Towards Trustworthy AI: Blockchain-based Architecture

1st Vipul Popat
School of Computing
Dublin City University
Dublin, Republic of Ireland
vipulpopat@gmail.com

2nd Vishal Padwal
School of Computing
Dublin City University
Dublin, Republic of Ireland
padwal.vishal@gmail.com

***Abstract*—Artificial intelligence (AI) systems are being rapidly adopted, yet concerns persist around reliability, transparency, and ethical alignment. Trustworthy AI that adheres to principles of lawfulness, ethics, and robustness is required. This paper explores integrating blockchain technology, given its decentralized and tamper-proof properties, to enhance AI trustworthiness. We propose an architecture that stores AI model information and metadata as hashes on blockchain, maintaining provenance records while preserving privacy. Smart contracts encode policies to add accountability. A proof-of-concept federated learning prototype demonstrates the architecture storing encrypted data and model hashes on Interplanetary File System (IPFS) and blockchain. Results indicate the architecture enhances transparency through the decentralized ledger and accountability via smart contracts. However, optimizations in scalability, fairness, and explainability are needed. This research provides valuable insights into using blockchain for secure and reliable AI development, though limitations around privacy and decentralization's implications remain. This pioneering architecture offers unprecedented transparency and accountability for federated learning, fostering collaboration and Trustworthy AI adoption.**

Keywords—Blockchain, AI, federated learning, Trustworthy AI, Model Aggregation

I. Introduction

Artificial intelligence (AI) systems are rapidly being adopted across various domains, yet concerns persist around their reliability, transparency, and alignment with ethical values. Incidents like biased algorithms that discriminate against certain demographics have highlighted the need for AI systems to be trustworthy, adhering to principles of lawfulness, ethics, and robustness. This paper explores the potential of blockchain technology, given its tamper-proof and decentralized properties, to enhance the trustworthiness of AI systems.

Blockchain provides a ledger for recording transactions in a secure, transparent, and immutable manner across a decentralized network. Smart contracts on blockchain enable rules-based execution and accountability. Integrating blockchain technology into AI development and deployment could address key challenges like data privacy, provenance tracking, and bias mitigation. Prior works have proposed blockchain-based approaches for explainable AI and data governance, but a holistic architecture encompassing the multiple facets of trustworthy AI remains unexplored.

This paper aims to develop an architecture that combines blockchain technology with AI to foster accountable and ethical systems. The architecture stores hashed information about AI models, training data, and relevant metadata on blockchain to maintain data provenance records while preserving privacy. Smart contracts encode policies to add accountability regarding data sharing and model usage. A proof-of-concept federated learning prototype demonstrates the feasibility of this architecture by storing encrypted data and model hashes on blockchain and IPFS (Interplanetary File System).

In summary, this paper makes important contributions towards trustworthy and responsible AI by proposing a blockchain-based architecture for provenance tracking, privacy preservation, and accountability. The demonstrated federated learning prototype is a stepping stone towards applying this architecture across real-world AI systems. Further enhancements to scalability, fairness and explainability would be vital to realize the full potential of this transformative approach. By fostering transparency and accountability, this architecture represents significant progress in establishing trustworthy AI ecosystems.

II. Technical Background

A. Preliminaries and definitions

Trustworthy AI is not a monolithic concept but rather a polylithic one [28]. There are several interpretations of different terms in this field. The definition and explanation of these terms are therefore essential before we can use them. There are several key definitions of terms relating to AI in this section.

- Black-box problem: This refers to a system that is opaque, and it is difficult to track its structure, internal workings, and implementation [28]. It is becoming increasingly difficult to comprehend AI systems as they become more complex [28]. As a result, the trustworthiness of the system is reduced since it is difficult to explain the reasoning behind the output system to different users that it interacts with.
- Explainable and interpretable AI: This refers to the development of models that can be explained and interpreted.
- Reliability: System reliability refers to ensuring that the system performs as it is intended to, that is, within specified limits and without any failures, producing the same outputs for the same inputs consistently.
- Fairness: A fair system ensures that there is no discrimination or favouritism toward individuals or groups based on their inherent or acquired characteristics that are irrelevant to the decision-making process [28].

- Trust: Technology defines trust as "the confidence that one element has in another, that that second element will behave as expected."[28]
- Acceptance: User acceptance of an AI system depends on its willingness to be used in service encounters.
- Trustworthy AI: Trustworthy AI refers to a framework which ensures that a system can be trusted based on the evidence provided by its stated requirements. By doing so, it ensures that users' and stakeholders' expectations are met in a verifiable manner [28]. Guidelines for achieving Trustworthy AI are based on [23] They encompass three components: Lawful AI, Ethical AI, and Robust AI. [23]

In summary, Trustworthy AI guidelines focus on Lawful, Ethical, and Robust AI, promoting legal compliance, ethical alignment, and safe development for reliable AI systems [23]

B. Requirement to make trustworthy AI

As a basis for fostering "responsible competitiveness," an assured approach ensures that all individuals impacted by AI systems can have confidence in their design, creation, and implementation adhering to lawful, ethical, and resilient standards. To foster responsible and sustainable AI advancement in Europe, these Guidelines have been formulated.

It is their goal to make ethics a core pillar of developing a distinctive approach to AI that benefits, empowers, and protects both the individual human flourishing and the common good. As a result, we believe Europe will be able to establish itself as a global leader in cutting-edge AI, worthy of our individual and collective trust. By ensuring trustworthiness, Europeans will be able to fully benefit from AI systems' benefits, secure in the knowledge that safeguards will be in place to protect them from potential risks [30] [31].

Listed below are several requirements that are not exhaustive. The concept encompasses systemic, individual, as well as societal factors:

- Human agency and oversight: AI systems should support human autonomy and decision-making, together with human oversight and intervention.
- Technical robustness and safety: AI systems must be resilient, reliable, and secure, and their development must be guided by a focus on preventing and minimizing unintended consequences.
- Data privacy and governance: AI systems should be governed appropriately in terms of data privacy and protection, as well as quality, integrity, and access.
- Transparency: AI systems should be transparent in terms of their data, systems, and business models. When humans interact with AI systems, they must be informed of both their capabilities and limitations. There is a need for AI systems to avoid unfair bias and provide accessibility and universal design to ensure diversity, non-discrimination, and fairness. There should be consideration and involvement of stakeholders who may be affected by AI systems.
- Social and environmental well-being: AI systems must be environmentally friendly and sustainable, considering the broader community and other sentient beings. It is also necessary to consider the impact on institutions and democracy.
- Accountability: Mechanisms must be implemented to ensure responsibility, accountability, and redress for AI systems and their outcomes. The adverse effects should be identified, assessed, documented, and minimized.[23]

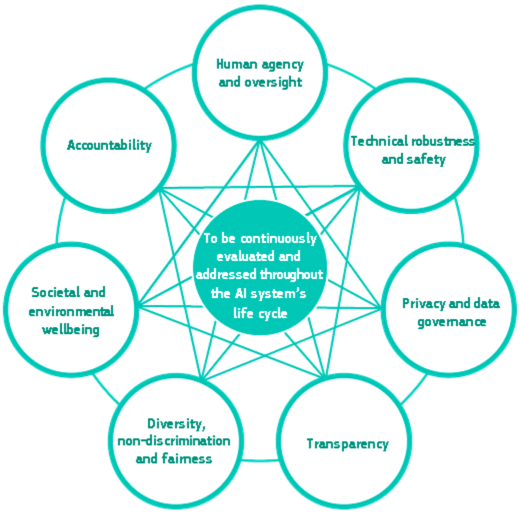

Human agency and oversight
Accountability
Technical robustness and safety
To be continuously evaluated and addressed throughout the AI system's life cycle
Societal and environmental wellbeing
Privacy and data governance
Diversity, non-discrimination and fairness
Transparency

Fig. 1. Interrelationship of the seven requirements [23]

C. Overview of blockchain and features

Blockchain is a decentralized digital ledger that securely records transactions across multiple computers. Each block contains multiple transactions and is linked to the previous block using cryptography, ensuring no single point of failure. Originally developed for Bitcoin, blockchain has found diverse applications in supply chain management, digital identity, and voting systems. It eliminates the need for intermediaries, resulting in faster, cheaper, and more secure, transparent, and accountable transactions. There are three types of blockchains: public, accessible to anyone; private, limited to select participants; and consortium, controlled by collaborating organizations. While poised to transform industries and disrupt traditional models, blockchain faces challenges of scalability, interoperability, and regulatory issues. [20] [19]

In order to make blockchain technology unique and valuable for a variety of applications, it has several key characteristics:

- Decentralization: Blockchain is decentralized in the sense that there is no single point of control or central authority. Due to the absence of a single point of failure, the system is more secure and resilient.
- Immutable records: Once a transaction has been recorded on a blockchain, it cannot be altered. As a result, a permanent and unalterable record of transactions is created, which can enhance transparency and accountability.
- Security of transactions: Blockchain technology utilizes cryptography to protect transactions and

prevent unauthorized access. As a result, transactions are kept private and secure.

- Transparency: All transactions recorded on the blockchain are visible to all participants, thus making the system transparent. As a result, trust in the system can be enhanced, and fraud and malicious activity may be more easily detected.
- Interoperability: Blockchains are interoperable, which means they can communicate with other systems. By doing so, different blockchain systems are able to exchange data and value.
- Smart contracts: Blockchain technology supports the use of smart contracts, which are self-executing code snippets that enforce a contract's terms. By eliminating intermediaries, agreements can be executed automatically and securely.
- Consensus mechanism: Blockchains use a consensus mechanism in order to ensure that all participants are in agreement with respect to the ledger's state. The integrity of the system is maintained by preventing disputes.[24][25][26]

A wide variety of use cases can be addressed by blockchain technology, including supply chain management, digital identity, voting systems, and financial services. Blockchain technology is well suited for applications where trust is critical due to its decentralized and secure nature.

III. Related work

This section explores the existing research and developments in the field of trustworthy AI. It delves into seminal works that laid the foundation for the current study. Various approaches and methodologies adopted by prior researchers are critically analyzed, highlighting their strengths and identifying potential gaps for the current study to address.

A. *Trustworthy AI frameworks*

The principles of trustworthy AI can be enforced through a number of frameworks. The following are examples:

- A framework for ethical development, deployment, and use of AI developed by the EU, based on the seven principles of trustworthy AI. [23]
- MIT-IBM Watson AI Lab's AI Accountability framework, which provides a framework for ensuring AI systems are accountable.
- The Partnership on AI's principles for responsible AI, which outline best practices and guidelines for designing, developing, deploying, and using AI responsibly. [31]
- Algorithm Accountability Act, a proposed piece of legislation in the US that would ensure that AI algorithms will be transparent, explainable, and fair [32].
- The FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for scientific data management and stewardship can be applied to AI systems to ensure responsible data management. [30]
- The International Organization for Standardization (ISO) standards on AI provide guidelines for the ethical and responsible development and use of AI [30].
- The AI Now Institute's AI Policy and Practice Playbook provides a practical guide to organizations and policymakers seeking to implement ethical and responsible AI practices [1].
- A strategy for responsible development and use of AI is outlined in the Long-Term AI Strategy of the AI Alignment Forum, based on principles of transparency, accountability, and alignment of value.
- IEEE Advancing technology for Humanity, ethically aligned designed first edition. [35]

Frameworks like these play a crucial role in supporting individuals and organizations as they develop and implement AI applications that can be trusted. However, the actual effectiveness of these frameworks depends on the legal and regulatory landscape in each specific jurisdiction [27].

These frameworks offer valuable guidance to ensure AI systems are designed and utilized in accordance with the principles of trustworthy AI. Nevertheless, it's essential to recognize that the application of these frameworks will be contingent on the unique circumstances and context of each situation.

Considerable efforts have been devoted to creating frameworks for Trustworthy AI, but they often target specific use cases and may not encompass all seven principles. One proposal suggests that significant decisions in complex AI systems should involve a consensus among distributed AI and explainable AI (XAI) agents hosted by trusted oracles, presuming the majority of these agents are reliable [12]. Blockchain technology has emerged as a promising solution to meet trustworthiness requirements, guarding against biases and adversarial attacks. Some use cases have demonstrated the effectiveness of combining blockchain smart contracts with decentralized storage to achieve a trustworthy XAI system.

Amidst the era of extensive data collection and computing, a framework was introduced to develop trustworthy AI systems with a data-centric level of abstraction, addressing ethical concerns within AI and Data Science [17]. This framework focused on the ethics of data and algorithms, emphasizing direct integration into AI system design and development.

TABLE I: COMPARISON OF TRUSTWORTHY AI FRAMEORK FOR NEGLECTED PRINCIPLES.

Framework Reference	Neglected Principles (Full or Partial)
[12]	• Human agency and oversight • Technical robustness and safety • Privacy and data governance • Diversity, non-discrimination and fairness
[17]	• Privacy and data governance • Technical robustness and safety • Diversity, non-discrimination and fairness • Societal and environmental well-being
[16]	• Human agency and oversight • Privacy and data governance • Diversity, non-discrimination and fairness • Societal and environmental well-being
[15]	• Technical robustness and safety • Diversity, non-discrimination and fairness.

Another endeavour sought to unify existing approaches to trustworthy AI by considering the entire AI system lifecycle, offering actionable steps for practitioners and stakeholders to enhance AI trustworthiness[16]. Identifying key opportunities and challenges, this work emphasized the need for a paradigm shift towards comprehensively trustworthy AI systems. As we approach the 6G era, AI/ML mechanisms will be integral to systems, necessitating mechanisms to ensure trust, explainability, and reliability in their operations [15].In the context of Industry 5.0, a proposed architecture combines Industry 4.0 paradigms and AI-based decision support while maintaining human-centricity and trustworthiness elements [14]. It aims to strike a balance between the benefits of AI-centric digitalization and the role of humans in critical decision-making processes.

Lastly, trustworthiness is identified as a central requirement for the acceptance and success of AI centered around human needs [13]. To determine whether an AI system is trustworthy, its behaviour and characteristics are assessed against a gold standard of trustworthy AI, which may be in the form of guidelines, requirements, or expectations.

B. *How could blockchain enable trustworthy AI*

Through blockchain technology, trusted AI can be created by providing a decentralized and tamper-proof ledger for storing data and executing smart contracts. In this way, the data used to train AI models is secure, transparent, and immutable, preventing malicious actors from manipulating it and compromising the accuracy of the models. Further, smart contracts on the blockchain can enforce specific rules and conditions for the use and deployment of AI models, creating an additional layer of accountability and trust.

By implementing blockchain technology, anonymous users are able to conduct transactions without the involvement of a third-party intermediary [22].

Big data and standardization can be incorporated into the future scope of blockchain technology. The use of blockchain technology in conjunction with big data allows the creation of a secure cryptographic layer to protect confidential data and copyright forms.

AI is currently gaining popularity and applicability across a wide range of fields due to the proliferation of big data and the increasing computational power. Critical decision-making systems suffer from a major drawback when it comes to the lack of explanations for the decisions made by AI algorithms of today. AI algorithms that provide explanations for their AI decisions are called XAI [12].

Decentralized Application (DApp) users' technical trustworthiness is a function of consensus, economic models, and incentives for honesty, explainability, and robustness of predictors. A number of additional infrastructure requirements are also necessary, including security, privacy, reliability, usability, dependability, performance, and governance. Blockchain technology appears to be the most suitable option, if not the only one, to meet these requirements. [3] However, several challenges remain to be overcome, including reducing the presence of humans in the loop when validating explanations, and ensuring accurate timeliness for certain applications.

A combination of blockchain technology and trustworthy AI has the potential to bring new levels of security and transparency to a variety of industries. Blockchains can be used to analyze and process data in real-time using AI algorithms, providing valuable insights and reducing the risk of fraud. In addition, the decentralized nature of blockchain ensures that data is not controlled by a single entity. This reduces the risk of data manipulation and enhances overall trust in the system. Thus, blockchain technology and trustworthy AI can be combined to create a secure and transparent ecosystem that benefits individuals and organizations alike.

IV. PROPOSED ARCHITECTURE

This section introduces a pioneering architecture that distinguishes itself through its innovative approach. The subsequent section expounds on the components depicted in Fig. 2, elucidating their intricate details. By proposing this unique framework, the study seeks to push the boundaries of current knowledge and address prevailing challenges. Thoroughly elucidating the architecture's design and functionalities, the paper lays the foundation for further exploration and practical implementation. The comprehensive breakdown of each element ensures readers gain a clear comprehension of the system's inner workings and its potential applications. Ultimately, this research significantly contributes to the field by presenting a cutting-edge and promising solution.

A. *Centralized learning at server*

The central entity's orchestrator (C1) in Fig. 2. initiates a model training task, starting with the creation of a global model and training hyperparameters. These details are shared with the client's orchestrator (E6). On the central entity's side, the model combiner (C2) awaits the local model parameters from all clients for model aggregation. Once the aggregation is complete, the updated global model is pushed to IPFS. The hashed value of the global model version is uploaded to the blockchain for provenance purposes and shared with edge locations by the model deployer (C3) for the next training round. The model deployer selects edge locations based on performance or resource availability. By default, all edge locations receive updates for fairness.

B. *Decentralized learning process at edge*

The edge devices begin by collecting raw data, which is then pre-processed (e.g., scaled, noise reduced) by the data processor (E3). This training data is used by the model trainer (E5) for local model training. The training data version for each epoch is hashed and uploaded to the blockchain for data-model provenance. The model trainer sets up the local model training environment based on the training job received from the central location. After each local epoch, the local model is transferred to the model evaluator (E7) for performance assessment. The hashed value of the local model versions and their performance are recorded and uploaded to the blockchain for data-model provenance. The performance and other parameters are collected by the model combiner (C2) on the central location. The orchestrator (E6) then waits for updated global model parameters from the model deployer (C3). The client orchestrator (E6) checks if the required federation epochs are achieved and decides when to terminate the training job. If termination occurs, the last global model version is deployed to the model users. Otherwise, the process repeats until the designated federation epoch is reached. When the training process ends, the orchestrator on the edge location stops local training and deploys the last global model. The model evaluator (E7) then assesses the real-world data inference performance of the deployed global model.

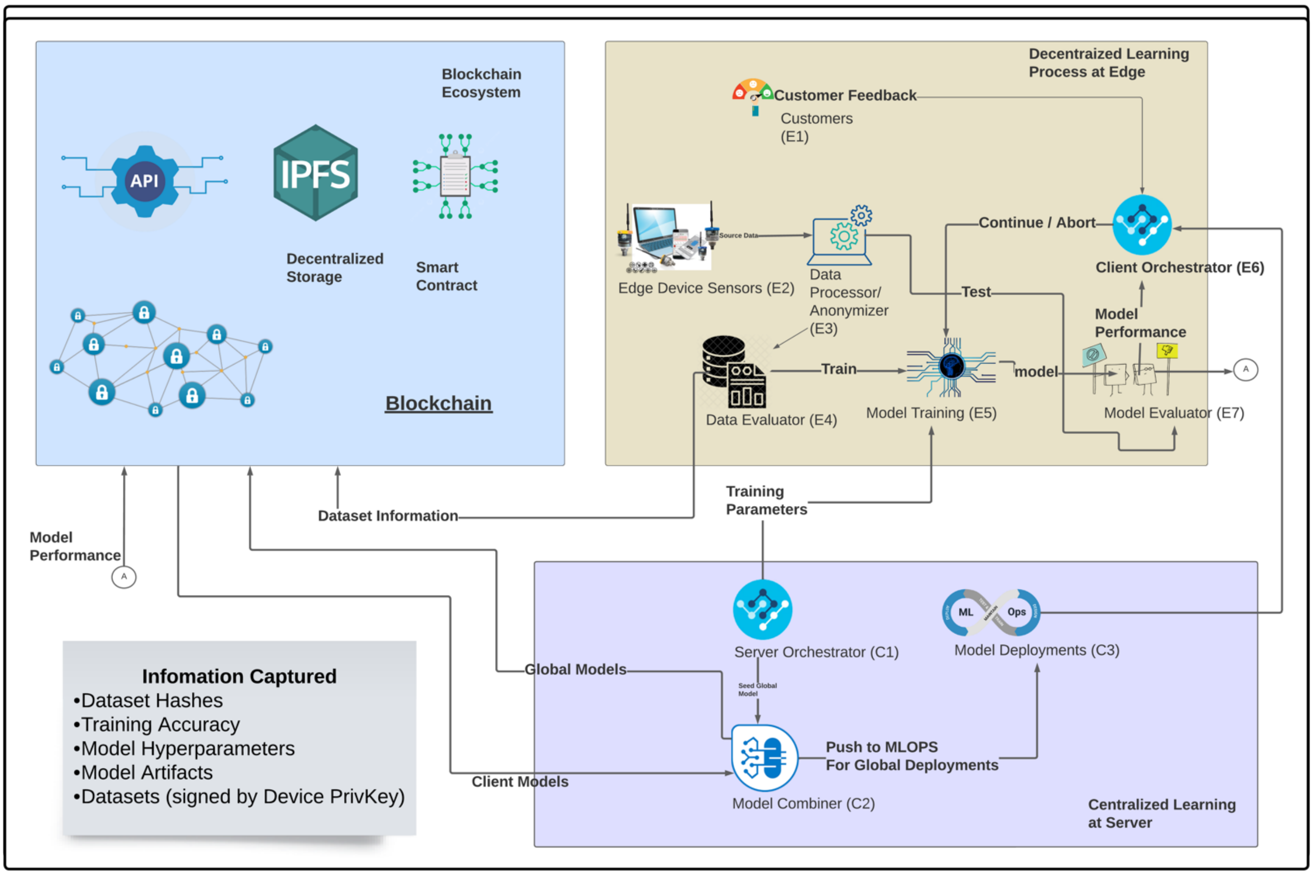


Fig. 2. Proposed Architecture

C. Blockchain ecosystem

Blockchain technology has the potential to address fairness and accountability concerns in federated AI/ML. By leveraging the decentralized and transparent nature of blockchain, it becomes possible to track and verify the integrity of data used in federated learning processes. Blockchain can ensure that data contributions from different participants are securely recorded, preventing unauthorized modifications or biases. Additionally, smart contracts on the blockchain can enforce fair and transparent rules for data sharing and model aggregation, providing a mechanism for consensus and accountability among the federated learning participants. This combination of transparency, immutability, and automated governance offered by blockchain can help establish a fairer and more trustworthy environment for federated AI/ML, reducing biases and ensuring accountable decision-making.

Model aggregation in federated AI/ML refers to the process of combining the locally trained models from multiple participating devices or edge nodes to create a global model that represents the collective knowledge. It is a crucial step in federated learning where the aim is to leverage the distributed data while maintaining privacy and minimizing communication costs.

D. Model aggregation

The model aggregation process typically follows a iterative approach. Initially, a global model is initialized, which can be a pre-trained model or a randomly initialized one. Then, the participating devices independently train their local models using their respective local datasets. The local models are trained using various machine learning algorithms, such as gradient descent, and are optimized based on local data.

The model aggregation process is an iterative approach commonly used in federated AI/ML. Initially, a global model is set up, either pre-trained or randomly initialized. Participating devices independently train their local models using their own datasets and machine learning algorithms like gradient descent. Once local training is complete, the local models are sent to a central server for aggregation. The server combines their parameters through techniques like model averaging, weighted averaging, or federated averaging. In federated averaging, the server calculates a weighted average based on the number of samples or the importance of local data.

The updated global model is then shared with participating devices, and the process is repeated for several rounds until convergence. This way, local models learn from their data while preserving privacy. Privacy-preserving techniques like differential privacy or secure multi-party computation can be employed during aggregation to protect individual data confidentiality. Federated AI/ML allows distributed devices to contribute their knowledge while respecting privacy and minimizing data transfer, making it a powerful approach for decentralized machine learning.

E. Architecture connections

The proposed architecture for federated learning incorporates several design choices that aim to address the challenges of data privacy and centralization while leveraging the collective intelligence of distributed devices. These choices are driven by the need to protect sensitive user data, improve model performance through diverse datasets, and ensure efficient and scalable training.

The decentralized nature of federated learning is a key design choice. By allowing training to occur on local devices or edge nodes, the architecture ensures that sensitive user data remains on the devices where it is generated. This addresses privacy concerns by minimizing the risk of data breaches and unauthorized access. Keeping data decentralized also reduces reliance on a central server, leading to faster and more efficient training as computations can be performed locally.

Another important design choice is the aggregation of model updates in a decentralized manner. Instead of centralizing the data and training process, the proposed architecture aggregates the model updates from participating devices. This approach enables the incorporation of knowledge from diverse datasets, capturing real-world variations and improving model performance. It also allows for the preservation of data privacy since only aggregated model updates are shared, rather than the raw data.

The architecture also emphasizes accountability and fairness through the integration of blockchain technology. Blockchain provides transparency and immutability, enabling the tracking and verification of data contributions in federated learning. Smart contracts on the blockchain enforce fair and transparent rules for data sharing and model aggregation, ensuring accountability among participants. This design choice addresses concerns regarding biased or manipulated data and promotes a more trustworthy and equitable learning environment.

Overall, the design choices in the proposed architecture prioritize data privacy, model performance, efficiency, accountability, and fairness. By decentralizing the training process, aggregating model updates, and leveraging blockchain technology, the architecture provides a solution that empowers devices to learn collaboratively while maintaining privacy, reducing biases, and advancing AI/ML in a decentralized and ethical manner.

In summary, the proposed architecture for federated learning addresses the challenges of data privacy and centralization in AI/ML. It leverages the decentralized nature of local devices or edge nodes to perform training while ensuring sensitive user data remains on the devices, reducing privacy risks. The aggregation of model updates from participating devices allows for the incorporation of diverse datasets, improving model performance and capturing real-world variations.

The integration of blockchain technology adds transparency, immutability, and accountability to the architecture. Blockchain enables the tracking and verification of data contributions, enforcing fair and transparent rules for data sharing and model aggregation. This helps address biases and ensures trustworthy decision-making processes.

However, the architecture also has limitations, including communication constraints, data heterogeneity, lack of central control, limited access to global knowledge, privacy and security risks, and scalability challenges. These limitations necessitate ongoing research and development to overcome them and enhance the effectiveness of federated learning.

Overall, the proposed architecture prioritizes data privacy, model performance, efficiency, accountability, and fairness. By decentralizing the training process, aggregating model updates, and leveraging blockchain technology, it enables collaborative learning while maintaining privacy, reducing biases, and advancing AI/ML in a decentralized and ethical manner. With continued innovation, federated learning holds significant promise for shaping the future of decentralized and privacy-preserving machine learning.

V. PROTOTYPE

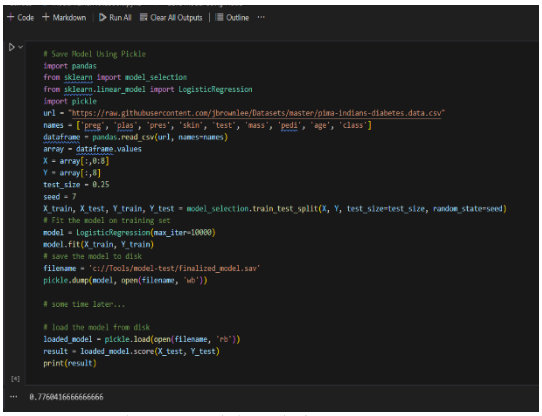
This section outlines the prototype that was developed in accordance with the architecture described earlier.

A. Technology stack

- Python Notebook – Training/Testing/Creating an AI Model
- JSON: Documentation of metadata
- Python – Scripting Language
- PKI – Public Key Infrastructure
- IPFS –Immutable Decentralized Storage
- Solidity – Smart Contracts on EVM
- Matic Mumbai Testnet – Deployment and execution of smart contracts
- Node/NPM – Development of Web3 UI for smart contract invocation
- Blockchain Explorer: Tool to identify blockchain transactions and its associated metadata.

B. Order of execution

- Model training/testing: Execute the python notebook to generate a model based on train/test data

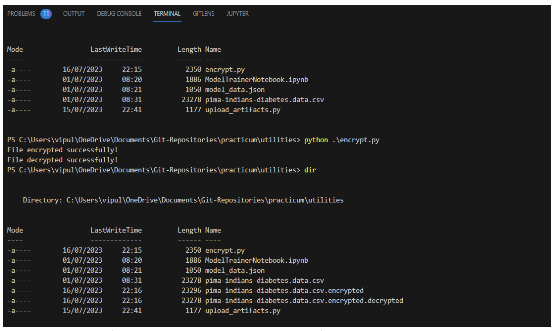


```
# Save Model Using Pickle
import pandas
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
import pickle
url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
dataframe = pandas.read_csv(url, names=names)
array = dataframe.values
X = array[:,0:8]
Y = array[:,8]
test_size = 0.25
seed = 7
X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X, Y, test_size=test_size, random_state=seed)
# Fit the model on training set
model = LogisticRegression(max_iter=10000)
model.fit(X_train, Y_train)
# save the model to disk
filename = 'c://Tools/model-test/finalized_model.sav'
pickle.dump(model, open(filename, 'wb'))

# some time later...

# load the model from disk
loaded_model = pickle.load(open(filename, 'rb'))
result = loaded_model.score(X_test, Y_test)
print(result)
```

```
0.7760416666666666
```

Fig. 3. Model Train/test

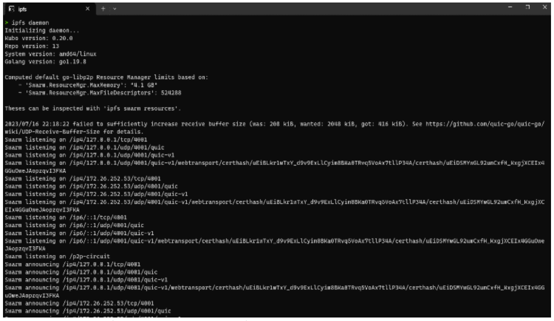
- Dataset encryption: Encrypt the Dataset using the python script provided in order to preserve the privacy of the dataset



```
Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        16/07/2023     22:15           2350 encrypt.py
-a----        01/07/2023     08:20           1886 ModelTrainerNotebook.ipynb
-a----        01/07/2023     08:21           1050 model_data.json
-a----        01/07/2023     08:31          23278 pima-indians-diabetes.data.csv
-a----        15/07/2023     22:41           1177 upload_artifacts.py


PS C:\Users\vipul\OneDrive\Documents\Git-Repositories\practicum\utilities> python .\encrypt.py
File encrypted successfully!
File decrypted successfully!
PS C:\Users\vipul\OneDrive\Documents\Git-Repositories\practicum\utilities> dir


    Directory: C:\Users\vipul\OneDrive\Documents\Git-Repositories\practicum\utilities


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        16/07/2023     22:15           2350 encrypt.py
-a----        01/07/2023     08:20           1886 ModelTrainerNotebook.ipynb
-a----        01/07/2023     08:21           1050 model_data.json
-a----        01/07/2023     08:31          23278 pima-indians-diabetes.data.csv
-a----        16/07/2023     22:16          23296 pima-indians-diabetes.data.csv.encrypted
-a----        16/07/2023     22:16          23278 pima-indians-diabetes.data.csv.encrypted.decrypted
-a----        15/07/2023     22:41           1177 upload_artifacts.py
```

Fig. 4. Dataset Encryption

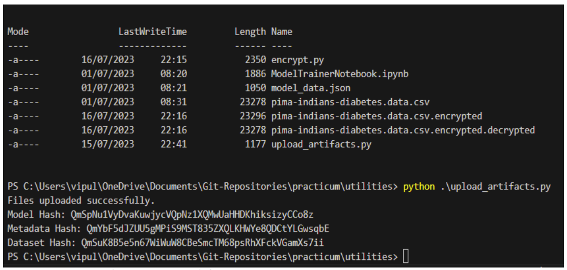
- IPFS node: Operate a local IPFS node in order to consistently push files to IPFS



```
> ipfs daemon
Initializing daemon...
Kubo version: 0.20.0
Repo version: 13
System version: amd64/linux
Golang version: go1.19.8

Computed default go-libp2p Resource Manager limits based on:
    - 'Swarm.ResourceMgr.MaxMemory': "4.1 GB"
    - 'Swarm.ResourceMgr.MaxFileDescriptors': 524288

Theses can be inspected with 'ipfs swarm resources'.

2023/07/16 22:18:22 failed to sufficiently increase receive buffer size (was: 208 kiB, wanted: 2048 kiB, got: 416 kiB). See https://github.com/quic-go/quic-go/
wiki/UDP-Receive-Buffer-Size for details.
Swarm listening on /ip4/127.0.0.1/tcp/4001
Swarm listening on /ip4/127.0.0.1/udp/4001/quic
Swarm listening on /ip4/127.0.0.1/udp/4001/quic-v1
Swarm listening on /ip4/127.0.0.1/udp/4001/quic-v1/webtransport/certhash/uEiBLkr1wTxY_d9v9ExLlCyim8BKa0TRvq5VoAx7tllP34A/certhash/uEiDSMYwGL92umCxfH_KxgjXCEIx4
GGuOweJAopzqvI3FKA
Swarm listening on /ip4/172.26.252.53/tcp/4001
Swarm listening on /ip4/172.26.252.53/udp/4001/quic
Swarm listening on /ip4/172.26.252.53/udp/4001/quic-v1
Swarm listening on /ip4/172.26.252.53/udp/4001/quic-v1/webtransport/certhash/uEiBLkr1wTxY_d9v9ExLlCyim8BKa0TRvq5VoAx7tllP34A/certhash/uEiDSMYwGL92umCxfH_KxgjXC
EIx4GGuOweJAopzqvI3FKA
Swarm listening on /ip6/::1/tcp/4001
Swarm listening on /ip6/::1/udp/4001/quic
Swarm listening on /ip6/::1/udp/4001/quic-v1
Swarm listening on /ip6/::1/udp/4001/quic-v1/webtransport/certhash/uEiBLkr1wTxY_d9v9ExLlCyim8BKa0TRvq5VoAx7tllP34A/certhash/uEiDSMYwGL92umCxfH_KxgjXCEIx4GGuOwe
JAopzqvI3FKA
Swarm listening on /p2p-circuit
Swarm announcing /ip4/127.0.0.1/tcp/4001
Swarm announcing /ip4/127.0.0.1/udp/4001/quic
Swarm announcing /ip4/127.0.0.1/udp/4001/quic-v1
Swarm announcing /ip4/127.0.0.1/udp/4001/quic-v1/webtransport/certhash/uEiBLkr1wTxY_d9v9ExLlCyim8BKa0TRvq5VoAx7tllP34A/certhash/uEiDSMYwGL92umCxfH_KxgjXCEIx4GG
uOweJAopzqvI3FKA
Swarm announcing /ip4/172.26.252.53/tcp/4001
Swarm announcing /ip4/172.26.252.53/udp/4001/quic
```

Fig. 5. Leveraging IPFS

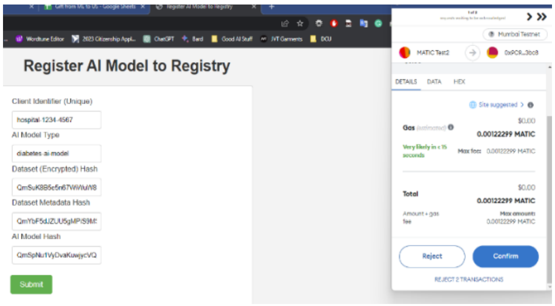
- Push artifacts to IPFS: Push the encrypted dataset, dataset metadata and AI model to IPFS using the script provided and retrieve the hashes generated in this process.



```
Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        16/07/2023     22:15           2350 encrypt.py
-a----        01/07/2023     08:20           1886 ModelTrainerNotebook.ipynb
-a----        01/07/2023     08:21           1050 model_data.json
-a----        01/07/2023     08:31          23278 pima-indians-diabetes.data.csv
-a----        16/07/2023     22:16          23296 pima-indians-diabetes.data.csv.encrypted
-a----        16/07/2023     22:16          23278 pima-indians-diabetes.data.csv.encrypted.decrypted
-a----        15/07/2023     22:41           1177 upload_artifacts.py


PS C:\Users\vipul\OneDrive\Documents\Git-Repositories\practicum\utilities> python .\upload_artifacts.py
Files uploaded successfully.
Model Hash: QmSpNu1VyDvaKuwjycVQpNz1XQMwUaHHDKhiksizyCCo8z
Metadata Hash: QmYbF5dJZUU5gMPiS9MST835ZXQLKHWYe8QDCtYLGwsqbE
Dataset Hash: QmSuK8B5e5n67WiWuW8CBeSmcTM68psRhXFckVGamXs7ii
PS C:\Users\vipul\OneDrive\Documents\Git-Repositories\practicum\utilities> 
```

Fig. 6. Artifacts storage on IPFS

- Push Artifacts Hashes to Blockchain: Using the web3 UI, push the artifact hashes and client identifiers to blockchain using the smart contract deployed on Matic Test net. The method used for registering the AI model to the registry is `registerModel`



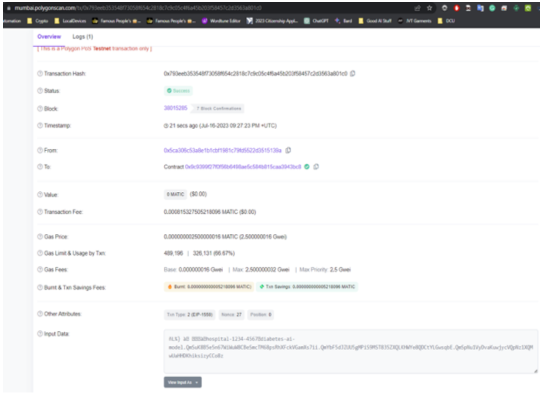
Register AI Model to Registry

Client Identifier (Unique)
hospital-1234-4567
AI Model Type
diabetes-ai-model
Dataset (Encrypted) Hash
QmSuK8B5e5n67WiWuW8
Dataset Metadata Hash
QmYbF5dJZUU5gMPiS9M
AI Model Hash
QmSpNu1VyDvaKuwjycVQ

Gas (estimated) $0.00 0.00122299 MATIC
Very likely in < 15 seconds Max fee: 0.00122299 MATIC
Total $0.00 0.00122299 MATIC
Amount + gas fee Max amount: 0.00122299 MATIC

Fig. 7. Smart contract Invocation

- Verify the submitted information on blockchain using the blockchain explorer.



[This is a Polygon PoS Testnet transaction only]

Transaction Hash: 0x793eeb353548f73058f654c2818c7c9c05c4f6a45b203f58457c2d3563a801c0
Status: Success
Block: 38015285 7 Block Confirmations
Timestamp: 21 secs ago (Jul-16-2023 09:27:23 PM +UTC)
From: 0x5ca306c53a8e1b1cbf1981c79fd5522d3515139a
To: Contract 0x9c9399f27f0f56b6498ae5c584b815caa3943bc8
Value: 0 MATIC ($0.00)
Transaction Fee: 0.000815327505218096 MATIC ($0.00)
Gas Price: 0.000000002500000016 MATIC (2.500000016 Gwei)
Gas Limit & Usage by Txn: 489,196 | 326,131 (66.67%)
Gas Fees: Base: 0.000000016 Gwei | Max: 2.500000032 Gwei | Max Priority: 2.5 Gwei
Burnt & Txn Savings Fees: Burnt: 0.000000000005218096 MATIC) Txn Savings: 0.000000000005218096 MATIC
Other Attributes: Txn Type: 2 (EIP-1559) Nonce: 27 Position: 0
Input Data: hospital-1234-4567 diabetes-ai-model.QmSuK8B5e5n67WiWuW8CBeSmcTM68psRhXFckVGamXs7ii.QmYbF5dJZUU5gMPiS9MST835ZXQLKHWYe8QDCtYLGwsqbE.QmSpNu1VyDvaKuwjycVQpNz1XQMwUaHHDKhiksizyCCo8z

Fig. 8. Blockchain Explorer

The Solidity contract `AIModelRegistry` includes the following methods:

- `registerModel`: Registers a new AI model with the provided client identifier, model type, and hashes.
- `getModelByIndex`: Retrieves the AI model details at the given index.
- `getModelsByType`: Retrieves an array of model indexes that match the given model type.
- `getAIModelByClientIdentifier`: Retrieves the AI model details with the given client identifier.

The method signatures offered present a defined interface specification for interacting with the AIModelRegistry contract. This smart contract interface is utilized to procure the model hashes from the blockchain, which subsequently facilitates the acquisition of model and dataset information from IPFS. Ultimately, this data is employed by the Model Combiner (C2) to consolidate the model data. The consolidated models are then pushed via MLOps to the edge for additional training.

VI. EVALUATION

Trustworthy AI and Blockchain Architecture, while distinct domains, share common principles that can create synergies. Trustworthy AI focuses on building AI systems that are lawful, ethical, transparent, robust, overseen by humans, non-discriminatory, and accountable. Blockchain Architecture, on the other hand, is about designing decentralized, immutable ledgers with consensus mechanisms, cryptography, smart contracts, tokens, and scalability. Both domains emphasize transparency, accountability, and robustness. While not all principles directly correlate, understanding these overlaps can be beneficial. For instance, blockchain could enhance AI's transparency and accountability by recording its decision-making process, while AI could improve blockchain network efficiency. As technology advances, we can expect more interactions between these areas.

A. *What has be acheived via prototype and expansions*

The presented prototype designed in a way that prioritizes data privacy, model performance, efficiency, accountability, and fairness. In terms of data privacy, blockchain uses cryptographic measures to protect data from unauthorized access.

- Privacy

In Fig. 4, the dataset at edge devices undergoes encryption, enabling federated learning to aggregate models without accessing the actual data. This process ensures data privacy and security, as sensitive information remains protected. The central server can improve the global model using encrypted data from multiple edge devices without compromising individual user data. This privacy-preserving approach fosters trust between users and the federated learning system, encouraging valuable data contributions to enhance the model's performance without data exposure concerns.

- Transparency

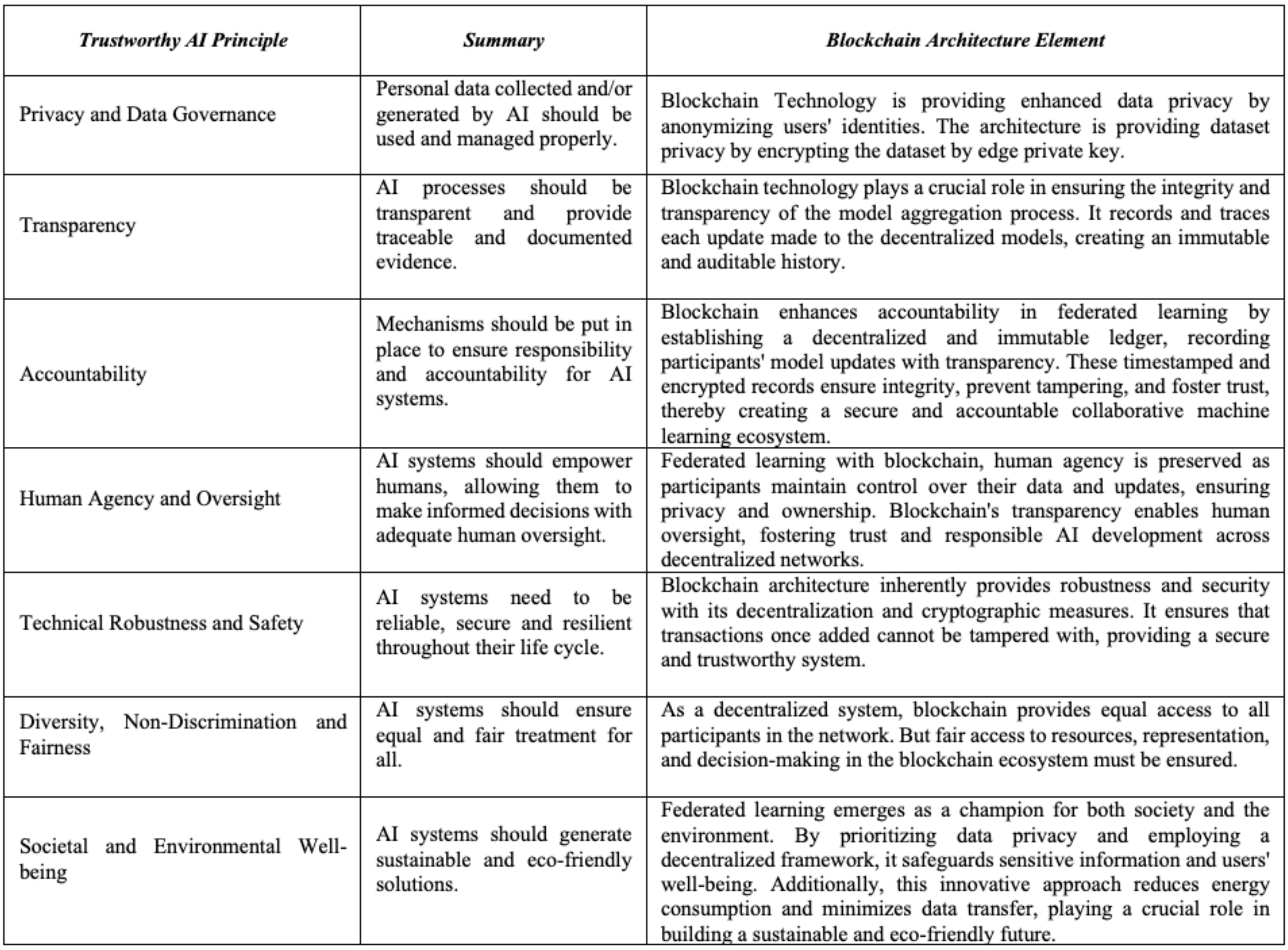
Transparency in federated learning is crucial for trust and accountability. Fig. 5. and Fig. 6. of protype shows utilization of IPFS and blockchain which enhances transparency by storing model metadata hash, model hash, and dataset hash. Model metadata and model hashes are uploaded to IPFS, allowing public access and verification. Dataset hashes are also stored on IPFS before training begins. Blockchain records these hashes, ensuring an immutable and transparent ledger. IPFS and blockchain provide decentralization, data integrity, accessibility, and auditing capabilities. Transparency enhances trust and accountability among federated learning participants, and stakeholders can verify the process's fairness and compliance.

- Accountability

Transparent nature of blockchain allows all participants to verify transactions, thereby ensuring accountability. Smart contracts also automate the enforcement of rules and agreements, further enhancing accountability as shown in Fig. 7. and Fig. 8.

The Table II summarizes the implications of the EU's principles of Trustworthy AI against the developed prototype. The proposed architecture is designed to encompass key elements of trustworthy AI, such as privacy, transparency, and accountability. It is equipped to safeguard user data, maintain clear visibility into decision-making processes, and hold responsible parties accountable for their actions. Moreover, the architecture extends its capabilities to uphold all other principles essential for trustworthy AI, ensuring fairness, robustness, and inclusivity. By encompassing these fundamental tenets, the proposed framework aims to instill confidence and reliability in AI systems, fostering trust between users and the technology they engage with.

TABLE II
EVALUATION OF THE EU PRINCIPLE OF TRUSTWORTHY AI AGAINST PROPOSED ARCHITECTURE



Trustworthy AI Principle	*Summary*	*Blockchain Architecture Element*
Privacy and Data Governance	Personal data collected and/or generated by AI should be used and managed properly.	Blockchain Technology is providing enhanced data privacy by anonymizing users' identities. The architecture is providing dataset privacy by encrypting the dataset by edge private key.
Transparency	AI processes should be transparent and provide traceable and documented evidence.	Blockchain technology plays a crucial role in ensuring the integrity and transparency of the model aggregation process. It records and traces each update made to the decentralized models, creating an immutable and auditable history.
Accountability	Mechanisms should be put in place to ensure responsibility and accountability for AI systems.	Blockchain enhances accountability in federated learning by establishing a decentralized and immutable ledger, recording participants' model updates with transparency. These timestamped and encrypted records ensure integrity, prevent tampering, and foster trust, thereby creating a secure and accountable collaborative machine learning ecosystem.
Human Agency and Oversight	AI systems should empower humans, allowing them to make informed decisions with adequate human oversight.	Federated learning with blockchain, human agency is preserved as participants maintain control over their data and updates, ensuring privacy and ownership. Blockchain's transparency enables human oversight, fostering trust and responsible AI development across decentralized networks.
Technical Robustness and Safety	AI systems need to be reliable, secure and resilient throughout their life cycle.	Blockchain architecture inherently provides robustness and security with its decentralization and cryptographic measures. It ensures that transactions once added cannot be tampered with, providing a secure and trustworthy system.
Diversity, Non-Discrimination and Fairness	AI systems should ensure equal and fair treatment for all.	As a decentralized system, blockchain provides equal access to all participants in the network. But fair access to resources, representation, and decision-making in the blockchain ecosystem must be ensured.
Societal and Environmental Well-being	AI systems should generate sustainable and eco-friendly solutions.	Federated learning emerges as a champion for both society and the environment. By prioritizing data privacy and employing a decentralized framework, it safeguards sensitive information and users' well-being. Additionally, this innovative approach reduces energy consumption and minimizes data transfer, playing a crucial role in building a sustainable and eco-friendly future.

B. How the architecture can be improved

- *Improved Fairness on AI Model*

Improving an AI model's fairness involves addressing data sampling biases. By carefully selecting and balancing training data, we can mitigate unfair outcomes. Firstly, identify underrepresented groups and oversample their instances to ensure adequate representation. Secondly, apply data augmentation techniques to create synthetic data, enhancing the model's understanding of diverse examples. Thirdly, consider incorporating fairness-aware algorithms that penalize biased predictions during training. Additionally, leveraging post-processing techniques to re-calibrate predictions can help correct disparities. Lastly, continuous monitoring of model performance and feedback loops with diverse stakeholders is essential for iterative improvements. By addressing data sampling biases, we foster a more equitable and unbiased AI model.

- *Scalability of blockchain platform*

A blockchain-based solution can enhance scalability and stability through various approaches. Implementing sharding allows the network to be divided into smaller segments, enabling parallel processing of transactions and enhancing scalability. By adopting a consensus algorithm like Proof-of-Stake (PoS), resource-intensive mining is reduced, improving efficiency and stability. Additionally, employing off-chain solutions, such as state channels or sidechains, can alleviate network congestion and enhance scalability. Continual protocol upgrades and optimizing smart contracts can further enhance stability, ensuring that the system remains resilient and secure. Combined, these measures create a more scalable and stable blockchain ecosystem, capable of handling increased transaction volumes and maintaining network integrity.

- *Adoption of generative AI*

To enhance a MLOps-based solution with generative AI and continuous learning, one can implement an iterative approach for model training. This involves integrating feedback loops from users, feeding data generated by the model back into the training process, and fine-tuning the model in real-time. Implementing active learning techniques can optimize data selection for model improvement. Furthermore, employing ensemble methods, such as model stacking, can boost performance and enhance generative capabilities. Employing version control for models and data ensures traceability and reproducibility. Adopting cutting-edge research in generative AI and integrating it into the pipeline helps maintain a state-of-the-art model for better adaptability and value delivery.

- *Explainable AI*

To adopt XAI in the proposed architecture for trustworthy AI, key steps must be implemented. Firstly, design the architecture to incorporate interpretable models, like decision trees or rule-based systems. Secondly, utilize feature attribution methods to understand model predictions. Thirdly, create an interface for users to access explanations. Fourthly, ensure transparency in data collection and processing, adhering to ethical guidelines. Fifthly, conduct rigorous testing and validation to ensure the explanations are accurate and reliable. Lastly, continuously update and improve the system based on user feedback and emerging research. By embracing XAI, the architecture can build trust and acceptance, making AI more accountable and reliable for users.

VII. CONCLUSION

The proposed architecture for federated learning using blockchain offers a groundbreaking approach with unprecedented transparency, security, and accountability. By leveraging blockchain technology, the architecture stores model metadata hash, model hash, and dataset hashes on a decentralized and tamper-proof ledger. This ensures the integrity and authenticity of information while maintaining participants' data privacy through hash representation. The decentralized nature of the blockchain fosters a trustless environment, with no single entity having control over the data. Auditing by researchers, regulators, and stakeholders becomes seamless, verifying model updates' fairness and regulatory compliance. This innovative architecture instills confidence in federated learning, unlocking its potential to address complex challenges across various industries and pushing the technology to new heights

The proposed architecture delivers substantial value in multiple key areas. In terms of model aggregation, the decentralized nature ensures that models from diverse sources are aggregated securely and efficiently, leading to robust and accurate global models. Decentralized learning empowers individual participants to train locally, preserving data privacy while contributing to the collective intelligence. The architecture's blockchain-based transparency enhances trust, enabling stakeholders to validate model updates and dataset contributions. Privacy is safeguarded by using hashes for datasets, preserving confidentiality while sharing essential information. Ultimately, this holistic approach in federated learning yields unparalleled value, fostering collaboration, protecting privacy, ensuring transparency, and producing high-quality models for solving complex challenges across various domains.

The future of AI and blockchain technologies hinges on four key areas: fairness in AI models, scalability of blockchain platforms, adoption of generative AI, and XAI.

Fairness in AI models can be improved by addressing data sampling biases, utilizing fairness-aware algorithms, and implementing continuous monitoring and feedback loops. This will foster more equitable and unbiased AI models. Scalability of blockchain platforms can be enhanced through sharding, consensus algorithms like Proof-of-Stake, off-chain solutions, and continual protocol upgrades. This will create a more scalable and stable blockchain ecosystem. Adoption of generative AI can be achieved through iterative model training, active learning techniques, ensemble methods, and version control. This will maintain a state-of-the-art model for better adaptability and value delivery. XAI can be adopted by incorporating interpretable models, feature attribution methods, user interfaces for explanations, transparency in data collection, and rigorous testing. This will make AI more accountable and reliable for users.

Future work should focus on refining these strategies, exploring new techniques, and integrating them into a cohesive framework. This will ensure the development of robust, fair, scalable, and XAI and blockchain systems.

REFERENCES

[1] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, 'Trustworthy AI: A Review', ACM Computing Surveys, vol. 55, no.2. Association for Computing Machinery, 2023.

[2] S. K. Lo et al., 'Towards Trustworthy AI: Blockchain-based Architecture Design for Accountability and Fairness of Federated Learning Systems', IEEE Internet of Things Journal, 2022.

[3] R. D. Garcia, G. Sankar Ramachandran, R. Jurdak, and J. Ueyama, A Blockchain-based Data Governance with Privacy and Provenance: A case study for e-Prescription. Institute of Electrical and Electronics Engineers Inc., 2022.

[4] YIANNAS FRANK, 'A NEW ERA OF FOOD TRANSPARENCY POWERED BY BLOCKCHAIN', 2019.

[5] M. K. Ahuja et al., 'Opening the Software Engineering Toolbox for the Assessment of Trustworthy AI', 2020.

[6] A. S. Rajasekaran, M. Azees, and F. Al-Turjman, 'A comprehensive survey on blockchain technology', Sustainable Energy Technologies and Assessments, vol. 52, 2022.

[7] Z. Yang, Y. Shi, Y. Zhou, Z. Wang, and K. Yang, 'Trustworthy Federated Learning via Blockchain', IEEE Internet of Things Journal, vol. 10, no. 1, pp. 92–109, 2023.

[8] K. Dey and U. Shekhawat, 'Blockchain for sustainable e-agriculture: Literature review, architecture for data management, and implications', Journal of Cleaner Production, vol. 316, 2021.

[9] S. K. Mangla, Y. Kazancoglu, E. Ekinci, M. Liu, M. Özbiltekin, and M. D. Sezer, 'Using system dynamics to analyze the societal impacts of blockchain technology in milk supply chainsrefer', Transportation Research Part E: Logistics and Transportation Review, vol. 149, 2021.

[10] W. Alshahrani and R. Alshahrani, Assessment of Blockchain technology application in the improvement of pharmaceutical industry. Institute of Electrical and Electronics Engineers Inc., 2021.

[11] R. Tian, L. Kong, X. Min, and Y. Qu, Blockchain for AI: A Disruptive Integration. Institute of Electrical and Electronics Engineers Inc., 2022.

[12] C. Baru, Institute of Electrical and Electronics Engineers, and IEEE Computer Society, 2019 IEEE International Conference on Big Data : proceedings : Dec 9 - Dec 12, 2019, Los Angeles, CA, USA. .

[13] Y. Li, 'Emerging blockchain-based applications and techniques', Service Oriented Computing and Applications, vol. 13, no. 4. Springer, pp. 279–285, Dec-2019.

[14] S. Barmpounakis and P. Demestichas, Framework for Trustworthy AI/ML in B5G/6G. Institute of Electrical and Electronics Engineers Inc., 2022.

[15] U. Wajid, A. Nizamis, and V. Anaya, 'Towards Industry 5.0-A Trustworthy AI Framework for Digital Manufacturing with Humans in Control', 2022.

[16] J. de V. Mohino, J. B. Higuera, J. R. B. Higuera, and J. A. S. Montalvo, 'The application of a new secure software development life cycle (S-SDLC) with agile methodologies', Electronics (Switzerland), vol. 8, no. 11, 2019.

[17] M. D. Wilkinson et al., 'Comment: The FAIR Guiding Principles for scientific data management and stewardship', Scientific Data, vol. 3, 2016.

[18] M. Nassar, K. Salah, M. H. ur Rehman, and D. Svetinovic, 'Blockchain for explainable and trustworthy AI', Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 1, 2020.

[19] M. C. N. Yates JoAnne, 'The International Organization for Standardization (ISO)', 2009.

[20] D. P. Srivasthav, L. P. Maddali, and R. Vigneswaran, Study of Blockchain Forensics and Analytics tools. Institute of Electrical and Electronics Engineers Inc., 2021.

[21] B. Liu, Overview of the Basic Principles of Blockchain. Institute of Electrical and Electronics Engineers Inc., 2021.

[22] O. Ali, A. Jaradat, A. Kulakli, and A. Abuhalimeh, 'A Comparative Study: Blockchain Technology Utilization Benefits, Challenges and Functionalities', IEEE Access, vol. 9, pp. 12730–12749, 2021.

[23] High-Level Expert Group on AI, 'HIGH-LEVEL EXPERT GROUP ON AISET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINES FOR TRUSTWORTHY AI', 2019

[24] F. Yang, L. Lei, and H. Zhu, Overview of Blockchain and Cloud Service Integration. Institute of Electrical and Electronics Engineers Inc., 2022.

[25] S. Utomo, A. John, A. Rouniyar, H. C. Hsu, and P. A. Hsiung, Federated Trustworthy AI Architecture for Smart Cities. Institute of Electrical and Electronics Engineers Inc., 2022.

[26] S. Ahmadjee, C. Mera-Gómez, R. Bahsoon, and R. Kazman, 'A Study on Blockchain Architecture Design Decisions and Their Security Attacks and Threats', ACM Transactions on Software Engineering and Methodology, vol. 31, no. 2, 2022.

[27] B. Li et al., 'Trustworthy AI: From Principles to Practices', ACM Computing Surveys, vol. 55, no. 9, pp. 1–46, 2023.

[28] R. Xu and J. Joshi, 'Trustworthy and transparent third-party authority', ACM Transactions on Internet Technology, vol. 20, no. 4, Nov. 2020.

[29] D. Murray-Rust, C. Elsden, B. Nissen, E. Tallyn, L. Pschetz, and C. Speed, 'Blockchain and Beyond: Understanding Blockchains Through Prototypes and Public Engagement', ACM Transactions on Computer-Human Interaction, vol. 29, no. 5, 2023.

[30] A. Kumar, T. Braud, S. Tarkoma, and P. Hui, 'Trustworthy AI in the Age of Pervasive Computing and Big Data; Trustworthy AI in the Age of Pervasive Computing and Big Data', 2020.

[31] Y. Xu, C. Zhang, Q. Zeng, G. Wang, J. Ren, and Y. Zhang, 'Blockchain-Enabled Accountability Mechanism against Information Leakage in Vertical Industry Services', IEEE Transactions on Network Science and Engineering, vol. 8, no. 2, pp. 1202–1213, 2021.

[32] A. Rustemi, V. Atanasovski, and A. Risteski, Overview of Blockchain Data Storage and Privacy Protection. Institute of Electrical and Electronics Engineers Inc., 2022.

[33] D. Almeida, K. Shmarko, and E. Lomas, 'The ethics of facial recognition technologies, surveillance, and accountability in an age of AI: a comparative analysis of US, EU, and UK regulatory frameworks', AI and Ethics, vol. 2, no. 3, pp. 377–387, 2022.

[34] THOMPSON KEN, 'Reflections on Trusting Trust', 1984.

[35] P. W. Staecker, "Research, education and training in the context of advancing technology for the benefit of humanity: An IEEE view," *2012 IEEE/MTT-S International Microwave Symposium Digest,* Montreal, QC, Canada, 2012, pp. 1-3, doi: 10.1109/MWSYM.2012.6259618.