European Journal of
Technology
(EJT)

Blockchain and Machine Learning Integration for Data Privacy and Security

Lavanya Shanmugam, Monish Katari and Kumaran Thirunavukkarasu

European Journal of Technology
ISSN 2520-0712 (online)
Vol.8, Issue 3, pp 1 - 11, 2024

www.ajpojournals.org

Blockchain and Machine Learning Integration for Data Privacy and Security

Lavanya Shanmugam[1*], Monish Katari[2] and Kumaran Thirunavukkarasu[3]
[1]Tata Consultancy Services, USA
[2]Marvell Semiconductor Inc, USA
[3]Novartis, USA

Article history
Submitted 26.02.2024 Revised Version Received 02.04.2024 Accepted 06.05.2024

Abstract

Purpose: In today's world, data is gathered without any particular objective in mind; every action taken by a computer, or a human being is documented, and the data is studied in the future, if it is considered important to do so. The data will be subjected to several steps for analysis by a variety of parties, which raises the issue of trust in this context. There is a risk that the organizations involved in the analytical stages may abuse the data, which might contain private or sensitive information. Consequently, data privacy considerations should be carefully considered at this time.

Methodology/Findings: A definition of "data privacy" is the practice of limiting access to information according to how important it is. People are usually very comfortable giving out their names to strangers, but they'll wait to give out their mobile phone numbers until they're more familiar with the individual. In this era of digital technology, important personal information is often the target of individuals' efforts to protect their data. From a business's point of view, data privacy encompasses more than just employees' and consumers' private information. Data privacy issues are often believed to be a barrier to the widespread adoption of AI and ML-driven technology. The reason for this is that ML can only be trained and tested on very large data sets.

Implications to Theory, Practice and Policy: Imagine a world where trust is impossible to establish; here is where Blockchain technology might be useful. Blockchain uses the data anonymously. In this study, we provide a solution that ensures data security by integrating Blockchain technology with machine learning (Alfandi et al., 2020).

Keywords: *Machine Learning, Data Analysis, Blockchain, Integration, Data Privacy, Encryption*

1.0 INTRODUCTION

Competition for future markets and bigger market shares is increasing across all sectors thanks to machine learning and artificial intelligence. Whatever the case may be, data privacy is widely believed to be a major obstacle preventing machine learning and AI from reaching their full potential. The word "data privacy" has me confused. Can you explain it? A simple definition of data privacy would be the usage or exploitation of data for its intended purpose. Data privacy refers to people's right to decide how and for what reasons their data is collected and used. The correct management of data, including compliance with regulatory requirements, notice, and permission, is ensured by data privacy, also known as information protection, a subsection of data security.

Any unauthorized use of personally identifiable information (PII) gives rise to the possibility of a data breach or violation. Consequently, how can we remove the substantial obstacle that data privacy poses? Everyone knows that training and evaluating machine learning models requires a huge amount of data. The accuracy of the output is directly proportional to the amount of data used for training. Data that may be used broadly and data that is deemed private are often mixed up when massive amounts of data are used. Following the completion of the data-gathering process, it will be subjected to several steps of analysis; hence, it is quite probable that personal or private data will be misused by an individual or group. There is a straightforward approach to resolving this data privacy issue, and that is to refrain from collecting the data at all. But is it feasible for this to happen? Nothing is feasible in the absence of data, thus the answer is no. If there isn't any data, how can we do an analysis? At this juncture, the integration of blockchain with machine learning will provide data privacy without sacrificing the effectiveness of data analysis.

The reliability of the data analysis procedure will suffer without sensitive information. Hence, without completely disclosing personal data, before analysis starts, what happens to the private data if it is anonymized? The use of blockchain technology makes it easy to get. Additionally, there is the problem of how to distinguish between public and private data, as no data source knows how to do this. This is one scenario where machine learning might be useful for sorting out public from private data. Thus, we reasoned that by integrating Blockchain with ML, we could ensure data privacy without sacrificing analysis (Ali, Karimipour and Tariq, 2021).

Using privacy-preserving techniques like federated learning or differential privacy is another way to address data privacy concerns while still allowing for effective data analysis. To ensure that no one data point can be traced back to an individual, differential privacy adds noise to each data point before collecting them for analysis. However, with federated learning, model training can take place locally on devices or servers, with only aggregated model updates being shared centrally. By minimising the need to share raw data, this approach effectively reduces the risk of data exposure. Data masking and tokenization are strong data anonymization techniques that can further protect sensitive information while still enabling meaningful analysis. Data analysis workflows that incorporate privacy-enhancing technologies allow organisations to derive valuable insights from datasets while maintaining data privacy.

Body

Organizational and individual data privacy has grown in importance in the modern digital era. Ensuring data privacy is of utmost importance due to the ever-growing number of online platforms and the substantial amount of sensitive information being transmitted. This essay delves into how

hybrid technologies like blockchain and artificial intelligence (AI) might improve data privacy and security, providing fresh approaches to protecting sensitive information (Ekramifard et al., 2020).

AI-Enabled Data Encryption: Encrypting data is a basic security measure to prevent unauthorized individuals from accessing critical information. Conventional methods of data encryption, which depend on cryptographic algorithms, are susceptible to cryptographic vulnerabilities and brute force assaults. Contrarily, encryption algorithms driven by AI use machine learning to improve encryption protocols in response to new threats and attack trends. Ensuring the security and integrity of data throughout transmission, storage, and integration across a decentralized network is made possible by the integration of blockchain technology with AI-driven encryption (Jebamikyous et al., 2023).

Protecting Personal Information: Theft of personal information and identity theft are major worries in the digital world. To authenticate users and identify suspicious actions in real time, AI-driven identity management systems integrate biometric authentication, behavioral analytics, and anomaly detection. Individuals may safeguard their digital identities while retaining control by integrating blockchain-based identity management systems with AI-based identity verification. Identity credentials may be securely stored on the immutable blockchain, which improves data privacy and decreases the likelihood of identity fraud (Liu et al., 2020).

Data Storage and Access Control Decentralized: Cyber threats and single points of failure are inherent dangers in centralized data storage infrastructures. By distributing data over several nodes in a network, blockchain technology makes decentralized data storage possible, doing away with the need for a central authority or middleman. By combining AI-powered access control with blockchain-based data storage, organizations can maintain data privacy and confidentiality while dynamically managing data access permissions based on user roles, preferences, and behavioral patterns. This allows for granular control over data access (Mohanta et al., 2020).

Sharing Data While Keeping Privacy: It is a tough undertaking to share sensitive data across multiple organizations while keeping privacy. Data may be handled and analyzed using AI methods like homomorphic encryption and federated learning without disclosing the raw data to other parties. To ensure that data is only accessed and used according to preset rules and circumstances, blockchain smart contracts may permit safe data-sharing agreements. Collaborative data analysis and knowledge sharing are made possible with the use of AI-driven privacy-preserving methods and blockchain-based data exchange protocols. This setup also protects individual privacy rights (Waheed et al., 2021).

Regulatory Compliance and Auditability: Two crucial components of data privacy and security governance are regulatory compliance and auditability. Artificial intelligence systems can automate auditing and monitoring of compliance by constantly examining data access records, finding outliers, and producing real-time compliance reports. Blockchain technology allows for audits of regulatory compliance and accountability by providing an unchangeable and transparent record of data exchanges. Organizations may show they are following data privacy laws and industry standards by combining audit trails based on blockchain technology with AI-driven compliance monitoring. This will promote trust and openness (Alfandi et al., 2020).

Every day, the scope of technology that relies on machine learning grows at an exponential rate. The major focus is on using a key component of data analysis, which is why this is the case. A

more precise and accurate prediction will be possible if we can make better use of the data; this will lead to fewer mistakes and better overall outcomes. Therefore, making good use of data and machine learning is of utmost importance. Since the authors firmly think that the data's quality is closely related to its correctness, they focused on methods to enhance the data's quality in this research.

In this case, they raise the doubtful topic of how reliable a data set is that was collected by an organization or person without sufficient experience in theme-based data collection. Finally, they have offered the idea that a conscientious society may greatly contribute to scientists having more trust in the collected facts. The researchers behind this work laid forth a paradigm that may guarantee precision even while using an unreliable private database. This approach works best when there are many stakeholders engaged in a long process and when there are various steps to the process (Ali, Karimipour and Tariq, 2021).

Current Scenario

As artificial intelligence and machine learning-based business opportunities continue to expand and become more unavoidable. There are likely to be a great deal of obstacles to overcome for data protection and security specialists who come from a variety of backgrounds. Challenges and concerns arise when trying to establish a line between social standards and the protection of personal information while still being forthright and honest about its usage. But today, more shades of gray are likely to appear than before. Regardless, there was never a simple yes or no. Before data can be collected for analysis, a question has to be asked to establish the aim of the data collection (see Figure 1) (Ekramifard et al., 2020).

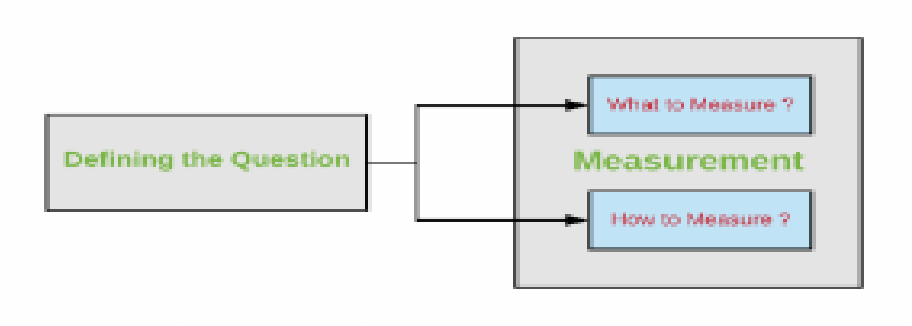


Figure 1: Data Analysis First Phase

Even though artificial intelligence and machine learning make the challenges of today's issues easier to solve, concerns about the privacy and security of data have been growing ever since the beginning of data utilization. Different parts are included in databases that include information on individuals. There are a few ways to group these parts, which usually deal with privacy and security:

i. Data Collected on Individuals (PII) Such components may almost directly identify or link to a person (e.g., an Ad-haar number, a phone number, an email address).
ii. Identifying Substitutes (QI) Such parts could be useless on their own, but when combined with other quasi-identifiers, inquiry outcomes, and other external data (such as a PIN, age, sex, etc.), they can be used to identify a person.

iii. Columns with Sensitivity These traits do not fit neatly into any of the preceding categories, but they do include personally identifiable information that has to be kept secret for several reasons (such as a person's income, HIV status, bank account information, geographic location, and a whole lot more).
iv. Gentle Columns stand for the other attributes (country, college, etc.) that don't fit neatly into the first three groups (Liu et al., 2020).

When seen through the lens of QIs, it is obvious that removing personally identifiable information (PII) portions from a dataset is not adequate for ensuring data security. To provide an example, if a dataset has crucial statistical information (which is referred to as QI), then this information might be combined with other open information sources, such as a voter registration list, to identify individuals with complete precision. Therefore, before we go into the specifics of the data privacy situation, let's have a look at the many steps that the data goes through (see Figure 2). The first stage, which is called "Defining the Goal," may not be considered a stage since it does not participate in the analysis process. However, it is the most crucial step because it determines the primary purpose of the data-collecting process. The real data-collecting process starts with the second step, which is called Data collecting. Online and offline surveys, random data collected from various locations (e.g., vehicles driving through a toll plaza), and tracking employees' arrival and departure times from the workplace are all potential ways to obtain the necessary information (Mohanta et al., 2020) .

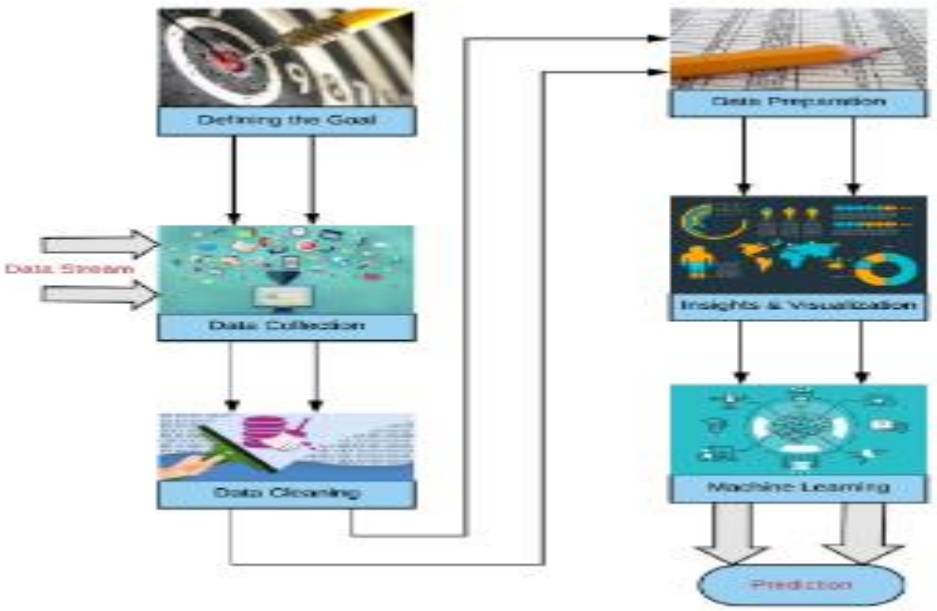


Figure 2: Phases of Data Analysis

It makes no difference what data was collected. It would be very time-consuming and provide useless results to analyze the whole dataset while including unnecessary data. Step three of data cleansing involves removing any unwanted information. The data is created for the analysis during the preparation stage, which is the fourth phase. To avoid inaccurate visualizations in the Insights and Visualization stage (the fifth level), which follows, the data must be appropriately prepared.

The penultimate step is machine learning, which is responsible for doing the analysis and making the O/P forecast. For the sake of clarity, we will keep the fifth and sixth phases separate, even if they may be combined into one. The situation of data privacy will now be examined (Waheed et al., 2021).

To ensure that the development of technology does not compromise the privacy protections of individuals, responsible data analysis is required. Avoiding the collection of information is, without a doubt, the most straightforward method for ensuring complete safety. With this technique, the major premise of the data-driven basic approach to data analysis is rendered null and void. Approaches to ensuring and protecting data privacy have, for the most part, attempted to achieve a balance between data consumption and any potential risks. There have emerged three broad methods, including the following:

i. The control of access. The strategy that is being taken here describes who is entering the database, how they are doing so, and the reason why they need the data.
ii. Data Anonymization. Altering the information to conceal the identities of individuals and the data is accomplished via the use of this technology.
iii. Privacy-protecting Data Sharing. Secure Multiparty Computing (SMC), which guarantees that private information may be obtained remotely while maintaining a specified and controlled level of security, is the fundamental component of this technique (Alfandi et al., 2020).

Motivation

Data science aims to create experiences that include the use of several technologies, such as Big Data, Machine Learning, Probability Statistics, Data Mining, Data Visualization, etc., to identify patterns and designs about the environment. Several examples show how many companies are improving their efficiency every day by including a data-driven decision-making component. This led to further data exploitation, which in turn affected society greatly due to worries over careless data use.

Blockchain technology has developed into a system that securely stores and transfers data in an easy-to-understand, transparent, and fault-tolerant way. The technology that relies on distributed records makes this a real possibility. With the blockchain, any company might be able to be decentralized, secure, efficient, and transparent. If we could merge the two most talked-about technologies right now, blockchain and machine learning, it would be a giant leap forward in keeping personal information private. As a consequence, a framework for integration has been created that may deliver valuable data without endangering people's privacy or security. Before delving into the integration of blockchain technology with machine learning, it is crucial to address the methods that safeguard user privacy in machine learning (Ali, Karimipour and Tariq, 2021).

Machine Learning in Data Privacy

The preparation of machine learning models requires a variety of inputs to be worked together to avoid missing out on the opportunity to use private data or information in their distinctive structure. This was accomplished via the utilization of cryptographic methods or the utilization of information that is classified as differently private (annoyance systems). The use of differential protection is especially effective when good effects are anticipated (Ekramifard et al., 2020).

Cryptographic Approaches

The cryptography approach to privacy in machine learning often employs three distinct types of techniques. Secret sharing, homomorphic encryption, and jumbled circuits are all examples of such methods.

i. Homomorphic Encryption: It is possible to make the computation of encrypted data more complex by using the homomorphic encryption method.
ii. The technique known as "Garbled Circuits" enables two parties to securely compute concurrently. Without the intervention of a third party, the two unknown participants evaluate their respective roles and responsibilities in this situation over an encrypted channel.

iii. Secret Sharing: This is a technique of cryptography strategy that involves dividing the data into many parts and then sharing those parts with multiple parties. Thus, the data may only be obtained after the parties involved combine their respective interests inside the firm. In some cases, the data possessor, sometimes called the dealer, is also the one who sets and decides the requirements for multiple shares needed to reconstruct the whole file (Liu et al., 2020).

Approaches from the Perturbation

This approach uses a series of mathematical processes to provide an approximate value for a given problem. It does this by starting with a precise answer to a similar but relatively easy issue. Differential privacy (DP) strategies use perturbation methods within the domain of privacy-preserving machine learning (PPML). The following are two categories into which the Perturbation Approach could be classified:

One such architecture is Differential Privacy (DP), which enables the unrestricted exchange of dataset data. It does this by displaying the dataset's collection instances while protecting the personal information of the individuals contained in the dataset.

Secondly, there is differential privacy on a local level, which is a subset of differential privacy that has certain limitations. If someone has access to someone's private data, they still won't be able to see all of that person's information (Mohanta et al., 2020).

Integrating with Blockchain

The data processing procedures, starting with data collection, and ending with the production of the forecast, were addressed in the preceding part of the present scenario. Numerous phases are there to carry out the processing, and there are several parties involved in the data process flow. The consequence is a high risk of data theft, leakage, or improper use. The biggest and most pressing problem that needs fixing is the misuse of personal data. So, they propose a new framework for the data analysis process that was previously employed to fix this problem. The newly proposed architecture uses blockchain and machine intelligence to protect user privacy. Figure 3 shows not just the whole process but also the integration of the Blockchain module into the data processing cycle.

The data collection phase follows the goal stage, which follows the determination of the data analysis process's purpose. Data is gathered in two ways: first, at random, and second, from

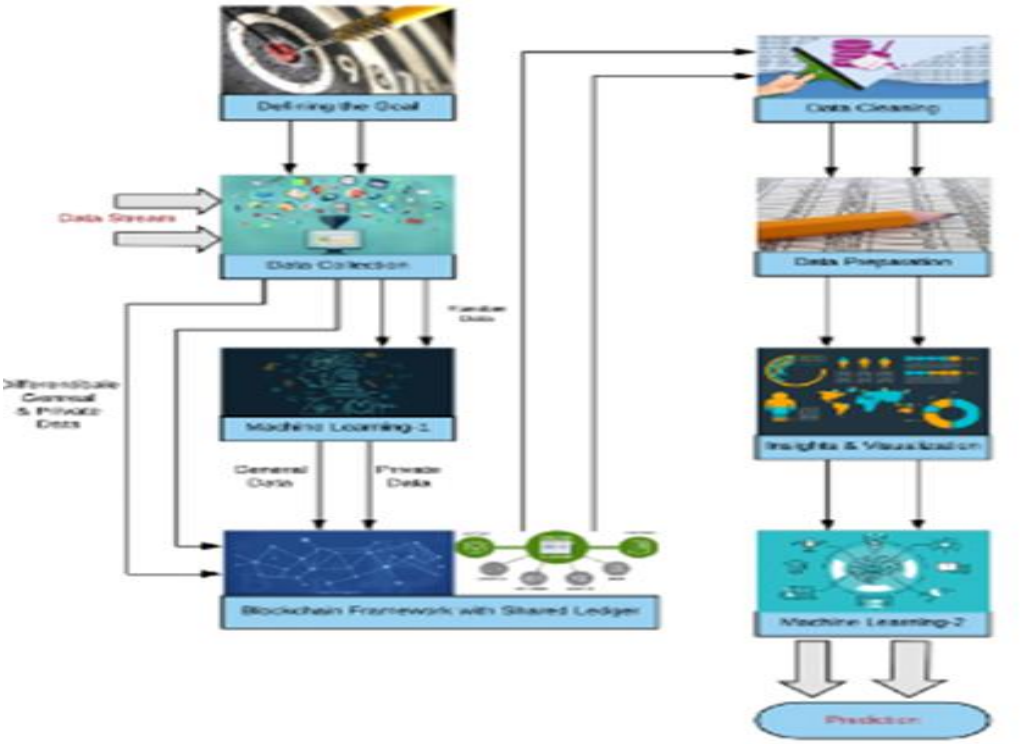
customers themselves. When the data is collected arbitrarily, it is impossible to determine which data should be classified as private and which data should not be classified as private until the data has been collected. The information is gathered directly from the users, with the users having the ability to choose by themselves which information should be deemed private and which information should not be considered private. After the data collection step, two distinct data streams are generated, and they are as follows:

i. General and private data may be sorted and organized.
ii. Because it is entirely random, there is no private or universal way to classify random data. Machine learning 1, which has been taught to distinguish between private and general data, is tasked with classifying and categorizing the second stream of data that follows the first random data stream. The present stage of Machine Learning-1 is followed directly by the blockchain module, which receives a direct feed of the first category of the data stream. Here we shall go into further detail about the Blockchain module (Waheed et al., 2021).

When data is collected from several users, each user has the power to decide which data should be treated as private and which should be treated as public, provided they have a clear knowledge of what data is considered general and what is considered private. to guarantee the immediate transmission of data to the blockchain module. The Machine Learning-1 module will be loaded with personal data if the user lacks the knowledge to distinguish between public and private data.

The objective of the Machine Learning-1 module is to group and label public and private data simultaneously. We may train the computer by feeding it the first kind of data stream information gathered directly from responsible and educated humans. Users may choose to operate their node or insert a machine-generated node into the blockchain network when data is put into the Blockchain module. Users can also classify their own private and public data. The Machine Learning-1 Module makes it possible for those who can independently generate their Blockchain network nodes to do so, and then the Blockchain network module will include these nodes.

Every node's private data is encrypted, and the only way to determine who owns it is to use a hash value. The future of machine intelligence and blockchain integration is almost impossible to foresee. Any and every data may be found in the shared ledger. After this is finished, the data analysis will continue as previously described. Afterward, the data is recorded in the shared ledger and moved on to the data purification stage (Alfandi et al., 2020).



2.0 CONCLUSION AND RECOMMENDATIONS

The issue of appropriate data usage has become more pressing in light of the exponential growth in data collection rates over the last few years. As the number of people using data continues to rise, the respectability of personal information is also growing. This essay explains why the current hot technologies machine learning and blockchain must work together. This research provides a comprehensive description of the data flow throughout the whole analytical process, from goal formation to the prediction stage. At that location, it became apparent that data corruption and inappropriate use of sensitive information were concerns. It goes much beyond by outlining the specific ways in which machine learning may be used independently to address the issue of data privacy. Then it gets down to the article's meat and potatoes: talking about how blockchain technology and machine learning go hand in hand. Thanks to the integration, a necessary fix for the most delicate issue of data privacy may be executed. Consequently, data analysis as a whole will be made safer, more efficient, intelligent, resilient, and decentralized.

The real-time analysis and implementation of the framework for integrating machine learning with blockchain technology have not been finished yet, but it has been provided in this study. This might serve as a model for many other industries, such as healthcare, banking, electronic voting, electronic voting, and military, to name a few. Put simply, there are still many challenges to be solved in the vast field of blockchain technology combined with machine intelligence. By incorporating all data sources into the blockchain infrastructure as nodes, processing time may be reduced. One of the limitations of the proposed model is this. To tackle this issue, you may either employ a high-performance processor or a parallel processing approach. Also, we expect that a

complex algorithm to overcome the limitations will be created soon. This study aims to examine the integration at a fundamental level and to explore possible future research areas that might be impacted by our findings.

Given the present state of affairs, blockchain technology has shown great promise due to its ability to eliminate trust issues. The problem of trust emerges whenever data flows between many parties participating in the analytic process. By establishing a decentralized trust-less environment, blockchain technology eliminates that issue. In order to safeguard sensitive information, the authors of this study recommended a privacy-aware public key infrastructure (PKI) system and went over the steps to build a permissioned blockchain network. This study discusses how blockchain may be used to promote trustworthiness. In this article, the writers touched on the potential uses of a blockchain architecture in situations where data tampering is a possibility. Also, they didn't include real-time analysis in their decentralised blockchain architecture proposal. This paper outlined the authors' vision for a transparent and tamper-proof use of the blockchain architecture that complies with all applicable data privacy standards. Data transmission for a networking system is shown by the authors of this study using a mix of Machine Learning and Blockchain. In addition, they mapped out all the places where Blockchain and Machine Learning have found use and laid up a detailed outline of the intersections between the two. This paper offers a look into how blockchain technology is going to progress in the future.

REFERENCES

Alfandi, O., Khanji, S., Ahmad, L. and Khattak, A. (2020). A survey on boosting IoT security and privacy through blockchain. *Cluster Computing*, 08(07). doi:https://doi.org/10.1007/s10586-020-03137-8.

Ali, M., Karimipour, H. and Tariq, M. (2021). Integration of Blockchain and Federated Learning for Internet of Things: Recent Advances and Future Challenges. *Computers & Security*, 08(09), p.102355. doi:https://doi.org/10.1016/j.cose.2021.102355.

Ekramifard, A., Amintoosi, H., Seno, A.H., Dehghantanha, A. and Parizi, R.M. (2020). A Systematic Literature Review of Integration of Blockchain and Artificial Intelligence. *Advances in Information Security*, 07(9), pp.147–160. doi:https://doi.org/10.1007/978-3-030-38181-3_8.

Jebamikyous, H., Li, M., Suhas, Y. and Kashef, R. (2023). Leveraging machine learning and blockchain in E-commerce and beyond: benefits, models, and application. *Discover Artificial Intelligence*, 3(1). doi:https://doi.org/10.1007/s44163-022-00046-0.

Liu, Y., Yu, F.R., Li, X., Ji, H. and Leung, V.C.M. (2020). Blockchain and Machine Learning for Communications and Networking Systems. *IEEE Communications Surveys & Tutorials*, [online] 22(2), pp.1392–1431. doi:https://doi.org/10.1109/COMST.2020.2975911.

Mohanta, B.K., Jena, D., Satapathy, U. and Patnaik, S. (2020). Survey on IoT Security:Challenges and Solution using Machine Learning, Artificial Intelligence and Blockchain Technology. *Internet of Things*, 11(07), p.100227. doi:https://doi.org/10.1016/j.iot.2020.100227.

Waheed, N., He, X., Ikram, M., Usman, M., Hashmi, S.S. and Usman, M. (2021). Security and Privacy in IoT Using Machine Learning and Blockchain. *ACM Computing Surveys*, 53(6), pp.1–37. doi:https://doi.org/10.1145/3417987.