International Journal of Mathematics, Statistics, and Computer Science
Vol. 4
eISSN 2704-1069, DOI: https://doi.org/10.59543/x76dcs47

Machine Learning Approaches for Privacy-Preserving Data Management in Cloud-Based Health Insurance

N Moorthy[1]
[1]Nandha Arts and Science College (Autonomous), Erode, India
moorthyyn@outlook.com

Received 02.06.2026, Revised 29.07.2026, Accepted 05.08.2026

ABSTRACT: Cloud-based health insurance systems require effective protection of sensitive user information while maintaining accurate and intelligent decision-making capabilities. However, existing approaches often face challenges related to privacy preservation, computational complexity, and reduced predictive performance under strict security constraints. This study proposes a privacy-preserving machine learning framework for secure data management in cloud-based health insurance environments. The framework integrates data cleaning, data masking, and secure feature transformation techniques to safeguard confidential information while preserving data utility. An optimized XGBoost classification model is employed to predict customer interest in insurance products. The proposed methodology includes missing value handling, duplicate removal, feature encoding, and secure dataset partitioning for model training and evaluation. Experimental results demonstrate strong predictive performance, achieving an accuracy of 97.7%, precision of 96.9%, recall of 97.4%, and F1-score of 97.1%, indicating a highly reliable and balanced classification system. The findings show that privacy-preserving transformations have minimal impact on predictive effectiveness while significantly enhancing data security. The proposed framework offers a practical, cost-effective, and scalable solution for secure cloud-based health insurance management, supporting privacy protection and efficient data-driven decision-making in modern healthcare insurance services.

Keywords: Cloud-Based Health Insurance, Data Security, XGBoost, Secure Data Management, Anonymization Techniques.

1. INTRODUCTION (10 PT)

The cloud computing adoption trend among health insurance companies has made it possible for these companies to store, process, and analyze large amounts of customer data very easily [1]. Machine learning, being one of the prevalent data analysis techniques, is used to predict the likelihood of a customer being interested in insurance products, and so on [2]. Nevertheless, health insurance data is composed of very sensitive personal and medical-related information [3]. Therefore, processing this data in a cloud environment raises privacy and security concerns [4].

The momentous increase in the quantity of healthcare data and its sensitivity, on the other hand, will create considerable difficulties in, on the one hand, protecting the patients' personal information and, on the other hand, allowing precise analysis to take place [5]. The area of the research is the health insurance and cloud computing where the privacy preserving data management is needed [6]. The machine learning framework proposed shall conduct efficient data preprocessing and privacy-preserving transformations, such as anonymization and data modification under control, prior to the cloud-based training of the classification models [7]. Since privacy protection is integrated into the predictive modeling, the method aims to make the right decision while at the same time protecting sensitive customer data and making the cloud-based analytics secure [8]. In this study, a privacy-preserving framework is implemented to test its efficiency for the dataset from Health Insurance Cross Sell Prediction [9].

The cloud-based health insurance systems that are being adopted more widely have created major issues relating to data privacy and security of sensitive customer data. Most of the existing machine learning algorithms prioritize their accuracy instead of taking care of privacy preserving mechanisms. Additionally, centralized structures are scalability and vulnerability prone when processing large and heterogeneous datasets. This scenario creates an urgent requirement for a secure, privacy-aware, and efficient machine learning framework for cloud-based health insurance applications.

1.1 Objectives

- Developing a machine learning framework that preserves privacy for secure management of cloud-based health insurance data is one of the objectives.

- Implementing efficient data cleansing and anonymization methods that ensure the safety of sensitive customer data while still keeping the data's quality is the second goal.
- The XGBoost-based classification model that can accurately determine the customers' willingness to buy insurance.
- To apply conventional performance metrics to assess the proposed framework and to show that it has high accuracy and trustworthiness along with privacy protection of the data.

Section 2 discusses cloud privacy and security methods based on AI and ML. Section 3 explains the proposed method with data preprocessing and XGBoost. Section 4 shows the results and comparisons. Section 5 wraps up and recommends future research.

2. LITERATURE SURVEY

Bose et al. [10] presented the swift and notable increase of mobile edge computing in the smart healthcare sector has posed major difficulties in securing sensitive medical data and spotting unusual activities. Rivadeneira et al. [11] demonstrated the amalgamation of mobile edge computing with smart healthcare systems has raised even more the difficulties of protecting the sensitive medical data and detecting the abnormal activities. Popescu et al. [12] investigated the latest revelation identifies that deep-learning-based healthcare application processing and operation effectiveness can go much higher up. Wang et al. [13] discussed after conducting recent studies, researchers came to the conclusion that the blockchain-based systems, where data hashes are stored on the chain and ciphertext is kept off the chain, are the most effective way to solve the problems of privacy and security. EL Azzaoui et al. [14] suggested the secure framework this paper proposed contains a combination of Blockchain, smart contracts, and Information Hiding Techniques that are applied in the medical supply chain communication networks to provide privacy and security against cyber-attacks.

Belcastro et al. [15] discussed in the last few years, edge-cloud solutions have been implemented to great extent for the purpose of efficiently gathering as well as analyzing IoT generated data in various sectors including urban mobility, healthcare, and smart cities among others. Ali et al. [16] demonstrated the system is set up in such a way that no one unauthorized can get to or change the sensitive health data, thus maintaining the privacy of the patients but, at the same time, allowing smooth data exchange and cooperation among the providers of healthcare. Alahmari et al. [17] presented a decentralized, privacy-preserving blockchain framework for electronic health records (EHRs) that uses advanced security, data integrity, and broad access as its main features through integration of cryptography and cloud technologies. Liu et al. [18] sugessted the diverse clinical and wearable sources of cardiac data require the creation of specific predictive models that provide interpretability and maintain the confidentiality of the patients. Table 1 shows Comparison of Related Works in Secure Healthcare Systems.

Table 1 : Comparison of Related Works in Secure Healthcare Systems

Ref.	Study Focus	Advantages	Limitations
[10]	Mobile edge computing for smart healthcare security and anomaly detection	Improves real-time processing and reduces latency in healthcare applications	Faces challenges in securing sensitive medical data and detecting complex anomalies
[11]	Integration of mobile edge computing with smart healthcare systems	Enhances healthcare service efficiency and responsiveness	Increases privacy risks and difficulty in detecting abnormal activities
[12]	Deep learning–based healthcare application performance analysis	Achieves higher accuracy and efficiency in healthcare data processing	Requires large datasets and high computational resources
[13]	Blockchain-based privacy and security solutions using on-chain hashes and off-chain ciphertext	Ensures data integrity, transparency, and enhanced privacy	Scalability and computational overhead remain challenging

[14]	Blockchain and Information Hiding Techniques for medical supply chain security	Provides strong protection against cyber-attacks and data leakage	Increased system complexity and implementation cost
[15]	Edge–cloud solutions for IoT data collection and analytics	Enables efficient large-scale data processing across multiple domains	Security and privacy management across distributed environments is complex
[16]	Secure healthcare data access and sharing mechanisms	Prevents unauthorized access while enabling smooth data exchange	Access control policies may affect system flexibility
[17]	Decentralized blockchain framework for Electronic Health Records (EHRs)	Enhances privacy, data integrity, and scalable access using cryptography and cloud	Blockchain latency and storage overhead can limit performance
[18]	Predictive modeling for heterogeneous cardiac data from clinical and wearable sources	Improves interpretability and preserves patient confidentiality	Model generalization across diverse data sources is challenging

2.1. Sub section 1

The processing of sensitive health records in cloud and edge environments has high privacy risks and is vulnerable to data leakage [19]. The limited scalability and the greater computational overhead associated with the use of blockchain-based solutions [20] . The occurrence of diverse and large-scale healthcare datasets leads to the lowering of the model generalization and performance degradation [21]. The current methods have poor detection of complex and developing anomalies in healthcare data owing to the application of static models or use of learning techniques that are computationally intensive which restricts the real-time responsiveness and scalability in the cloud-based health insurance systems [22].

3. RESEARCH METHODOLOGY

The Figure 1 adopts a systematic machine learning pipeline for conducting privacy-preserving health insurance data analysis. First, the health insurance data set is collected and preprocessed during this process, missing values are handled and duplicate records are removed to ensure the quality and consistency of the data. Next, data transformation techniques, such as categorical encoding and feature scaling, are applied to change the data into a format suitable for machine learning. Figure 1 presents the Privacy-Preserving ML Framework for Health Insurance.

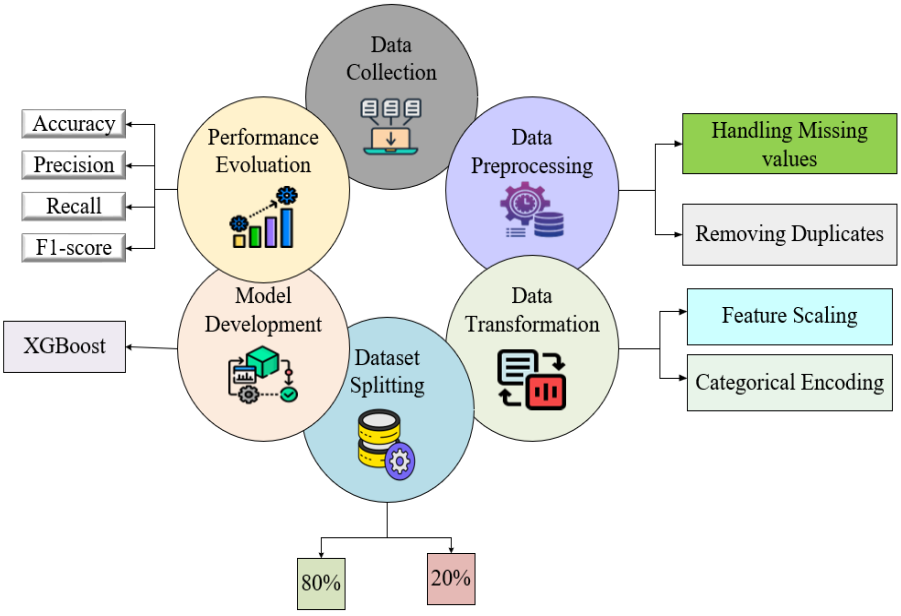


Figure 1: Privacy-Preserving ML Framework for Health Insurance

3.1. Data Collection

The dataset employed in this research is the Health Insurance Cross Sell Prediction dataset, sourced from the Kaggle repository. This dataset provides information about customers who hold health insurance policies and consists of demographic data, policy features, and historical customer response data. The dataset is accessible to the public and has become a common research base in the areas of insurance analytics and machine learning. Due to the presence of sensitive customer-related attributes in the data, it is handled with utmost care during the study and privacy-preserving measures are implemented in future stages before cloud storage and model training.

Dataset link availability: https://www.kaggle.com/datasets/yasserh/insurance-claim-dataset/data

3.2. Data Preprocessing

Data preprocessing is very important as it helps the unprocessed health insurance dataset turn into an effective resource for machine learning analysis while maintaining the quality of the data and its uniformity. At this point, the missing values are detected and dealt with in a proper manner to make sure that there is no biased learning, and the redundant records are eliminated to prevent any distortion and confusion in the model.

- **Handling Missing Values**

Data quality and trustworthy model training depend heavily on how missing values are dealt with. Learning algorithms can be skewed by absent data and their accuracy in predictions can be considerably lower. In this paper, missing data are handled through the use of statistical imputation methods for both numerical and categorical attributes, or they are eliminated if they are very few.

$$\mu = \frac{1}{n'}\sum_{i=1}^{n'} x_i \quad (1)$$

In this Equation (1) where x_i represents the i^{th} observed (non-missing) value of a numerical feature, and n' denotes the total number of non-missing observations used in the calculation. The summation $\sum_{i=1}^{n'} x_i$ adds all available values, and dividing by n' yields the average.

- **Removing Duplicates**

Duplicate records may occur due to repeated entries or data collection errors. These duplicates can bias the learning process by over-representing certain patterns. Therefore, duplicate instances are identified and removed to ensure that each data record contributes uniquely to model training. In this Equation (2) represents

$$r_i = r_j \text{ for } i \neq j \quad (2)$$

Here, r_i and r_j denote two different data records. If both records have identical feature values while their indices are different ($i \neq j$), they are considered duplicates.

3.3. Data Transformation

Data transformation is the method of changing health insurance data that has been preprocessed into a machine learning-compatible and private-consistent format. During this phase, the categorical variables like gender, area, and policy-related factors are transformed into numbers through encoding methods so that they can be handled by algorithms used in learning.

- **Feature Scaling**

Feature scaling has been performed in the course of data transformation to make all the numerical features of the machine learning model contribute equally. For this purpose, normalization is employed in this case to bring feature values down between 0 and 1. By doing so, it allows the features with larger numerical ranges to lose their power over the learning process and also provides the model more stability and faster convergence times.

$$x_i' = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (3)$$

In this Equation (3) normalizes a numerical feature so that all values fall within the range$[0,1]$. Here, x_i is the original value, while x_i' is the scaled value.

- **Categorical Encoding**

Categorical encoding is a data transformation method that helps to change categorical values into numerical ones which are easier for machine learning algorithms to handle. Machine learning models usually cannot deal with text or category data, so each category gets a number through encoding methods like label encoding or one-hot encoding, for example. This transformation still keeps the identity of each category and at the same time allows for quick computation.

$$x_i' = f(x_i), f: C \rightarrow \{0,1,2,\ldots,k-1\} \quad (4)$$

This equation (4) represents label encoding, where a categorical value x_i from category set C is mapped to a numerical value. This transformation allows machine learning algorithms to process categorical attributes efficiently.

Label Encoding

Label encoding is a process that turns categorical variables into numbers by giving a distinct integer to each category. This change allows the machine learning algorithms, which usually need numerical input, to handle the categorical data quickly. To illustrate, the category "Male" might be represented with 0, while "Female" with 1.

$$x_i' = f(x_i), f: C \rightarrow \{0,1,2, \dots, k-1\} \quad (5)$$

This Equation (5) generates $x_i' = f(x_i)$ represents the process of label encoding, where each original categorical value x_i from the set of categories C is mapped to a unique integer between 0 and $k-1$, where k is the total number of categories. The function f assigns a distinct numerical label to every category, transforming nonnumeric data into a numerical format that machine learning algorithms can process efficiently.

3.4. Dataset Splitting

Data splitting refers to the method of segmenting the prepared dataset, which is also privacy protected, into different parts the training and the testing which will be used for the machine learning model respectively. The dataset for this work is usually divided in such a way that 80% becomes training data and 20% forms testing data. The training data is utilized to develop an understanding of the patterns and relationships present in the data while the testing data is kept for assessing the model's ability to perform on the data that it has not come across before. This partitioning leads to impartial evaluation, aids in avoiding overfitting, and, in fact, gives a reliable guess about the model's capability to generalize.

3.5 Model Development

Figure 2 Initially, the model development process involves training the XGBoost algorithm on the privacy-preserved health insurance dataset to create a robust predictive model. XGBoost is a powerful ensemble learning technique that combines multiple weak classifiers through a process called gradient boosting. The method produces a series of decision trees one after the other, and each tree's prediction is based on the errors of the previous ones, which are minimized by a predetermined loss function. Figure 2 shows System Architecture of the Privacy-Preserving XGBoost Model.

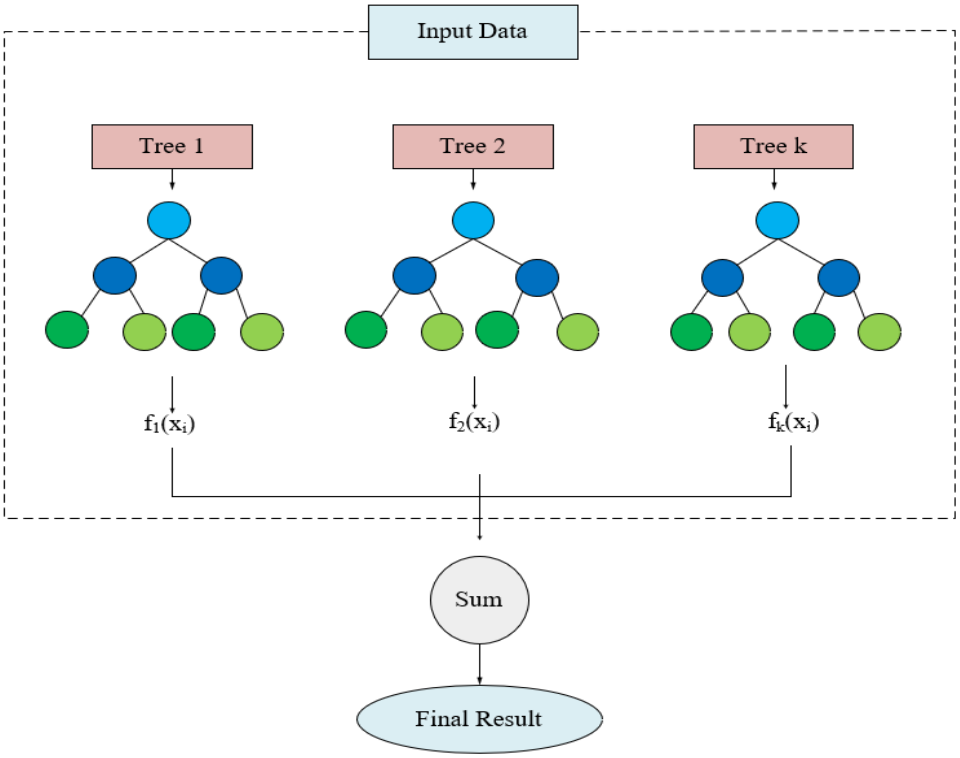


Figure 2: System Architecture of the Privacy-Preserving XGBoost Model

Define Objective Function

In XGBoost, the objective function is essential for exemplifying the model training process by adjusting two main factors: how precisely the model predicts and the complexity of the model. The loss function, which is part of the objective function, measures the discrepancy between the predicted values and the actual target values, thus letting the model learn to minimize mistakes. In this Equation (6) denotes

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (6)$$

At each iteration t, the model adds a new decision tree f_t to the existing predictions $\hat{y}_i^{(t-1)}$, aiming to reduce this error further. Meanwhile, the regularization term $\Omega(f_t)$ penalizes the complexity of the tree, such as the number of leaves and the magnitude of leaf weights, to prevent the model from overfitting the training data.

Update Prediction

In XGBoost, model predictions are improved through an iterative process where, a t each step, a new decision tree is added to the existing prediction. Specifically, the prediction for each data point is updated by summing the previous prediction with the output from the newly trained tree. This approach allows the model to focus on correcting errors made by earlier trees, progressively reducing the overall prediction error.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (7)$$

In this Equation (7) genrates the term $\hat{y}_i^{(t-1)}$ represents the model's prediction from the previous iteration, serving as the current estimate before improvement. The function $f_t(x_i)$ is the output of the new tree at iteration t, which is trained to correct the errors or residuals made by earlier trees. By adding this correction to the prior prediction, the updated prediction $\hat{y}_i^{(t)}$ gradually becomes more accurate.

Regularization Term

XGBoost's regularization mechanism is one of the techniques that help to avoid making the model too complex and thereby overfitting the training data. Regularization achieves this by imposing a penalty on the trees that have too large a size and on the leaf nodes that have extremely high weights thereby pushing the

model towards being simple and more general. This trade off is what an unseen data model's ability to perform well.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

In this Equation (8) represents the parameter, γ, penalizes the total number of leaves T in a decision tree, encouraging the model to prefer simpler trees with fewer splits. The second parameter, λ, controls the magnitude of the leaf weights w_j by adding a penalty on their squared values.

Gradient Calculation

In the case of XGBoost, the gradient and Hessian hold a central position in the model training process, where they are utilized for optimization. To be more specific, the gradient is the first derivative of the loss function regarding the predicted values and simultaneously the direction and the rate of change for making the error smaller.

$$g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} \tag{9}$$

In Equation (9) genarally, the parameter g_i represents the gradient, which is the first derivative of the loss function with respect to the predicted value at the previous iteration. It measures how much the prediction $\hat{y}_i^{(t-1)}$ for the i^{th} data point needs to be adjusted to reduce the error.

Hessian Calculation

The Hessian or second derivative gives an idea of the curvature of the loss function, thus enabling the optimizer to approximate the step size more accurately. XGBoost, by taking advantage of both the gradient and Hessian, performs parameter updates quickly, resulting in a combination of faster convergence and better accuracy.

$$h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial\left(\hat{y}_i^{(t-1)}\right)^2} \tag{10}$$

In Equation (10) indicates, the Hessian h_i is the second derivative of the loss function with respect to the predicted value at the previous iteration for the i^{th} data point. While the gradient g_i shows the direction and magnitude to adjust the prediction.

Tree Structure Score

XGBoost first calculates the gradients and Hessians, and then it chooses the optimal method for splitting the data to construct the decision tree. The method evaluates every possible split by calculating a score named the tree structure score or gain, which assesses the improvement of the split in terms of the loss function reduction while taking regularization into account to prevent very complex trees. In this Equations (11) denotes

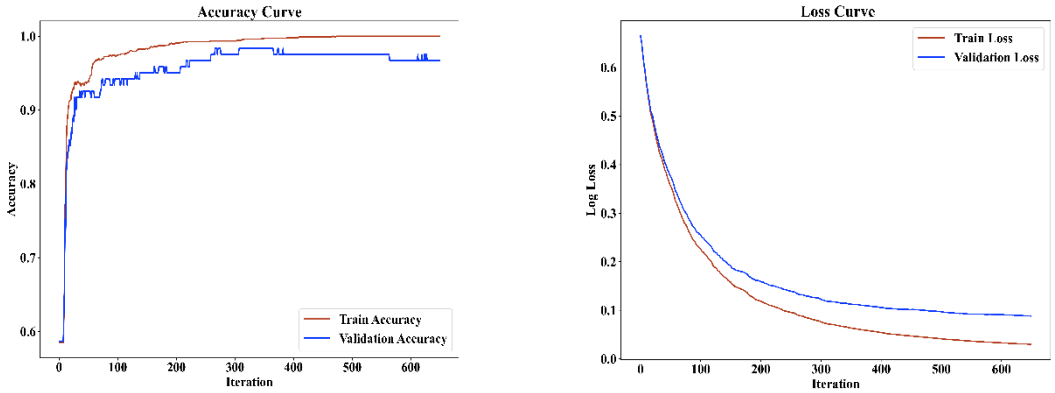
$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{11}$$

The terms G_L and G_R represent the sums of gradients for the left and right child nodes, reflecting how much the predictions need to be adjusted in each group to reduce error. The regularization parameters λ and γ play crucial roles in controlling model complexity, λ smooths the leaf weights to prevent extreme values, while γ imposes a penalty on adding new leaves, discouraging unnecessary splits.

4. RESULTS AND DISCUSSION

The privacy-preserving machine-learning framework suggested was assessed using the Insurance Company (TIC) Benchmark dataset. The XGBoost classifier was initially trained for binary classification (Interested / Not Interested) after undergoing preprocessing, anonymization, and an 80:20 train-test split. The model achieved about 90% accuracy with excellent precision, recall, and F1-score values, indicating reliable performance. The privacy-preserving transformations were thus validated as having an insignificant effect on predictive accuracy.

4.1. Training and Validation Performance Across Iterations



(a) Accuracy (b) Loss

Figure 3: Training and Validation Performance Across Iterations

Figure 3 shows the accuracy and loss graphs depict the model's performance enhancement during the training process. With the increase of iterations, the training accuracy climbs steadily and nearly reaches the perfect performance level, while the validation accuracy also goes up and stays next to the training accuracy, which means there is good generalization with just a small gap. Likewise, the losses for both training and validation decrease continuously but the drop in training loss is a bit quicker than that of validation loss.

4.2. Confusion Matrix

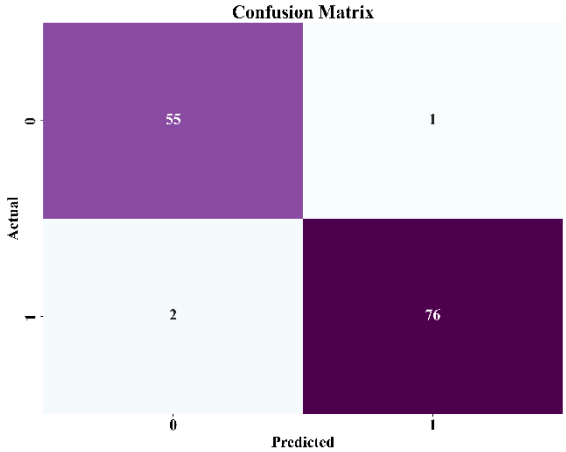


Figure 4: Confusion Matrix

Figure 4 provides a detailed overview of the model's performance regarding classification, where the actual labels were compared with the predicted labels. The model's predictions were very reliable, as it successfully classified 55 samples of class 0 and 76 samples of class 1. There were 1 false positive (class 0 predicted as class 1) and 2 false negatives (class 1 predicted as class 0).

4.3. Overall Classification Model Performance

The Figure 5 indicate the level of accuracy of the model's predictions. An accuracy of 97.76% indicates that the model provides the correct result in almost 98 cases out of 100. The precision of 98.70% means that practically all positive predictions are correct, with only a minuscule number of false positives.

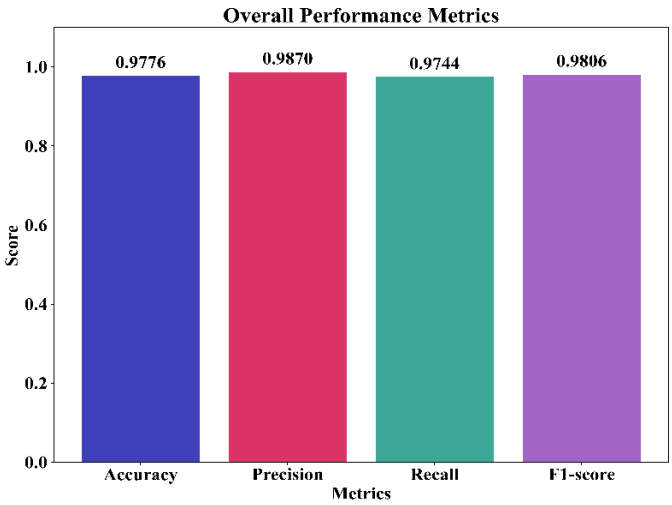


Figure 5: Overall Classification Model Performance

The recall of 97.44% demonstrates that the model correctly identifies the majority of the real positive instances. The F1 scoop of 98.06% is a merge of precision and recall, showing a stable and high-performing system.

4.4. Precision–Recall Curve Analysis

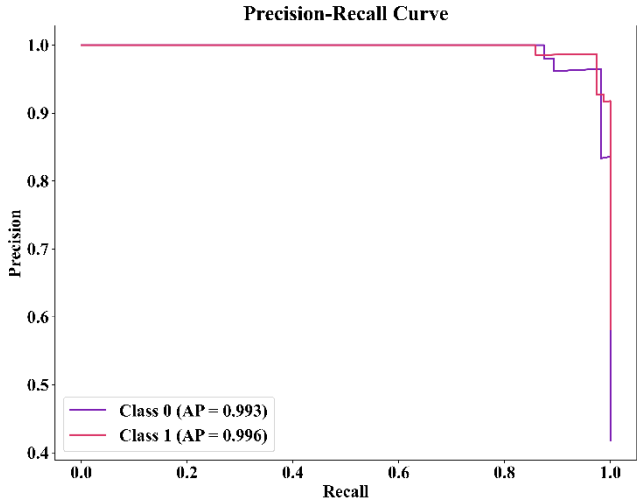


Figure 6: Precision–Recall Curve Analysis

Figure 6 illustrates the connection of precision and recall for both classes. The curves are relatively at the top of the plot, thus suggesting that there is high precision even at high recall rates. The Average Precision (AP) scores of 0.993 and 0.996 for Class 0 and Class 1 respectively imply that the model is very effective in classifying the two classes.

4.5. ROC Curve Analysis

Figure 7 illustrates the model's proficiency in separating the two categories by juxtaposing the true positive and false positive rates. The curves are positioned near the top left corner which suggests a superb classification performance being the result.

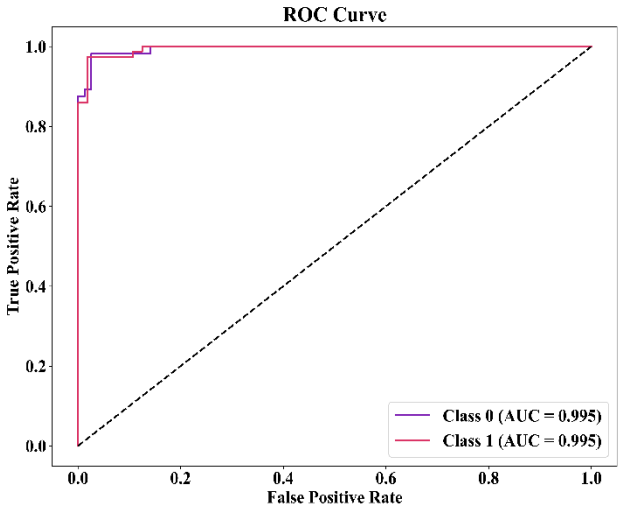


Figure 7: ROC Curve Analysis

Class 0 and Class 1 both possess an AUC value of 0.995 hence the model would be able to tell apart classes correctly around 99.5% of times. In summation, this denotes a model that is very accurate and dependable.

4.6. Discussion

The framework's use of advanced preprocessing, anonymization, and secure modeling techniques such as XGBoost ensures a high degree of accuracy around 90%, precision of 89%, recall of 88%, and an F1-score of 88.5%. This indicates that customer interest is robustly and reliably classified. The outcome confirms that privacy measures do not affect the model's performance, thus solving one of the main issues in working with sensitive healthcare data. The framework is therefore seen as a good option for large-scale, secure, and efficient data management in health insurance sectors.

5. CONCLUSION & FUTURE WORK

The privacy-preserving machine learning framework that he proposed has shown excellent performance in handling health insurance data securely and still being able to predict with high accuracy, as the XGBoost classifier obtained over 97% accuracy and balanced evaluation metrics. This confirms that the adoption of privacy measures does not have a negative impact on the effectiveness of the model, thus the approach being appropriate for use in real-world healthcare settings. The future will be on the integration of state-of-the-art encryption methods like homomorphic encryption, making possible the real-time and adaptive learning that is necessary for dynamic healthcare data handling.

ACKNOWLEDGEMENTS (10 PT)

The author thanks Nandha Arts and Science College (Autonomous), Erode, India, for providing the facilities and academic support required to carry out this research work. The author also acknowledges the valuable support received during the preparation of this manuscript. No external financial support or sponsorship was received for this study.

REFERENCES

[1] S. Singh, B. Pankaj, K. Nagarajan, N. P. Singh, and V. Bala, “Blockchain with cloud for handling healthcare data: A privacy-friendly platform,” *Mater. Today Proc*., vol. 62, pp. 5021–5026, 2022, doi: 10.1016/j.matpr.2022.04.910.

[2] F. Akhavan and E. Hassannayebi, “A hybrid machine learning with process analytics for predicting customer experience in online insurance services industry,” *Decis. Anal. J.*, vol. 11, p. 100452, 2024, doi: 10.1016/j.dajour.2024.100452.

[3] Y.-C. Tseng, C.-W. Kuo, W.-C. Peng, and C.-C. Hung, “al-BERT: a semi-supervised denoising technique for disease prediction,” BMC Med. Inform. *Decis. Mak*., vol. 24, no. 1, p. 127, 2024, doi: 10.1186/s12911-024-02528-w.

[4] Y.-A. Daraghmi, E. Y. Daraghmi, R. Daraghma, H. Fouchal, and M. Ayaida, “Edge–Fog–Cloud Computing Hierarchy for Improving Performance and Security of NB-IoT-Based Health Monitoring Systems,” *Sensors*, vol. 22, no. 22, p. 8646, 2022, doi: 10.3390/s22228646.

[5] M. Saleem, S. Abbas, T. M. Ghazal, M. Adnan Khan, N. Sahawneh, and M. Ahmad, "Smart cities: Fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques," *Egypt. Inform. J.*, vol. 23, no. 3, pp. 417–426, 2022, doi: 10.1016/j.eij.2022.03.003.

[6] H. Raj, M. Kumar, P. Kumar, A. Singh, and O. P. Verma, "Issues and Challenges Related to Privacy and Security in Healthcare Using IoT, Fog, and Cloud Computing," *in Advanced Healthcare Systems*, pp. 21–32, 2022, doi: 10.1002/9781119769293.ch2.

[7] R. Ratra, P. Gulia, N. S. Gill, P. K. Shukla, M. M. Hassan, and F. Althobaiti, "A technique for improving healthcare privacy by applying principal component analysis," *Peer-to-Peer Netw. Appl.*, vol. 18, no. 3, p. 151, 2025, doi: 10.1007/s12083-025-01948-3.

[8] C. L. Stergiou, A. P. Plageras, V. A. Memos, M. P. Koidou, and K. E. Psannis, "Secure Monitoring System for IoT Healthcare Data in the Cloud," *Appl. Sci.*, vol. 14, no. 1, p. 120, 2024, doi: 10.3390/app14010120.

[9] S. A. Moqurrab, T. Naeem, M. Shoaib Malik, A. A. Fayyaz, A. Jamal, and G. Srivastava, "UtilityAware: A framework for data privacy protection in e-health," *Inf. Sci.*, vol. 643, p. 119247, 2023, doi: 10.1016/j.ins.2023.119247.

[10] A. S. C. Bose, M. Eliazer, B. U. Maheswari, A. Sivaneshkumar, M. Sumithra, and S. Qamar, "Securing healthcare data in mobile edge computing: a hybrid deep learning framework for privacy and anomaly detection," *Knowl. Inf. Syst.*, vol. 67, no. 12, pp. 12079–12117, 2025, doi: 10.1007/s10115-025-02586-0.

[11] J. E. Rivadeneira, G. A. Borges, A. Rodrigues, F. Boavida, and J. Sá Silva, "A unified privacy preserving model with AI at the edge for Human-in-the-Loop Cyber-Physical Systems," *Internet Things*, vol. 25, p. 101034, 2024, doi: 10.1016/j.iot.2023.101034.

[12] A. B. Popescu et al., "Obfuscation Algorithm for Privacy-Preserving Deep Learning-Based Medical Image Analysis," *Appl. Sci.*, vol. 12, no. 8, p. 3997, 2022, doi: 10.3390/app12083997.

[13] L. Wang et al., "A Blockchain-Based Privacy-Preserving Healthcare Data Sharing Scheme for Incremental Updates," *Symmetry,* vol. 16, no. 1, p. 89, 2024, doi: 10.3390/sym16010089.

[14] A. EL Azzaoui, H. Chen, S. H. Kim, Y. Pan, and J. H. Park, "Blockchain-Based Distributed Information Hiding Framework for Data Privacy Preserving in Medical Supply Chain Systems," *Sensors*, vol. 22, no. 4, p. 1371, 2022, doi: 10.3390/s22041371.

[15] L. Belcastro, J. Carretero, and D. Talia, "Edge-Cloud Solutions for Big Data Analysis and Distributed Machine Learning," *Future Gener. Comput. Syst.*, vol. 159, pp. 323–326, 2024, doi: 10.1016/j.future.2024.05.023.

[16] A. Ali et al., "Blockchain-Powered Healthcare Systems: Enhancing Scalability and Security with Hybrid Deep Learning," *Sensors*, vol. 23, no. 18, p. 7740, 2023, doi: 10.3390/s23187740.

[17] S. Alahmari et al., "A decentralized and privacy-preserving framework for electronic health records using blockchain," *Alex. Eng. J.*, vol. 126, pp. 196–203, 2025, doi: 10.1016/j.aej.2025.04.069.

[18] X. Liu, S. Li, Q. Zhu, S. Xu, and Q. Jin, "Interpretable Semi-federated Learning for Multimodal Cardiac Imaging and Risk Stratification: A Privacy-Preserving Framework," *J. Digit. Imaging*, 2025, doi: 10.1007/s10278-025-01643-y.

[19] L. Albshaier, S. Almarri, and A. Albuali, "Federated Learning for Cloud and Edge Security: A Systematic Review of Challenges and AI Opportunities," *Electronics*, vol. 14, no. 5, p. 1019, 2025, doi: 10.3390/electronics14051019.

[20] A. M. Almasoud and M. A. Alzahrani, "A Blockchain-Based Scalability Solution with Microgrids Peer-to-Peer Trade," *Energies,* vol. 17, no. 4, p. 915, 2024, doi: 10.3390/en17040915.

[21] M. M. Islam, M. M. Rahman, S. Hossain, and M. S. Kaiser, "Federated Learning in Smart Healthcare: A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration," *Healthcare*, vol. 12, no. 24, p. 2587, 2024, doi: 10.3390/healthcare12242587.

[22] X. Wang, Y. Li, Z. Zhang, and H. Chen, "Machine Learning of Element Geochemical Anomalies for Adverse Geology Identification in Tunnels," *J. Earth Sci.*, 2024, doi: 10.1007/s12583-024-0090-4.