[1]Manjula K Pawar
[2]Prakashgoud Patil

Enhancing the Scalability Of Blockchain System Using Machine Learning Algorithms

Journal of Electrical Systems

Abstract: Blockchain is a distributed ledger that stores blocks called records that are linked with cryptography in a decentralized environment. It has many characteristics, such as immutability, security, and transparency. Hence the, blockchain technology is widely used in various real time applications such as industry and business. Though it has many good characteristics, it faces challenges such as Scalability and Privacy. The scalability of blockchain systems is an ongoing challenge in the pursuit of efficient transaction processing and widespread adoption. To address this issue, this study examines how machine learning algorithms can optimize blockchain systems' scalability. Specifically, we investigate the effectiveness of four popular machine learning algorithms: linear regression, logistic regression, random forest, and XGBoost gradient boosting. Through empirical analysis and experimentation, we evaluate the performance of these algorithms in improving the scalability of blockchain systems. We use real-world data sets and simulation environments to measure each algorithm's ability to predict transaction throughput, latency, and network congestion under various loads and conditions. Our research provides practical recommendations for integrating machine learning techniques into blockchain infrastructure to achieve enhanced scalability and improved transaction processing efficiency. The study advances how machine learning can be leveraged to optimize blockchain scalability. It paves the way for more robust and scalable decentralized systems in the future.

Keywords: Scalability, Linear Regression, Logistic Regression, Random Forest, XG Boost.

I. INTRODUCTION

Blockchain technology represents a robust, decentralized, and trustworthy database system that ensures data integrity and security through various mechanisms such as data encryption, data linking, multi-copy storage, and distributed consensus [4]. it discusses how blockchain technology solves the Byzantine attack problem in digital cash systems[14]. Blockchain allows transactions to be conducted without intermediaries or trusted third parties. This paper explores the principles and mechanisms of blockchain and its potential applications. Improving the performance of blockchain is vital for its use in various sectors.

The significant impact of blockchain technology [7] was initially propelled into the spotlight by the success of Bitcoin. Its immutability, security, and transparency attributes have garnered widespread interest across various sectors, offering a new paradigm for decentralized systems in inherently untrusted environments. However, despite its potential, blockchain adoption has faced obstacles primarily related to security and performance issues[19]. The surge in transactions coupled with restrictions on block sizes has exacerbated these challenges, prompting concerns about scalability and efficiency [12].

Beyond crypto currencies, blockchain technology holds promise for revolutionizing numerous sectors, including healthcare, manufacturing, distribution, and governance [8].Blockchain technology has many applications, including secure sharing of health records, logistics agreements, and financial transactions. However, as transaction volumes and network sizes increase, the limitations of the technology become more evident, resulting in reduced efficiency and sluggish processing. The current performance of blockchain technology poses limitations, particularly concerning cryptocurrencies [13].

In order to improve the scalability issue, several techniques are developed such as onchain, and offchain techniques[16].Sharding is one onchain technique that paralyzes transaction execution by creating shards[21][25][26]. Some solutions contribute to scalability improvement by reducing the latency of transactions execution [15].

[1]*Corresponding author: Assistant Professor, KLE Technological University, Hubli(KAR), India, manjulap@kletech.ac.in

[2] Head MCA, KLE Technological University, Hubli(KAR), India, prakashpatil@kletech.ac.in

In order to overcome the challenges[11], the research explores the integration of machine learning techniques to optimize the performance of Proof-of-Work (PoW) consensus protocols, a cornerstone of blockchain technology. By leveraging machine learning algorithms, we aim to develop a robust solution capable of enhancing the scalability and efficiency of blockchain networks without sacrificing security or decentralization [18][20].

II. LITERATURE SURVEY

The intersection of machine learning (ML) and blockchain Technology presents promising avenues for enhancing identity verification and security protocols [1].ML algorithms precisely validate users' identities by analyzing biometric data, such as fingerprints or facial recognition, which are inherently difficult for hackers to replicate.By securely storing this sensitive information in a decentralized manner using Blockchain, the risks associated with centralized data repositories are mitigated, providing an added layer of security. This integration of ML-driven identity verification and blockchain-based security measures represents a formidable approach to safeguarding digital systems and protecting sensitive information.

Bitcoin Bitcoin-NG [2] introduces a significant performance enhancement by two phases of operations in Bitcoin: leader election and transaction serialization.This structural segmentation optimizes the transaction process, enabling improved scalability and throughput within the network.By decoupling these essential functions, Bitcoin-NG streamlines transaction validation and block creation, reducing latency and enhancing overall efficiency.In contrast, Federated Blockchain diverges from the conventional centralized model by distributing control across multiple organizations rather than a single entity. Federated Blockchain decentralized governance structure promotes transparency, resilience, and equitable participation, offering a promising alternative to traditional centralized systems.

The paper [3] introduces the "Blockchain Reputation-Based Consensus"(BRBC) mechanism as a novel approach to achieving consensus within blockchain networks, leveraging the reputation and behavior of participating nodes. By incorporating reputation metrics, BRBC aims to enhance the reliability and security of the consensus process.Furthermore, the paper addresses a comprehensive threat model for the system, acknowledging potential attacks where adversaries can intercept, manipulate, or discard transactions within the network. To evaluate the effectiveness and correctness of the proposed system, the paper conducts a thorough analysis, including an examination of the system state, security assessment, and simulation studies. Notably, the BRBC mechanism operates without assuming inherent trust among network participants.

Paper [4] introduces a decentralized FL (Federated Learning) framework for data evaluation, utilizing a private blockchain infrastructure and a model owner responsible for managing the global model.Users and miners play distinct roles within this framework, with users granted access to the model while the owner specifically selects miners.The two datasets used for evaluating the FL framework are the Brain Tumor dataset and the Medical MNIST, aiming to simulate real-world FL applications closely.Notably, the framework employs a custom-built blockchain rather than relying on existing ones. The baseline accuracy for the datasets is approximately 95 and 99.7, respectively, without blockchain integration. However, when blockchain is introduced, accuracy hovers around 90, with a slight decrease observed as the number of malicious nodes increases. Additionally, the study highlights the necessity of creating separate blockchains for different models as part of the FL process.

III. PROPOSED METHODOLOGY

The proposed methodology involves developing machine learning algorithms for improving the scalability of blockchain systems without compromising security and decentralization.It involves several steps, including data collection, preprocessing, model implementation.The methodology leverages a combination of statistical and machine learning models to improve the accuracy of predictions.

A. Dataset Description

The Bitcoin dataset is taken from https:/gz.blockchair.com/bitcoin/blocks/.The dataset sample is given in Fig 1.The fields of the dataset are:

newdf

	id	size	input_count	output_count	weight	transaction_count	DIFF	nonce
0	1471	216	1	1	864	1	0.025486	0.024271
1	1472	216	1	1	864	1	0.024201	0.202071
2	1473	216	1	1	864	1	0.017905	0.360509
3	1474	216	1	1	864	1	0.022465	0.538936
4	1475	216	1	1	864	1	0.025683	0.270224
...	...	...	...	...	...	...	...	...
196	1667	216	1	1	864	1	0.017627	0.911027
197	1668	216	1	1	864	1	0.018669	0.908004
198	1669	216	1	1	864	1	0.015590	0.440122
199	1670	216	1	1	864	1	0.027905	0.142081
200	1671	216	1	1	864	1	0.030486	0.761414

Fig.1. Dataset

- **Id** : denotes each transaction unique value
- **Size :** Actual amount of data
- **Input count** : total number of inputs given at a time
- **Diff** : time taken to execute transaction .
- **Output count** : number of outputs generate by machine
- **Nounce** : probability finding element .

The system design of the proposed methodology is shown in Fig 2.The steps to be carried are as below.

- **Dataset Collection:** The first step involves collecting blockchain data from various sources. This data may include transaction records, block information, smart contract data, or any other relevant information stored on the Blockchain.
- **Preprocessing**: For blockchain data, preprocessing involve handling missing values, removing duplicates, and converting data types as needed and also parsing raw data structures, converting them into a structured format.
- **Normalization:**Normalization is the procedure of rescaling numerical variables to a consistent range, usually between 0 and 1. The main purpose of normalization is to ensure that each variable contributes equally to the analysis, thereby preventing variables with larger values from dominating the analysis.
- **Partitioning:**After the data has been preprocessed and normalized, it is usually divided into the training and testing sets. This assesses how well the models will generalize to new, unseen data.
- **Setting Parameters:** Before training the models, it's essential to set parameters such as the learning rate, regularization strength, and model architecture. These parameters control how the models learn from the data and can significantly impact their performance.
- **Implementing Various Models:** Once the parameters are set, various machine learning models can be implemented and trained using the training data.
- **Training and Testing:** The training phase involves feeding the training data into the machine learning models and adjusting the model parameters iteratively to minimize the error between the predicted and actual values.
- **Evaluation and Tuning:** Finally, parameters such as accuracy, precision, recall, or F1-score are measured, depending on the specific task.This measures the training model performance.

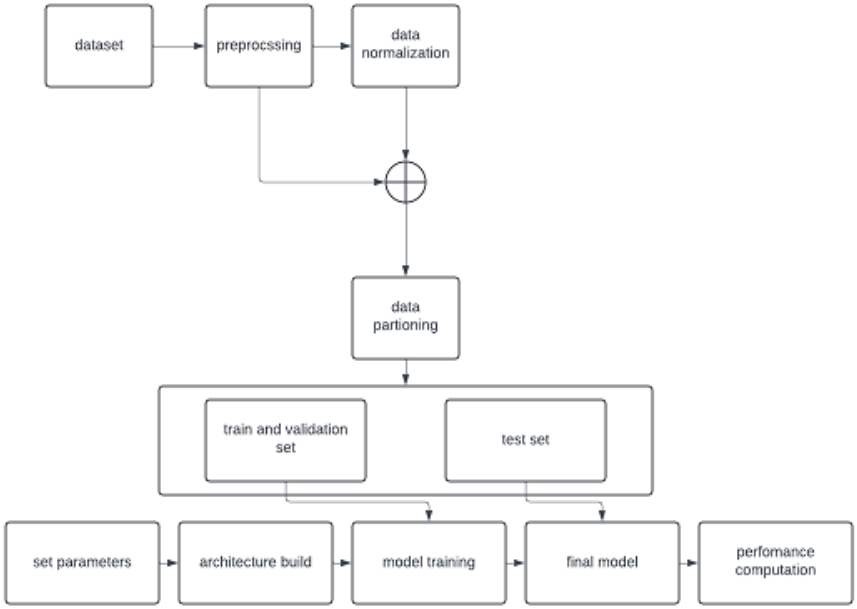

Fig 2. System Design

The dataset, crucial for training and evaluating the model, is divided into three main sets: training, testing, and validation, as shown in Fig 3.

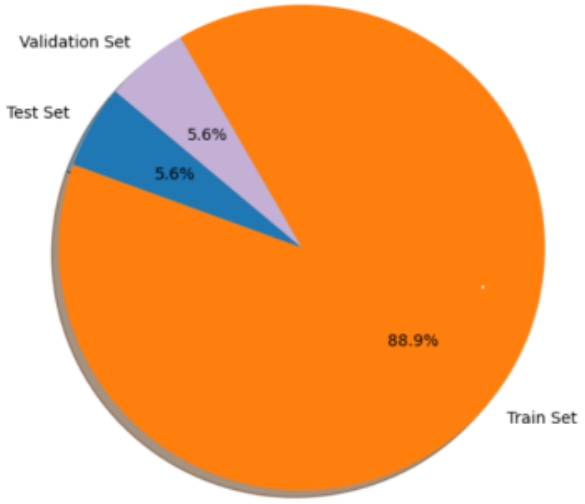

Fig 3. Dataset distribution into train, test, and validation set

B. Implementation

The implementation section explains the details of the system by describing each component with its code skeleton in terms of algorithm.Here the accuracy of each model is measured. The model with highest accuracy is chosen for transactions executions in blockchain system.

a) Linear regression

Linear regression model [9] shown in Fig 4 describes the relationship between a dependent variable (Y) and independent variables (X).It assumes a linear relationship described by $Y= +1\ x^1+2\ x^2 + 3\ x^3.....+n\ x^n$ The goal is to find coefficients that minimize the difference between predicted and observed values. Metrics like R-squared and mean-squarederror assess the performance of the model.

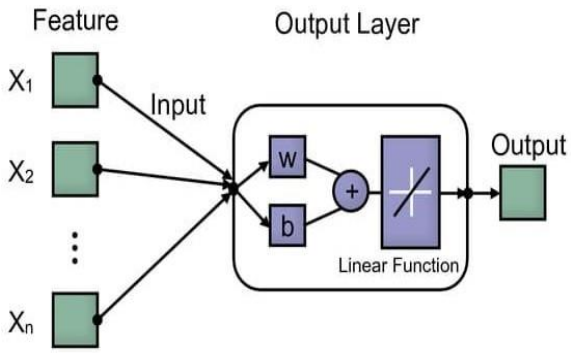

Fig.4. Linear Regression Architecture

b) Logistic Regression

Logistic regression [9] as shown in Fig.5 is a supervised learning algorithm used for classification tasks. It predicts the probability of an instance belonging to a specific class using a sigmoid function that maps real-valued inputs to a value between 0 and 1. Unlike linear regression, it predicts categorical or discrete values.

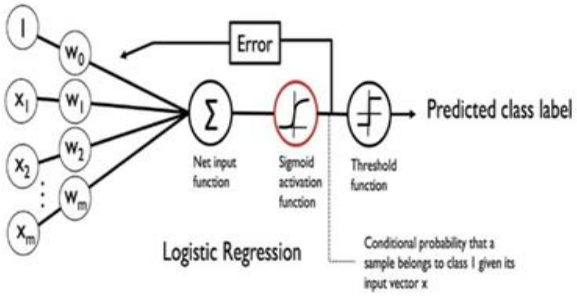

Fig. 5 Logistic Regression Architecture

c) XG Boost

XG Boost[9] combines decision trees sequentially to build a strong predictive model is shown in Fig 6. It reduces computation time and memory usage by randomly sampling a subset of the training data at each iteration. Regularization techniques are employed to prevent over fitting. XG Boost supports various loss functions and is evaluated using metrics like mean squared error, log loss, accuracy, precision, recall, and F1-score. It's widely used in regression and classification tasks across domains like finance, health- care, marketing, and recommendation systems, offering efficiency for large datasets with high predictive accuracy.

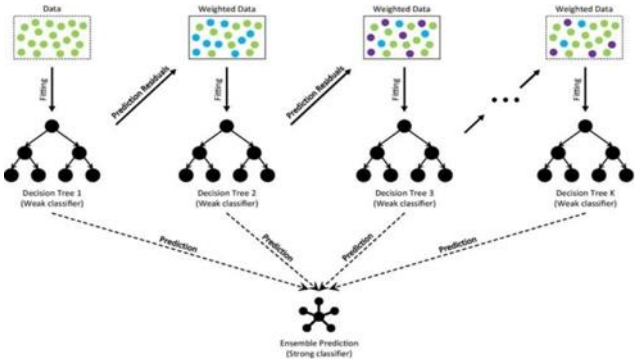

Fig.6 XG Boost architecture

d) LSTM

LSTM[9] is a type of RNN that overcomes the limitations of traditional RNNs in cap- turing long-range dependencies and handling vanishing gradients is shown in Fig 7. LSTM uses memory cells with input,

forget, and output gates to selectively retain or forget information over long sequences. It's flexible, scalable, and has become a fundamental tool in natural language processing, time series analysis, and speech recognition.

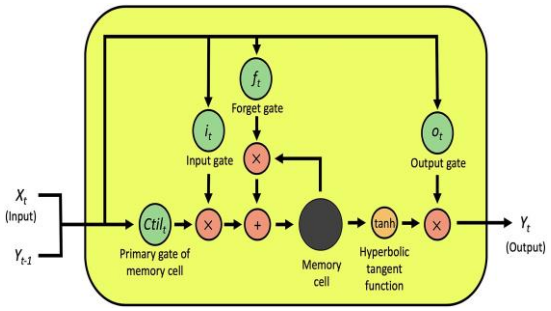

Fig.7. LSTM Architecture

IV. RESULTS AND DISCUSSION

The evaluation metrics[10], including make span, response time, throughput, size, and swarm calculation follow the same formulas for different algorithms. These metrics provide a standardized and consistent way to assess the performance of each algorithm in dynamic load balancing, ensuring a comprehensive and comparable evaluationacross the diverse architectures employed in the study.

1. Accuracy:

$$\text{Accuracy}=\frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ Predictions} \quad (1)$$

2. Precision:

$$\text{Precision}=\frac{True\ Positives}{True\ Positives+False\ Positives} \quad (2)$$

3. F1Score:

$$\text{F1 Score}=2*\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Table 1 shows the results of the analysis of the models. Five different models are used, and their training and testing accuracy are compared. The Fig 8. shows the accuracy results of the various machine learning models, providing valuable insights into their suitability for addressing the problem of enhancing blockchain scalability without compromising security and decentralization.

XGBoost [7] demonstrates the highest accuracy of 82.3, indicating its potential effectiveness in optimizing blockchain scalability while maintaining security and decentralization.This suggests that XGBoost's ensemble learning approach may offer robust performance in managing the complexities of blockchain networks, potentially improving transaction throughput and overall scalability without sacrificing security measures.

Table1.Performance of Machine Learning Models

Model	**Accuracy**
Linear Regression	14.5
Logistic Regression	14

Model	**Accuracy**
XGBoost	82.3
Random Forest	10
LSTM	10

On the other hand, Linear Regression[7] and Logistic Regression exhibit similar accuracy of 14.5 and 14 respectively. While these models may provide some insights into the problem, their lower accuracy's suggest that they may not capture the intricate relationships within blockchain data as effectively as XGBoost.

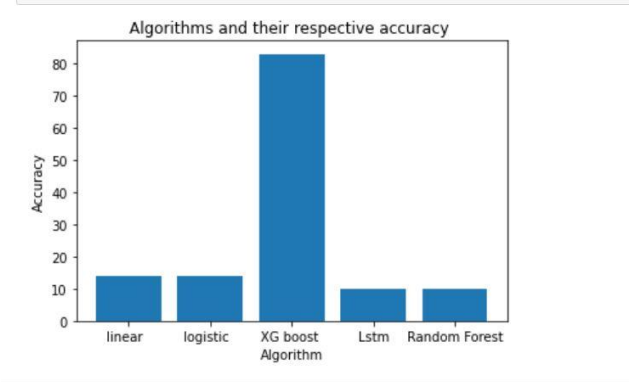

Fig.8. Algorithms and their Accuracy

Random Forest and LSTM[7], with accuracy's of 10 each, also indicate limited effectiveness in addressing the problem statement. Random Forest's ensemble learning approach and LSTM's recurrent neural network architecture may not be well-suited for capturing the nuanced patterns and dynamics of blockchain data in a scalable and secure manner.

V. CONCLUSION

The comparison of machine learning algorithms revealed distinct variations in their predictive accuracy. XG Boost and Random Forest emerged as the top performers, achieving an impressive accuracy rate of 82.3%. Conversely, Linear Regression, Logistic Regression, and LSTM demonstrated considerably lower accuracy rates, ranging from 10% to 14.5%. These findings show that employing XGBoost or Random Forest would yield the most favorable outcomes for tasks requiring precise predictions. Conversely, Linear Regression, Logistic Regression, and LSTM may not be optimal due to their comparatively lower accuracy rates. Hence, the XGBoost model can applied for transaction executions, which improves the scalability. There is the future scope for determining accuracy with DNN and Reinforcement Learning models..

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

[1] Taherdoost, H. Blockchain and Machine Learning: A Critical Review on Security. Information 2023, 14, 295. https://doi.org/10.3390/info14050295S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” http://Www.Bitcoin.Org, p. 9, 2008.

[2] Journal Article Taher doost, Hamed 2023 Information 2078-2489 14 5 295 Blockchain and Machine Learning: A Critical Review on Security doi:10.3390/info14050295 https://www.mdpi.com/2078-2489/14/5/295

[3] Marcela T. de Oliveira, Lúcio H.A. Reis, Dianne S.V. Medeiros, Ricardo C. Carrano, Sílvia D. Olabarriaga, Diogo M.F. Mattos,Blockchain reputation-based consensus: A scalable and resilient mechanism for distributed mistrusting applications,Computer Networks,Volume 179,2020,107367,ISSN 1389-1286.

[4] Laveen Bhatia, Saeed Samet,A decentralized data evaluation framework in federated learning,Blockchain: Research and Applications,Volume 4, Issue 4,2023,100152,ISSN 2096-7209.

[5] G. Othman Alandjani, "Blockchain Technology and Impacts on Potential Industries," 2023 IEEE 2nd International Conference on AI in Cybersecurity (ICAIC), Houston, TX, USA, 2023, pp. 1-4, doi: 10.1109/ICAIC57335.2023.10044170.

[6] A. Singh, A. P. Srivastava, P. Choudhary, H. Pandey and A. K. Singh, "Blockchain in Healthcare," 2021 International Conference on Technological Advancements and Innovations (ICTAI), Tashkent, Uzbekistan, 2021, pp. 168-172, doi: 10.1109/ICTAI53825.2021.9673187.
[7] OpenAI. (2024). ChatGPT (3.5) [Large language model]. https://chat.openai.com\\
[8] According to Towards Data Science, the article "8 Common Evaluation Metrics for Machine Learning Models" provides insights into various evaluation metrics used in machine learning algorithms (Towards Data Science, 2018).
[9] P. Tasatanattakool and C. Techapanupreeda, "Blockchain: Challenges and applications," 2018 International Conference on Information Networking (ICOIN), Chiang Mai, Thailand, 2018, pp. 473-475, doi: 10.1109/ICOIN.2018.8343163.
[10] Hemlata Kohad, Sunil Kumar, Asha Ambhaikar. ” Scalability Issues of Blockchain Technology” International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-3, February, 2020
[11] Leslie Lamport et al. 2001. Paxos made simple. ACM Sigact News (2001)
[12] Sheng Wang, Tien Tuan Anh Dinh, Qian Lin, Zhongle Xie, Meihui Zhang, Qingchao Cai, Gang Chen, Wanzeng Fu, Beng Chin Ooi, and Pingcheng Ruan. 2018. ForkBase: An Efficient Storage Engine for Blockchain and Forkable Applications. In VLDB.
[13] Xinan Yan, Linguan Yang, Hongbo Zhang, Xiayue Charles Lin, Bernard Wong, Kenneth Salem, andTim Brecht. 2018. Carousel: low-latency transaction processing for globally-distributed data. InSIGMOD.
[14] Mahdi Zamani, Mahnush Movahedi, and Mariana Raykova. 2018. RapidChain: Scaling Blockchain via Full Sharding. In CCS.
[15] Eyal, I., Gencer, A.E., Sirer, E.G. and Van Renesse, R. (2016) ‘Bitcoin-ng: a scalable Blockchain protocol’, Proceedings of 13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16), Santa Clara, CA, USA, pp.45–59
[16] Deval V, Dwivedi V , Dixit A, Norta A , Shah S, Sharma R and Draheim D . Mobile Smart Contracts: Exploring Scalability Challenges and Consensus Mechanisms. IEEE Access. 10.1109/ACCESS.2024.3371901. 12. (34265-34288). https://ieeexplore.ieee.org/document/10453564/
[17] Kandpal M, Goswami V , Priyadarshini R and Barik R . (2023). Towards Data Storage, Scalability,and Availability in Blockchain Systems: A Bibliometric Analysis. Data. 10.3390/data8100148. 8:10. (148).https://www.mdpi.com/2306-5729/8/10/148
[18] Monrat A, Schelén O and Andersson K . (2023). Addressing the Performance of Blockchain by Discussing Sharding Techniques 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). 10.1109/ICECCME57830.2023.10252191. 979-8-3503-2297-2. (1-9).https://ieeexplore.ieee.org/document/10252191/
[19] Li X, Luo H and Duan J. Security Analysis of Sharding in Blockchain with PBFT Consensus. The 2022 4th International Conference on Blockchain Technology. (9-14).https://doi.org/10.1145/3532640.3532642
[20] Kohad H, Kumar S and Ambhaikar A. (2022). Scalability of Blockchain based E-voting system using Multiobjective Genetic Algorithm with Sharding 2022 IEEE Delhi Section Conference (DELCON). 10.1109/DELCON54057.2022.9753019. 978-1-6654-5883-2. (1-4).https://ieeexplore.ieee.org/document/9753019/
[21] Pawar M.K., Patil P., Hiremath P.S. (2021) A Study on Blockchain Scalability. In: Tuba M., Akashe S., Joshi A. (eds) ICT Systems and Sustainability. Advances in Intelligent Systems and Computing, vol 1270. Springer, Singapore. https://doi.org/10.1007/978-981-15-8289-9_29
[22] Pawar, Manjula K., Prakashgoud Patil, P. S. Hiremath, Vaibhav S. Hegde, Shyamsundar Agarwal, and P. B. Naveenkumar. ”Scalable blockchain framework for a food supply chain.” In Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 1, pp. 467-478. Springer Singapore, 2021
[23] M. K. Pawar, P. Patil, M. Sharma and M. Chalageri, ”Secure and Scalable Decentralized Supply Chain Management Using Ethereum and IPFS Platform,” 2021 International Conference on Intelligent Technolo- gies (CONIT), 2021, pp. 1-5, doi: 10.1109/CONIT51480.2021.9498537.
[24] Pawar, Manjula K., Prakashgoud Patil, and Amit Singh Patel. ”Distributed and Scalable Healthcare Data Storage Using Blockchain and KNN Classification.” In Smart Trends in Computing and Communica- tions: Proceedings of SmartCom 2022, pp. 741-750. Singapore: Springer Nature Singapore, 2022.
[25] Pawar, Manjula K., Prakashgoud Patil, Ridham Sawhney, Prem Gumathanavar, Shraddha Hegde, and Kavya Maremmagol. ”Performance analysis of e-certificate generation and verification using blockchain and ipfs.” In 2022 International Conference on Inventive Computation Technologies (ICICT), pp. 345-350. IEEE, 2022.
[26] R. R. Mahatungade, P. Patil and M. K. Pawar, ”Performance analysis of Reinforcement Learning for Miner Selection in Blockchain,” 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-4, doi: 10.1109/INCET57972.2023.10170107.
[27] Pawar, Manjula K., and Prakashgoud Patil. ”Performance Enhancement of Blockchain Systems using AI based Consensus Mechanism.” Grenze International Journal of Engineering & Technology (GIJET) 9, no. 1 (2023).
[28] Kalwad, Triveni, Vinuta Aramani, Needa Fatima Attar, Neha Raikar, D.G. Narayan, Manjula Pawar, and Pooja Shettar. ”Performance Evaluation of Consensus Algorithms for Permissioned Blockchain Networks.” In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-7. IEEE, 2023.
[29] Manjula K. Pawar, Prakashgoud Patil,” Miner Selection in Blockchain using Proof of Artificial Intelligence”,Procedia Computer Science,Volume 230,2023,Pages 838-845,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2023.12.049.
[30] M. K. Pawar, H. S. Hiregowdar and S. Joshi, "Land Registry System using Blockchain With Multiple Nodes," 2023 2nd International Conference on Futuristic Technologies (INCOFT), Belagavi, Karnataka, India, 2023, pp. 1-3, doi: 10.1109/INCOFT60753.2023.10425604