SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V13I8P107

Volume 13 Issue 8, 70-82, August 2026

Original Article

Blockchain-Enabled AI-Driven Big Data Analytics for Secure and Scalable IoT Ecosystems

Hema Malini G.B[1], Agnes Sheila S.P[2], Anitha S[3], Subalakshmi K[4], Jeevitha S[5]

[1,2,3,4,5]*Department of Computer Applications, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Tamil Nadu, India.*

[1]*Corresponding Author : gbhmalini@yahoo.co.in*

Received: 02 March 2026 Revised: 10 April 2026 Accepted: 24 July 2026 Published: 25 August 2026

Abstract - *The rapid evolution of Internet of Things (IoT) ecosystems produces large quantities of heterogeneous data streams, resulting in major issues related to security, scalability, privacy, and real-time data analytics. This paper suggests a novel framework of a blockchain-based and AI-oriented big data analytics system that supports secure and scalable IoT-based smart grid environments. The proposed architecture combines data acquisition using IoT, big data processing using a distributed system, intelligent predictions using machine learning techniques such as LightGBM, Random Forest, and XGBoost, and integrity preservation using a blockchain-based system, such as Hyperledger Fabric. In the proposed system, the Energy Efficiency Scores are predicted using advanced regression techniques, and LightGBM performs better with the least MAE of 0.4512, MSE of 0.325, RMSE of 0.5701, and the highest* R^2 *of 0.9988. To guarantee data immutability and trust, prediction records are cryptographically hashed using SHA-256 hashing and stored on a blockchain ledger, while sensitive payloads are kept off-chain for privacy preservation purposes. Experimental results show that accuracy is improved, residual variance is reduced, and system stability is improved. This proposed framework successfully leverages decentralized trust, intelligent analytics, and scalable processing to deliver a powerful tool for secure, transparent, and privacy-preserving IoT-based energy management systems.*

Keywords - *Big data analytic, Blockchain, Energy efficiency prediction, IoT, LightGBM, RandomForestRegressor, Smart grid, XGBRegressor.*

1. Introduction

Scalability, privacy, interoperability, energy consumption, data integrity, and legal issues are challenges that may be faced when integrating blockchain technology, which enables AI-driven big data analytics, with an IoT ecosystem. Due to the resource constraints of IoT devices, scalability becomes an issue when integrating blockchain and AI-driven big data analytics with an IoT ecosystem [1, 2]. Concerns about privacy derive from the nature of data in the Internet of Things, which is considered sensitive, and the unalterable nature of the ledger in the blockchain [3]. Research on federated learning and decentralized data sharing. Other ethical issues that need consideration in the use of the Internet of Things include [4].

The combination of blockchain, AI-driven analytics, and IoT ecosystems is considered a revolutionary concept for modern digital infrastructure that addresses pressing requirements for security, scalability, and intelligent automation. The blockchain technology has the essential elements which ensure data integrity while establishing trust through its decentralized and unalterable ledger system, which ensures the protection of vital Internet of Things data and device identification information. The artificial intelligence analytics tools have the essential functions which ensure that the Internet of Things systems achieve their full potential through their capabilities to enable immediate decision-making and maintenance and operation improvements, particularly through the latest advancements in edge intelligence and federated learning frameworks.

The main goal of this particular research is to propose and assess a blockchain-based and AI-based big data analytics system that promotes secure, scalable, and efficient management of IoT ecosystems. Moreover, the proposed research attempts to resolve some of the major issues associated with data manipulation, scalability, and privacy using the security benefits of blockchain technology, intelligent processing of data by AI, and big data analytics for better decision-making. Moreover, the proposed research attempts to provide solutions that promote trust, transparency, and resiliency in IoT networks, paving the way for its successful implementation in various industry applications.

Due to rapid development in IoT technologies, it is now possible to develop interconnectable devices capable of gathering and exchanging vast amounts of real-time data in different environments. There are numerous applications of IoT solutions that include smart cities, health monitoring, industrial processes, agriculture, and transportation solutions. In all of the aforementioned applications, there are continuous data-generating processes supporting intelligent monitoring processes and automation techniques. The large-scale implementation of IoT solutions brings up several issues associated with data security, user privacy, scaling processes, and efficient distributed data streams processing [1].

In order to resolve these issues and bring extra benefits to IoT systems, blockchain technology can be utilized. First, a blockchain system allows for recording transaction history using decentralized ledgers, making transactions immutable and resistant to any changes. As a result, blockchain ensures safe and reliable device and data verification.

Moreover, blockchain-based consensus algorithms allow for increasing the security and authenticity of stored data since no one can make any changes in the existing data records. Therefore, blockchain appears to be beneficial for IoT environments where several devices continuously exchange their data in different and potentially malicious environments. Nevertheless, integrating blockchain with IoT creates its own problems.

On the other hand, AI-powered big data analytics has recently gained popularity due to their capability to analyze large volumes of IoT-generated data in order to derive valuable insights. Machine learning models facilitate predictive maintenance, detect anomalies, make smart decisions, and optimize performance.

With the emergence of edge computing and federated learning techniques, data processing can take place close to data sources, thus minimizing latencies, increasing system response time, and ensuring privacy of analyzed information [3]. Despite the above-mentioned capabilities of using AI to analyze IoT-generated data, data-sharing and data-storage capabilities should be ensured for trustworthy data collection.

The integration of blockchain, AI-powered analytics, and big data solutions within IoT environments is currently seen as a novel approach to creating reliable and intelligent digital infrastructures. The combination of the mentioned innovations helps achieve transparency of information exchange, enhanced security of data transmission, and intelligent automation of processes. Nonetheless, there still exist a number of open questions in terms of interoperability of heterogeneous devices, high power consumption by blockchain technologies, insufficient storage resources of IoT nodes, and user privacy issues [4].

This study aims to achieve the following objectives.

- Develop a secure and scalable integrated IoT–AI Blockchain big data framework.
- Ensure data integrity and privacy using blockchain-based security mechanisms.
- Apply and evaluate machine learning models for accurate energy efficiency prediction in smart grid systems.
- To develop a unified framework that enhances privacy protection and enables real-time threat detection using AI-driven techniques.

2. Related Work

2.1. Literature Review

The rapid growth of IoT, which is predicted to exceed 40 billion devices by 2030, creates vast amounts of heterogeneous data, which is a challenge for traditional infrastructure in terms of scalability, security, and privacy [5]. The convergence of blockchain, AI, and big data analytics is considered to be a revolutionary approach for addressing these issues, as blockchain technology provides decentralized integrity for data, AI provides intelligent automation, anomaly detection, and prediction, and big data analytics provides edge-cloud processing for large-scale, high-velocity data [6]. Studies have shown performance improvements in terms of throughput and accuracy using optimized architectures for IoT edge, cloud, and deep reinforcement learning [7, 8].

Moreover, AI-based IoT technologies are further improving predictive maintenance and adaptive decision-making for smart cities, 5G networks, UAV-based networks, and sustainable infrastructures. In addition, blockchain technology is being integrated into IoT technologies for enhanced trust, privacy, and integrity using smart contracts and lightweight consensus protocols such as DPoS, PoA, and PoC [9]. Blockchain technology has challenges associated with it regarding latency, overheads, and interoperability in federated learning [10, 11]. The literature demonstrates that AI, blockchain technology, and big data can function together to improve security and scalability in IoT technologies. Table 1 shows the performance and Functional Comparison between Blockchain and AI-Oriented Big Data Frameworks in IoT Environments.

The paper [12] presents a proposed blockchain-based AI and ML-powered big data analysis framework that aims to improve the security and stability of CIoT networks. The work highlights the importance of securing IoT data against data poisoning attacks that may negatively affect machine learning algorithms, causing inaccuracies in predictions.

In [13], an IBBA-based approach was suggested to improve the food safety monitoring systems. It was developed to support real-time monitoring of environmental factors like temperature, humidity, and contamination. Its goal was to increase transparency, traceability, and food safety within the food supply chain. This framework could tackle problems associated with food quality management, late detection of contamination, and unreliable tracking mechanisms in conventional food safety systems.

In [14], the author has performed an extensive survey of the integration of Blockchain Technology and Federated Learning (FL).

It has been analyzed how the integration of blockchain technology with federated learning ensures safe and decentralized training of models without having to store the data centrally.

It has been pointed out by the authors that such integration proves beneficial, especially for the Internet of Things (IoT) scenarios, where large amounts of data are generated in a distributed manner.

Table 1. Comparison of blockchain and AI-Based big data frameworks for IoT

No.	Paper (Authors, Year)	Main Focus	Tech Stack / Methods Used	IoT Domain Target	Key Strengths	Limitations / Gaps Identified
1	Mitra et al., 2023 - Blockchain-based AI/ML-enabled Big Data for Cognitive IoT [12]	Blockchain-anchored AI pipelines for Cognitive IoT	Blockchain for provenance, ML analytics, big-data streaming	Cognitive IoT	Strong integration of provenance + analytics; real architecture	Needs real-world deployment; blockchain overhead not optimized
2	Peddareddigari et al., 2025 - IBBA Framework for Food Safety (IoT+Blockchain+AI)[13]	End-to-end IBBA framework	IoT sensors + blockchain + AI analytics	Food safety monitoring	Practical demonstration of complete pipeline	Application-specific; limited generalizability
3	Ning et al., 2024 - Blockchain-Based Federated Learning [14]	Latest BCFL directions	FL framework + blockchain orchestration	Resource-constrained IoT	Detailed analysis of consensus vs device constraints	Gaps in lightweight BCFL; no large-scale experiments
4	Bothra et al., 2023 - AI + Blockchain Synergy (Survey) [15]	Synergy of AI + Blockchain	AI-enhanced blockchain; blockchain-trust for AI	Multi-domain inc.IoT	Conceptual clarity; useful taxonomy	Lacks IoT-specific evaluation datasets
5	Ashwini et al., 2025 - Adaptive Edge Blockchain Architecture for Big Data [16]	Real-time threat detection using blockchain at edge	Edge computing + blockchain + interpretable ML	Edge IoT / Smart environment	Concrete architecture addresses real-time analytics	No end-to-end scalability analysis; moderate dataset

The Study [15] discussed an extensive review to analyze how the use of Blockchain and AI technologies will enhance the effectiveness of Internet of Things systems. In this context, the study focused on analyzing the recent innovations and advances made in blockchain-based IoT and AI-based IoT systems, emphasizing increased system reliability, security, and effective decision-making capabilities.

As pointed out in [16], an adaptive edge blockchain framework was suggested in order to enhance the performance, scalability, and security capabilities of big data processing platforms in IoT settings. The research work concentrated on solving issues related to latency, data security, and insufficient bandwidth capacity through the use of both edge computing and blockchain technology. With the new approach, big data could be processed not only at the cloud layer but also at the edge level.

There have been numerous empirical studies that have established the benefits associated with the integration of these technologies in practical settings. For example, a study conducted to evaluate the impacts of incorporating artificial intelligence in IoT and blockchain concluded that the use of sensors for real-time data collection, coupled with transaction systems supported by blockchains, had significant implications for enhancing the efficiency and effectiveness of supply chains. The study further found that the incorporation of predictive algorithms and automation played an important role in ensuring system reliability [17].

Applications within specific domains provide another set of examples that highlight the efficiency of technology integration. For instance, studies in the area of logistics and intelligent transport systems suggested that integration of Internet-of-things-based sensor technology, blockchain verification, and artificial intelligence led to increased visibility and efficient infrastructure management. However, latency issues, increased energy consumption, and other problems associated with such systems prevent large-scale application of these solutions [18].

2.2. Research Gap

The current state of AI-powered Internet of Things devices and blockchain-based security systems has reached substantial progress, yet existing research studies have investigated these technologies as separate entities without establishing a complete solution framework. The current solutions create challenges when they attempt to process real-time Internet of Things data streams because they require security mechanisms and storage systems and authentication methods, and privacy protection measures, especially in resource-limited environments. The use of AI-based decision systems together with blockchain technology for establishing trust and transparency and tracking data origins still remains at an early stage of development. The lack of empirically validated architectures that jointly exploit AI, blockchain, and big data analytics highlights the need for scalable, secure, and efficient IoT ecosystem models.

3. Proposed Framework

The proposed architecture employs a four-layer layered architecture that supports secure, intelligent, and scalable energy efficiency prediction in smart grid systems. The proposed architecture layers include the IoT Data Layer, Big Data Processing Layer, AI Intelligence Layer, and Blockchain Security Layer. These four layers constitute the overall architecture of the proposed architecture, as depicted in Figure 1. The first layer of the proposed architecture layers is the IoT Smart Grid Layer, which comprises heterogeneous devices such as smart meters, load sensors, weather sensors, and distributed renewable energy sources such as solar and wind energy.

These devices continuously provide data on energy consumption and weather conditions. The second layer of the proposed architecture layers is the Big Data Processing Layer, which comprises data acquisition and ingestion using Apache Kafka, data cleaning and normalization using Apache Spark, and data processing using Hadoop. The third layer is the AI Prediction Engine, which uses Feast to combine feature engineering and batch as well as real-time pipelines; here, several machine learning models, including LightGBMRegressor, Random Forest Regressor, and XGBoostRegressor, are trained and tested to produce a prediction for the Energy Efficiency Score; the output includes a prediction result, metadata, and a cryptographic hash; lastly, the Blockchain Security Layer, which uses Hyperledger Fabric, ensures trust and data integrity through smart contracts (chaincode), which include prediction hashes, timestamps, and types of devices/users, with a guarantee of authenticity, traceability, and resistance to tampering of AI-generated energy efficiency results.

The proposed architecture for the development of an IoT-based smart grid infrastructure includes a unified integration of four major layers to ensure intelligent grid management in a secure, efficient, and scalable way, as depicted in Figure 1 below. The data layer of the proposed architecture includes heterogeneous data sources such as smart meters, environmental sensors, and renewable energy sources, which enable real-time monitoring of grid consumption using IoT technologies, as discussed in [19, 20].

The big data processing layer includes Apache Spark, Hadoop, and other big data processing tools to ensure efficient handling of high-velocity data from smart grids, as discussed in [21]. In addition, the AI intelligence layer includes deep learning for efficient load forecasting, federated learning for decentralized model training, and reinforcement learning for efficient grid stability, as discussed in [22]. To ensure enhanced trustworthiness of the proposed smart grid infrastructure, a blockchain-based security layer has been integrated to ensure efficient data sharing using PBFT or PoA algorithms, as discussed in [23]. Finally, a privacy and compliance layer has been integrated to ensure efficient handling of GDPR regulations in India, as discussed in [24].

3.1. Comparative Advantages of the Proposed Framework

The uniqueness of the proposed framework is due to the design of a holistic architecture that incorporates IoT-based data acquisition, big data processing, AI-based prediction, and blockchain-based security for IIoT smart-grid scenarios. Most of the prior literature has concentrated only on a few aspects of these technologies. In [12] presented a blockchain enabled AI/ML framework for Cognitive IoT with a focus on data security and resilience to poisoning attacks, but have not provided an entire energy analytics pipeline of a smart grid. Likewise, the IBBA framework designed by [13] combined IoT, blockchain, big data, and AI technologies for food safety monitoring, but this framework had a specific domain application and did not consider energy efficiency prediction and blockchain-based prediction provenance. Also, [14] gave an extensive review of blockchain-based federated learning techniques with a particular focus on privacy and trust improvement, but did not propose a validated end-to-end architecture for real-time IoT analytics.

However, in the case of the proposed model, a four-layer system architecture, which consists of the following layers, has been introduced: the IoT Smart Grid Layer, the Big Data Processing Layer, the AI Prediction Layer, and finally the Blockchain Security Layer. The major contribution made in this research paper is the use of Hyperledger fabric along with SHA-256 for the creation of immutable records for AI-generated energy efficiency predictions. In contrast to traditional models based on blockchain, where the entire dataset is stored using blockchain technology, the proposed approach uses only prediction hashes on the blockchain.

4. Proposed Methodology

4.1. Dataset Description and IoT Data Acquisition Process

The IIoT Smart Grid Dataset is an example of live operational data that can be obtained from an industrial Internet of Things smart power grid. This dataset contains data on electricity measurements, weather conditions, production of renewable energy, and the usage of electricity by consumers.

The data collection process takes place using smart meters and IoT devices placed within the smart grid system. These data contribute to intelligent management of energy, load prediction, and anomalies detection, as well as other types of analytics. Preprocessing steps include data cleaning, normalization, and feature engineering and transform the data into machine learning and deep learning problems. IIoT Smart Grid Dataset is a Kaggle dataset simulating live operations of the IoT smart grid.

The setting up of the experimental setup for the implementation of the AI-Enabled Smart Grid with Blockchain-based Verification using Hyperledger Fabric was done with a computer system that has Windows 11 OS with Intel Core i7, 16 GB RAM, and Google Colab being the main cloud computing environment used for software development. The machine learning process was done using Python 3 within Google Colab and used libraries like Pandas, NumPy, Scikit-learn, LightGBM, Matplotlib, Seaborn, SciPy, Requests, and OpenPyXL.

A total of 17 features were used as input variables to build our models, and Energy_Efficiency_Score was chosen as the target variable. Selected features include essential operational and power quality, renewable energy generation, environmental, and consumer-based features of the smart-grid ecosystem. Operational and power quality features include Power_Consumption_kWh, Voltage_V, Current_A, Power_Factor,Grid_Frequency_Hz, eactive_Power_kVAR, Active_Power_kW, and Demand_Response_Event, indicating electricity consumption pattern and grid efficiency. Environmental factors are captured through Temperature_C and Humidity_%, while renewable energy generation is captured through Solar_Power_Generation_kW and Wind_Power_Generation_kW. Historical consumption behaviors are taken into account through Previous_Day_Consumption_kWh, Peak_Load_Hour, and Normalized_Consumption, while consumer and energy-source-based factors are captured using Energy_Source_Type and User_Type.

In combination, these features provide an all-round coverage of energy utilization behaviors, power quality, renewable energy contribution, environmental factors, and consumer profile in smart grids. Through such a diverse selection of features, our model is able to capture the complex interactions in the smart-grid system and predict Energy_Efficiency_Score.

4.1.1. IoT Sensor Configuration

The IIoT Smart Grid Data is collected using a series of IoT-enabled sensors and monitors installed throughout the smart grid system, providing live data on its operations. The sensors usually consist of smart energy meters used to measure the consumption of energy in kilowatts per hour (kWh), voltage sensors used to monitor the levels of voltage, current sensors used to monitor the electrical currents flowing in the grid, and frequency sensors that help to keep track of the stability of the frequency levels in the grid.

Power quality sensors are used to measure active power, reactive power, and power factors to ensure effective energy usage and to detect any power disturbances. Environmental monitoring can be done using temperature and humidity sensors that help determine the impact of weather conditions on the grid operations.

Solar irradiation sensors, photovoltaic monitors, and wind energy sensors are used in the smart grids that incorporate renewables to monitor renewable energy production. The sensors are connected to the smart grid via the IoT communication protocols like MQTT, Zigbee, WiFi, and LoRaWAN, where they feed data for analysis and decision-making at edge or cloud platforms.

4.1.2. Data Preprocessing

Prior to the training process, several processing steps are performed to improve the accuracy of the dataset. At first, any missing values are determined and treated to eliminate possible contradictions while performing calculations. After that, any duplicate entries and noisy data are eliminated to preserve the integrity of data used for analysis.

Then, appropriate methods of encoding are applied for categorical features to transform them into numerical values. Further, feature engineering is executed to determine relevant parameters that will be useful for energy efficiency analysis. Then, the dataset is normalized with the help of the Standard Scaler method. Finally, the data is split into training and testing subsets.

The proposed framework combines IoT technology-based smart grids' data and advanced machine learning models for the assessment and improvement of energy efficiency for different categories of consumers. It collects electrical parameters in real-time, such as voltage, current, power factor, active and reactive power, and smart grids' frequency. It also collects environmental factors such as weather conditions, temperature, humidity, and renewable energy sources. It collects historical factors as well. The heterogeneous nature of the collected data requires a Big Data Processing Layer for its processing.

It involves data ingestion, cleaning, categorical encoding, feature engineering, and normalization. The features are standardized using StandardScaler. The dataset is then fed into machine learning models for prediction.

The machine learning models are random forest, XGBoost, and LightGBM. Early stopping is implemented in the models for overfitting avoidance. The proposed framework can be useful for the accurate prediction of energy efficiency in smart grids. It can be useful for accurate demand forecasting and better allocation of resources, and better integration of renewable energy in smart grids.

4.2. Hashing Technique

Security considerations span far past SHA-256 hash values along four axes. Privacy is maintained by way of Standard Scaler and categorical encoding, using Energy Efficiency Score instead of sensor data in the on-chain store. Identity-based authentication is established with an X.509 MSP identity, Spydra API key, and versioning based on the modelVersion parameter.

Data poisoning defense takes shape by removing missing data via dropna (), and only applying scaler transformations for training data, while prediction modification attacks can be countered with immutable hashes linked in a chain on the distributed ledger, such that any modified scores would compromise blocks, and replay attacks via unique keys created by the composite transaction key, in addition to Fabric duplicates by way of nonce.

The proposed architecture addresses security not only via the SHA-256 hash but also offers mechanisms to protect the privacy of data, authentication, and resistance to attacks. Confidential sensitive IoT data and predictions are saved in an off-chain way, whereas only cryptographic hashes of data and metadata are saved in the Hyperledger Fabric blockchain, which allows keeping data confidential and saving memory.

Authentication and access management are performed using Hyperledger Fabric CA, which grants an X.509 certificate to authenticated devices and users. Integrity of the data is achieved by means of immutable blockchain transactions and SHA-256 hashes, as well as any alterations can be detected. The architecture is also protected against replay attacks due to unique transaction identifiers and timestamps, while the permissioned approach in Hyperledger Fabric mitigates Sybil and unauthorized access attacks. In addition, data cleaning, normalization, and provenance help to mitigate the problem of data poisoning attacks for AI models.

4.3. Hyperledger Fabric

Hyperledger Fabric utilizes an execute-order-validate model whereby transactions are simulated by endorsing peers, ordered by the ordering service, and then validated and committed into the ledger by committing peers. The Raft consensus protocol ensures sub-second finality, fitting well for high-frequency commits of energy score.

The Hyperledger Fabric network, deployed on Spydra (AWS ap-south-1), involves one endorsing peer, one committing peer, an orderer implementing the Raft consensus protocol, and the Certificate Authority for identity management.

The energy channel, which provides an exclusive ledger for the energy efficiency data, only grants access to its participants. The energy-score-contract smart contract, developed with Node.js v2.5.0 using the fabric-shim, implements three chaincode methods. StoreEnergyScore serializes per-device JSON records, including device ID, predicted energy score and timestamp into the blockchain. StoreModelMetrics commits run-level aggregated metrics.

GetEnergyScore performs deterministic read-only queries for post-commit verification purposes. The chaincode deployment was accomplished through the Spydra console by following a four-step workflow that took approximately three minutes to accomplish. Each block contains the previous block's hash, creating an unalterable hash-chain structure.

4.4. LightGBM Regression Model for Energy Efficiency Prediction

4.4.1. Feature Normalization and Data Standardization

The features of the smart grid datasets are heterogeneous in nature and Include Voltage (V), Current (A), Power Consumption (kWh), Temperature (°C), and Renewable Energy Generation (kW). The features are measured in different units. To avoid the dominance of a particular feature in the learning process, z-score normalization is performed as follows:

$$X_{norm} = \frac{X-\mu}{\sigma} \tag{1}$$

Here, (X) represents the original value of the feature, (μ) represents the mean of the feature values, and (σ) represents the standard deviation of the feature values over the training data set. In code, this operation is carried out by using the StandardScaler class of the scikit-learn library. This is an important function of the Big Data Processing Layer, which is necessary for the stable convergence of the LightGBM model and for increasing its prediction accuracy.

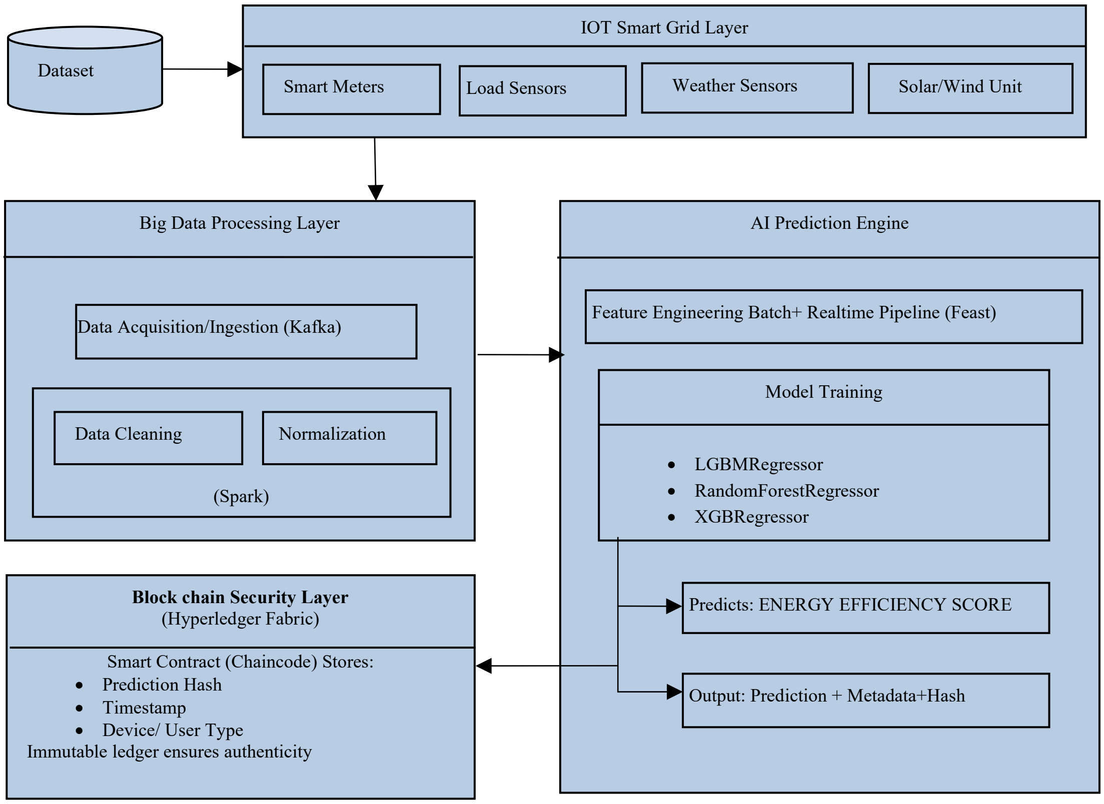


Fig. 1 Flowchart for architecture layers

4.4.2. Energy Efficiency Score Prediction using LightGBM

The prediction of the Energy Efficiency Score is done using a Light GBM regression model. This model forms an ensemble of decision trees in a sequential manner. The predicted score for a given data instance is defined as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (2)$$

The predicted energy efficiency score is represented by (y^_i), and (K) stands for the total number of trees, while (f_k) represents the function of a decision tree, and xi denotes the normalized feature vector. This process operates like its real-world version because multiple trees work together to understand how electrical and environmental factors interact with past consumption data.

The loss function that is optimized throughout the training process is defined as:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

The loss function calculates prediction error on l (y_i,y^_i) while the regularization term Ω (f_k) imposes penalties on the complexity of the model. The statement establishes a direct link between early stopping and the two regularization parameters num_leaves and learning_rate, which the code uses to maintain model generalization while stopping overfitting.

4.4.3. Model Performance Evaluation Metrics

To quantitatively evaluate the accuracy and robustness of the predictions made by the Energy Efficiency Score model, multiple evaluation metrics are used.

The Mean Absolute Error (MAE) measures the average magnitude of prediction errors:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (4)$$

This metric reflects the average deviation between actual and predicted energy efficiency values.

The Mean Squared Error (MSE) is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (5)$$

By squaring the errors, MSE penalizes larger deviations more heavily, making it useful for identifying significant prediction inaccuracies. The Root Mean Squared Error (RMSE) is obtained as:

$$RMSE = \sqrt{MSE} \quad (6)$$

RMSE preserves the original unit of the Energy Efficiency Score, making the results easier for stakeholders to interpret.

The coefficient of determination (R^2) can be defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2} \quad (7)$$

This metric quantifies how well the model explains the variance in energy efficiency behavior, directly validating the effectiveness of the AI Prediction Layer.

4.4.4. Energy Efficiency Score Interpretation

To enhance decision-making, the predicted Energy Efficiency Score can be categorized into qualitative levels:

$$H = \begin{cases} Low \quad \hat{y} < 50 \\ Moderate \quad 50 \leq \hat{y} < 75 \\ High \ \hat{y} \geq 75 \end{cases} \quad (8)$$

The established thresholds allow grid operators to find inefficient energy consumers who need immediate attention, while they work to optimize energy efficiency, which helps them achieve the demand response and sustainability goals mentioned in the first part of the article.

4.4.5. Blockchain Hash Generation for Prediction Integrity

The process of achieving trustworthiness and permanent record maintenance requires all Energy Efficiency Score forecasts to undergo cryptographic protection before their blockchain storage. The hash generation process gets described through mathematical equations as follows:

$$H = SHA - 256\big(JSON(R)\big) \quad (9)$$

The prediction record (R) holds timestamp information together with the predicted score and feature summary, and model version details.

The Python code implementation, which uses hashlib library functions, operates according to this specific mathematical formulation. The hash functions as a unique identifier which lets users see all modifications that occurred in the prediction records.

4.4.6 Blockchain Record Formalization

Every blockchain transaction creates a permanent record that maintains its structured data format.

$$BlockRecord = \{ID, H, t, y_{pred,} V_m\} \quad (10)$$

The system uses unique record identifiers for its IDs, which identify all records. The system uses the cryptographic hash as its H, while the timestamp is represented by the t value. The system uses the predicted Energy Efficiency Score through the variable (y_(pred,)) and the model version is represented through the variable (V_m), which functions as the system's model version. The system creates an official framework which uses the Hyperledger Fabric chaincode as its basis to establish complete traceability and auditability, together with the original AI-generated predictions.

5. Results & Discussion

The current research offers a comprehensive model that is designed based on a multi-level structure. The proposed solution involves IoT-based information gathering, big data analysis, AI-enabled prediction, and blockchain technology-based protection in one integrated package. Contrary to previous works that use only on-chain or off-chain data storage methods, the new system uses both options via SHA-256 encryption and Hyperledger Fabric. Furthermore, an extensive analysis of various AI models suggests LightGBM as the best choice for energy efficiency prediction. The study was conducted using an IIoT Smart Grid dataset containing 8,738 observations with 18 attributes was split into training and testing partitions using an 90% - 10% split. The number of observations utilized for training the models was 7864, while the number of observations employed for testing the models was 874.

5.1. Experimental Environment and System Configurations

Implementation of the suggested framework involved its realization within the distributed computing architecture. It allowed for the analysis of IoT-enabled data in a scalable and secure manner. As part of the implementation process, an IoT data acquisition layer responsible for receiving smart grid data on voltage, current, power consumption, power factor, frequency, temperature, humidity, and renewables generation was designed. Data ingesting and streaming operations involved the use of Apache Kafka, whereas the data processing, cleaning, normalization, and feature engineering phases leveraged the functionality of Apache Spark and Hadoop tools. Machine learning operations were carried out in Python, where LightGBM, Random Forest Regressor, and XGBoost Regressor algorithms were used to train models and predict energy efficiency levels. Feature scaling was done with the help of the Standard Scaler model provided by the Scikit-learn package. Blockchain security was realized with the help of the Hyperledger Fabric framework, whereas SHA-256 cryptographic hash function was used to create immutable records of the prediction outcomes. Overall, the system relied on the Feast platform for managing the features in batch and real-time prediction pipelines.

5.2. Blockchain Performance Metrics

The performance of Fabric deployment was tested on four parameters. The transaction latency took 420-650ms for each invocation, which is good for batch-commit operations. The throughput is between 50 to 100 TPS with default settings, adequate for storing after inference. Scalability can be done through horizontal replication of peers and sharding channels. The storage overheads are negligible; for instance, 2,000 device records require less than 1 MB, while a year's ledger increases only by 1.2 GB.

The performance measurement of blockchains was done through the use of a private blockchain network that consisted of different nodes. For the blockchain architecture, spydra hyperledger was used.

For throughput, it was defined by the number of transactions per second. Latency was defined by the amount of time taken from the transaction generation to its confirmation. Multiple experiments were run, and the results are averages of all runs.

In order to examine the effectiveness of the blockchain-based framework, the following four main performance indicators of blockchain technology have been considered: latency, throughput, scalability, and storage overhead. The latency is the time gap between the submission and confirmation of transactions.

The throughput was calculated in terms of the number of successful Transactions committed Per Second (TPS). Scalability was considered by analyzing the system performance with an increase in the number of nodes and transaction volume. The storage overhead has been estimated in terms of the accumulated size of the blockchain.

$$Throughput = \frac{T_n}{t} \tag{11}$$

where Tn represents the number of successful transactions, and t denotes the total execution time. Throughput is measured in Transactions Per Second (TPS) and indicates the transaction processing capability of the blockchain network.

$$Latency = T_c - T_s \tag{12}$$

where Tc represents the transaction confirmation time, and Ts denotes the transaction submission time. Latency measures the time required for a transaction to be successfully confirmed on the blockchain.

$$Scalability = \frac{\mathrm{T_N}}{\mathrm{T_{iN}}} \tag{13}$$

where TN denotes the throughput achieved with NNN nodes, and TiN represents the throughput obtained with the initial number of nodes. This metric evaluates the ability of the blockchain network to maintain or improve its performance as the network size increases.

$$Storage\ Overhead = \frac{M}{T_n} \tag{14}$$

where M represents the total blockchain storage size, and Tn denotes the total number of transactions. This metric indicates the average storage required for each transaction recorded on the blockchain.

5.3. Metric-Based Comparative Analysis of Regression Models

The process starts with obtaining the prediction payload, which includes the record identifier, timestamp, predicted energy efficiency score, and model version. Following that, the payload is serialized to ensure consistency when creating a hash. A hash of the payload is generated using the SHA-256 hashing algorithm. After that, the entire payload is stored securely using off-chain storage mechanisms to maintain data privacy and prevent unauthorized exposure of sensitive data.

Following that, a blockchain transaction is generated with the hash of the record, along with other minimal metadata such as the record type and block timestamp. Finally, the transaction is propagated to the blockchain network, where it is validated by multiple nodes before being permanently stored in a block. Following that, a confirmation of insertion is generated once it is successfully stored in the blockchain ledger. Table 2 presents the assessment of Regression Models through Comparative Performance Analysis.

Figure 2 shows the performance evaluation metrics of the proposed model. It is clear that LightGBM performs better than other models in all performance evaluation metrics, as it provides the lowest MAE, MSE, and RMSE of 0.4512, 0.325, and 0.5701, respectively, along with the highest value of R^2, which is 0.9988, indicating highly accurate predictions. XGBRegressor is also performing well, as its error rate is low, i.e., MAE = 0.7539, RMSE = 1.0137, along with the second highest value of R^2, which is 0.9961. On the other hand, the performance of the Random Forest Regressor is poor, as its error rate is higher, i.e., MAE = 1.3167, RMSE = 1.7251, along with the lowest value of R^2, which is 0.9888.

The results, in general, indicate that the LightGBM model is the best for the prediction task. The figure show Performance Comparison of Regression Models. From the MAE graph, it is evident that the LightGBM model has the lowest error value, implying that it is the most accurate in predicting the energy efficiency score compared to the others.

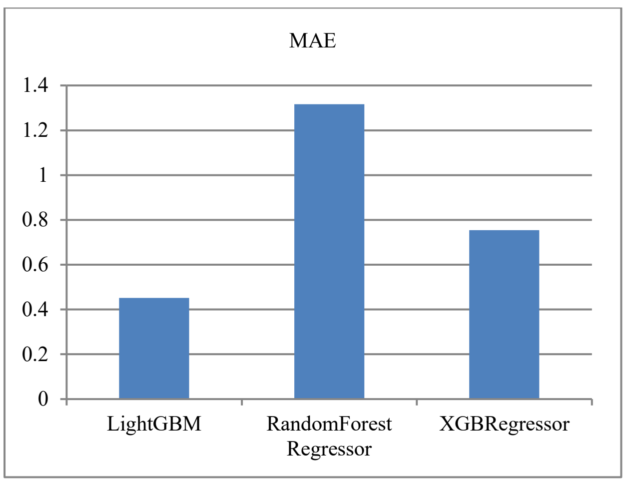
On the other hand, the error value for the Random Forest Regressor is the highest, implying that it is the least accurate in predicting the energy efficiency score. XGBoost has a moderate performance, outperforming the Random Forest Regressor but performing slightly worse than the LightGBM model.

Similarly, in the MSE and RMSE graphs, a trend can be established. It can be noted that LightGBM has the lowest MSE and RMSE values. This indicates that the variance and prediction errors are minimal. The values for the Random Forest Regressor are considerably higher. This indicates that the model is not as stable. Again, XGBoost has a balanced performance.

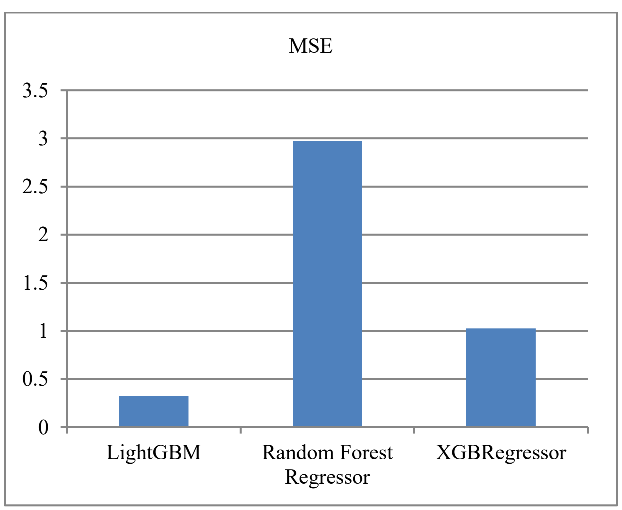
The model provides better accuracy than the Random Forest Regressor but lacks the performance of LightGBM. The R^2 values for each model are as follows. LightGBM has the highest R^2 value close to unity. This indicates a good relationship between the predicted and actual values. Another observation is that the coefficient of determination is also high in the case of the XGBoost model. On the other hand, the Random Forest model gives a low R^2 score compared to the other models. Based on the graphical analysis, it is confirmed that the LightGBM model performs better than the other models in terms of all the metrics used for evaluation. Therefore, the model is the most suitable for the prediction of energy efficiency in the proposed framework.

Table 2. Comparative performance analysis of regression models

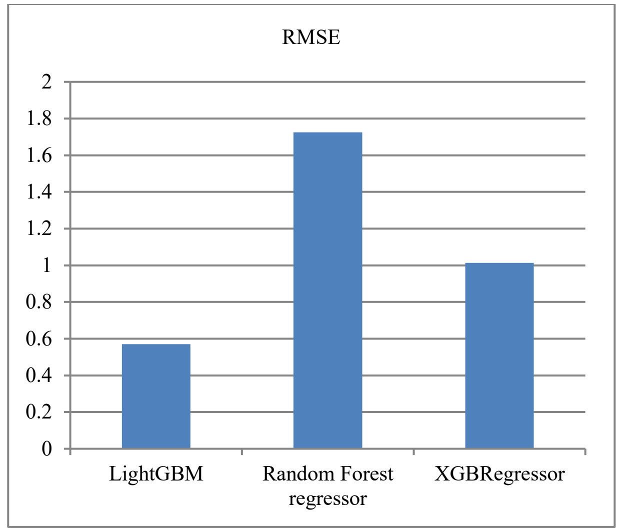
Model	MAE	MSE	RMSE	R^2
LightGBM	0.4512	0.325	0.5701	0.9988
Random Forest Regressor	1.3167	2.9759	1.7251	0.9888
XGBRegressor	0.7539	1.0275	1.0137	0.9961



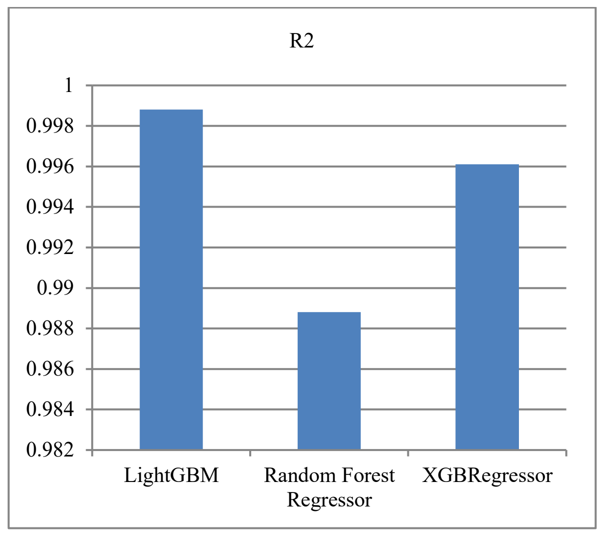
(a)



(b)



(c)



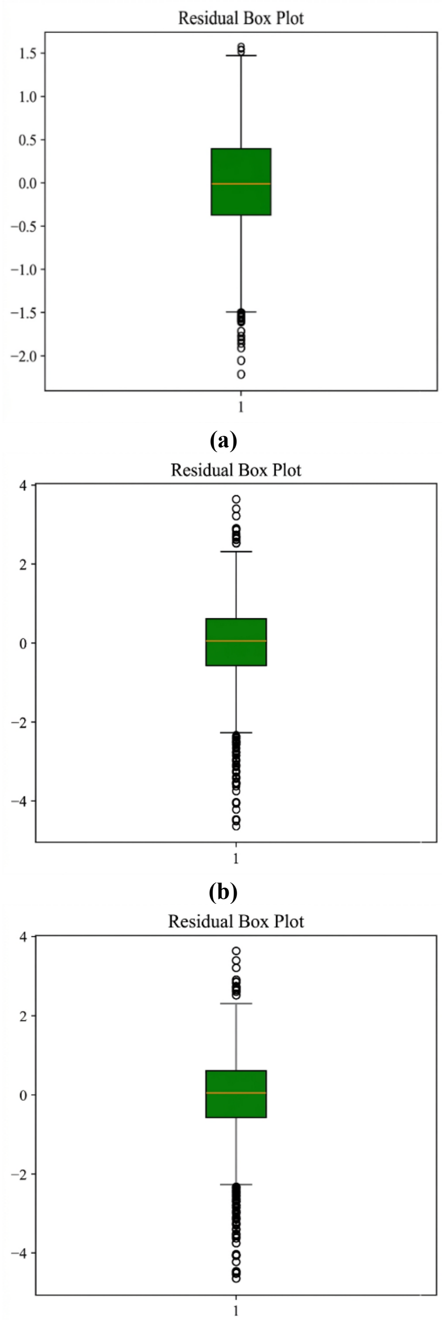
(d)

Fig. 2 (a) MAE, (b) MSE, (c) RMSE, and (d) R2.

5.4. Residual Analysis and Model Comparison

Residual plots were analyzed to check the predictive potential of the Gradient Boosting Machine (GBM) model and the Random Forest Regression model. Box plots of the residual plots were analyzed to check for the central tendencies, variability, and outliers of the data. Figure 3 shows residual plots of the models that were evaluated.

It is seen that the residual plot of the GBM model shows that the median is around zero, with no systematic error in the predictions. It is also seen that the Interquartile Range (IQR) is small, showing that there is less variability in the predictions. Although there are some outliers, they are not too high, and the range of extreme residual values is from -2.2 to 1.5.



(a)

(b)

(c)

Fig. 3(a) Light GBM residual, (b) Random forest residual, and (c)XGBoost.

On the other hand, the median of the residuals for the Random Forest model was close to zero. This indicates that the predictions made by the model were unbiased. The range of the spread of the residuals was wider for this model. The negative values of the residuals ranged up to -4.5. The positive values of the residuals ranged up to 3.7. This indicates that the predictions made by this model were not as stable as those made by the GBM model. The GBM model was better than the Random Forest model.

5.5. Hyperparameter Tuning and Model Optimization

To increase the prediction accuracy and guarantee the ability to generalize, hyperparameter tuning was done for the three regression models (LightGBM, Random Forest, and XGBoost). Hyperparameter tuning was carried out using a structured approach whereby the best parameter set was obtained based on the training set. In the case of the LightGBM regression model, parameters including the number of leaves, learning rate, maximum tree depth, and the number of estimators were optimized. On the other hand, the number of trees, maximum depth, minimum samples needed for node split, and minimum samples needed at leaf node were tuned for the Random Forest regression model. The same process was done for XGBoost, where optimization of parameters included learning rate, maximum depth, subsample ratio, colsample_bytree, and number of estimators. The effectiveness of each model was assessed through cross-validation. The best parameter set, with the minimum prediction error and maximum R^2 score, was selected.

Hyperparameter optimization was done for improving the predictive capacity of the LightGBM regression model. The initial settings for the baseline model were set to include 500 estimators, a learning rate of 0.05, 31 leaves, a subsample of 0.8, and a features ratio of 0.8, whereas all other parameters remained on their default levels. In the result of optimization, there were significant improvements in the configuration of the optimal model, where the number of estimators was boosted to 1,444, and the learning rate was raised to 0.0914, suggesting that the model required larger trees and a faster learning rate. There was a significant increase in the number of leaves, which grew from 31 to 94, providing for capturing more nonlinear relations in the data, while the maximum depth was limited to 3 to avoid growing trees too much and thus reducing the probability of overfitting.

The minimum child samples parameter was raised from 20 to 29. Moreover, the value of the subsampling ratio was reduced from 0.8 to 0.5763, and the feature sampling ratio was reduced from 0.8 to 0.6162. The reduction in these parameters adds more randomness to training phase. This increased randomness increases the robustness of model. In addition, the regularization parameters used in LightGBM, which include reg_alpha (L1 regularization) and reg_lambda (L2 regularization), were initially zero in the base case setup. The above regularization terms have been optimized along with other hyperparameters used in LightGBM. The values selected for reg_alpha (L1) and reg_lambda (L2) were 0.000357 and 0.000189, respectively. The role of L1 regularization is that of reducing any unnecessary complexity of the model, whereas L2 regularization reduces the possibility of having very large weights in the model. These values helped control model complexity and mitigate the risk of overfitting while preserving predictive performance. Thus, the selected regularization coefficients were not arbitrarily assigned but were obtained as part of the optimized

LightGBM configuration. From above, it is evident that the optimized hyperparameter values represent a trade-off between complexity and generalization of the model.

Table. 3 K-Fold Cross-Validation results (K = 10)

Metric	Mean	Standard Deviation
MAE	0.4310	0.0076
RMSE	0.5489	0.0104
R^2	0.9989	0.0001

As for the prediction of the Energy Efficiency Score from the IIoT Smart Grid dataset, the LightGBM regression model demonstrated excellent prediction accuracy. The summary statistics of 10 Fold Cross Validation is shown in Table 3.

LightGBM regression model was able to predict the Energy Efficiency Score very well. During Ten-fold cross-validation, the model attained the average MAE of 0.4231, RMSE of 0.5436, and R^2 of 0.9989 with low standard deviations per fold, showing that the model is stable and robust. For preventing over-fitting and leakage of data into the model, all processes related to data preparation and training the model were carried out solely using the training dataset without any leakage from the test set. This further shows the generalization ability of the model since the test set had not been seen by the model before.

Besides, the residuals analysis showed that there were no error patterns, meaning that the model was able to capture the relationships between independent variables and the dependent variable. It should be noted that the high accuracy of the model is due to the high correlations and informativeness of the chosen operational parameters of the smart grid. Hence, LightGBM is a robust approach for assessing energy efficiency in IIoT-enabled smart grids.

Bottom of Form the contribution of this research lies not in technologies themselves, but rather in a comprehensive approach to ensure their security and scalability for analysis of the smart grid. In contrast to previous research that focused on no more than two components of this system, our approach is holistic and integrates IoT for real-time data acquisition, big data analytics via Apache Kafka, Spark, and Hadoop, AI for efficient energy prediction via LightGBM, Random Forest, and XGBoost models, and blockchain technology with Hyperledger Fabric and SHA-256 for integrity verification of these predictions.

Implementation of this system has been proven through a smart grid energy efficiency prediction application in which the LightGBM model demonstrated superior performance with an MAE of 0.4512, RMSE of 0.5701, and R^2 of 0.9988. In addition, blockchain integration has been implemented in order to securely store prediction hashes and other metadata and guarantee their immutability, traceability, and resistance to tampering. The novel aspects of our paper and its implementation have been further emphasized by expanding the sections of Research Gap, Proposed Framework, and Results, which describe interactions between different technologies and implementation and performance results.

6. Conclusion & Future Work

This work introduces a framework for privacy preservation in the secure recording of energy efficiency prediction data using blockchain technology. The proposed framework employs a hash-based data storage technique in such a way that the data is not revealed on the blockchain, and still gets the benefits of blockchain technology. Its immutability and transparency. The mechanism requires collecting and serializing the prediction payload, creating a cryptographic hash using SHA-256 hashing, and storing the entire data off-chain in a safe environment. Only the generated hash and limited non-sensitive metadata are stored on the blockchain ledger. The validation and permanent storage of the transaction on the blockchain ledger ensure integrity, traceability, and trustworthiness. The proposed mechanism has successfully achieved a balance between preserving privacy and managing data securely and reliably. Therefore, it can be used for energy analytics and other data-intensive applications.

However, future studies can build upon this model by utilizing sophisticated cryptographic techniques like Zero Knowledge Proofs to verify the results of predictions without disclosing any information. This system model can also be improved upon by utilizing permission-based Blockchain technologies like Hyperledger Fabric to ensure scalability for business needs. The use of decentralized storage systems, such as IPFS, enables organizations to improve their data protection through encryption capabilities which these systems provide. Future studies can also build upon this model to utilize smart contracts for access control, as well as performance optimization to test latency for transactions and storage for large-scale systems.

References

[1] Ehtisham Ul Haque et al., “A Scalable Blockchain based Framework for Efficient IoT Data Management using Lightweight Consensus,” *Scientific Reports*, vol. 14, no. 1, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Shengdong Zhang et al., “A Novel Ultrathin Elevated Channel Low-Temperature Poly-Si TFT,” *IEEE Electron Device Letters*, vol. 20, no. 11, pp. 569-571, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[3] Alina Granwehr, and Verena Hofer, “Blockchain Enabled Big Data Analytics Framework for Secure Cyber Defense,” *Elaris Computing Nexus*, pp. 50-60, 2025. [CrossRef] [Publisher Link]

[4] Elham Shammar, Xiaohui Cui, and Mohammed A.A. Al-qaness, "Swarm Learning: A Survey of Concepts, Applications, and Trends," *arXiv preprint*, pp. 1-37, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Md Eshrat E. Alahi et al., "Integration of IoT-Enabled Technologies and Artificial Intelligence (AI) for Smart City Scenario: Recent Advancements and Future Trends," *Sensors*, vol. 23, no. 11, pp. 1-36, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Muhammad Babar et al., "An Optimized IoT-enabled Big Data Analytics Architecture for Edge-Cloud Computing," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 3995-4005, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Alaelddin F. Y. Mohammed et al., "Deep-Reinforcement-Learning-based IoT Sensor Data Cleaning Framework for Enhanced Data Analytics," *Sensors*, vol. 23, no. 4, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Jason Moss et al., "Explainable AI in IoT: A Survey of Challenges, Advancements, and Pathways to Trustworthy Automation," *Electronics*, vol. 14, no. 23, pp. 1-20, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ehtisham Ul Haque et al., "Performance Enhancement in Blockchain based IoT Data Sharing using Lightweight Consensus Algorithm," *Scientific Reports*, vol. 14, no. 1, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Lang Wu et al., "A Survey on Blockchain-based Federated Learning," *Future Internet*, vol. 15, no. 12, pp. 1-22, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Yanna Jiang et al., "Blockchained Federated Learning for Internet of Things: A Comprehensive Survey," *arXiv preprint*, pp. 1-37, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ankush Mitra et al., "Impact on Blockchain-based AI/ML-Enabled Big Data Analytics for Cognitive Internet of Things Environment," *Computer Communications*, vol. 197, pp. 173-185, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Siva Peddareddigari, Sri Vigna Hema Vijayan, and Manickavasagan Annamalai, "IoT, Blockchain, Big Data and Artificial Intelligence (IBBA) Framework-for Real-Time Food Safety Monitoring," *Applied Sciences*, vol. 15, no. 1, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Weiguang Ning et al., "Blockchain-based Federated Learning: A Survey and New Perspectives," *Applied Sciences*, vol. 14, no. 20, pp. 1-35, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Priyanka Bothra et al., "How can Applications of Blokchain and Artificial Intelligence Improve Performance of Internet of Things? - A Survey," *Computer Networks*, vol. 224, pp. 1-30, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] N. Ashwini et al., "ThreatFedChainAI: An Adaptive Edge Blockchain Architecture for Big Data-Driven Threat Analytics in IoT Networks," *Scientific Reports*, vol. 16, no. 1, pp. 1-47, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Bhanu Prakash Sah, and Shofiullah Shaikh, "AI-Driven IoT and Blockchain Integration in Industry 5.0: A Systematic Review of Supply Chain Transformation," *Innovatech Engineering Journal*, vol. 1, no. 1, pp. 99-116, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Arya Kharche, Sanskar Badholia, and Ram Krishna Upadhyay "Implementation of Blockchain Technology in Integrated IoT Networks for Constructing Scalable ITS Systems in India," *Blockchain: Research and Applications*, vol. 5, no. 2, pp. 1-27, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Ahmad Enaya, Xavier Fernando, and Rasha Kashef, "Survey of Blockchain-based Applications for IoT," *Applied Sciences*, vol. 15, no. 8, pp. 1-37, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Vehbi C. Gungor et al., "Smart Grid Technologies: Communication Technologies and Standards," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 4, pp. 529-539, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[21] Fisayo Sangoleye et al., "Reinforcement Learning-based Demand Response Management in Smart Grid Systems with Prosumers," *IEEE Systems Journal*, vol. 17, no. 2, pp. 1797-1807, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Zainab Iftikhar et al., "Privacy Preservation in Resource-Constrained IoT Devices using Blockchain-A Survey," *Electronics*, vol. 10, no. 14, pp. 1-26, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Seetah Almarri, and Ahmed Aljughaiman, "Blockchain Technology for IoT Security and Trust: A Comprehensive SLR," *Sustainability*, vol. 16, no. 23, pp. 1-35, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Vinay Maurya et al., "Blockchain-Driven Security for IoT Networks: State-of-the-Art, Challenges and Future Directions," *Peer-to-Peer Networking and Applications*, vol. 18, no. 1, pp. 1-35, 2025. [CrossRef] [Google Scholar] [Publisher Link]