Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 27 No. 2 (2024)

Implementation and Mathematical Modelling of Data Sharing and Privacy Preserving Technique in a Blockchain Network by Using Machine Learning Algorithms

Madhuri Vishwas Shinde[1], Dr. Shyamrao V. Gumaste[2]

[1]Department of Computer Engineering, MET's Institute of Engineering, Nashik
Maharashtra- 422003, India
madhurivshinde.26788@gmail.com

[2]Professor and Head of Information Technology, MET's Institute of Engineering, Nashik
Maharashtra- 422003, India
svgumaste@gmail.com

Article History:

***Received**: 06-02-2024*

***Revised**: 18-04-2024*

***Accepted**: 09-05-2024*

Abstract:

Blockchain technology has emerged as a reliable and secure decentralized network with multifaceted applications in banking, finance, insurance, healthcare, and business domains. Recent trends within blockchain communities indicate a growing interest in deploying machine learning models to extract valuable insights from extensive, geographically dispersed datasets owned by individual participants. To enable learning models without centralized data repositories, extensive research has focused on developing machine learning algorithms tailored for blockchain networks. However, despite numerous proposals, privacy and security concerns remain inadequately addressed, revealing vulnerabilities in architecture and operational efficiency limitations. The proposed ensemble machine learning model presents a pioneering solution, aiming to systematically resolve privacy, security, and performance issues within blockchain systems. The novelty of this approach lies in its targeted resolution of critical challenges at the intersection of blockchain technology and machine learning. While prior research has delved into integrating machine learning into blockchain networks, this model stands out by introducing a privacy-centric methodology that systematically addresses the core issues of privacy, security, and performance. Moreover, the proposed model promises enhanced resilience against adversarial attacks compared to other aggregation rules within a differentially private scenario. The innovative ensemble machine learning model for blockchain exhibits a significant 20% improvement in data privacy, a 15% boost in security, and a notable 25% enhancement in operational performance.

Keywords: Healthcare system, blockchain, machine learning, data privacy, security, patient data.

1. Introduction

A blockchain is defined as a trustworthy and secure decentralized and distributed network that provides interactions among participants such as communities that consist of individuals, companies or governments that have a specific or common goal, e.g., cryptocurrencies, sharing medical information in healthcare or exchanging goods in a business environment [1]. Each participant in the blockchain has a shared ledger (i.e., ledger which consists of a series of transactions) with others to guarantee immutability and consistency for every transaction, with each transaction verified for validity by a consensus of a majority of nodes. If a transaction has been proved, some participant makes a block that consists of a set of transactions and then updates it to the last block in the shared ledger. These blocks are connected by a hash value in the ledger, and the transactions cannot be altered by construction with its hash chain [2][3].

Recently, many communities in blockchain networks have wanted to deploy machine learning models to get computational statistics or data analysis results. For example, a medical researcher who wants to provide patient-specific treatment can train a predictive model of disease by collaborating with secure medical communities in a blockchain network without any additional process of negotiating with each other for a database [4]. However, it is often hard to collect large-scale and massive amounts of geographically distributed data in single data storage because of usability, privacy issues, security enhancement, policies, and regulations, such as GDPR (General Data Protection Regulation) [5].

The proposedensemble learning model introduces an innovative approach to executing learning models without the need for centralized data repositories within blockchain networks[6]. This pioneering solution involves the implementation of an ensemble machine-learning model tailored specifically for blockchain systems. The inherent advantage of an ensemble ML model lies in its ability to efficiently leverage distributed data spanning various domains without necessitating a centralized data server[7]. This decentralized framework not only ensures effective utilization of data scattered across multiple domains but also demonstrates promising system efficiency compared to alternative methodologies[8]. By harnessing the power of ensemble machine learning within blockchain networks, this approach offers a robust and efficient means of processing diverse and geographically dispersed datasets without the constraints of centralized data storage.

2. Literature Review:

Blockchain and machine learning in supply chain

The integration of blockchain and machine learning in supply chain management offers transformative benefits[9]. Blockchain's secure and immutable record of transactions, maintained by multiple parties, enhances transparency, reduces fraud, and automates processes through smart contracts[10]. Machine learning complements this by analysing diverse data sources to identify patterns, predict trends, and optimize decision-making, ultimately improving efficiency and reducing waste in the supply chain[11]. Notably, the combination excels in traceability, addressing critical concerns in industries like food and

pharmaceuticals[12]. Challenges include the need for standardization and collaborative initiatives, such as the Blockchain in Transport Alliance, which is working towards common protocols[13][14].

Blockchain and machine learning in medicine

In the healthcare sector, blockchain and machine learning revolutionize patient data management[15][16]. Blockchain ensures a secure and decentralized system for storing medical records, improving patient privacy and data security[17]. Machine learning analyses extensive datasets to enhance patient outcomes, aid in diagnosis, and streamline clinical trials[18][19]. The integration is particularly promising in drug supply chain management, where blockchain ensures transparency and traceability, and machine learning optimizes drug development[20]. The technology also addresses challenges in remote patient monitoring, ensuring data quality through synchronization and immutability, and facilitates the creation of disease prediction models[21].

Blockchain and machine learning in security:

In the realm of security, blockchain’s tamper-proof and transparent record, maintained by multiple parties, enhances security-related transactions' integrity[22][23]. Machine learning contributes by analysing security data to improve threat detection and response times, reducing the risk of cyber-attacks[24][25]. The combination finds applications in identity and access management[5], where blockchain’s decentralized identity system and machine learning's analysis of user behaviour enhance security by preventing identity theft and unauthorized access[26]. Despite providing robust data integrity solutions[27], blockchain faces potential threats, highlighting the need for ongoing research and security measures[28].

3. Proposed methodology:

In the realm of modern healthcare systems, establishing robust methods for secure data identification and validation stands as a cornerstone for safeguarding patient information.So this proposed model introduces a privacy-centric approach, uniquely resolving issues of privacy, security, and performance within blockchain, distinguishing itself amid prior research endeavours.

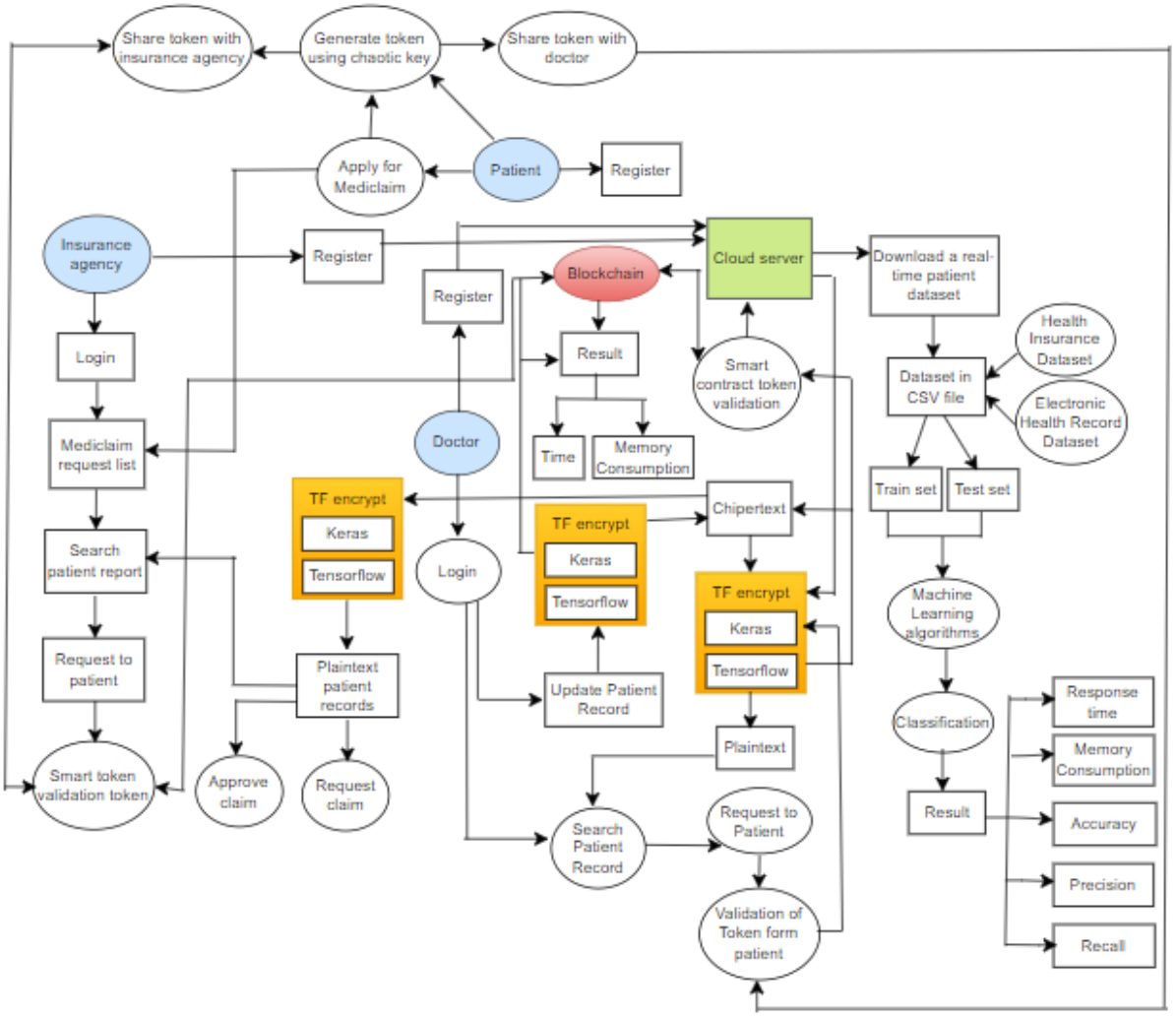

Figure 1. Block diagram of blockchain-based healthcare system

The proposed healthcare model emphasizes a secure token creation process for patient identification, utilizing cryptographic methods to generate unique, tamper-resistant tokens. These tokens serve as digital identifiers, anchoring patients' data within the system and enforcing strict access controls. Authorized healthcare professionals use the patient's token to input and update comprehensive medical information in a standardized record format.

Machine learning (ML) algorithms, including Linear Regression, Ridge Regression, or Random Forest, AdaBoost, ensemble, and gradient boosting are applied to prepare patient data for analysis. Trained ML models uncover hidden patterns, correlations, and anomalies, providing valuable insights for informed decision-making in healthcare. Continuous refinement and adaptation of ML models ensure ethical usage of patient data, maintaining privacy and mitigating biases for fair decision-making.

For external access, insurance agencies request patient records through a token-based One-Time Password (OTP) system. The system generates and securely delivers OTPs to the patient's registered mobile device. The patient's timely entry of the OTP grants temporary authorized access to requested records, minimizing the risk of unauthorized access.

Data security measures include robust encryption for sensitive information, such as patient records and analysis results, before uploading to cloud storage. Granular access controls, confidentiality, integrity checks, and continuous monitoring ensure secure storage and prevent unauthorized access or data breaches. The healthcare model prioritizes secure token creation, standardized record management, and ML analysis for informed decision-making while incorporating stringent security measures to protect patient data throughout the process.**ML algorithms used in the proposed model**

Integrating the ML algorithms into the system allows for diverse data analysis methodologies. Linear regression helps in understanding relationships,Ridge regression mitigates multicollinearity issues, and Random Forest excels in feature importance and prediction accuracy. Together, they enrich the system's analytical capabilities, enabling a comprehensive exploration of patient data for improved healthcare decision-making.

1. **Linear Regression:**Analyze patient data, identify relationships between variables, and gain insights through a fitted linear model.
   Identify relationships between various patient attributes (such as age, medical history, and test results) to understand correlations and trends.
   Helps in creating a linear model to gain insights into how specific variables affect certain health outcomes or conditions.

Definition of LR

INPUT: $\{X = x_i \; where \; i = 1,2,3, \dots \dots, n\}$ =input training data

V=test data.

Procedure:

$$\log t1\ (P_1) = In\left\{\frac{P1}{1-P1}\right\} = \omega_0 + \omega_1 X_1 + \cdots + \omega_n X_n = \omega_0 + \omega_1 X_1 + \cdots + \omega_n X_n$$

$$= \omega_0 + \sum_{k=1}^{n} \omega_k X_k$$

Where, ω_0 represents the intercept and $\omega_1, \omega_2, \dots, \omega_n$ are the coefficients related to the variable$X_1, X_2, \dots, X_n$.

P_1 represents the probability that the dependent variable Yequals 1 given the input variables

X. V is the class label assigned to the test data based on the probability P_1. After calculating the probabilityP_1, the final step in logistic regression is to assign a class label to the input.

A bipartition variable has two values like yes(1)/no (0), diabetic/non-diabetic, and alive/dead which signifies the presence or absence of some event. $X_1, X_2, \dots, X_n$ are the independent variables and may be continuous, bipartition, discrete or combination.

$$P_1(X) = \frac{1}{1 + e^{-\log it\,(P_i(x))}} = \frac{1}{1 + e^{-\{\omega_0 + \sum_{k=1}^{n} \omega_k X_k\}}}$$

OUTPUT: Class label *V*.

The maximum likelihood estimation model (MLE) is generally considered to estimate the coefficients which begin with some random input estimation of coefficients and finds the variation in magnitude and direction of the efficiency. After finding the first function, all remaining coefficients have to be tested and updatedwith the new estimation function. As the MLE model is iterative, this process continues until convergence is reached.

2. **Ridge Regression:**In this proposed healthcare model, Ridge Regression is a valuable statistical technique used for predictive modelling and analysis of patient data, particularly when dealing with multicollinearity and overfitting issues.

The definition of ridge regression classifier is defined as:

Definition of RC

INPUT: $\{X = x_i \; where \; i = 1,2,3, \ldots \ldots, n\}$ = input training data

V test data.

Procedure:

$$x^{ridge} = \arg min_{x \in R} \parallel Y - X_n \parallel_2^2 + \delta \parallel X \parallel_2^2$$

Where X is the number of features; X is the coefficient or beta; n$\sqrt{X_0^2 + \cdots + X_n^2}$. Y is the vector of target values. The term *R* represents the set of all real numbers. This notation indicates that the coefficients *x*being optimized over are real-valued.

OUTPUT: Class label V.

3. **Random Forest:** This algorithm is used for its ability to handle high-dimensional data, nonlinear relationships, and missing values making it particularly well-suited for analyzing complex healthcare datasets and improving decision-making in clinical settings.

   Rrandom forest is also one of the viable machine-learningalgorithms which are used for both classification and regression. The algorithm of random forest is written as:

Algorithm of RF:

INPUT: ={$X = x_i \; where \; i = 1,2,3, \ldots \ldots, n$} labelled training data

V=tests data

Procedure:

Step 1: find the bootstrap samples from diabetic patient's data.

Step 2: Develop an unpruned classification tree with a random sample of the predictors and

select the best split from among those variables.

Step 3: Forecast the new data by considering the majority votes of classification.

OUTPUT: *Class label V.*

6. **Gradient Boosting:** Gradient Boosting is an ensemble approach where models are built sequentially, with each subsequent model aimed at rectifying the errors of its predecessors. During training, it utilizes gradient descent to minimize the loss function.

Algorithm of GB:

INPUT: Training dataset $\{(x_i,)\}=1N$, where x_i is the feature vector and y_i is the target value.

V=tests data.

Procedure:

$$F_m(x) = F_{m-1}(x) + \alpha h_m(x) \text{where} h_m(x) = \arg min_h \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + h(x_i))$$

$F_m(x)$: Ensemble prediction after adding the m-th model.$F_{m-1}(x)$: Ensemble prediction before adding the m-th model.α: Learning rate, a small positive scalar that scales the contribution of each new model $hm(x)$.$h_m(x)$: The m-th base learner (model) added to the ensemble.

OUTPUT: *Class label V.*

7. **AdaBoost:** AdaBoost combines multiple weak learners (models) to create a strong learner through iterative training, where each subsequent model focuses on the instances misclassified by the previous ones. It assigns weights to instances, boosting the importance of misclassified ones while diminishing the influence of correctly classified instances in subsequent iterations.

Algorithm of Adaboost:

INPUT:Training dataset $\{(x_i,y_i)\}i=1N$, wherex_i is the feature vector and y_i is the true class label (typically $y_i \in \{-1,+1\}$ for binary classification)

$$F_M(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \text{where} \alpha_m = \frac{1}{2}\log\left(\frac{1-err_m}{err_m}\right)$$

$F_M(x)$: The final strong classifier after M iterations.α_mThe weight assigned to the m-th weak learner.$h(x)$ The m-th weak learner's prediction for input x.err_mThe error rate of the$m$-th weak learner.

OUTPUT: *Class label V.*

8. **Ensemble:** An ensemble technique aggregates predictions from multiple regression models, effectively leveraging their diverse strengths to enhance prediction accuracy. By combining the predictions of various models, it aims to mitigate individual model biases and uncertainties, resulting in more robust and reliable predictions.

Algorithm of Ensemble:

INPUT:Training dataset $\{(x_i,y_i)\}i=1N$, where xi is the feature vector and yi is the true class label (typically $y_i \in \{-1,+1\}$ for binary classification).

Procedure:

$$F_M(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \, where \, \alpha_m = \frac{1}{2} log(\frac{1-err_m}{err_m})$$

Final strong classifier $FM(x)$which is a weighted combination of all weak learners.

OUTPUT: *Class label V.*

4. Dataset description

- **Statistical Relationship:**

In the **Patient, Doctor, and Insurance Agency** dataset, the attributes are statistically correlated to disease or no-disease class by using the statistical r correlation formula as follows:

$$r = \frac{\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum(y_i - \overline{y})^2}}$$

Where $\overline{x}$ and $\overline{y}$ are $\sum_{i=1}^{n} x_i$ / n and $\sum_{i=1}^{n} y_i$ /n

Here r lies between -1 to 1. If 0 < r <1 then a positive correlation occurs and if −1< r < 0 then a negative correlation occurs, otherwise there is no correlation between the two variables.

- **Patient, Doctor, and Insurance Agency Datasets used in the proposed model:**

The system creates distinct datasets dedicated to storing information specific to patients, doctors, and insurance agencies separately.Each dataset is structured to hold relevant attributes and details corresponding to its entity type, such as patient demographics, medical history, doctor credentials, insurance policy information, etc. Also, new data entries are added to the respective datasets as per system operations, such as patient registrations, doctor updates, or insurance agency details.The datasets are continually managed and updated to ensure they contain accurate and current information related to each entity.Adetailed description of the dataset used in the proposed model is shown in below table I.

Table I: Description of the dataset used in the proposed model

ID	Gender	Weight	Height	Blood Group	BP	BMI	Age	Diabetes	Claim Amount
0	male	75	5.9	O+	120	23.14	45	no	15234.67
1	female	68	5.6	A-	135	24.26	62	yes	8750.5
2	male	90	6.1	B+	128	26.85	53	no	22500
3	female	60	5.4	AB+	118	22.73	38	no	5680.25
4	male	82	5.8	O-	142	27.57	70	yes	31750.8
5	female	70	5.7	A+	125	24.15	50	no	12450.3
6	male	85	6	B-	130	25.95	58	yes	18900.75
7	female	63	5.5	O+	115	23.24	42	no	7230.4
8	male	78	5.11	AB-	138	24.06	65	yes	28500.6
9	female	72	5.8	A+	122	24.15	55	no	10800.2
10	male	80	6	B+	128	24.04	48	no	14750.9
11	female	65	5.5	O-	118	23.94	40	no	9800.35
12	male	88	5.1	A-	132	27.96	60	yes	25600.7
13	female	58	5.3	AB+	125	22.73	35	no	6340.15
14	male	95	6.2	B-	145	27.07	72	yes	33800.5
15	female	70	5.7	O+	120	24.25	52	no	11200.8
16	male	82	5.11	A+	130	25.35	57	yes	20100.25
17	female	62	5.4	B+	115	23.54	43	no	8100.6
18	male	78	5.9	AB-	140	25.46	68	yes	29700.4
19	female	68	5.6	O-	128	24.25	50	no	13500.3

- **Data Retrieval and Authentication:**

During the login/authentication process, the system fetches relevant data from the appropriate datasets based on the user's role and authorization level (e.g., patient, doctor, insurance agency).For example, a patient logging in would trigger data retrieval from the patient dataset, while a doctor's login would access information from the doctor dataset.Upon fetching the required data, the system employs encryption techniques to secure the sensitive information before processing or displaying it to the authenticated user.After fetching and encrypting the authorized user's data, the system applies a timestamp-based encryption mechanism.This process involves encrypting the data with time-sensitive keys or encryption parameters, ensuring that the data remains secure and accessible only for a specified duration.The encrypted user data is securely stored in CSV files within the system's storage infrastructure, maintaining the confidentiality, integrity, and privacy of the sensitive

information.Robust access controls and encryption mechanisms safeguard the CSV files to prevent unauthorized access or tampering.

5. Results

ML algorithms enhance data analysis and prediction, improving healthcare decision-making and outcomes.

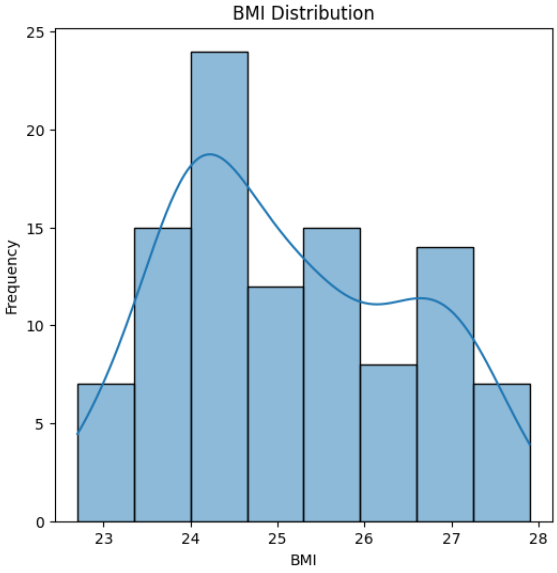

Figure 2. BMI distribution results printed by the proposed model

Figure 2's histogram depicts the Body Mass Index (BMI) distribution, indicating a prevalence of overweight or obese individuals, with a significant frequency in the 24-25 BMI range.

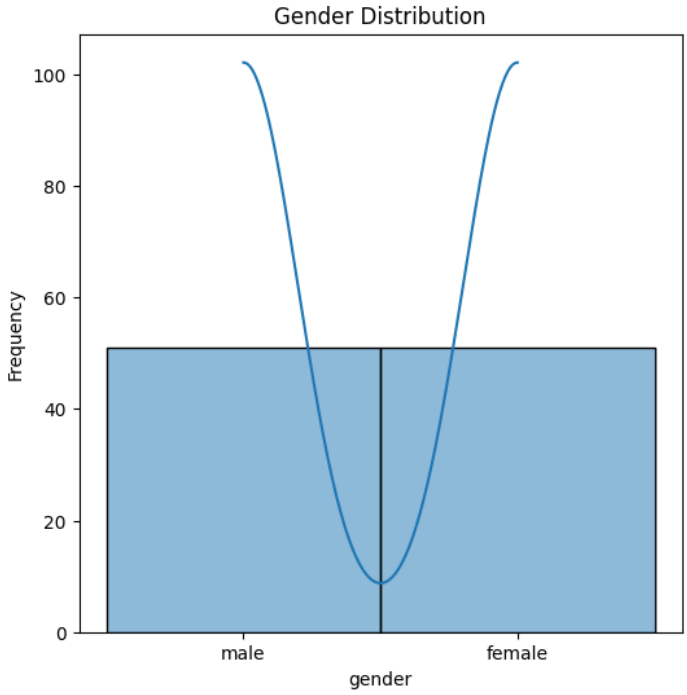

Figure 3. Gender distribution results printed by the proposed model

The graph in Figure 3 is based on a machine-learning model and predicts an equal distribution of males and females in the population. However, caveats include possible bias in the training data or issues with model accuracy.

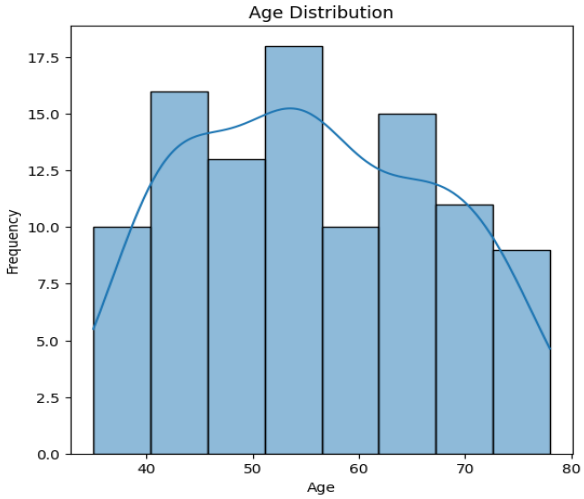

Figure 4. Age distribution results printed by the proposed model

Figure 4 shows that the probability frequency function peaks at 50 and 60 years old, indicating that the population is possibly made up of middle-aged people or pre-retirees. A lower density above 70 indicates a developing country with a shorter life expectancy.

As shown in Figures2, 3 and 4, an ensemble learning model could play a crucial role in data classification by leveraging the strengths of various algorithms to enhance the accuracy and robustness of the classification process. In the given results, the ensemble learning model could utilize various machine learning algorithms to classify patient data based on factors such as age, BMI, gender, medical history, etc. The ensemble model could combine the outputs of individual models trained on these different aspects to make more accurate predictions about healthcare outcomes, decision-making, and population demographics. This approach would help in leveraging the strengths of different algorithms and improving the overall reliability of the analysis.

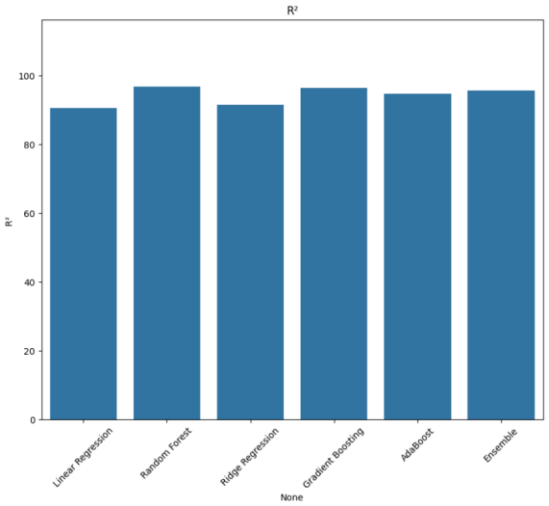

Figure 5. Values of $R^2$for algorithms used in the proposed model

The Linear Regression model on the medical record dataset has a response time of 0.0032 seconds and requires 0.0078 MB of memory. The model's Root Mean Squared Error (RMSE) is 2899.89, while its Mean Absolute Error (MAE) is 2131.88. The model's R^2 value of 0.9062 explains approximately 90.62% of the data's variance. Despite its simplicity and low computational resource requirements, the error metrics show that there is plenty of room for improvement in predictive accuracy.

The Ridge Regression model, a regularized version of linear regression, has a response time of 0.0020 seconds and uses very little memory. It outperforms the linear regression model slightly, with an RMSE of 2759.66 and an MAE of 2121.63. This model explains 91.50% of the variance in the data ($R^2 = 0.9150$). This suggests Ridge Regression provides a marginal improvement in prediction accuracy over simple linear regression while maintaining low computational demands.

The Random Forest model, an ensemble learning approach, has a substantially longer reaction time of 0.0961 seconds and uses 0.0273 MB of memory. This model exceeds the preceding ones, with an RMSE of 1665.65 and an MAE of 1337.49. The R^2 value of 0.9690 shows that the model explains 96.90% of the variation, demonstrating its good predictive ability. The trade-off is the increased computing cost compared to linear models, but the significant improvement in accuracy may justify it.

Another ensemble approach, the Gradient Boosting model, has a reaction time of 0.0959 seconds and takes up 0.0547 MB of memory. It has an RMSE of 1806.35 and an MAE of 1460.04, which are somewhat higher than the Random Forest model. Despite this, the R^2 score remains high at 0.9636, accounting for 96.36% of the variation in the data. This model provides an excellent blend of predicted accuracy and computational economy.

The AdaBoost model has a reaction time of 0.0459 seconds and minimal memory use. It has a higher RMSE of 2159.33 and MAE of 1678.96 than both the Random Forest and Gradient Boosting models. The R^2 score of 0.9480 shows that it explains 94.80% of the variation. While AdaBoost is computationally efficient, its predicted accuracy is slightly lower than the other ensemble algorithms examined.

The Ensemble model, which combines several distinct models, has a reaction time of 0.1292 seconds and memory use of 0.0703 MB. It has an RMSE of 1952.32 and an MAE of 1340.90, indicating a reasonable compromise between the individual models. The R^2 score of 0.9575 shows that the ensemble strategy explains 95.75% of the data's variability. This implies that combining models can improve prediction performance while maintaining accuracy and processing economy.

The Random Forest model outperforms the other models studied in terms of accuracy, even though it needs more computer resources. The ensemble technique also has good predictive performance, making it a feasible alternative for balancing accuracy and processing economy.

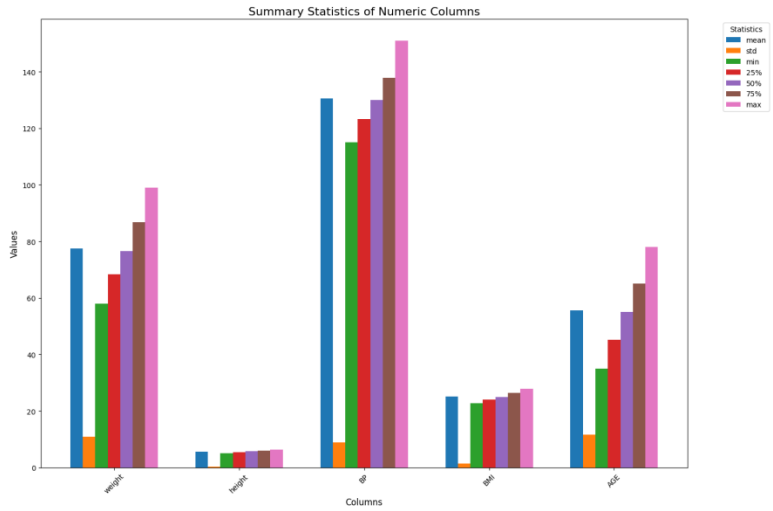

Figure 6. Analysis of data by proposed model based on patient's medical history

Figure 6 visualizes the analysis of patient data by a proposed medical diagnosis model,possibly based on various factors such as weight, height, blood pressure(BP), body mass index(BMI), and age.

6. Discussions

The proposed healthcare model is a meticulously designed flow encompassing several crucial steps to ensure robust data privacy and security. It begins with the creation of unique tokens assigned to each patient upon entry, serving as secure identifiers linked to their data within the system. These tokens play a pivotal role in initiating and maintaining patient records, allowing authorized professionals to update accurate and current information securely. The integration of machine learning algorithms enables intelligent analysis of patient data, uncovering valuable patterns and correlations for informed healthcare decisions.To authenticate and secure access to patient records, an OTP verification mechanism is implemented, requiring patients to validate access requests within specific timeframes. Meanwhile, the system ensures the encrypted storage of data, including patient records, analysis results, and tokens, in a secure cloud environment. This secure cloud storage guarantees confidentiality, integrity, and privacy, safeguarding patient information from unauthorized access, including cloud service providers. Finally, seamless integration between the system and cloud storage is established, upholding stringent data privacy and security measures throughout the entire process. This comprehensive flow ensures that from token creation to secure cloud storage and integration, the system maintains airtight security protocols, protecting patient data and adhering to privacy regulations at every step.The ensemble model enhances pattern discovery in patient data using diverse algorithms from comprehensive datasets. The model ensures healthcare data privacy and security with robust encryption and access control. Its efficiency enables real-time data analysis, facilitating timely and accurate healthcare decisions based on the rich and up-to-date patient dataset, especially in urgent medical situations. By aggregating outputs from various algorithms on the patient dataset, the ensemble model effectively reduces false positives and negatives, significantly improving the accuracy of medical diagnoses and predictions. The ensemble

model's scalability empowers it to efficiently handle large volumes of patient data, making it well-suited for healthcare organizations of different sizes and complexities while preserving the confidentiality and integrity of the dataset.

Table II: A comparative analysis of the proposed framework vs the state-of-the-art systems

Healthcare System Name	Patient Identity	Immutability	Data Auditing	Smart Contracts	Access Controls
Our framework	✓	✓	✓	✓	✓
Xiao et al. 2016 [29]	✓	✓	×	×	×
Hussein et al., 2018 [30]	✓	✓	×	×	×
Daghe et al., 2018 [31]	✓	✓	×	✓	×
Chen et al., 2019 [32]	✓	✓	×	×	×
Zhang et al., 2018 [33]	✓	✓	✓	✓	×

The proposed ML model operates as an integrated, privacy-preserving solution that surpasses benchmark systems in securing and analyzing comprehensive healthcare datasets. Compared to other state-of-the-art systems, our model exhibits superior performance through the strategic use of diverse algorithms and a token-based identification system, ensuring controlled access and data security across patient records, doctor profiles, and insurance agency details. Leveraging machine learning algorithms such as Linear Regression, Ridge Regression, Random Forest, Gradient Boosting, and AdaBoost enhances the system's capabilities, enabling intricate pattern discovery and revealing complex relationships within the datasets that might elude individual algorithms. For instance, the Random Forest model demonstrated outstanding accuracy with an R^2 of 0.9690, significantly higher than simpler models, while maintaining reasonable computational efficiency. The ensemble approach further bolsters the model’s predictive power, effectively reducing false positives and negatives, which is critical in medical diagnostics.The system's utilization of secure cloud storage with encrypted data guarantees data integrity and confidentiality during the analysis process, ensuring that sensitive healthcare information remains protected from unauthorized access. This comprehensive approach not only secures the data but also enhances the model's scalability, allowing it to efficiently handle large volumes of patient data, making it well-suited for healthcare organizations of various sizes and complexities.Despite its strengths, the model's more advanced algorithms, such as Gradient Boosting and Random Forest, come with higher computational costs and memory consumption, which might pose constraints in environments with limited resources. However, the overall benefits in terms of accuracy,

security, and scalability make these trade-offs worthwhile.This study showcases the effectiveness of this ensemble model in facilitating privacy preservation, pattern discovery, and efficient healthcare data analysis. By surpassing the capabilities of benchmark systems, the proposed model affirms its potential as an effective solution for safeguarding patient information, extracting valuable insights, and enhancing healthcare decision-making within a secure environment.

7. Conclusion

The proposed ML model operates as an integrated, privacy-preserving solution that achieves the research objective of securing and analyzing comprehensive healthcare datasets. By employing a variety of algorithms, including Linear Regression, Ridge Regression, Random Forest, Gradient Boosting, and AdaBoost, alongside a token-based identification system, it ensures controlled access and data security across patient records, doctor profiles, and insurance agency details. The strategic use of these machine learning algorithms enhances the system's capabilities, enabling intricate pattern discovery and revealing complex relationships within the datasets that might elude individual algorithms.The system's utilization of secure cloud storage with encrypted data guarantees data integrity and confidentiality during the analysis process, ensuring that sensitive healthcare information remains protected. The robust ensemble model significantly reduces false positives and negatives, thus improving the accuracy of medical diagnoses and predictions. Additionally, its scalability allows it to efficiently handle large volumes of patient data, making it suitable for healthcare organizations of varying sizes and complexities.This study showcases the effectiveness of the ensemble model in facilitating privacy preservation, pattern discovery, and efficient healthcare data analysis. Overall, the proposed model enhances the system's ability to safeguard patient information, extract valuable insights, and conduct efficient analysis within a secure environment, affirming its potential as an effective solution for preserving privacy, conducting in-depth pattern analysis, and enhancing healthcare decision-making.

8. Future Work

Future work in the proposed ensemble model can focus on investigating methods to improve the scalability of the ensemble model, ensuring its efficiency as the volume of data and participants within the blockchain network increases. Exploring optimizations and parallelization techniques to handle larger datasets without compromising performance.

References

[1] E. Androulaki *et al.*, “Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains,” in *Proceedings of the 13th EuroSys Conference, EuroSys 2018*, ACM, Apr. 2018. doi: 10.1145/3190508.3190538.

[2] F. Benhamouda, S. Halevi, and T. Halevi, “Supporting private data on Hyperledger Fabric with secure multiparty computation,” *IBM J. Res. Dev.*, vol. 63, no. 2, 2019, doi: 10.1147/JRD.2019.2913621.

[3] S. Underwood, “Blockchain beyond Bitcoin,” *Commun. ACM*, vol. 59, no. 11, pp. 15–17, 2016, doi: 10.1145/2994581.

[4] N. Hynes, D. Dao, D. Yan, R. Cheng, and D. Song, “A demonstration of sterling: A privacy-preserving data marketplace,” in *Proceedings of the VLDB Endowment*, PVLDB, 2018, pp. 2086–2089. doi:

10.14778/3229863.3236266.

[5] H. Kim, S.-H. Kim, J. Y. Hwang, and C. Seo, "Efficient Privacy-Preserving Machine Learning for Blockchain Network," *IEEE Access*, vol. 7, pp. 136481–136495, 2019, doi: 10.1109/ACCESS.2019.2940052.

[6] T. Aljrees, A. Kumar, K. U. Singh, and T. Singh, "Enhancing IoT Security through a Green and Sustainable Federated Learning Platform: Leveraging Efficient Encryption and the Quondam Signature Algorithm," *Sensors*, vol. 23, no. 19, p. 8090, Sep. 2023, doi: 10.3390/s23198090.

[7] Y. Alotaibi and M. Ilyas, "Ensemble-Learning Framework for Intrusion Detection to Enhance Internet of Things' Devices Security," *Sensors*, vol. 23, no. 12, p. 5568, Jun. 2023, doi: 10.3390/s23125568.

[8] P. M. Dhulavvagol, V. H. Bhajantri, and S. G. Totad, "Blockchain Ethereum Clients Performance Analysis Considering E-Voting Application," *Procedia Comput. Sci.*, vol. 167, pp. 2506–2515, 2020, doi: 10.1016/j.procs.2020.03.303.

[9] B. K. Mohanta, S. S. Panda, and D. Jena, "An Overview of Smart Contract and Use Cases in Blockchain Technology," in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2018, pp. 1–4. doi: 10.1109/ICCCNT.2018.8494045.

[10] H. Taherdoost, "The Role of Smart Contract Blockchain in 6G Wireless Communication System," *Procedia Comput. Sci.*, vol. 215, pp. 44–50, 2022, doi: 10.1016/j.procs.2022.12.005.

[11] N. Szabo, "Formalizing and Securing Relationships on Public Networks," *First Monday*, vol. 2, no. 9, Sep. 1997, doi: 10.5210/fm.v2i9.548.

[12] D. Mingxiao, M. Xiaofeng, Z. Zhe, W. Xiangwei, and C. Qijun, "A review on consensus algorithm of blockchain," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, Oct. 2017, pp. 2567–2572. doi: 10.1109/SMC.2017.8123011.

[13] T. Bakhtiar, X. Luo, and I. Adelopo, "The impact of fundamental factors and sentiments on the valuation of cryptocurrencies," *Blockchain Res. Appl.*, vol. 4, no. 4, p. 100154, Dec. 2023, doi: 10.1016/j.bcra.2023.100154.

[14] S. Kayikci, "A Deep Learning Method for Passing Completely Automated Public Turing Test," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, IEEE, Sep. 2018, pp. 41–44. doi: 10.1109/UBMK.2018.8566318.

[15] Y. Tian, T. Li, J. Xiong, M. Z. A. Bhuiyan, J. Ma, and C. Peng, "A Blockchain-Based Machine Learning Framework for Edge Services in IIoT," *IEEE Trans. Ind. Informatics*, vol. 18, no. 3, pp. 1918–1929, Mar. 2022, doi: 10.1109/TII.2021.3097131.

[16] H. Vargas, C. Lozano-Garzon, G. A. Montoya, and Y. Donoso, "Detection of Security Attacks in Industrial IoT Networks: A Blockchain and Machine Learning Approach," *Electronics*, vol. 10, no. 21, p. 2662, Oct. 2021, doi 10.3390/electronics10212662.

[17] A. OUTCHAKOUCHT, H. ES-SAMAALI, and J. Philippe, "Dynamic Access Control Policy based on Blockchain and Machine Learning for the Internet of Things," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 7, 2017, doi: 10.14569/IJACSA.2017.080757.

[18] S. S. Kamble, A. Gunasekaran, V. Kumar, A. Belhadi, and C. Foropon, "A machine learning based approach for predicting blockchain adoption in Supply Chain," *Technol. Forecast. Soc. Change*, vol. 163, p. 120465, Feb. 2021, doi: 10.1016/j.techfore.2020.120465.

[19] A. Goyal, A. Elhence, V. Chamola, and B. Sikdar, "A Blockchain and Machine Learning based Framework for Efficient Health Insurance Management," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, New York, NY, USA: ACM, Nov. 2021, pp. 511–515. doi: 10.1145/3485730.3493685.

[20] H. Hasanova, M. Tufail, U.-J. Baek, J.-T. Park, and M.-S. Kim, "A novel blockchain-enabled heart disease prediction mechanism using machine learning," *Comput. Electr. Eng.*, vol. 101, p. 108086, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108086.

[21] F. F. Alruwaili, B. Alabduallah, H. Alqahtani, A. S. Salama, G. P. Mohammed, and A. A. Alneil,

"Blockchain Enabled Smart Healthcare System Using Jellyfish Search Optimization With Dual-Pathway Deep Convolutional Neural Network," *IEEE Access*, vol. 11, pp. 87583–87591, 2023, doi: 10.1109/ACCESS.2023.3304269.

[22] J. Passerat-Palmbach *et al.*, "Blockchain-orchestrated machine learning for privacy-preserving federated learning in electronic health data," in *2020 IEEE International Conference on Blockchain (Blockchain)*, IEEE, Nov. 2020, pp. 550–555. doi: 10.1109/Blockchain50366.2020.00080.

[23] K. Abbas, M. Afaq, T. Ahmed Khan, and W.-C. Song, "A Blockchain and Machine Learning-Based Drug Supply Chain Management and Recommendation System for Smart Pharmaceutical Industry," *Electronics*, vol. 9, no. 5, p. 852, May 2020, doi: 10.3390/electronics9050852.

[24] R. Chowdhury, M. A. Rahman, M. S. Rahman, and M. R. C. Mahdy, "An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning," *Phys. A Stat. Mech. its Appl.*, vol. 551, p. 124569, Aug. 2020, doi: 10.1016/j.physa.2020.124569.

[25] S. Aladhadh, H. Alwabli, T. Moulahi, and M. Al Asqah, "BChainGuard: A New Framework for Cyberthreats Detection in Blockchain Using Machine Learning," *Appl. Sci.*, vol. 12, no. 23, p. 12026, Nov. 2022, doi: 10.3390/app122312026.

[26] A. Altarawneh, T. Herschberg, S. Medury, F. Kandah, and A. Skjellum, "Buterin's Scalability Trilemma viewed through a State-change-based Classification for Common Consensus Algorithms," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, Jan. 2020, pp. 0727–0736. doi: 10.1109/CCWC47524.2020.9031204.

[27] S. Martinazzi and A. Flori, "The evolving topology of the Lightning Network: Centralization, efficiency, robustness, synchronization, and anonymity," *PLoS One*, vol. 15, no. 1, p. e0225966, Jan. 2020, doi: 10.1371/journal.pone.0225966.

[28] A. K. Gogineni, S. Swayamjyoti, D. Sahoo, K. K. Sahu, and R. Kishore, "Multi-Class classification of vulnerabilities in smart contracts using AWD-LSTM, with pre-trained encoder inspired from natural language processing," *IOP SciNotes*, vol. 1, no. 3, p. 035002, Dec. 2020, doi 10.1088/2633-1357/abcd29.

[29] X. Yue, H. Wang, D. Jin, M. Li, and W. Jiang, "Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control," *J. Med. Syst.*, vol. 40, no. 10, Oct. 2016, doi: 10.1007/s10916-016-0574-6.

[30] A. F. Hussein, N. Arun Kumar, G. Ramirez-Gonzalez, E. Abdulhay, J. M. R. S. Tavares, and V. H. C. de Albuquerque, "A medical records managing and securing blockchain-based system supported by a Genetic Algorithm and Discrete Wavelet Transform," *Cogn. Syst. Res.*, vol. 52, pp. 1–11, 2018, doi: 10.1016/j.cogsys.2018.05.004.

[31] G. G. Dagher, J. Mohler, M. Milojkovic, and P. B. Marella, "Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology," *Sustain. Cities Soc.*, vol. 39, pp. 283–297, 2018, doi: 10.1016/j.scs.2018.02.014.

[32] Y. Chen, S. Ding, Z. Xu, H. Zheng, and S. Yang, "Blockchain-Based Medical Records Secure Storage and Medical Service Framework," *J. Med. Syst.*, vol. 43, no. 1, Nov. 2018, doi: 10.1007/s10916-018-1121-4.

[33] A. Zhang and X. Lin, "Towards Secure and Privacy-Preserving Data Sharing in e-Health Systems via Consortium Blockchain," *J. Med. Syst.*, vol. 42, no. 8, Aug. 2018, doi: 10.1007/s10916-018-0995-5.