curve cryptography is implemented using FPGA. For various application needs, three architectures are proposed: High-performance, area-time efficient, and lightweight. The performance architecture, which is synthesized on a Xilinx Zynq 7020 FPGA, achieves a 40% efficacy boost in clock cycles x area and a 12% throughout increase (1,219 multiplications per second) over previous work. Operating at 250 MHz, the lightweight design preserves performances while conserving 96% of resources, by balancing time and resource trade-off, the area tie efficient design boosts efficiency by 48% while using 52% fewer resources, Additionally, side channel countermeasures that are effective are integrated, outperforming earlier versions in terms of performance and security metrics.

Cintas et al., [146] focuses on finite field arithmetic fault detection in cryptographic schemes, which is essential for safeguarding against man-made or natural errors. It presented fault detection algorithms for finite field multipliers in post-quantum and conventional cryptography that are based on cyclic codes. These techniques, when implemented on AMD/Xilinx FPGAs, provide great error coverage with low overhead, which makes them appropriate for embedded systems with severe constraints.

Karam et al., [147] discusses the difficulties that the COVID-19 pandemic has made to laboratory-based, hands-on computer engineering courses like Computer Logic Design and FPGA Design. It highlights how simulators fall short of providing the same hands-on experience as dealing with hardware, especially when it comes to hardware security education. The study outlines their creative process for creating practical hardware security that can be delivered via HyFlex.

11. Validation and verification

To make AI trustworthy, privacy, accountability, explainability, and fairness will make the AI ethical thus, more trust will be built in the system. Several methods have been developed additionally to make AI more trustworthy. It can be a complex situation to make

AI trustworthy without compromising the performance of the system.

AI systems consist of non-deterministic and deterministic elements. Both elements require to be examined for the validation and verification of the system. Deterministic elements are those that can be predicted clearly. Non-deterministic elements which cannot be predicted [148]. Hence, substantial human involvement is required to check the output produced by the system. To control this issue, various techniques have been developed as follows:

11.1. Benchmarking

It is a method that is utilized to compare, measure, and test the AI system performance of the datasets that are designed [148]. Ngan et al. [149] developed the face recognition vendor test FRVT to examine and compare the performance algorithms of automatic gender classification. Jiang et al. [150] proposed a benchmark suite HPCAI500 for high-performance scientific computing (HPC) systems for AI.

11.2. Expert panels

It can be utilized when conventional testing is not possible. This approach can be considerable when AI is developed to help experts [128]. This panel which is independent is accountable for giving possible recommendations and diagnoses outputs of the system [151].

11.3. Metamorphic testing

This testing is utilized to prevent the Oracle problem. It depends on the system's output and input relationship in the form of multiple iterations. The system test is based on the outputs of several related inputs [152]. Many researchers have utilized this method to examine systems of A. Lindvall et al. [153] utilized it to examine the software control for autonomous drones.

11.4. Testing in a simulated environment

This testing is beneficial when the system of AI is proposed to do physical actions [148]. This testing is executed on drones and robots made of AI. In simulated testing, a controlled environment is utilized to examine the performance of the system under various conditions.

11.5. Field trials

This approach is utilized to examine the durability and performance of the operating system. It is a significant method because it reports how the individuals will associate with the AI system. This method of testing is employed when the testing environment is purely dissimilar from the actual environment [148]. Field trials are an efficient way to crosscheck the acceptability of AI system individuals. The United Kingdom is utilizing field trials to examine self-driving cars [154]. Bundesamt [155] employed field trials to examine systems for face recognition.

12. Applications of AI

AI applications have been significantly tested in several problems in various areas. Cybersecurity, healthcare, social media, business, robotics, and several other areas are frequent these days. Several AI methods, like deep learning, machine learning, natural language processing, and many others known as powerful AI techniques are utilized in these domains. In Table 3 we have listed several AI applications utilized in the real world.

13. Ethical consideration

AI is a transformative force with the potential to revolutionize various aspects of society. However, the rapid advancement of AI technology has raised significant ethical concerns, necessitating careful consideration of issues such as bias and fairness, responsible deployment, and compliance with regulations

13.1. Bias and fairness

One of the primary ethical challenges in AI revolves around bias and fairness. AL systems learn from historical data, which can embed societal biases. This can result in discriminatory outcomes, disadvantaging certain groups. Addressing data bias is crucial requiring the curation of diverse and representative datasets. Additionally, proactive measures during the training process, such as bias detection and mitigation strategies are essential. To ensure fair outcomes, AI practitioners must conduct impact assessments and involve diverse stakeholders in the decision-making process.

13.2. Responsible AI deployment

It encompasses various considerations to ensure the ethical use of AI technologies. Explainability and transparency are critical to building trust in AI systems. The black-box nature of the complex algorithms poses challenges, emphasizing the need for interpretability models and transparent decision-making processes. Human-in-the-loop approaches acknowledge the importance of human oversight, preventing undue resilience on AI and allowing for human interaction when necessary. Establishing avoiding applications that may lead to social harm.

13.3. Regularity landscape and compliance

The regular landscape surrounding AI is evolving rapidly, reflecting socially growing awareness of the need for ethical oversight. Data protection and privacy concerns arise as AI often involves the processing of personal data. Adhering to the regulations such as the general protection regulation. Implementing privacy-preserving techniques and obtaining informed consent are crucial steps. Auditing and accountability mechanisms play a pivotal role in addressing ethical concerns. AL system should be subject to regular audits, and monitory records of decisions made by AI algorithms. Ethical standards and guidelines though not yet universal, provide a framework for ethical AI development. Organizations are encouraged to adhere to existing standard participation in the development of industry guidelines.

14. Future challenges

14.1. Scalability of privacy-preserving techniques

As AI systems scale up in complexity and data volumes, there's a challenge in ensuring that privacy-preserving techniques like homomorphic encryption and differential privacy remain scalable without compromising performance.

14.2. Interpretability across platforms

Achieving the interpretability of security and privacy models across diverse AI platforms and frameworks presents a challenge. Harmonizing standards and ensuring seamless collaboration in FL settings will be crucial.

14.3. User-friendly explainability

Enhancing the User-friendliness of explainable AI models is essential. Future efforts should focus on making complex security and privacy models interpretable and transparent to non-expert users.

15. Future applications

15.1. Healthcare data collaboration

Security and privacy techniques will play a pivotal role in enabling collaborative research and analysis of healthcare data across institutions. Federated learning and differential privacy can facilitate secure data sharing for medical advances.

15.2. Secure financial transaction

Homomorphic encryption can be applied to secure financial transactions allowing computation on encrypted data without exposing sensitive information. This implies secure and private financial services.

15.3. Resilient IoT ecosystem

Building robust and secure AI model for analyzing data from IoT devices using federated learning, ensuring the integrity of smart city initiatives and other IoT applications while preserving privacy through techniques like differential privacy.

15.4. Hybrid quantum resistant crypto

Merging current cryptography with post quantum algorithm for a smooth transition to a quantum future.

15.5. Evolving crypto for new tech

Designing lightweight crypto for resource constrained devices and exploring AI and quantum resistant solutions.

16. Future studies

AI security and privacy and several emerging technologies are expected to shape the future, offering both opportunities and challenges. Post-quantum threats are poised to become integral in securing AI systems. Advances in privacy-preserving machine learning, particularly through technologies like federated learning and homomorphic encryption are anticipated to facilitate collaborative model training without compromising individuals' data privacy.

Explainable AI (XAI) models are expected to play a vital role, however, challenges such as the continual evolution of adversarial attacks and the ethical implications of implementing security measures. Additionally, it ensures resilience against model extraction attacks manages cross-domain security concerns, and explains the compatibility of blockchain with AI computation domain key areas for future exploration.

17. Conclusion

In conclusion, the landscape of security, privacy, and robustness in trustworthy AI systems is rapidly evolving, presenting both unprecedented opportunities and formidable challenges. Our review article has traditional and emerging realms of safeguarding AI systems, recognizing the critical importance of balancing innovation with a proactive stance against adversarial threats.

From the fortification of conventional security measures such as access control and encryption to the exploration of cutting-edge technologies like federated learning and homomorphic encryption, the present article underscores the dynamic nature of the field. The ever-expanding capabilities of AI systems necessitate a comprehensive and adaptive security framework that goes beyond mere compliance and aims to foster user trust, ethical deployment, and societal benefit. The current study has delved into the intricate interplay between privacy preservation and AI advancement, acknowledging the imperative to navigate the delicate balance between utility and individuals' privacy rights.

Funding statement

This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2024/R/1445)

CRediT authorship contribution statement

Mozamel M. Saeed: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft. **Mohammed Alsharidah:** Data curation, Formal analysis, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mozamel Musa Saeed reports financial support was provided by Prince Sattam bin Abdulaziz University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The authors would like to express their deep thankful and grateful to all the associated personnel in any reference that contributed in/for the purpose of this research.

References

[1] Li B, Qi P, Liu B, Di S, Liu J, Pei J, Yi J, Zhou B. Trustworthy AI: from principles to practices. ACM Comput Surv 2023;55:1–46.
[2] Hao K. AI is sending people to jail—and getting it wrong. Technol Rev 2019;21.
[3] Singh R, Vatsa M, Ratha N. Trustworthy ai. InProc 3rd ACM India Joint Int Confer Data Sci Manage Data (8th ACM IKDD CODS & 26th COMAD) 2021:449–53.
[4] Liu H, Wang Y, Fan W, Liu X, Li Y, Jain S, Liu Y, Jain A, Tang J. Trustworthy AI: a computational perspective. ACM Trans Intell Syst Technol 2022;14:1–59.
[5] Liang W, Tadesse GA, Ho D, Fei-Fei L, Zaharia M, Zhang C, Zou J. Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell 2022;4(8):669–77.
[6] Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. CSUR 2022;55(2):1–38.
[7] Wing JM. Trustworthy AI. CACM 2021;64(10):64–71.
[8] Lau D, Samy GN, Rahim FA, Maarop N, Hassan NH. Review of the governance, risk and compliance approaches for artificial intelligence. Int J Med Inform 2023;11(2):25–35.

[9] Nagarajan SM, Devarajan GG, Ramana TV, Bashir AK, Al-Otaibi YD. Adversarial deep learning based Dampster–Shafer data fusion model for intelligent transportation system. Inf Fusion 2024;102:102050.
[10] Chakraborty C, Nagarajan SM, Devarajan GG, Ramana TV, Mohanty R (2023) Intelligent AI-based healthcare cyber security system using multi-source transfer learning method. TOSN.
[11] Nagarajan SM, Deverajan GG, Bashir AK, Mahapatra RP, Al-Numay MS. IADF-CPS: Intelligent anomaly detection framework towards cyber physical systems. Comput Commun 2022;188:81–9.
[12] Nagarajan SM, Devarajan GG, Mohammed AS, Ramana TV, Ghosh U. Intelligent task scheduling approach for IoT integrated healthcare cyber physical systems. IEEE Trans Netw Sci 2022.
[13] Gopal DG, Saravanan R. Selfish node detection based on evidence by trust authority and selfish replica allocation in DANET. Int J Info Comm Technol 2016;9(4):473–91.
[14] Deverajan GG, Muthukumaran V, Hsu CH, Karuppiah M, Chung YC, Chen YH. Public key encryption with equality test for industrial internet of things system in cloud computing. Transac Emerg Telecomm Technol 2022;33(4):4202.
[15] Selvaraj A, Patan R, Gandomi AH, Deverajan GG, Pushparaj M. Optimal virtual machine selection for anomaly detection using a swarm intelligence approach. App soft comp 2019;84:105686.
[16] Nagarajan SM, Deverajan GG, Kumaran U, Thirunavukkarasan M, Alshehri MD, Alkhalaf S. Secure data transmission in internet of medical things using RES-256 algorithm. IEEE Trans Industr Inform 2021;18(12):8876–84.
[17] Subramaniam EV, Srinivasan K, Qaisar SM, Pławiak P. Interoperable IoMT approach for remote diagnosis with privacy-preservation perspective in edge systems. Sensors 2023;23(17):7474.
[18] Sita Kumari K, Shivaprakash G, Arslan F, Alsafarini MY, Ziyadullayevich AA, Haleem SL, Arumugam M. Research on the quantum photonic convolutional neural network for artificial intelligence-based healthcare system security. Opt Quantum Electron 2024;56(2):149.
[19] Tabassum T, Toker O, Khalghani MR. Cyber–physical anomaly detection for inverter-based microgrid using autoencoder neural network. Appl Ener 2024;355:122283.
[20] Prakash N, Vignesh J, Ashwin M, Ramadass S, Veeranjaneyulu N, Athawale SV, Ravuri A, Subramanian B. Enabling secure and efficient industry 4.0 transformation through trust-authorized anomaly detection in cloud environments with a hybrid AI approach. Opt Quant Electron 2024;56(2):251.
[21] Suresh J, Sahayaraj JM, Rajakumar B, Jayapandian N. Hybrid Bayesian and modified grey PROMETHEE-AL model-based trust estimation technique for thwarting malicious and selfish nodes in MANETs. Wirel Netw 2024:1–22.
[22] Ilakkiya N, Rajaram A. Blockchain-enabled lightweight intrusion detection system for secure MANETs. J Elect Eng Technol 2024:1–5.
[23] Fathi F, Baghani M, Bayat M. Light-PerIChain: Using lightweight scalable blockchain based on node performance and improved consensus algorithm in IoT systems. Comput Commun 2024;213:246–59.
[24] Hande JY, Sadiwala R. Data security-based routing in MANETs using key management mechanism. SN comput sci 2024;5(1):155.
[25] Lin Q, Li X, Cai K, Prakash M, Paulraj D. Secure Internet of medical Things (IoMT) based on ECMQV-MAC authentication protocol and EKMC-SCP blockchain networking. Inf Sci 2024;654:119783.
[26] Sekar S, Jeyalakshmi S, Ravikumar S, Kavitha D. Modified light GBM based classification of malicious users in cooperative cognitive radio networks. CPS 2024;10(1):104–22.
[27] Yuan Z, Tian Y, Zhou Z, Li T, Wang S, Xiong J. Trustworthy federated learning against malicious attacks in Web 3.0. IEEE Trans Netw Sci 2024.
[28] Wang W, Abbasi O, Yanikomeroglu H, Liang C, Tang L, Chen Q. VHetNets for AI and AI for VHetNets: an anomaly detection case study for ubiquitous IoT. IEEE Netw 2024.
[29] Mitropoulou K, Kokkinos P, Soumplis P, Varvarigos E. Anomaly detection in cloud computing using knowledge graph embedding and machine learning mechanisms. J Grid Comput 2024;22(1):6.
[30] Sharma B, Sharma L, Lal C, Roy S. Explainable artificial intelligence for intrusion detection in IoT networks: a deep learning based approach. Exp Syst Appl 2024;238:121751.
[31] Hu Y, Kuang W, Qin Z, Li K, Zhang J, Gao Y, Li W, Li K. Artificial intelligence security: threats and countermeasures. CSUR 2021;55(1):1–36.
[32] Berghoff C, Neu M, von Twickel A. Vulnerabilities of connectionist AI applications: evaluation and defense. Front big Data 2020;3:23.
[33] Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H, Yang J, Toner H, Fong R, Maharaj T (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims.
[34] Hamon R, Junklewitz H, Sanchez I (2020) Robustness and explainability of artificial intelligence 207.
[35] Sharma A, Tyagi A, Bhardwaj M. Analysis of techniques and attacking pattern in cyber security approach: a survey. Int J Health Sci 2022:13779–98.
[36] Fischer EA (2014) Cybersecurity issues and challenges: in brief.
[37] Ibrahim UM. The impact of cybercrime on the Nigerian economy and banking system. NDIC Quarter 2019;34:1–20.
[38] Qi H, Di X, Li J. Formal definition and analysis of access control model based on role and attribute. J Inf Secur Appl J INF SECUR APPL 2018;43:53–60.
[39] Sarker IH, Furhad MH, Nowrozy R. Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. Comput Sci 2021;2:1–8.
[40] Xue Y, Meng G, Liu Y, Tan TH, Chen H, Sun J, Zhang J. Auditing anti-malware tools by evolving android malware and dynamic loading technique. Trans Inf Forens Secur 2017;12:1529–44.
[41] Hunt T, Zhu Z, Xu Y, Peter S, Witchel E. Ryoan: A distributed sandbox for untrusted computation on secret data. TOCS 2018;35:1–32.
[42] Li B, Feng Y, Xiong Z, Yang W, Liu G. Research on AI security enhanced encryption algorithm of autonomous IoT systems. Inf Sci 2021;575:379–98.
[43] Aloraini F, Javed A, Rana O, Burnap P. Adversarial machine learning in IoT from an insider point of view. J Inf Secur Appl 2022;70:103341.
[44] Lin J, Dang L, Rahouti M, Xiong K (2021) ML attack models: adversarial attacks and data poisoning attacks. arXiv preprint arXiv:2112.02797.
[45] Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
[46] Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. IEEE SP:39–57.
[47] Berghel H. Equifax and the latest round of identity theft roulette. Comput 2017;50(12):72–6.
[48] Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M. Decision support system using trust planning among food-energy-water actors. In: InAdvanced Information Networking and Applications: Proceedings of the 33rd International Conference on Advanced Information Networking and Applications (AINA-2019). Springer International Publishing; 2020. p. 1169–80.
[49] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression. Adv Neural Inf Process 2008;21.
[50] Garfinkel S. De-identification of Personal Information. US Department of Commerce, National Institute of Standards and Technology; 2015.
[51] Khalil M, Ebner M. De-identification in learning analytics. J Learn Anal 2016;3(1):129–38.
[52] Truex S, Baracaldo N, Anwar A, Steinke T, Ludwig H, Zhang R, Zhou Y. A hybrid approach to privacy-preserving federated learning. In: InProceedings of the 12th ACM workshop on artificial intelligence and security; 2019. p. 1–11.
[53] Kim Y, Hwang H, Shin J. Robust object detection under harsh autonomous-driving environments. IET Image Process 2022;16:958–71.
[54] Michaelis C, Mitzkus B, Geirhos R, Rusak E, Bringmann O, Ecker AS, Bethge M, Brendel W (2019) Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484.
[55] Ponn T, Kröger T, Diermeyer F. Performance analysis of camera-based object detection for automated vehicles. Sensors (Basel) 2020:20.
[56] Hamon R, Junklewitz H, Sanchez I (2020) Robustness and explainability of artificial intelligence. OP 207.
[57] Gade K, Geyik SC, Kenthapadi K, Mithal V, Taly A. Explainable AI in industry. In: InProceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining; 2019. p. 3203–4. https://doi.org/10.1145/3292500.3332281.
[58] Thampi A. Interpretable AI: Building explainable machine learning systems. Simon and Schuster; 2022.
[59] Khaleghi B (2019) The why of explainable AI. blog article https://www.elementai.com/news/2019/the-why-ofexplainable-ai.
[60] Steinruecken C, Smith E, Janz D, Lloyd J, Ghahramani Z. The automatic statistician. Automated machine learning: Methods, Syst, Challen 2019:161–73. https://doi.org/10.1007/978-3-030-05318-5_9.

[61] Kurakin A, Goodfellow I, Bengio S (2016) Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.http://arxiv.org/abs/1611.01236.
[62] He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. Knowl Based Syst 2021;212:106622.
[63] Srivastava G, Jhaveri RH, Bhattacharya S, Pandya S, Maddikunta PK, Yenduri G, Hall JG, Alazab M, Gadekallu TR (2022) XAI for cybersecurity: state of the art, challenges, open issues and future directions. arXiv preprint arXiv:2206.03585.
[64] Russell SJ, Norvig P (2010) Artificial intelligence a modern approach.
[65] Couteaux V, Nempont O, Pizaine G, Bloch I. Towards interpretability of segmentation networks by analyzing deepdreams. ininterpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. In: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019; 2019. p. 56–63.
[66] Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Sci Adv 2018;4:5580.
[67] Soares E, Angelov P, Biaso S, Cury M, Abe D. A large multiclass dataset of CT scans for COVID-19 identification. Evol Syst 2023:1–6.
[68] Camilleri MA. Artificial intelligence governance: Ethical considerations and implications for social responsibility. Expert Syst 2023:e13406.
[69] McMahan B, Ramage D. Federated learning: Collaborative machine learning without centralized training data. Google Res Blog 2017;3.
[70] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22(10):1345–59.
[71] Du W, Atallah MJ. Privacy-preserving cooperative statistical analysis. In: InSeventeenth Annual Computer Security Applications Conference. IEEE; 2001. p. 102–10.
[72] Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data. In: InProceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining; 2002. p. 639–44.
[73] Gascón A, Schoppmann P, Balle B, Raykova M, Doerner J, Zahur S, Evans D. Secure linear regression on vertically partitioned datasets. IACR Cryptol. ePrint Arch 2016:892.
[74] Karr AF, Lin X, Sanil AP, Reiter JP. Privacy-preserving analysis of vertically partitioned data using secure matrix products. J Off Stat 2009;25(1):125.
[75] Sanil AP, Karr AF, Lin X, Reiter JP. Privacy preserving regression modelling via distributed computation. In: InProceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining; 2004. p. 677–82.
[76] Du W, Han YS, Chen S. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In: InProceedings of the 2004 SIAM international conference on data mining; 2004. p. 222–33. Society for Industrial and Applied Mathematics.
[77] Wan L, Ng WK, Han S, Lee VC. Privacy-preservation for gradient descent methods. In: InProceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining; 2007. p. 775–83.
[78] Hardy S, Henecka W, Ivey-Law H, Nock R, Patrini G, Smith G, Thorne B (2017) Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint arXiv:1711.10677.
[79] Nock R, Hardy S, Henecka W, Ivey-Law H, Patrini G, Smith G, Thorne B (2018) Entity resolution and federated learning get a federated resolution. arXiv preprint arXiv:1803.04035.
[80] Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K. Practical secure aggregation for privacy-preserving machine learning. In: Inproceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security; 2017. p. 1175–91.
[81] Aono Y, Hayashi T, Wang L, Moriai S. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Trans Inf Forensics Secur 2017;13(5):1333–45.
[82] Ma X, Ma J, Li H, Jiang Q, Gao S. PDLM: Privacy-preserving deep learning model on cloud with multiple keys. IEEE Trans Serv Comput 2018;14(4):1251–63.
[83] Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, Jin S, Quek TQ, Poor HV. Federated learning with differential privacy: Algorithms and performance analysis. IEEE Trans Inf Forensics Secur 2020;15:3454–69.
[84] Li Y, Zhou Y, Jolfaei A, Yu D, Xu G, Zheng X. Privacy-preserving federated learning framework based on chained secure multiparty computing. IEEE Internet Things J 2020;8:6178–86.
[85] Li L, Fan Y, Tse M, Lin KY. A review of applications in federated learning. Comput Ind Eng 2020;149:106854.
[86] Wang X, Han Y, Wang C, Zhao Q, Chen X, Chen M. In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning. Ieee Netw 2019;33:156–65.
[87] Qian Y, Hu L, Chen J, Guan X, Hassan MM, Alelaiwi A. Privacy-aware service placement for mobile edge computing via federated learning. Inf Sci 2019;505:562–70.
[88] Pfohl SR, Dai AM, Heller K (2019) Federated and differentially private learning for electronic health records. arXiv preprint arXiv:1911.05861. Retrieved from http://arxiv.org/abs/1911.05861.
[89] Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed Inform 2019;99:103291.
[90] Silva S, Gutman BA, Romero E, Thompson PM, Altmann A, Lorenzi M. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In: In2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019); 2019. p. 270–4.
[91] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol (TIST) 2019;10(2):1–9.
[92] Bharati S, Podder P, Thanh DN, Prasath VS. Dementia classification using MR imaging and clinical data with voting based machine learning models. Multimed Tools Appl 2022;81(18):25971–92. https://doi.org/10.1007/s11042-022-12754-x.
[93] Bharati S, Podder P, Mondal MR. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In: In2020 IEEE region 10 symposium (TENSYMP). IEEE; 2020. p. 1486–9.
[94] Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S. The future of digital health with federated learning. NPJ Digit Med 2020;3(1):119.
[95] Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25(1):37–43.
[96] Lee JS, Darcy KM, Hu H, Casablanca Y, Conrads TP, Dalgard CL, Freymann JB, Hanlon SE, Huang GD, Kvecher L, Maxwell GL. From discovery to practice and survivorship: building a national real-world data learning healthcare framework for military and veteran cancer patients. Clin Pharmacol Ther 2019;106(1):52–7.
[97] Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. IEEE Sig Process Mag 2020;37:50–60.
[98] Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, Kiddon C, Konečný J, Mazzocchi S, McMahan B, Van Overveldt T. Towards federated learning at scale: system design. Proc Mach Learn 2019;1:374–88.
[99] Carlini N, Liu C, Erlingsson Ú, Kos J, Song D. The secret sharer: evaluating and testing unintended memorization in neural networks. In: In28th USENIX Security Symposium (USENIX Security 19; 2019. p. 267–84.
[100] Duchi JC, Jordan MI, Wainwright MJ. Privacy aware learning. Adv Neural Inf Process 2012;25:1430–8.
[101] Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 2014;9:211–407.
[102] Yang X, Deng C, Wei K, Yan J, Liu W. Adversarial learning for robust deep clustering. Adv Neural Inf Process Syst 2020;33:9098–108.
[103] Hassan MU, Rehmani MH, Chen J. Differential privacy techniques for cyber physical systems: a survey. IEEE Commun Surv Tutor 2019;22:746–89.
[104] An n. Alexandre Privacy. 2019 Jul 1. Available at: https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a.
[105] Zhu T, Ye D, Wang W, Zhou W, Philip SY. More than privacy: applying differential privacy in key areas of artificial intelligence. IEEE Trans Knowl Data Eng 2020;34:2824–43.
[106] Dwork C. Differential privacy. In: Proc. 33rd Int. Conf. Automata Lang. Program; 2006. p. 1–12.
[107] Zhao Y, Chen J. Survey on differential privacy for unstructured data content. CSUR 2022;54:1–28.
[108] Frome A, Cheung G, Abdulkader A, Zennaro M, Wu B, Bissacco A, Adam H, Neven H, Vincent L. Large-scale privacy protection in google street view. In: In2009 IEEE 12th international conference on computer vision; 2009. p. 2373–80.
[109] Nautsch A, Jasserand C, Kindt E, Todisco M, Trancoso I, Evans N (2019) The GDPR & speech data: reflections of legal and technology communities, first steps towards a common understanding. arXiv preprint arXiv:1907.03458.

[110] Justin T, Štruc V, Dobrišek S, Vesnicer B, Ipšić I, Mihelič F. Speaker de-identification using diphone recognition and speech synthesis. In: In2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 4; 2015. p. 1–7.
[111] Armknecht F, Boyd C, Carr C, Gjøsteen K, Jäschke A, Reuter CA, Strand M (2015) A guide to fully homomorphic encryption. Cryptology ePrint Archive.
[112] Rivest RL, Adleman L, Dertouzos ML. On data banks and privacy homomorphisms. Found Comput Decis Sci 1978;4:169–80.
[113] Alexandre G. Homomorphic encryption & machine learning: new business models. 8 Pct 2020. Available at: https://towardsdatascience.com/homomorphic-encryption-machine-learning-new-business-models-2ba6a4f185d.
[114] Alaya B, Laouamer L, Msilini N. Homomorphic encryption systems statement: trends and challenges. Comput Sci Rev 2020;36:100235.
[115] Parmar PV, Padhar SB, Patel SN, Bhatt NI, Jhaveri RH. Survey of various homomorphic encryption algorithms and schemes. Int J Comput Appl 2014:91.
[116] Armknecht F, Boyd C, Carr C, Gjøsteen K, Jäschke A, Reuter CA, Strand M (2015) A guide to fully homomorphic encryption. Cryptology ePrint Archive.
[117] Lo'ai AT, Saldamli G. Reconsidering big data security and privacy in cloud and mobile cloud systems. J King Saud Univ - Comput Inf Sci 2021;33:810–9.
[118] Mr MM, Dhote CA, Mr DH. Homomorphic encryption for security of cloud data. Procedia Comput Sci 2016;79:175–81.
[119] Sun X, Yu FR, Zhang P, Xie W, Peng X. A survey on secure computation based on homomorphic encryption in vehicular ad hoc networks. Sensors 2020;20(15):4253.
[120] Zhou S, Yu Z, Nasr ES, Mahmoud HA, Awwad EM, Wu N. Homomorphic encryption of supervisory control systems using automata. IEEE Access 2020;8:147185–98.
[121] Cominetti EL, Simplicio MA. Fast additive partially homomorphic encryption from the approximate common divisor problem. IEEE Transact Info Forens Secur 2020;15:2988–98.
[122] Cheon JH, Hong H, Lee MS, Ryu H. The polynomial approximate common divisor problem and its application to the fully homomorphic encryption. Inf Sci 2016;326:41–58.
[123] Post-Quantum Cryptography. https://csrc.nist.gov/projects/post-quantum-cryptography. Accesses on 22 November, 2023.
[124] Canto AC, Kaur J, Kermani MM, Azarderakhsh R (2023) Algorithmic security is insufficient: a comprehensive survey on implementation attacks haunting post-quantum security. arXiv preprint arXiv:2305.13544.
[125] Anastasova M, Azarderakhsh R, Kermani MM. Fast strategies for the implementation of SIKE round 3 on ARM Cortex-M4. IEEE Transac Circ Sys I: Reg Papers 2021;68(10):4129–41.
[126] Seo H, Azarderakhsh R. Curve448 on 32-bit ARM Cortex-M4. In: InInformation Security and Cryptology–ICISC 2020: 23rd International Conference, Seoul, South Korea, December 2–4, 2020, Proceedings 23 2021. Springer International Publishing; 2021. p. 125–39.
[127] Anastasova M, Azarderakhsh R, Kermani MM, Beshaj L. Time-efficient finite field microarchitecture design for curve448 and ed448 on cortex-M4. In: InInternational Conference on Information Security and Cryptology 2022 Nov 30. Cham: Springer Nature Switzerland.; 2022. p. 292–314.
[128] Anastasova M, Bisheh-Niasar M, Azarderakhsh R, Kermani MM. Compressed SIKE round 3 on ARM Cortex-M4. In: InSecurity and Privacy in Communication Networks: 17th EAI International Conference, SecureComm 2021, Virtual Event, September 6–9, 2021, Proceedings, Part II 17 2021. Springer International Publishing; 2021. p. 441–57.
[129] Sanal P, Karagoz E, Seo H, Azarderakhsh R, Mozaffari-Kermani M. In: Kyber on ARM64: Compact implementations of Kyber on 64-bit ARM Cortex-A processors. InInternational Conference on Security and Privacy in Communication Systems 2021 Sep 6. Cham: Springer International Publishing; 2021. p. 424–40.
[130] Bisheh-Niasar M, Azarderakhsh R, Mozaffari-Kermani M. Cryptographic accelerators for digital signature based on Ed25519. IEEE Trans Very Large Scale Integr VLSI Syst 2021;29(7):1297–305.
[131] Jalali A, Azarderakhsh R, Kermani MM, Jao D. Supersingular isogeny Diffie-Hellman key exchange on 64-bit ARM. IEEE Trans Depend Secure Comput 2017;16(5):902–12.
[132] Kaur J, Kermani MM, Azarderakhsh R. Hardware constructions for error detection in lightweight authenticated cipher ASCON benchmarked on FPGA. IEEE Transac Circ Sys II: Exp Briefs 2021;69(4):2276–80.
[133] Cintas-Canto A, Kaur J, Mozaffari-Kermani M, Azarderakhsh R. ChatGPT vs. lightweight security: First work implementing the NIST cryptographic standard ASCON. In: ; 2023. arXiv:2306.08178.
[134] Kermani MM, Azarderakhsh R, Xie J. Error detection reliable architectures of Camellia block cipher applicable to different variants of its substitution boxes. In2016 IEEE Asian Hardware-Oriented Security and Trust (AsianHOST). IEEE; 2016. p. 1–6. 2016 Dec 19.
[135] Aghaie A, Kermani MM, Azarderakhsh R. Fault diagnosis schemes for low-energy block cipher Midori benchmarked on FPGA. IEEE Trans Very Large Scale Integr VLSI Syst 2016;25(4):1528–36.
[136] Kaur J, Kermani MM, Azarderakhsh R. Hardware constructions for lightweight cryptographic block cipher QARMA with error detection mechanisms. IEEE Trans Emerg Top Comput 2020;10(1):514–9.
[137] Kermani MM, Bayat-Sarmadi S, Ackie AB, Azarderakhsh R. High-performance fault diagnosis schemes for efficient hash algorithm blake. In: In2019 IEEE 10th Latin American Symposium on Circuits & Systems (LASCAS). IEEE; 2019. p. 201–4. 2019 Feb 24.
[138] Canto AC, Kermani MM, Azarderakhsh R. CRC-based error detection constructions for FLT and ITA finite field inversions over GF (2 m). IEEE Trans Very Large Scale Integr VLSI Syst 2021;29(5):1033–7.
[139] Canto AC, Sarker A, Kaur J, Kermani MM, Azarderakhsh R. Error detection schemes assessed on FPGA for multipliers in lattice-based key encapsulation mechanisms in post-quantum cryptography. IEEE Trans Emerg Top Comput 2022;11(3):791–7.
[140] Kaur J, Canto AC, Kermani MM, Azarderakhsh R. Hardware constructions for error detection in WG-29 stream cipher benchmarked on FPGA. IEEE Transac Comp-Aid Des Integr Circ Sys 2023.
[141] Canto AC, Kermani MM, Azarderakhsh R. Reliable constructions for the key generator of code-based post-quantum cryptosystems on FPGA. ACM J Emerg Technol Comput Syst 2022;19(1):1–20.
[142] Koziel B, Jalali A, Azarderakhsh R, Jao D, Mozaffari-Kermani M. In: NEON-SIDH: Efficient implementation of supersingular isogeny Diffie-Hellman key exchange protocol on ARM. InCryptology and Network Security: 15th International Conference, CANS 2016, Milan, Italy, November 14-16, 2016, Proceedings 15. Springer International Publishing; 2016. p. 88–103.
[143] Kermani MM, Azarderakhsh R, Mirakhorli M. Multidisciplinary approaches and challenges in integrating emerging medical devices security research and education. In: In2016 ASEE Annual Conference & Exposition; 2016.
[144] Mozaffari-Kermani M, Azarderakhsh R, Ren K, Beuchat JL. Guest editorial: introduction to the special section on emerging security trends for biomedical computations, devices, and infrastructures. IEEE/ACM Trans Comput Biol Bioinform 2016;13(03):399–400.
[145] Niasar MB, Azarderakhsh R, Kermani MM. Optimized architectures for elliptic curve cryptography over Curve448. Cryptol ePrint Archive 2020.
[146] Cintas-Canto A, Kermani MM, Azarderakhsh R. Reliable architectures for finite field multipliers using cyclic codes on FPGA utilized in classic and post-quantum cryptography. IEEE Trans Very Large Scale Integr VLSI Syst 2022;31(1):157–61.
[147] Karam RA, Katkoori S, Kermani MM. Work-in-progress: Hyflex hands-on hardware security education during covid-19. In: In2022 IEEE World Engineering Education Conference (EDUNINE) 2022. IEEE; 2022. p. 1–4.
[148] ISO 24028:2020. Information Technology–Artificial Intelligence–Overview of Trustworthiness in Artificial Intelligence. Standard. International Organization for Standardization; 2020.
[149] Ngan M, Grother PJ, Ngan M. Face recognition vendor test (FRVT) performance of automated gender classification algorithms. Gaithersburg, MD, USA: US Department of Commerce, National Institute of Standards and Technology; 2015.
[150] Jiang Z, Gao W, Wang L, Xiong X, Zhang Y, Wen X, Luo C, Ye H, Lu X, Zhang Y, Feng S. HPC AI500: a benchmark suite for HPC AI systems. In: Benchmarking, Measuring, and Optimizing: First BenchCouncil International Symposium, Bench 2018, Seattle, WA, USA, December 10-13, 2018, Revised Selected Papers 1. Springer International Publishing; 2019. p. 10–22.
[151] Sojda Richard S. Empirical evaluation of decision support systems: Needs, definitions, potential methods, and an example pertaining to waterfowl management. Environ Model Softw 2007;22:269–77.

[152] Chen TY, Kuo FC, Liu H, Poon PL, Towey D, Tse TH, Zhou ZQ. Metamorphic testing: A review of challenges and opportunities. CSUR 2018;51(1):1–27.
[153] Lindvall M, Ganesan D, Árdal R, Wiegand RE. Metamorphic model-based testing applied on NASA DAT–An experience report. In: In2015 IEEE/ACM 37th IEEE International Conference on Software Engineering. 2. IEEE; 2015. p. 129–38.
[154] Department for Transport (UK). The pathway to driverless cars: A code of practice for testing.
[155] BUNDESAMT F. Study:" An investigation into the performance of facial recognition systems relative to their planned use in photo identification documents–BioP I"[DISK].
[156] Wang W, Zhao M, Wang J. Effective android malware detection with a hybrid model based on deep autoencoder and convolutional neural network. J Amb Intell Humaniz Comput 2019;10(8):3035–43.
[157] Islam MZ, Islam MM, Asraf A. A combined deep cnn-lstm network for the detection of novel coronavirus (covid-19) using x-ray images. Inform Med Unlocked 2020;20:100412.
[158] Anuradha J. Big data based stock trend prediction using deep cnn with reinforcement-lstm model. Int J Syst Assur Eng Manage 2021;2:1–11.
[159] Dhyani M, Kumar R. An intelligent chatbot using deep learning with bidirectional rnn and attention model. Mater Today Proc 2021;34:817–24.
[160] Sarker H. Cyberlearning: effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multiattacks. Int Things 2021;100:393.
[161] Blumenstock J. Machine learning can help get covid-19 aid to those who need it most. Nature 2020;20:20.
[162] Saharan S, Kumar N, Bawa S. An efficient smart parking pricing system for smart city environment: a machine-learning based approach. Future Gener Comput Syst 2020;106:622–40.
[163] Hamamoto AH, Carvalho LF, Sampaio LDH, Abrão T, Proença Jr ML. Network anomaly detection system using genetic algorithm and fuzzy logic. Expert Syst Appl 2018;92:390–402.
[164] Reddy GT, Reddy MPK, Lakshmanna K, Rajput DS, Kaluri R, Srivastava G. Hybrid genetic algorithm and a fuzzy logic classifer for heart disease diagnosis. Evol Intel 2020;13(2):185–96.
[165] Kang X, Porter CS, Bohemia E. Using the fuzzy weighted association rule mining approach to develop a customer satisfaction product form. J Intell Fuzzy Syst 2020;38(4):4343–57.
[166] Fatima A, Maurya R, Dutta MK, Burget R, Masek J. Android malware detection using genetic algorithm based optimized feature selection and machine learning. In: 2019 42nd international conference on telecommunications and signal processing (TSP). IEEE; 2019. p. 220–3.
[167] Onah JO, Abdullahi M, Hassan IH, Al-Ghusham A. Genetic algorithm based feature selection and naïve bayes for anomaly detection in fog computing environment. Mach Learn Appl 2021;6:100156.
[168] Sarker IH, Khan AI, Abushark YB, Alsolami F. Mobile expert system: exploring context-aware machine learning rules for personalized decision-making in mobile applications. Symmetry (Basel) 2021;13(10):1975.