Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FedXGB-OptDP: A Privacy-Optimised Federated XGBoost Framework with Differential Privacy for IID and Non-IID healthcare data

B. SASIREKHA
Vellore Institute of Technology
C. GUNAVATHI
gunavathi.cm@vit.ac.in

Vellore Institute of Technology

Article

Keywords: Machine Learning, Federated Learning, Privacy, XGBoost, Optimisation techniques, Independent and Identically Distributed data (IID) and Non-IID data, Depth Adaptive Differential privacy, Noise Aware Regularisation, Genetic Algorithm, Tree Structured Parzen Estimator

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8425166/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

FedXGB-OptDP: A Privacy-Optimised Federated XGBoost Framework with Differential Privacy for IID and Non-IID healthcare data

B. SASIREKHA [1] and C. GUNAVATHI[2]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.
[2]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.
Corresponding author: C. Gunavathi (gunavathi.cm@vit.ac.in)
This work was supported in part by the Vellore Institute of Technology, Vellore, India.

Abstract

The rapid growth of sensitive healthcare data results in a significant need for machine learning systems capable of providing accurate predictions while safeguarding patient privacy. Due to rapid growth, current privacy-preserving federated tree models face significant computational expenses, inadequate noise allocation methodologies, and losses in accuracy while maintaining a trade-off between privacy and utility in both IID and non-IID scenarios. To overcome the challenges, a privacy-focused extension of the Federated XGBoost architecture, FedXGB-OptDP, has been developed. It integrates Hybrid optimisation techniques along with the regularisation. A Depth-Adaptive Differential privacy (DAD), Noise–Aware Regularisation (NAR), and a hybrid optimisation technique such as Genetic Algorithm (GA) and Bayesian TPE search. The DAD-NAR is essential for adaptively regulating the allocation of privacy budgets across tree depths, using calibrated Laplace Noise, and implementing noise-aware node dropout-ensures that model stability throughout training while safeguarding privacy. Each client executes GA-driven federated feature selection when combined with TPE–based hyperparameter optimisation, facilitating efficient learning while maintaining data privacy. Global aggregation is achieved through consensus-driven feature voting and weighted averaging of hyperparameters, eliminating the necessity for complex cryptographic techniques such as Homomorphic Encryption (HE) or Secure Multi-Party Computation (SMPC). Experiments performed on five datasets across both IID and Non-IID configurations demonstrate that our model consistently achieves high levels (up to 95-96%) while ensuring robust privacy safeguards. It exceeds the performance of centralised XGBoost and prominent federated baselines, including PrivaTree, FedXHDP, and FedBoost. Overall, the results show that adaptive differential privacy, when integrated with optimisation, substantially enhances the trade-off between privacy and utility, as well as the reliability and scalability of federated decision-tree models. Hence, it provides a practicable, efficient, and highly precise solution for privacy-preserving collaborative learning within real-world environments.

Keywords: *Machine Learning, Federated Learning, Privacy, XGBoost, Optimisation techniques, Independent and Identically Distributed data (IID) and Non-IID data, Depth Adaptive Differential privacy, Noise Aware Regularisation, Genetic Algorithm, Tree Structured Parzen Estimator.*

Introduction:

Nowadays, enormous amounts of data are collected and stored in databases all around the world, reaching terabytes and growing quickly every year. The healthcare industry is one of the world's largest and most-intensive sectors, with hospitals and, clinics, and biomedical research facilities all producing massive volumes of health-related data on daily basis. Data privacy and security are essential in healthcare as they highlight the confidence between a patient and their healthcare provider. When a patient feels trust in their healthcare provider, they are more likely to continue treatment and follow medical treatment. High-quality

healthcare services depend heavily on accurate, timely, and error-free diagnosis and treatment planning. To accomplish this, advanced technology capable of processing huge amounts of medical data are required. They introduced data mining concept, in which it is a techniques that is more important for translating raw healthcare data into useful knowledge that can help healthcare professionals in making better decisions and improving patient outcomes [1]. The following part will discuss data mining in healthcare, machine learning in healthcare, and privacy in healthcare.

1.1 Data Mining in Healthcare:

Data mining is the process of discovering meaningful patterns, trends, and insights from large volumes of data These techniques are used in business area like marketing, risk management, fraud detection, cybersecurity, medical diagnosis and mathematics [2]. It enables the complete retrieval of information from previously unrecognised correlations within patient data and supports the identification of key diagnostic patterns. The review identified Neural Networks (NN) as the primary research area, closely combined with techniques such as Support Vector Machines (SVM) and Decision Trees (DT), highlighting the importance of Machine Learning(ML) in medical data mining. Moreover, in biomedical research [3][4], it is used to identify the relationship between patient attributes and disease prediction.

Many researchers have utilised data mining to mine complex healthcare relationships, which can produce disease diagnosis and predictive care in past decades, but have also faced some problems. Increasing amount of big data in healthcare now a days, leads to enhanced utilization of sophisticated datamining techniques to improve service delivery and optimize decision making in healthcare [5]. In this techniques, predictive analytics, an essential application of data-driven techniques, analysis past patient data to predict diseases progression, hospital readmission or therapy efficiency. Thereby enhancing patient care and resource distribution. Additionally, it highlight the use of Electronic Health Records (EHR's) to predict clinical situations by categorizing a predictive analytics as a substantial advancement of datamining in healthcare [6].

1.2 Machine Learning in Healthcare:

Machine Learning is a subfield of Artificial Intelligent that allows computers learn from the mistakes without being explicitly programmed. It uses algorithm to analyse the data, identify the patterns and make predictions on new data. Methodologies enables computer to identify complex and non-linear correlations within large medical datasets for diagnosis, prognosis, and treatment enhancement. Thereby it reduces the human error. Prediction based analysis used to determine suitable sample sizes, mitigate errors in data and enhance the reliability of clinical research [7] [8]. Rule-based system historically constituted the basis for artificial intelligence in medicine around the mid-20th century. These algorithms depended on human-defined logical rules to assess electrocardiogram (ECGs), identify conditions and propose solutions [9].

However, they faced obstacles due to increased maintenance expenses, stringent regulations, and limited scalability. The ML algorithms overcome these constraints by autonomously extracting knowledge from data and revealing hidden patterns without the necessity traditional development [10]. They are utilised in disease classification, medical imaging, risk assessment, and treatment recommendation [11]. ML combined clinical expertise with computational intelligence, delivering personalised, dynamic and data-driven healthcare solutions. Moreover it is better than traditional diagnostic scenarios, covers advanced fields such as bioinformatics, IoT-integrated healthcare, and reinforcement learning for clinical decision-making. Moreover, machine-learning algorithms have significantly improved healthcare efficiency and cost effectiveness by frequently analysing patient data, including genetic information [9], medical history, and life style attributes, to provide personalised healthcare goal. In terms of optimising medical dosage and personalising lifestyle recommendations for ongoing medical conditions such as hypertension and diabetes. Hence, improving patient's quality of life [13].Even medical imaging has also benefited greatly from ML's revolutionary impact. To improve diagnosis accuracy and streamline image processing. AIML have become essential in handling prostate cancer [14]. This technology advancement is supported by extensive datasets such as the National Health and Nutrition Examination Survey (NHANES), which provides exhaustive health and socio-demographic data for healthcare [15].

1.3 Privacy in Healthcare:

Sharing healthcare data is important of medical study because it can move forward and for the creation of more advanced diagnostic systems, Industries can only share their data with third party cloud or data tools if they are sure that privacy will be protected. Because it is very important to protect right things, many healthcare providers are afraid to share patient information, even though it is helpful for research. To overcome this problem, Privacy Preserving Data Publication (PPDP) was introduced to protect people privacy while keeping data for useful analysis [16]. Anonymisation and encryption are at heart of the most traditional PPDP methods. Authentication, password management and encryption are all cryptographic techniques, that provide strong security, but they are very hard to use with large amounts of healthcare data [17]. So it not good for ongoing medical analysis, which needs be fast and able to handle large amount of data. Differential Privacy (DP) and Homomorphic Encryption (HE) have also have been studied recently to keep data safe. However, it may create a problem with keeping the quality of data good. To protect privacy DP uses random noise, which can make calculations less accurate if not set up correctly. On the other hand HE, it do calculations on protected data but takes a long time to process [18].

The arrival of healthcare 5.0 has pushed the area of Healthcare Informatics (HI) towards intelligent data-driven decision-making. This combines various resources, technologies, and methodologies to improve medical data gathering and retrieval. However, handling such data while maintaining patient anonymity is problematic. Healthcare data must be anonymised before distribution, with access limited to authorised institutions, due to privacy rules and patient permission requirements. This makes mistakes to create models that offer both high predicted accuracy and privacy protection.

IBM Cost of Data Breach Report says that the prices of healthcare data breaches are currently increased by 53.3% since 2020. The industry continues to have the highest cost of data breaches of any industry, with losses of $10.93 million in 2023. Making sure healthcare organisations follow rules like HIPAA and GDPR will lower the chance of data breaches, cut down the financial losses, protect Patient privacy, and build trust with their patients. Hence, healthcare organisations have to follow best practices, such as running training and education program and education programs to teach patients and healthcare workers about their rights. It includes ensuring that people give permission and know what information is being gathered and how it will be used. Data minimisation is nothing but only collecting the data that is needed and keeping it for as long as it is needed. Internal and external audits should be done on a regular basis to identify and eliminate possible risks.

In this context, standard centralised ML systems have considerable disadvantages. Collecting all patient data on a single server introduces single points of failure and data loss due to network failure. Centralised systems are also vulnerable to cyberattacks, which may damage sensitive patient data [19]**.** It frequently yields less accurate models since it is based on a single data source, restricting generalisation across different patient populations. In order to overcome these drawbacks, a Federated Learning (FL) has developed in 2016 , has a promising approach that can allow multiple client/ Organisation to collaboratively train ML without sharing the client's original information in order to protect from unknown or triparty theft. Each client trains models locally and only model updates or model weight is transformed to server, and server side it aggregate the updates to develop the global model. This process is called decentralisation which can improve data privacy, minimise the risk related to data transmission, and facilitates model personalisation while maintaining the robust predictive performance [20]. Even though it protects privacy, it suffers from privacy breaches such as shared gradients during aggregation. To avoid this an Adaptive DP and secure optimisation techniques is used to maintain optimal equilibrium between privacy and accuracy.
The main problem in centralised ML is protecting sensitive information while maintaining model accuracy is most important. That is overcome by FL, distributed model training without data exchange. Traditional period an privacy is maintained using methodologies like Anonymisation, Homomorphic Encryption, Secure Multi-Party Computation (SMPC) but it takes computational and communication cost. Hence, Differential Privacy (DP) has established a lightweight non cryptographic technique and efficient alternative by employing an optimised privacy-aware federated architecture that enhances both privacy preservation and model efficiency in healthcare applications.

II RELATED WORK:

Privacy-Preserving Machine Learning (PPML) has emerged as an essential area of research due to the growing dependence of ML models on sensitive healthcare data, including patient medical information. Numerous studies have investigated privacy strategies for ML models, concentrating mostly on inference rather than training due to computational expense and low scalability of cryptographic primitives with huge datasets [21] [22]**,**[23]**.** Early

research used safe two-party computation (2PC), additive secrete sharing, and Homomorphic Encryption (HE) to facilitate privacy-preserving inference in decision trees and random forests [24], [25] [26]. Although these approaches-maintained data security, their significant computational and transmission overheads restricted their feasibility for real-time or large-scale biological applications. To address the issue, academics are exploring federated learning (FL) as a substitute for centralised in safe distributed training [27], [28][29], [30]

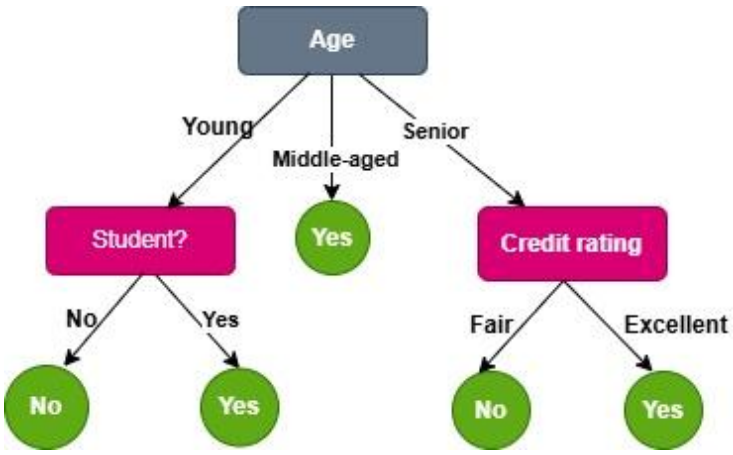


Figure 1: Decision Tree

FL allows numerous data properties to jointly train a common model without the transfer of raw data, presenting an appealing balance between data utility and privacy. [28] Established the concept of federated averaging, where clients locally train models and transmit only parameter updates to a central aggregator. This paradigm significantly diminishes data exposure risk, rendering it appropriate for privacy-sensitive sectors such as healthcare [31]. Subsequent research expanded FL to tree-based models due to their interpretability and efficiency in structured healthcare data [28] ,[27] & [32]. However, existing privacy-preserving federated decision tree models faces challenges in effectively balancing accuracy, computational efficiency, and privacy protection. Integrated DP with HE for multiparty decision trees [33]; however, their approach was restricted to categorical data and experienced accuracy deterioration due to the induction of noise. Similarly, [34] author utilised Shamir secrete-sharing, which guaranteed secure distributed training but resulted in significant communication cost and scalability issues.

In Traditional privacy-preserving methods such as K-anonymity, HE, and Secure Multi-Party Computation (SMPC) [27],[35]&[36] gives theoretical assurance but show inefficiency when implemented on extensive healthcare datasets. DP, initially introduced in [37], has developed into a fundamental privacy-preserving framework and has been implemented in numerous ML models, such as SVM (Support Vector Machine) and NN (Neural Network) [38],[39] & [40]. Even though, DP protects privacy of individual by introducing noise, in terms of trade-off between privacy vs utility, it reduces the accuracy because noise is added. In [41], proposed a ID3 algorithm in Decision Tree (DT) that gives

formal guarantees but incorporates significant noise into split criteria, resulting in reduced model quality. DP is also applied in Gradient-Boosted Decision Trees (GBDT) in distributed environments; it produces accuracy loss due to the training of each tree using only local data [42].

A federated ensemble tree model was proposed to address these difficulties. Presented in [43], Federated GBDT (Gradient Boost Decision Tree) with gradient-based model selection (F-GBDT-G), enhancing prediction performance but lacks adaptive noise allocation. Conversely, [44]utilised a sequential update strategy that established a uniform privacy threshold across all trees, leading to inconsistent privacy-utility trade-offs. Further, FL methodology was extended to blockchain for transparent contribution monitoring, its computing cost remain high[45]. Likewise, SecureGBM [44][33] and Secure DT [46] integrated SMPC with DP to improve security; nevertheless, these approaches remained best by significant inefficiencies at scale and demonstrated inadequate adaptability for non-IID healthcare data.

Recent research has investigated federated random forest and XGBoost models for distributed healthcare knowledge [47], [48],[49]. However, the majority of these methodologies choose secure inference over training or suffer significant processing overhead due to cryptographic processes. DPRF models shown by [48],[49] strive to balance privacy and accuracy through the introduction of calibrated noise into the tree split; however, additional noise frequently diminishes utility. Likewise, smooth-sensitivity algorithms [50] adjust noise according to dataset sensitivity but do not completely resolve the privacy-utility trade-off. Revocable Federated Random Forest (RevFRF) models [51]provide multi-party training with adjustable privacy management; however, they exhibit significant complexity and instability in noisy environments. The summary of this related work is explained in Table 1.

To enhance privacy and scalability, numerous researchers have suggested hybrid frameworks that integrate FL with optimisation or blockchain technology. In [52], HE for GBDTs, but the datasets to categorical features. [53] Addressed cryptographic limitations by integrating FL with partial DP to improve computing performance. Still, both methodologies encountered challenges with model interpretability and modification flexibility in medical settings. Research conducted by [54] and [55] further diminished communication overheads by additive secret sharing, however remained primarily focused on inference.

Even though considerable advancement has been achieved in privacy-preserving federated decision tree learning, current methods focus on inference attacks, either suffer significant computational expenses, or lack the ability to dynamically adjust privacy across model layers. Moreover, non-IID healthcare data distributions and varied client resources increases model imbalance and performance decrease.

The proposed Federated XGBoost with Adaptive Differential Privacy and Optimisation (FedXGB-OptDP) presents an innovative hybrid framework that combines Depth-Adaptive Differential Privacy (DAD) and Noise-Aware Regularization (NAR) with Genetic Algorithm (GA) and Tree-structured Parzen Estimator (TPE) optimization to tackle these challenges. In previous research [54][33] & [42] that utilise fixed privacy budgets, FedXGB-OptDP allocates the privacy budget (ε) dynamically across tree depths-imposing less noise on shallow nodes are more noise on deeper nodes-thereby improving split accuracy while maintaining privacy. Additionally, GA and TPE collaboratively optimise features and hyperparameters at the client

level to alleviate non-IID impacts and enhance convergence efficiency. The combination of federated aggregation with adaptive DP guarantees equitable privacy-utility trade-offs, scalability across diverse healthcare datasets, and protection against inference attacks, thus addressing the drawbacks of previous privacy-preserving ML and FL-based decision tree frameworks [56], [46]& [37].

III BACKGROUND

The background section highlights the basic concept necessary for understand our proposed FedXGB-OptDP architecture. It discusses traditional DT algorithms, current privacy-preserving machine learning techniques, the concept of FL, and the importance of DP in distributed environments. It also includes optimisation algorithms that enhance accuracy and stability under privacy limitations, providing the theoretical framework for the proposed model.

3.1 Traditional Decision Tree Algorithm:

The main goal of traditional ML is to find trends in data that can be used to make predictions about things like classification and regression. DT are very popular ML models because they are easy to understand, don't cost much to run, and can handle different types of data [57]. A decision tree has root node, several internal decision nodes, and leaf nodes. Each internal node tests a characteristic and divides the data based on the result until the leaf node assigns a class label [58] shown in figure 1.

There are a number of well-known methods for building DTs. Quinlan created the ID3 (Iterative Dichotomizer 3) algorithm, which chooses attribute based on Information Gain. However, it only works with categorical attributes and might be too good for small datasets [59]. To fix these problems, an add-on called C4.5 was created that works with both continuous and categorical attributes, handles missing values, and uses pruning methods to make generalisation better [60].

Table 1: Summary of the related research work

Ref. No:	Method	Limitations	Privacy Techniques	Attacks focused
[21]	Secure computation framework	High computation cost, not scalable	SMPC	Protect against data leakage during collaborative learning
[22]	HE to enable the encrypted evaluation of DT and classification models.	Computationally expensive, large ciphertexts increase network overhead	HE	Prevents information leakage in inference attacks
[23]	Additive secret sharing protocol	Requires continuous communication and synchronisation between clients.	Additive Secret Sharing	Attackers attempt to get knowledge from shared computations

[24] [25] [26]	HE and SS (Secret Sharing) for secure inference in DT and random forests	Works only on inference not training.	HE+ Additive Secrete Sharing	Model output is protected
[33]	DP+ HE for multi-party Decision trees	Only for categorical features; accuracy loss due to noise.	DP+HE	Membership inference, attribute inference
[34]	Shamir Secret Sharing for distributed trees.	High communication cost, poor scalability.	Shamir Secret Sharing	Collusion attacks, reconstruction attacks
[43]	Federated GBDT with gradient-based model selection	No adaptive privacy budget allocation	FL+ Gradient-based selection	Update inference attacks
[44]	Federated sequential update	Uniform ϵ distribution	FL+DP	Privacy leakage through imbalance updates
[45]	Blockchain-supported Federated GBDT	Very high computational cost	FL Blockchain	Tampering malicious updates
[44][33]	SecureGBM with DP +SMPC	Still heavy due to SMPC; accuracy decreases	SMPC+DP	Gradient leakage, collusion attacks
[46]	Secure DT using SMPC	Large communication overhead; not scalable	SMPC	Split inference attacks
[47] [48][49]	Federated XGBoost	Mostly secure inference; training underexplored	FL+ basic DP	Model inversion, Membership inference
[48] [49]	Differentially Private Random Forest (DPRF)	Noise reduces accuracy	DP	Reconstruction, attribute inference
[50]	Smooth sensitivity DP	Fail to solve privacy-utility fully	Smooth sensitivity DP	Sensitive value inference
[51]	Revocable Federated Random Forest	Complex design; unstable under noise	DP+ revocable keys	Inference attacks
[52]	HE for federated GBDTs	Supported only for categorical features	HE	Inference and feature leakage
[53]	Partial DP in FL for gradient sharing	Limited interpretability, complex design	Partial DP	Gradient leakage
[54]	Additive SS+ DP for low communication	Primarily inference, not suitable for training	SS+ DP	Inference attacks during prediction
[55]	Additive sharing for FL	Focus only on inference, limited scalability	Additive SS	Reconstruction attacks
[35] [36]	SMPC, HE K-Anonymity for PPML	Inefficient for a large dataset.	SMPC, HE K-Anonymity	Linkage attacks, reconstruction attacks.
[41]	DP-based ID3 algorithm	High noise on split criterion, - lower accuracy.	DP	Membership inference, attribute inference
[42]	DP for GBDT in distributed.	Accuracy drops, each tree trained only.	DP	Gradient-based leakages.

Breiman developed CART (Classification and Regression Trees), which employs the Gini Index for classification tasks and Mean Squared Error for regression analysis. It constructs binary trees and prunes them to simplify their structure [61]. Although traditional DT algorithms are efficient in centralised systems, they imply that all data is stored in a single location. HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulations) are major data privacy laws, in which HIPAA protects health information where GDPR covers all personal data, because in real-world healthcare environments, data is scattered not only within organisation, outside the organisation also , hence data cannot be easily shared due to strict standards [62].

3.2 Privacy-Preserving Machine Learning:

Since ML is used in all sectors, protecting data privacy while maintaining model efficacy is a more critical problem. Many techniques and models have been developed to address these challenges, including HE, SMPC, DP and FL. These techniques provide various approaches for protecting data and ensuring confidentiality throughout processing and analysis [63]. Anonymisation is the traditional method that was used initially to generalise, suppress and ensure K-anonymity of the data.

Initially, privacy-preservation techniques focused mainly on anonymisation methods such as generalisation, suppression, and k-anonymity. A lot of research shows that even anonymised data can be exploited to identify patient original information, especially if the attackers have other relevant information [64]. As a result, ways of anonymisation alone are not thought to be good enough for modern healthcare systems. To make privacy more secure, encryption methods like HE and SMPC were created. These methods enable users to perform encrypted or secretly shared data without revealing confidential information [65] [66]**.** These methods theoretically protect privacy very well, but they require a lot of processing power, which means they cannot be used for large-scale or real-time healthcare applications.

DP (Differential Privacy) has evolved as a more computationally efficient approach that offers quantifiable privacy assurance by including carefully controlled noise into learning processes [67]**.** It guarantees that the impact of any individual contribution does not substantially affect the model's output, therefore preventing attackers from obtaining confidential patient information. Even though DP effectively gives a utility-privacy trade-off, where enhancing privacy generally reduces model accuracy. However, the majority of PPML methodologies assume centralised data storage, which is sometimes difficult in healthcare, as data is distributed across various institutions due to legal constraints like HIPAA and GDPR. This limitation has caused interest in FL, which allows multiple data owners to collaboratively train models without revealing their original data [68]. When combined with DP and FL provides both data governance and computational privacy guarantees, making it highly suitable for real-world healthcare environments.

3.3 Federated Learning

Federated Learning (FL) is a collaborative and privacy-preserving ML approach where a group of clients U = {1, N} collectively train an integrated global model under the control of a central server, without transmitting their raw data. Each client trains s model locally with its private dataset and only exchanges model parameters or updates, thus safeguarding data confidentiality throughout the training process shown in Figure 2.
The FL training generally comprises three consecutive stages:

1. **Parameter Initialization:** Initially, the central server initialises the global model parameters w_0, including additional hyperparameters such as total number of communication rounds T and number of participating clients per round K. The server subsequently disseminates the initialised model w_0 to the chosen clients to initiate the collaborative training process.

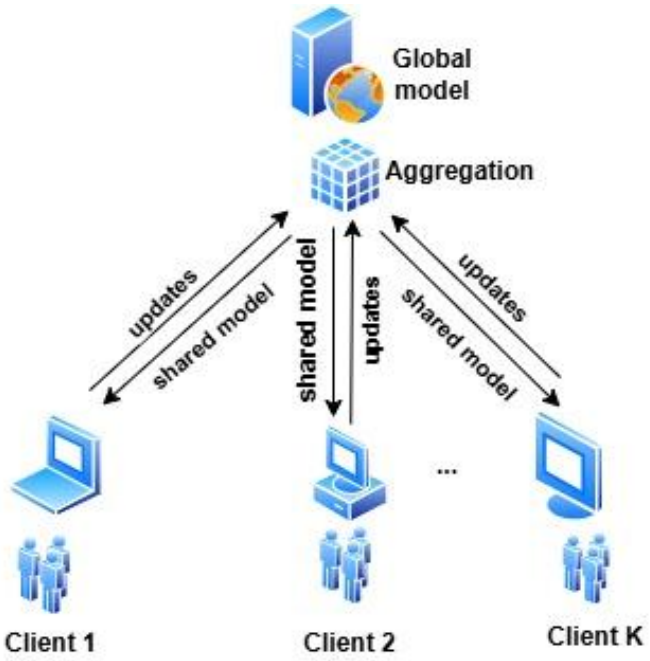


Figure 2: Federated learning model

2. **Local Training:** Upon obtaining the global model w_{t-1} each selected client K conducts training on its local dataset D_K. Utilising stochastic gradient descent (SGD) or similar optimisation algorithms. The procedure for local updates is given by:

$$w_t^{(k)} = w_{t-1} - \eta \nabla_w \left(\frac{1}{|B|} \sum_{i \in B} L(F(x_i, w_{t-1}), y_i) \right) \quad (1)$$

Where η is the learning rate, B is a mini-batch sampled from D_K, and L (.) is the loss function. After completing local training, each client sends only the updated model parameters $w_t^{(k)}$ back to the server.

3. **Global aggregation:** The server combines the information obtained from local models to update the global model. The main aggregation method used is Federated Averaging (FedAvg), as mentioned below:

$$w_t = \sum_{k=1}^{K} \frac{|D_k|}{\sum_{k=1}^{K} |D_k|} w_t^{(k)} \quad (2)$$

The modified effect of each client is weighted based on the size of its dataset. The revised global model w_t is subsequently redistributed among the clients for the next training round. This iterative procedure continues until convergence is achieved.

3.4 Differential Privacy in Federated learning:

DP has become a prevalent privacy-preserving approach in FL, offering robust theoretical assurances against the exposure of sensitive information via shared model updates. In the traditional DP method, privacy safeguards guarantee that the inclusion or exclusion of singular record has minimal impacts on the models output. A mechanism M satisfies ϵ-Differential Privacy if, for any pair of adjacent datasets D1 and D2 for any output S within the range of M, the probability ratio is satisfied:

$$\Pr [M (D_1) \in S] \leq e^{\epsilon} \Pr [M(D_2) \in S] \quad (3)$$

Where ϵ denotes the privacy budget that regulates the balance between privacy and utility [37]. The privacy loss is dependent upon the sensitivity of the function f to change in the dataset, defined as $\Delta S = max_{D_1, D_2} \| f (D_1) - f (D_2) \|$. To achieve DP during model training, methodologies like the Exponential Mechanism introduce calculated noise in propagation to sensitivity, expressed as:

$$M (D) \propto \exp \left(\frac{q(D,r)}{2\Delta S} \right) \quad (4)$$

Where q (D, r) is a quality score that quantifies the extent to which output r aligns with the data [37][69]. In FL, DP is implemented by introducing noise to local model gradients or parameter updates before aggregation. This safeguards the sensitive information from being reconstructed by the server while facilitating collaborative model training [70]. However, the system accuracy may be affected by noise injection; thus, it is critical to find a balance between privacy and efficiency. Personalised noise scale optimisation frameworks and DP optimisation frameworks have been developed to collaboratively minimise global loss and privacy leakage, thereby improving the overall performance of model comparison to the fixed-noise approach [71]. Furthermore, clustered and adaptive DP-FL strategies offers significant benefit in non-IID data environments that are found in healthcare and IoT applications, as they modify noise level according to data heterogeneity and client-specific features [72][73].

3.5 Optimization in Privacy-Aware Federated Learning:

Optimisation is essential for enhancing the efficiency and accuracy of privacy-aware FL, especially in the concept of constraints like heterogeneous data, varying device capabilities, and performance decoration results from DP noise. Meta-heuristic optimisation techniques have shown improvement in feature selection and model stability in privacy-restricted environments. Hybrid methodologies, such as SOA-TEO, significantly diminish redundant features while improving classification accuracy [74]. Similarly, hybrid algorithms such as GWO-CSA (Grey Wolf Optimisation – Crow Search Algorithm) and SHO-SA (Spotted Hyena Optimiser – Simulated Annealing) improve the search efficiency for discovering appropriate feature subsets in high-dimensional biological datasets [75] [76]. In FL, improving customer involvement is essential, as the quality of each client's inputs affects global model convergence. Some methods, like FedBoost, use Bayesian estimates to give more weight to clients who are likely to provide important updates [77]. Another method, FedCSGA, uses genetic algorithms to deal with system and data variations in client selection [78].

Additionally, models like IGA-XGBoost concurrently optimise hyperparameters and feature selection to improve prediction efficacy [79]**.** This study shows that integrating adaptive optimisation into FL can maintain privacy while improving model utility, thereby providing the justification for our proposed FedXGB-OptDP framework, which combines DP with optimised gradient boosting and adaptive aggregation to achieve an improved equilibrium between privacy and performance.

Key Contribution:

FedXGB-OptDP, a privacy-aware and optimisation-enhanced federated XGBoost framework specifically designed for healthcare data analysis, is introduced in this work. The following is an overview of key contributions:

1. **Federated XGBoost with Differential Privacy:** To enhance the standard XGBoost model for FL by integrating Hierarchical Differential Privacy (HDP) into the tree-building process. The global decision structure is preserved while fine-grained patient-specific patterns are protected by adaptively inserting privacy noise at the root, intermediate, and leaf nodes.
2. **DAD-NAR regularisation for stability under noise:**
   A combined Depth-Adaptive Dropout (DAD) and Noise-Aware Regularisation (NAR) mechanism is introduced to address training instability induced by DP noise. DAD prevents deeper divides that are vulnerable to overfitting, while NAR modifies the regularisation strength in accordance with the noise magnitude, thereby guaranteeing enhanced generalisation and model robustness.
3. **Optimisation-based client-side learning:**
   Each federated client performs Genetic Algorithm (GA)-based feature selection to train only the most informative attribute, and Tree-Structured Parzen Estimator (TPE) hyperparameter tuning to optimise XGBoost parameters locally, thereby improving effectiveness without sharing raw data.
4. **Privacy-Preserving Model Aggregation:**

We aggregate only model parameters, selected feature indices, and optimised hyperparameters, rather than relying on computationally high-cost cryptographic methods (HE, SMPC).Hence, this approach guarantees that both scalable and lightweight, making it appropriate for real-world healthcare environments.

5. **Consensus-Driven Global Model Development: T**he server combines important features, average modified hyperparameters, and re-train a global XGBoost model, ensuring a consistent model architecture among clients while preserving the advantages of personalized insights from distributed training data.
6. **Privacy-Utility Trade-Offs:**
   A complete analysis of the datasets indicates that FedXGB-OptDP frequently achieves superior accuracy and F1-scores within strong privacy constraints when compared to baseline federated boosting and DP-based decision tree models. Finally, our model is compared with IID and Non-IID (Independent and Identically Distributed & Non-Independent Identically Distributed) data.

IV PROCEDURE:

FedXGB-OptDP, a privacy-preserving federated learning framework using optimised XGBoost and adaptable DP. To improve learning efficiency without sharing raw data the model uses GA-based feature selection and TPE hyperparameter adjustment at each client. A DAD-NAR regularization system stabilises training under DP noise using depth-adaptive dropout and noise-aware filtering. Then the trained XGBoost model is moved to server. The server does global model reconstruction and consensus feature aggregation without requiring cryptographic techniques. Compared to the previous base model, our model shows better accuracy and privacy utility trade-offs. Our model flow is shown in figure 3.

A. Data Pre-processing:

The initial phase of the FedXGB-OptDP framework development is data pre-processing and normalisation, which guarantees that the data has been cleaned and transformed into a usable format. The EEG (Electroencephalogram) Eye state dataset, the Diabetic dataset, and the Obesity dataset are the datasets that we are utilising in this research. These datasets contain both categorical and numerical attributes and features. The dataset is initially imported and examined to identify numerical and categorical attributes. In order to convert nominal values into discrete integer representations, categorical features are transformed using Label Encoding. Strategic partitioning is employed to ensure class balance by dividing the dataset into two subsets: Train-Validation (85%) and Global Test (15%). The feature distributions are normalised by fitting a Standard Scaler to the training features and subsequently applying it to both the training and test sets. Allow the pre-processed dataset to be represented as:
$X \in R\ n \times d$, $y \in \{1, 2 \ldots C\}$ X is a point in the n× d ring, and y is a point in the set {1, 2, C}. Where n represents the number of samples, d represents the number of features, and C represents the number of classes.

B. Federated Client simulation:

In client, the training set is divided into three non-overlapping client datasets via Stratified Shuffle Split, hence maintaining the class distribution within each client in alignment with the overall dataset. Each client $c \in \{1, 2, 3\}$ is allocated a unique local dataset denoted as $(X^{(c)}, y^{(c)})$ ensuring balanced label distributions among clients. No raw data is transmitted or centralised in this process; each client maintains its local dataset and conducts training independently. This configuration simulates a federated learning environment, ensuring privacy is maintained as no data exists in the client domain and no central aggregation of raw samples takes place was shown in architecture diagram figure 4.

C. Differential Privacy Configuration:

The proposed model FedXGB-OptDP includes a DAD-NAR++ privacy control module that manages the privacy-utility balance during client-side model training. The configuration has four important parameters.

Table 2: Parameters and their roles

Parameter	Role
ϵ	Local differential privacy budget controlling noise intensity
α	Depth-adaptive decay factor controlling how privacy changes across tree layers
β	Node dropout coefficient influencing the likelihood of skipping deeper splits
λ_0	Regularization scaling factor applied to stabilize training under noise

i. Laplace noise injection:

During the federated training process, each client calculates local statistical metrics, such as gradients, split gains, or class distributions, to inform the design of decision nodes. Nonetheless, these numbers may unknowingly disclose sensitive information regarding the underlying data. DAD-NAR++ prevents leakage by applying Laplace noise to the data points before sending them to the server. Specifically, if g denotes the locally computed parameter, the transmitted value is expressed as $g\sim = g + (1/\epsilon)$. Where ϵ represents the client's local differential privacy budget. The noise protects small variations in individual data points, hence minimising the probability that an outsider might deduce personal information. The additional noise is carefully controlled to maintain privacy without significantly affecting the model's learning capacity. The important parameters used for our model are shown in Table 2.

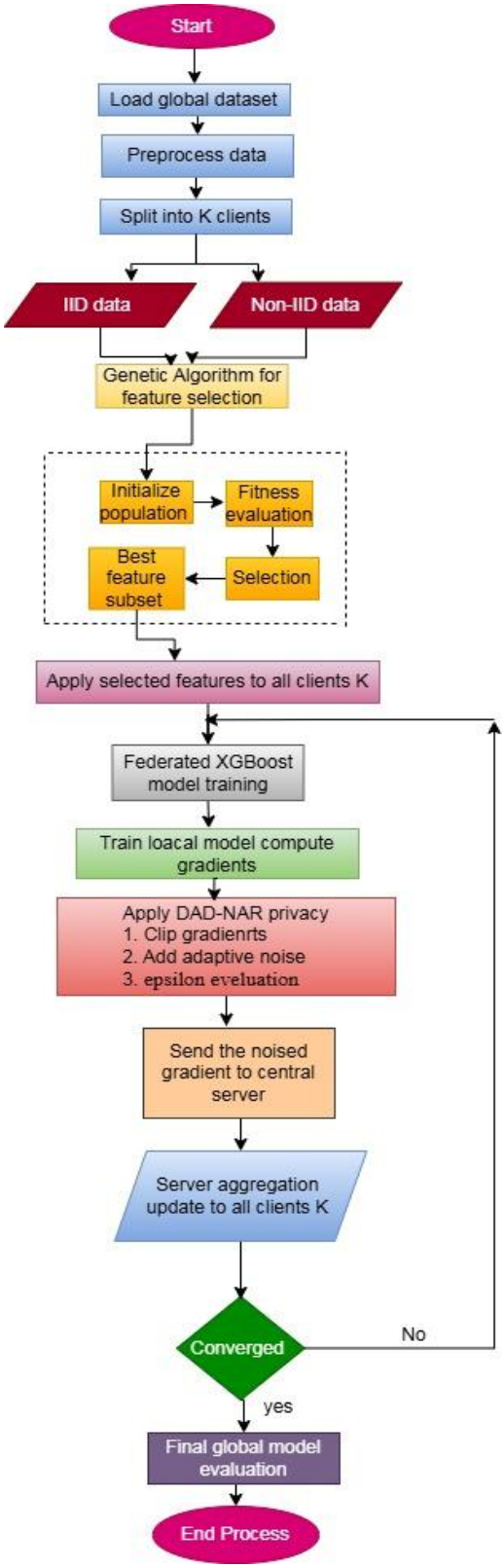


Figure 3: Flow diagram of FedXGB-OptDP

ii. Depth-Adaptive privacy control:

The privacy requirements vary at different levels for the decision tree. Deeper nodes reflects smaller, more particular subsets within the population and are hence more sensitive, whereas nodes nearest to the root reflect broad data features and present comparatively low privacy risk. DAD-NAR provides a depth-adaptive privacy technique to resolve this issue. The effective privacy budget decreases as the depth of the tree increases, as expressed by a mathematical equation: $\epsilon\, d = \epsilon \cdot e - \alpha d$. The privacy budget decreases as the depth d increases, which leads to a more robust noise injection. This exponentially increasing privacy protection resulting in more effective noise injection. This rapid improvement in privacy protection ensures that complex pattern, capable of disclosing private attribute, or identifying person, are safeguard with more accuracy. Therefore, the model produces a reasonable balance between providing enhanced privacy at deeper, more sensitive levels and maintaining predictive performance at higher levels of the classification system in tree.

iii. Noise-aware node dropout:

As the tree gets deeper, noise may get worse, making split choices in deeper nodes less reliable. In addition, deep splits are more likely to lead to overfitting and may show sensitive patterns that are unique to a grouping. Noise-aware node dropout is used by DAD-NAR to stop both overfitting and privacy leaks. This approach involves the removal of a node at depth d with a probability of $P_{drop}() = 1 - e^{-\beta d}$. This mechanism essentially stops the tree from growing into unnecessary portions of the data by making it more likely that deeper node will not be reached. Not only does dropping lower nodes protect privacy, but also it makes generalisation better because the model stops learning unstable patterns that come from noisy data. Hence, this creates a tree structure that is safe and strong, as well as smaller and more aware of privacy

D. FEATURE SELECTION THROUGH GENETIC ALGORITHM:

In order to mitigate feature redundancy and improve the selective capabilities of the model, each client conducts Genetic Algorithm (GA)-based feature selection locally, without disclosing intermediate evaluation metrics or feature significance scores to the server. In this representation, each prospective feature subset is encoded as a binary chromosome: $z = [z_1, z_2 \ldots, zd]$, where $z_i=1$. The value of 1 indicates that the i^{th} feature of the it is included, while $zi = 0$.

The absence is indicated by the value of 0. The GA evolves populations of these chromosomes through mutation, crossover, and selection. The validation accuracy of an XGBoost classifier trained on the specified subset of features is used to calculate the fitness of each chromosome. Fitness is synonymous with accuracy (v [:,]). Fitness = accuracy val (X[:,z]).

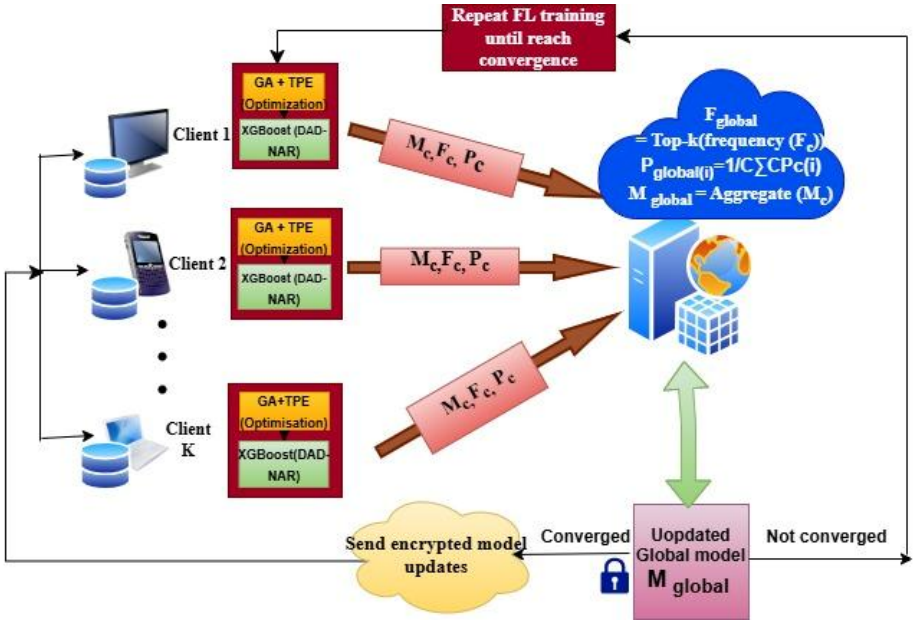


Figure 4: Architecture of FedXGB-OptDP

The GA indirectly attempts compressed subsets of features that optimise predictive performance while minimising redundancy over successive generations. Each client retains its top-k most informative features based on the highest-fitness chromosomes once convergence is achieved or a stopping condition is met. By making sure that each client only sends the most useful set of features to the federated training process, the computational complexity is lowered while efficiency is increased, although strict confidentiality and privacy are maintained.

E. HYPERPARAMETER OPTIMIZATION:

Each client adapts its own hyperparameters to make the model run better after choosing features, before global aggregation take place. The optimisation is mainly done on important XGBoost factors like learning_rate, Max_depth, and estimators. The Tree-structured Parzen Estimator (TPE) method from the Hyperopt library is used to quickly look through the hyperparameter search field. TPE creates a numerical model of the desired function and chooses parameter settings that are likely to improve validation performance. In each trial, an XGBoost model is trained with given hyperparameters and then tested on the local validation group. The implementation that works highest will be stored as the most effective local hyperparameter setting for that client. The global hyperparameter tuning makes sure each client's model is adaptively tuned to how data is distributed in its area. This improves the quality of downstream aggregation while protecting data privacy.

F. LOCAL MODEL TRAINING:

As part of the DAD-NAR privacy framework, each client starts local model training once hyperparameter optimisation and feature selection are done. Because the GA process choose a

subset of features, the model only looks at the most distinguishing ones. This cuts down on duplication and improves the decision limits. After that, the optimum hyperparameters that are found through TPE tuning are used to set up an XGBoost classifier so that it fits the local data distribution. To keep private information from getting out though model structure or gradient information and to protect sensitive local patterns, Laplace noise injection, depth-adaptive privacy scaling, and probabilistic node dropout are used during the training process. When training s over, each client creates a pair $M_{(c)}$, $F_{(c)}$ and $P_{(c)}$. Where $M_{(c)}$ is the trained privacy-preserving model, $F_{(c)}$ is the set of chosen features indices used in training, and $P_{(c)}$ is the final set optimised hyperparameters. It is very important to remember that server only gets these model errors. Any unprocessed data stays locally, which protects the privacy of all data.

G. FEDERATED AGGREGATION:

After local training is done, the server manages the federated aggregation process, ensuring that no raw data is transmitted or centralised from the client. Instead, each client transmits only its selected feature subset $F_{(c)}$ and optimised hyperparameter set $P_{(c)}$ along with obtained model parameters.

a) Feature Consensus Aggregation:

The server implements a consensus-based feature voting strategy to establish a unified global feature space. The top-K most frequently selected features are retained as the global set, and each feature index is counted across client selections.

$$F_{\text{global}} = \text{Top-K} \left(\bigcup_{c=1}^{C} F^{(c)} \right)$$

This ensures that the final global model consists purely of highly informative and mutually accepted features, hence improving generalisability and preventing client-specific bias.

b) Hyperparameter Averaging:

The server subsequently calculates the global hyperparameter configuration by aggregating the corresponding hyperparameters from client-trained models:

$$P_{\text{global}}(i) = \frac{1}{C} \sum_{c=1}^{C} P^{(c)}(i) \qquad (5)$$

The total number of clients is represented by c. Hyperparameters that must remain integer-constrained (e.g., n_estimators, Max_depth, Min_child_weight) are adjusted carefully to avoid overfitting while maintaining model accuracy. The step guarantees that strict privacy are maintained while establishing an integrated global setup that can be used for server-side model retraining by aggregating purely model information and not data itself.

H. GLOBAL MODEL RETRAINING:

After the creation of the global feature set F_{global} and aggregation of hyperparameters P_{global} after completion, the server generates the global XGBoost model. The model is trained exclusively on the standard training subset of the original dataset, restricted to the consensus-

selected feature indices, specifically, X train [: F global] with corresponding labels y_train. Subsequently, add up all hyperparameters, they show how all the clients usually learn, making it a fair setup that works with any local data distribution. A smaller learning rate and more boosting rounds are used during global retraining to improve stability and prevent overfitting. This makes the preparation phase easier and more efficient. The aggregated hyperparameters reflect the average learning behaviour of all clients, providing a balanced configuration that generalises across local data distributions. To enhance stability and mitigate overfitting during global retraining, a diminished learning rate and an augmented number of boosting rounds are utilised, facilitating a slower and more regulated optimisation process.

The resultant global model thereby includes a privacy-preserving majority among all clients, attaining enhanced performance without necessitating direct data sharing. Algorithm shows step by step procedure how federated XGBoost with optimized differential privacy using GA and Bayesian achieve efficient and secure training.

Algorithm 1 FedXGB-OptDP

Input:

$D_c = \{X_c, y_c\}$ →Local dataset for each client c

C → Number of clients

N_T → Number of trees

d → Maximum depth

ϵ,α,β → Privacy parameters

GA → Genetic Algorithm

TPE → Hyperparameter Optimiser

Output:

Global Model M_{global}

Phase 1: Data Processing (Client-Side)

For each client c=1…C:

1. Pre-process D_c → encode categorical features and normalise numeric values.
2. Apply Genetic Algorithm to select optimal feature subset:
   F_c = GA_ Select (X_c, y_c)
3. Apply TPE optimisation to tune model hyperparameters:
   P_c = TPE_Optimize (X_c, y_c)

Phase 2: Local model training with DAD-NAR++ privacy

1. Initialise local XGBoost model.
2. For each tree t =1→ N_T:
3. For each node level K(depth):
   (i) Depth-Adaptive Privacy Allocation: $\epsilon_k = \epsilon.e^{-\alpha k}$
   (ii) Laplace Noise Injection to split statistics: $\tilde{g}$ = g + Laplace ($\frac{1}{\epsilon_k}$)
   (iii)Noise-Aware Node Dropout: $P_{drop}(K) = 1\text{-}e^{-\beta k}$
   (iv) If rand () < $P_{drop}(K)$ THEN convert node to leaf and stop expansion.
   (V) Otherwise, perform a split using noisy gain with selected features F_c
4. Store trained local model M_c
5. Each client sends only (M_c, F_c, P_c) to the server.
6. No raw data is shared.

Phase 3: Federated Aggregation (Server-Side)

1. Global Feature Consensus: F_{global} = Top-k (frequency (F_c))
2. Hyperparameter Aggregation: $P_{global}(i) = \frac{1}{C}\sum_{c=1}^{c} P_c(i)$
3. Global Model Parameter Aggregation: M_{global} = Aggregate (M_c)

Phase 4: Find Global Training and Prediction

1. Retrain using global features and parameters:
   M_{global} – Train (X_{train} [: F_{global}], $y_{train, P\ global}$)
2. Predict on global test set: $\hat{y} = M_{global}(X_{test})$

End

V EXPERIMENTAL RESULT

The experimental results show that our model is effective on several kinds of datasets, such as EEG Eye State, Diabetes, Obesity, Breast Cancer, and Housing Price.

5.1 IMPLEMENTATION:

In the FedXGB-OptDP approach, the training data is initially pre-processed by encoding categorical attributes and normalising numerical values to guarantee steady learning. Then, the Genetic Algorithm (GA) is employed to discover a compact and discriminative subset of features by evaluating different feature combinations recorded as binary chromosomes, where fitness corresponds to validation accuracy. This stage removes irrelevant and noisy features, so the model learns only from useful inputs. On the smaller feature space, TPE-based hyperparameter optimisation fine-tunes learning rate, tree depth, and boosting rounds, allowing the model to react to the underlying data features rather than depending on fixed or human parameter values. After both feature selection and parameter tuning are completed, training proceeds under the DAD-NAR privacy mechanism, which includes Laplace noise injection to gradient-based split statistics, depth-adaptive privacy scaling that increases noise deeper in the tree, and noise-aware node dropout to prevent overfitting and avoid revealing sensitive subgroup patterns. The process provides an XGBoost model that was trained locally while protecting privacy, along with its chosen feature indices and optimised hyperparameters. The outputs, which include the trained model parameters, the chosen feature subset, and the optimised hyperparameters, are sent to the aggregation process. The original data, on the other hand, remains local and confidential. This approach makes users to share the information they have successfully learnt, which helps them to work together to build models, while protecting private data from being viewed or exchanged.

5.1.1 EEG DATASET RESULT COMPARISON:

The dataset comprises of brainwave data obtained through Electroencephalography (EEG). The main objective of this EEG eye state is whether an individual's eyes are open or closed. Our model EEG data is spread across multiple devices or organisation, comprises of

14,981 records, and make classification to predict accuracy in terms of privacy. Accuracy is given in horizontally portioned data shown in Figure 5- 9.

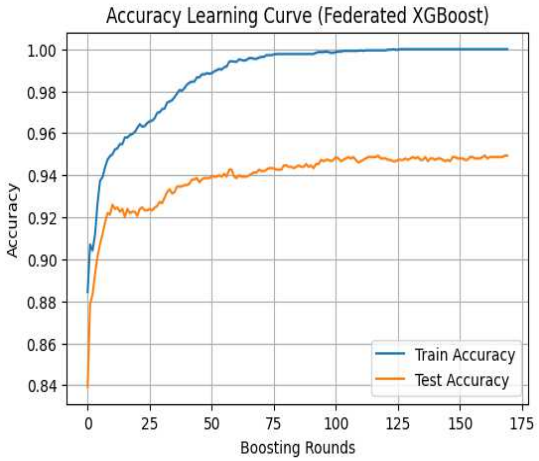


Figure 5: FedXGB-OptDP accuracy curve on EEG Dataset

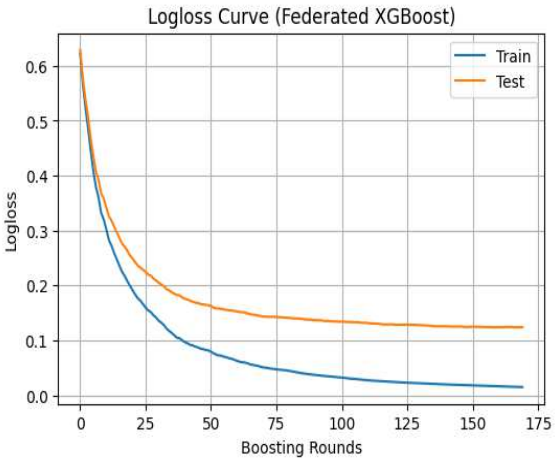


Figure 6: Log loss curve on EEG Dataset

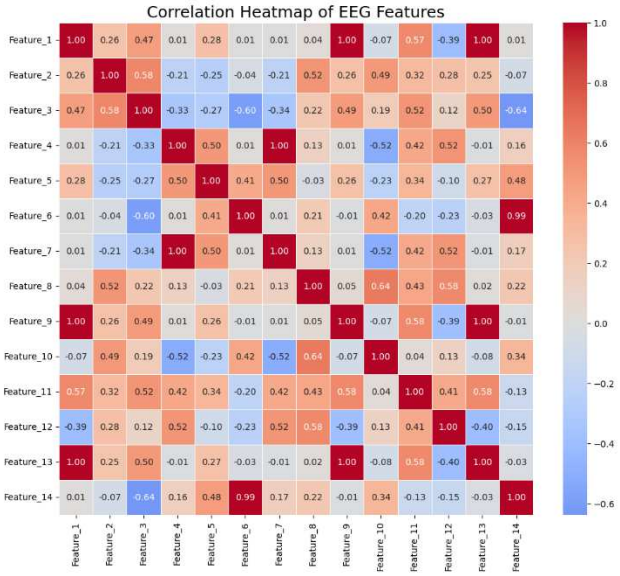


Figure 7: Correlation Heat map on EEG Dataset

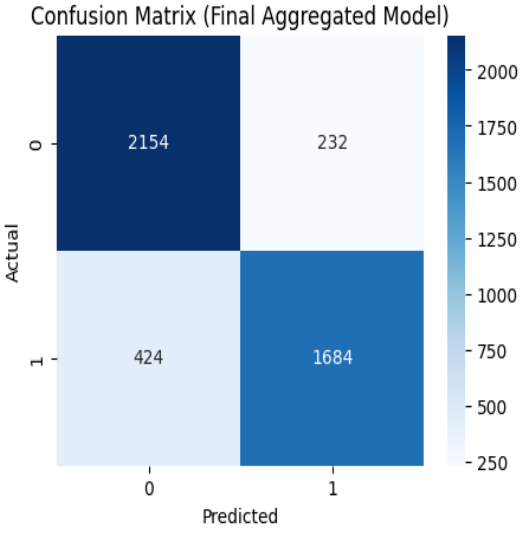


Figure 8: Confusion matrix on EEG Dataset

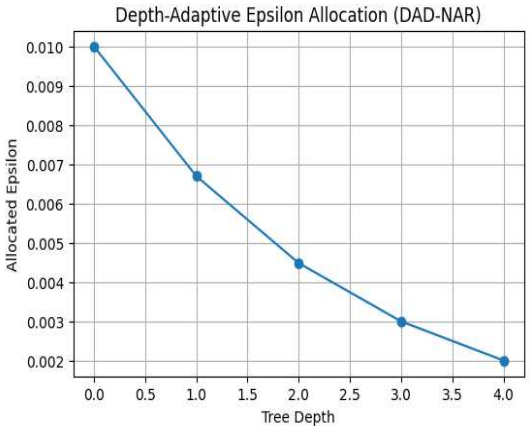


Figure 9: Depth Adaptive Epsilon Allocation

The accuracy in Figure 5. Shows that the model converged effectively, achieved 0.9377 with an F1 score of 0.9314, and approximately 100 boosting rounds. The log loss curve shown in fig.6 shows that substantial decrease during early boosting stages, which indicates that the model achieved a proper balance between learning and generalisation, to prevent overfitting. The confusion matrix in fig.8 produced 2154 true negatives, 1684 true positives, 232 false positives, and 424 false negatives, indicating significant selective capacity. In the regularisation process DAD-NAR, a consistent reduction in epsilon values as tree depth increased, thereby improving privacy preservation for deeper and more sensitive splits. This adaptive technique shows balances between privacy and utility, hence preventing a significant decrease in accuracy in accordance with DP standards. The results shows that the model FedXGB-OptDP produce strong privacy, making it appropriate for secure and distributed EEG data signal processing

5.1.2 DIABETES DATASET RESULT COMPARISON:

In early stage of prediction of diabetes is crucial for improving patient outcomes , which can make use of disease management and appropriate treatment. Many research developing this better and better now days in terms of accuracy for diabetes patients. Our dataset comprises of 70,693 patient records. In traditional centralised system privacy is maintained with only one organisation no more that one organisation in that. In FedXGB-OptDP model enables multiple clients or organisation utilises shared model while protecting a personal data through hybrid optimised (GA+TPE) differential privacy. The evaluation analyses trade of between privacy and utility across various privacy configuration (ε values). The data set was divided into three client, each of which used XGBoost, hyperparameter optimisation, and GA-driven feature selection to find the post predictive clinical features. The procedure used optimised DP for model updates. Only model parameters were safely collected at the server after local training, without patient data. Classification where made at the dataset, where model performed uniformly across all three regions, achieving individual accuracy rates of 0.7392, 0.7542 and 0.7311 respectively. After aggregation accuracy is 0.7415 and F1-score of 0.7513, indicating strong predictive performance.

The consensus feature selection identified ten frequency-selected clinical signals, demonstrating the models, interpretability and usability for medical decision-making. FedXGB-OptDP protects patient privacy and makes accurate predictions, making it feasible option for collaborative healthcare analytics in multi-institutional organisation. The results are shown in Figure 10-14.

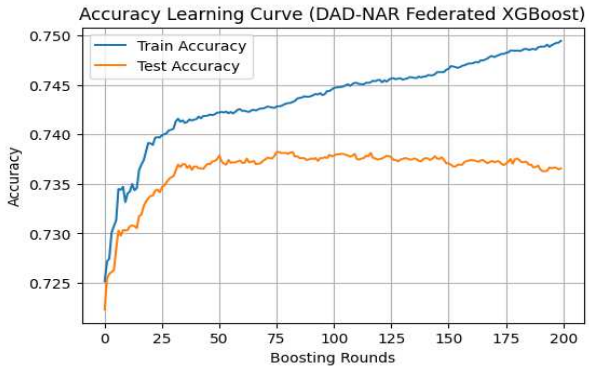


Figure 10: Accuracy curve on Diabetes Dataset

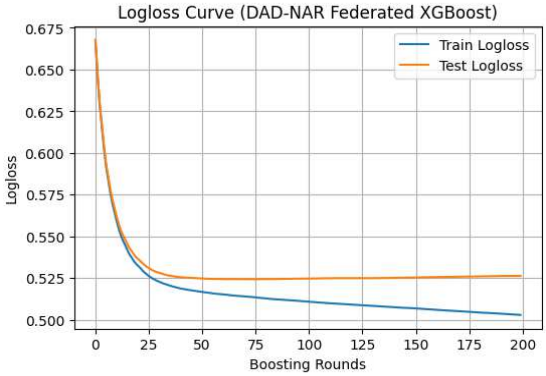


Figure 11: Log loss curve on Diabetes Dataset

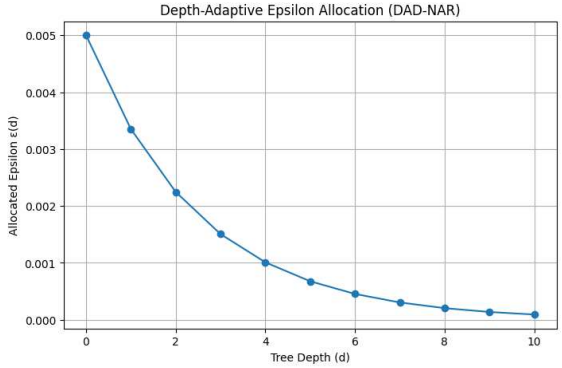


Figure 12: Depth Adaptive Epsilon Allocation

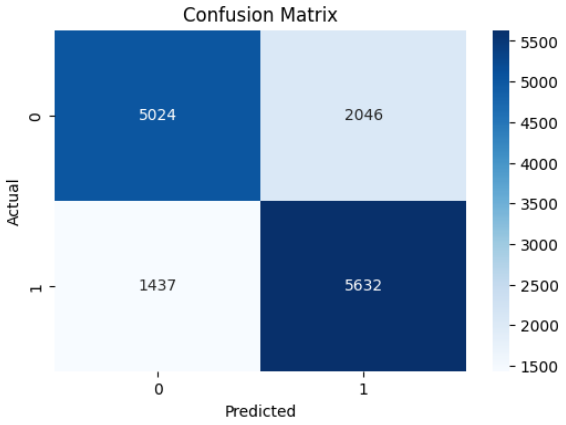


Figure 13: Confusion matrix on Diabetes Dataset

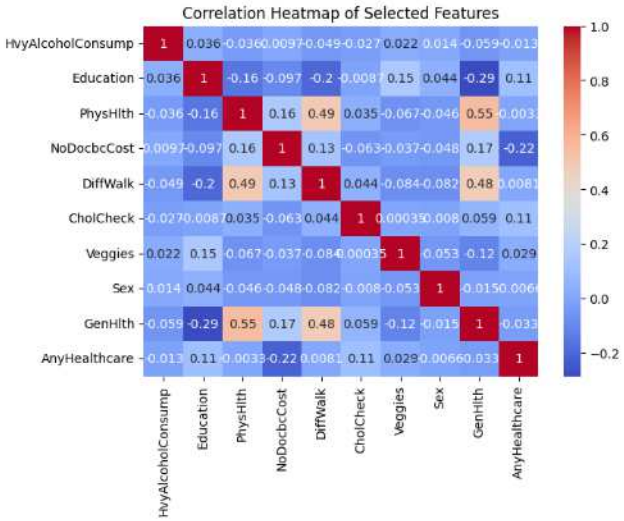


Figure 14: Feature correlation Heat map on Diabetes Dataset

5.1.3 OBESITY DATASET RESULT COMPARISON:

This dataset uses 2,112 obesity records with various medical and lifestyle variables. This classification splits class 0 –not obesity, which includes people with insufficient weight or normal weight, whereas class 1 includes those with overweight levels I, II, or III. Preventive healthcare requires early obesity risk diagnosis, which this binary classification provides. The suggested federated DAD-NAR model used local hyperparameter optimisation and feature selection, while consensus-based feature aggregation ensured the system's exclusive use of the most informative features. The final FL global model outperformed client models with 95.27% accuracy and strong performance. The depth adaptive epsilon decay balances privacy and model utility, while accuracy and log-loss curves converged efficiently. The confusion matrix shows the model's strong discrimination with little misclassification. These results confirm that the privacy-conscious federated strategy is highly effective for early obesity detection in practical remote healthcare settings. The results are shown in Figure 15-19.

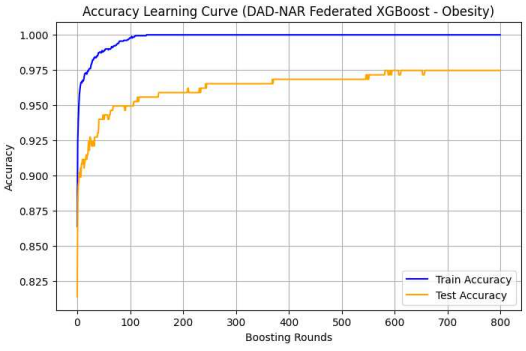


Figure 15: Accuracy curve on Obesity Dataset.

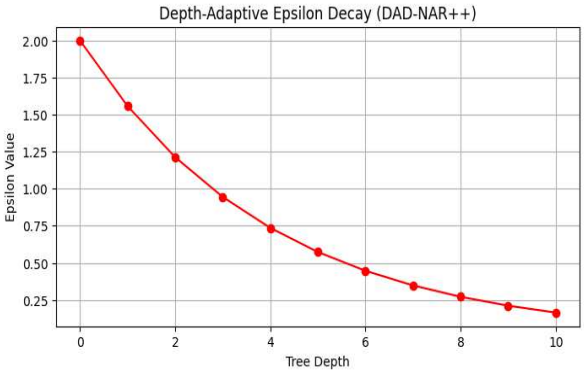


Figure 16: Depth Adaptive Epsilon Allocation

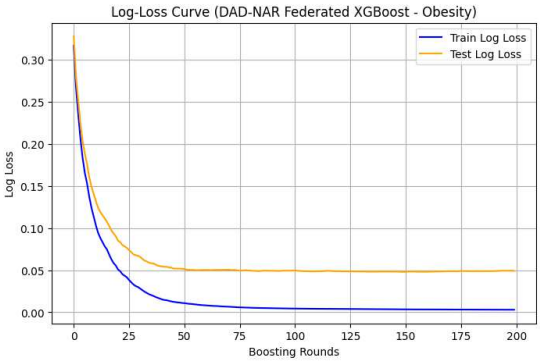


Figure 17: Log loss curve on obesity dataset

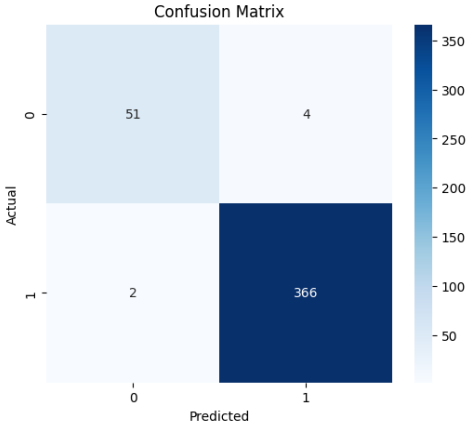


Figure 18: Confusion matrix on obesity dataset

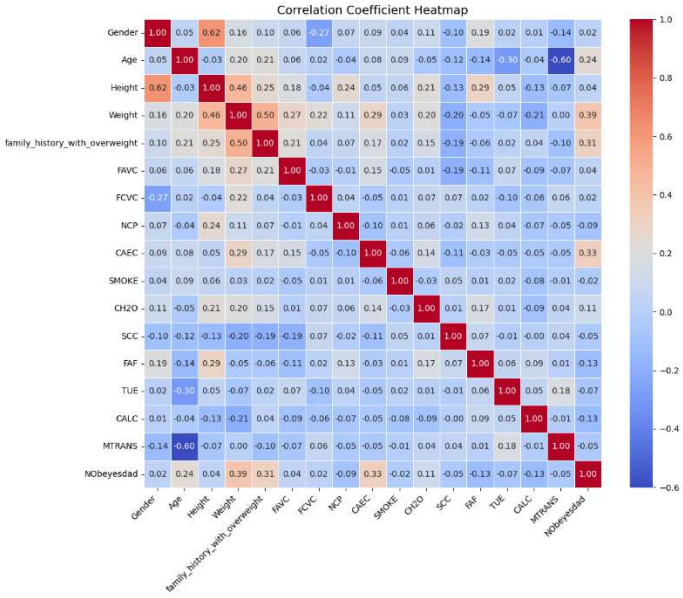


Figure 19: Feature correlation Heat map on Obesity Dataset

5.1.4 BREAST CANCER DATASET RESULT COMPARISON:

The Breast Cancer Wisconsin dataset comprises 569 samples and 32 features, providing a generally huge binary classification task. The proposed model demonstrates robust predictive accuracy across federated clients. The accuracy curve indicates that the model converges frequently, achieving a global test accuracy of approximately 95-96% within the initial 15-20 boosting cycles. The confusion matrix further confirms this, accurately classifying 70 benign and 39 malignant samples, with only a minimal number of misclassifications. Further, compared to convolutional Federated XGBoost, lacking privacy mechanisms or optimisation, the proposed approach preserves nearly equivalent accuracy while incorporating depth-adaptive ε-decay. Thereby enhancing privacy safeguards. The DAD-NAR epsilon decay pattern indicates that the noise introduced for privacy purposes diminishes at deeper levels of tree, thereby mitigating training imbalance. The Feature-correlation heat maps also shows substantial relationship between genetic feature selection qualities, including radius, perimeter, area, and concavity. The suggested model achieves optimal utility-privacy protections, outperforming current methods. Along with this, tested with non-healthcare scenarios also.

House price dataset to test our federated system in non-healthcare settings. The dataset has 545 entries and 13 characteristics from three federated clients. Each customer used a genetic algorithm to optimise hyperparameters and choose features. Due to differences in local data, client 1 had robust result with an accuracy of 0.8909, client 2 moderates with 0.7091, and client 3 inferior with 0.5455. After aggregation, the global model achieved an accuracy of 0.7152 and an F1-score of 0.7516, indicating variations in data quality among the clients. The primary features identified-[1, 3, 0, 2, 4], representing area, bathrooms, bedrooms, and were uniformly selected by all clients, showing their significance in predicting house categories. Hence, dataset shows the adaptability and strength of our model beyond healthcare contexts, validating that the federated methodology with GA-based feature selection is suitable for various real-world applications.

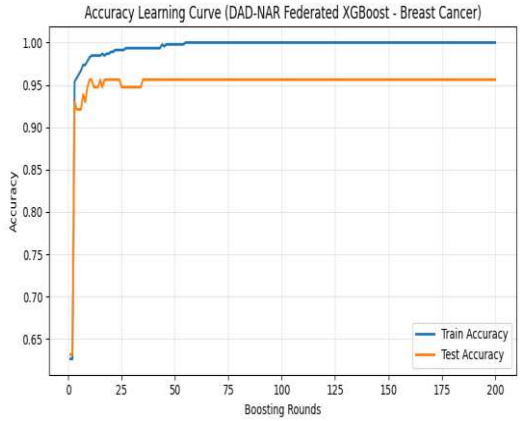


Figure 20: Accuracy curve on Breast Cancer dataset

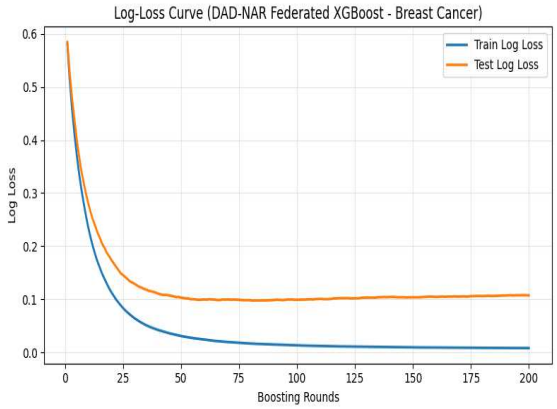


Figure 21: Log loss curve on Breast Cancer dataset

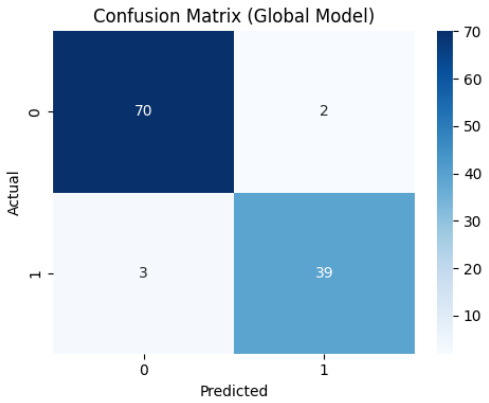


Figure 22: Confusion matrix on Breast cancer

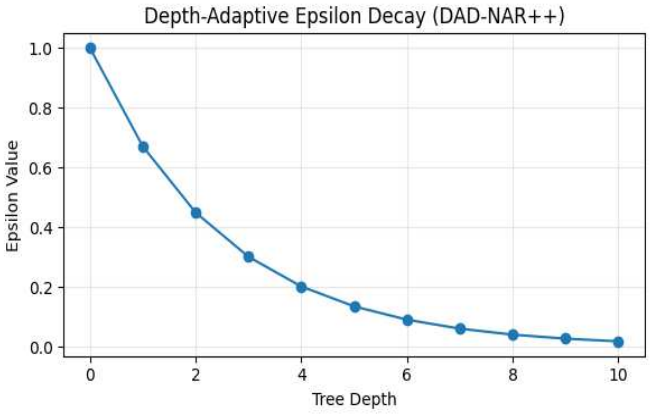


Figure 23: Depth Adaptive Epsilon Allocation

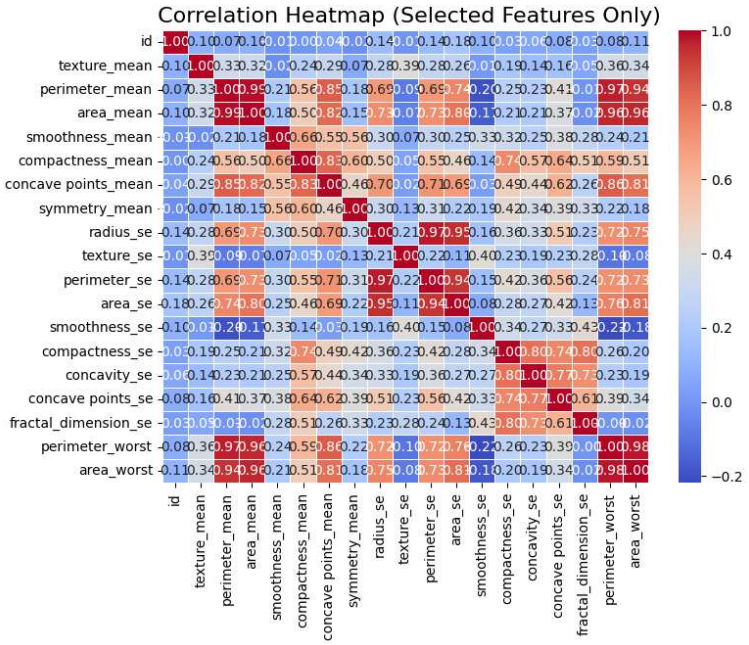


Figure 24: Feature correlation Heat map on Breast Cancer dataset

In Figure 25 shows the accuracy curves for multiple optimisation strategies, demonstrating how each optimiser affects the model's convergence behaviour with datasets such as ACO (Ant Colony Optimisation), CS (Cuckoo Search), FPA (Flower Pollination Algorithm), PSO (Particle Swarm Optimisation), SCA (Sine Cosine Algorithm), SSA (Slap Swarm Algorithm), and hybrid GA+BO (Genetic Algorithm) and Bayesian Optimisation) algorithm and the values are shown in table 3.

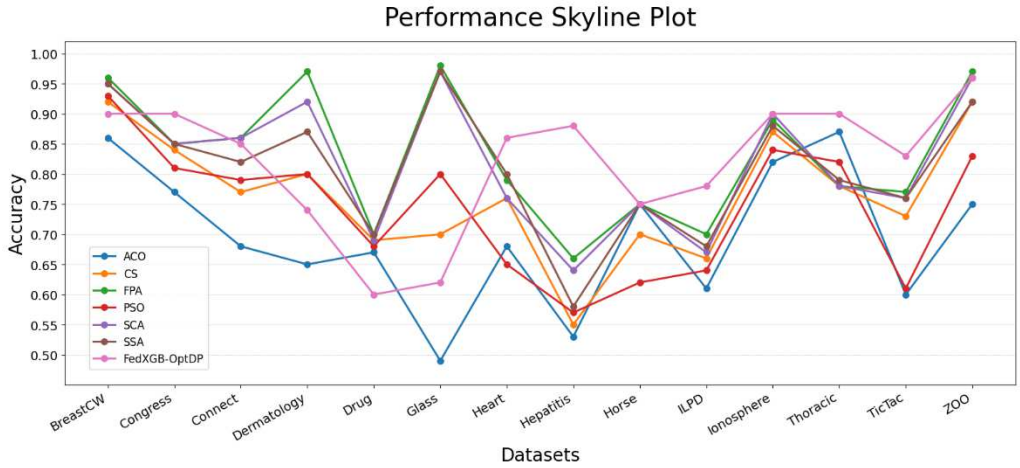


Figure 25: Comparison of Optimization Techniques and Hybrid FedXGB-OptDP Accuracy Across Multiple Datasets"

Table 4 summarises the performance of the proposed FedXGB-OptDP model on five datasets. The results compare centralised XGBoost, privacy-only XGBoost, federated XGBoost without DP, and our optimised DP-based federated approach. FedXGB-OptDP achieves the highest accuracy for EEG and obesity datasets, demonstrating significant utility even under privacy constraints. Breast Cancer and House Pricing show slight accuracy drops, but overall performance is stable. Overall, the analysis shows that FedXGB-OptDP balances

privacy and accuracy better than other model. Despite uniform experimental settings, the suggested model consistently outperforms the others by supporting federated learning and DP with improved accuracy. Random Forest, Adaboost, and Gradient Boosting work without privacy noise or federated constraints, however, FedXGB-OptDP performs better across datasets. The proposed model remains predictive even under stricter privacy limitations. As shown in table 5, its reliability under these situations illustrates its robustness, adaptability, and applicability for real-world applications.

Table 3: Comparison with other optimisation techniques with FedXGB-OptDP

Dataset	ACO	CS	FPA	PSO	SCA	SSA	**FedXGB-OptDP (GA+BO)**
BreastCW	0.8584	0.9503	0.9649	0.9326	0.9637	0.9557	**0.9298**
Congressional	0.7688	0.8349	0.8518	0.8119	0.8495	0.8528	**0.9470**
Connect Bench	0.6827	0.7692	0.8538	0.7885	0.8606	0.8183	**0.8413**
Dermatology	0.6501	0.7978	0.9716	0.8077	0.9590	0.8716	**0.7445**
Drug consumption	0.6691	0.7041	0.7075	0.6740	0.6883	0.7021	**0.6014**
Glass	0.4836	0.6953	0.9869	0.7972	0.9822	0.9860	**0.6212**
Heart	0.6761	0.7556	0.7993	0.6370	0.7704	0.7830	**0.8602**
Hepatitis	0.5337	0.5513	0.6628	0.5436	0.6462	0.5769	**0.8750**
Horse_colic	0.6625	0.6987	0.7573	0.6153	0.7533	0.7567	**0.7444**
ILPD	0.6093	0.6637	0.6938	0.6428	0.6966	0.6777	**0.7797**
Ionosphere	0.8272	0.8693	0.8949	0.8375	0.8938	0.8722	**0.9039**
Thoraric Surgery	0.7251	0.7736	0.8353	0.7579	0.8247	0.7940	**0.9022**
Tic-tac-toe	0.5995	0.7290	0.7843	0.6116	0.7568	0.7626	**0.8299**
ZOO	0.7517	0.9137	0.9725	0.8353	0.9647	0.9216	**0.9677**

Hence, the proposed architecture exhibits excellent overall performance, achieving the highest accuracy on the EEG, Obesity, and Breast Cancer datasets, while also yielding robust and reliable results on the Diabetes and House Pricing datasets. Table 6 shows that suggested approach outperforms PrivaTree, FedXHDP, and FedBoost in numerous usage, proving its robustness and efficacy in federated and privacy preserving systems.

Table 4: Ablation results of all 5 datasets.

FedXGB-OptDP	Centralised model (No privacy, only with XGBoost)		Privacy with XGBoost (Without FL)		XGBoost with FL (Without DP)		**Horizontal Partitioning (Accuracy) FedXGB-OptDP**	
Dataset	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	**Accuracy**	**F1-Score**
EEG dataset	0.9263	0.9262	0.8531	0.8283	0.9252	0.9252	**0.9377**	**0.9314**
Diabetes	0.7501	0.7621	0.7265	0.7367	0.7467	0.7460	**0.7415**	**0.7513**
Obesity	0.9621	0.9624	0.7161	0.7000	0.9117	0.9118	**0.9527**	**0.9526**

Breast cancer	0.9649	0.9726	0.9386	0.9510	0.9386	0.9504	**0.9415**	**0.9148**
House pricing	0.7956	0.7971	0.5036	0.4603	0.7737	0.7832	**0.7152**	**0.7516**

Table 5: Other model comparison:

Different models	EEG dataset		Diabetes		Obesity		Breast cancer		House pricing	
	Accuracy	F1-Score	**Accuracy**	F1-Score	**Accuracy**	F1-Score	**Accuracy**	F1-Score	**Accuracy**	F1-Score
FedXGB-OptDP	**0.9377**	**0.9314**	**0.7415**	**0.7513**	**0.9527**	**0.9526**	**0.9415**	**0.9148**	**0.7152**	**0.7516**
Random Forest	0.9157	0.9105	0.6956	0.6999	0.9464	0.9469	0.9415	0.9183	0.7988	0.7950
Ada-Boosting	0.8400	0.8062	0.7300	0.7323	0.8511	0.8825	0.9240	0.8965	0.7805	0.7831
Gradient Boosting	0.9413	0.9324	0.7334	0.7446	0.9621	0.9623	0.9123	0.8851	0.7134	0.7283

Table 6: Comparison of accuracy value with existing work

Model Name	EEG dataset	Diabetes	Obesity	Breast cancer	House pricing
FedXGB-OptDP(Proposed work)	**0.9377**	**0.7415**	**0.9587**	**0.9415**	**0.7152**
PrivaTree	0.8050	0.7220	0.8640	-	-
FedXHDP	0.8561	0.7561	0.9551	-	-
FedBoost	-	-	-	0.9400	$R^2 \approx 0.70 - 0.82$

Table 7 summarises our model hyperparameters for each datasets. All datasets show a consistent tuning pattern: moderate tree depth, balanced sampke ratios, and calibrated learning rates for stable independent convergence. Most datasets use DART boosting to redce overfitting, while depth-adaptive DP limits training noise. Each datasets shows the core consensus features identified during optimisation, stressing the traits that model influenced the models final judgements.

Table 7: Values of hyperparameters:

Parameter	**EEG Dataset**	**Diabetes dataset**	**Obesity dataset**	**Breast cancer**	**Housing price**
Max_depth	6	5	6	4	6
Objective	Binary: logistic	Binary: logistic	Multi-class classification	Binary: logistic	Binary: logistic

Min_child_weight	3	2	3	3	3
Rate drop	0.25	0.3	-	0.1	-
Skip drop	0.20	0.25	-	0.5	-
n-estimators	250	200	800	100	230
Learning rate	0.08	0.12	0.045	0.1	0.084
Subsample	0.85	0.85	0.82	0.8	0.86
Colsample_bytree	0.75	0.8	0.78	0.8	0.77
Alpha (L1 reg)	0.4	0.5	1.2	0.1	-
Reg_lambda (L2 reg)	1.2	1.1	2.1	1.1	1.24
Gamma (Split penalty)	0.1	0.2	0.6	0.1	0.1`8
Eval_metric	logloss	logloss	mlogloss	logloss	logloss
Booster	Dart (DAD-Dropout Additive Trees)	Dart (DAD-Dropout Additive Trees)	Gbtree (with DAD)	Dart	Dart
Epsilon-DP noise level	1.0(adaptive decreased during depth growth)	1.0(adaptive decreased during depth growth)	2.0 (Depth-Adaptive Differential privacy)	0.5	Root-5.0, Middle-2.0, Leaf- 0.5.
Top consensus features	[2,6,10,8,13,1, 5,3,9,12]	[2,5,7,9,11,13, 15,17,18, 20]	[1, 2, 3, 5, 7, 9, 11, 12, 13, 16, 18, 19, 21, 22]	[12,0,7,1,8,11, 10 ,4,16,15]	[1,3,0,2,4]

VI. Comparison of IID & Non-IID Data:

An extensive comparison was performed on IID (Independent and Identically Distributed Data) and Non-IID (Non-Independent and Identically Distributed Data) training settings to evaluate the durability of our federated model across realistic data distributions. because, in practical federated environments—particularly in healthcare—clients rarely have identically distributed data. Hence, divergence in patient demographics, hospital apparatus, collection methodologies, or regional health trends naturally results in non-IID datasets, this approach is essential. Hence, to assess the adaptability of our model in different environments, we evaluated its performance across five datasets shown in table 8. The findings indicate that our model regularly demonstrates reliability under non-IID situations. Notably, for both the Diabetes and Housing Price datasets, the accuracy improved in the non-IID scenario (0.8630 and 0.7956), suggesting that the model benefits from heterogeneous client data.

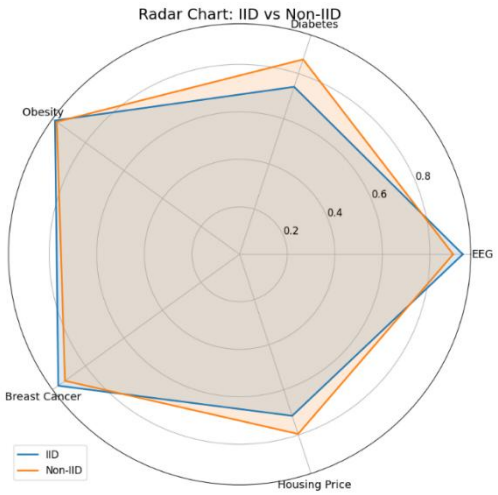


Figure 26: Accuracy comparison between IID and Non-IID data

Despite a slight reduction in accuracy for datasets such as EEG and Breast Cancer under non-IID conditions, the decrease was small, and overall performance continued to be strong. Hence, the results validate that our federated system is not limited to optimal IID conditions but is capable of managing practical, real-world non-IID data distributions shown in figure 26. This makes our concept appropriate for implementation in healthcare, multidisciplinary collaborations, and other fields where consistent data distribution across customers is not assured.

Table 8: IID and Non-IID comparison for the same dataset:

Dataset	IID data (Accuracy)	Non-IID data (Accuracy)
EEG Dataset	0.9377	**0.8972**
Diabetes dataset	0.7415	**0.8630**
Obesity dataset	0.9587	**0.9495**
Breast cancer	0.9415	**0.9066**
Housing price	0.7152	**0.7956**

VII CONCLUSION:

The proposed model provides a robust and feasible approach for privacy-preserving collaborative learning in healthcare by integrating XGBoost with optimised DP, DAD-NAR regularisation. This consistently provides high accuracy while safeguarding sensitive patient information, demonstrating reliable performance across EEG, Diabetes, Obesity, Breast Cancer, and Housing datasets under both IID and Non-IID conditions to achieve privacy in terms of accuracy. This ,given that real-world healthcare data frequently produces imbalance, inconsistency, and irregular distribution across different client environments, hence adaptability is essential. The optimised DP developed through algorithms and Bayesian optimisation ensures a robust equilibrium between privacy preservation and predictive accuracy. To enhance generalisation and mitigate overfitting, the model utilises DAD regularisation, which adaptively decreases necessary complexity. NAR regularisation enhances this process by stabilising updates through the application of normalised and altered learning

rates across clients, thereby increasing the model's robustness against biased data and different gradients. Together, DAD and NAR enhance the training process, minimise variations, and facilitate more stable convergence in federated environments. When compare to expensive cryptographic techniques, such as SMPC or HE, the proposed model gives significantly greater computational efficiency and is more appropriate for resource-limited healthcare settings. By integrating XGBoost, the framework advantages includes more rapid convergence, reduced Communication overhead and enhanced predictive accuracy on healthcare datasets.

The hybrid optimisation strategy further enables accurate tuning of the privacy-utility balance, enhancing the models adaptability to a broad spectrum of clinical solution for real-world healthcare monitoring, effectively handling the challenges of data heterogeneity, privacy preservation, and model stability within federated learning environments.
Future work might strengthen FedXGB-OptDP, enhancing communication efficiency to deal with larger federated networks. Specific privacy measures can be implemented, ensuring that each client receives alerts according to their data quality and sensitivity. Model accuracy can be enhanced by using advanced optimisation or ensemble-based federated boosting. Further, it can be developed to include heterogeneous healthcare data, including images, signals, and medical text.

ACKNOWLEDGMENT
All authors thank Vellore Institute of Technology, Vellore, India, for providing the facility.

FUNDING

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CONFLICT OF INTEREST

The authors of this work declare that they have no conflict of interest.

REFERENCES:

[1] Ekwonwune EN, Ubochi CI, Duroha AE. Data Mining as a Technique for Healthcare Approach. International Journal of Communications, Network and System Sciences [Internet]. 2022;15(09):149–165.

[2] Kolling ML, Furstenau LB, Sott MK, et al. Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development. Int J Environ Res Public Health [Internet]. 2021;18(6):3099.

[3] Eyal-Altman N, Last M, Rubin E. PCM-SABRE: a platform for benchmarking and comparing outcome prediction methods in precision cancer medicine. BMC Bioinformatics [Internet]. 2017;18(1):40.

[4] Urbanowicz RJ, Olson RS, Schmitt P, et al. Benchmarking relief-based feature selection methods for bioinformatics data mining. J Biomed Inform [Internet]. 2018;85:168–188.

[5] No Title.

[6] Avula, R. (2022). Data-driven decision-making in healthcare through advanced data mining techniques: A survey on applications and limitations. International Journal of Applied Machine Learning and Computational Intelligence, 12(4) 64-85. No Title.

[7] Chen IY, Joshi S, Ghassemi M, et al. Probabilistic Machine Learning for Healthcare. Annu Rev Biomed Data Sci [Internet]. 2021;4(1):393–415.

[8] Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med [Internet]. 2020;104:101822.

[9] Mohammad-Rahimi H, Nadimi M, Rohban MH, et al. Machine learning and orthodontics, current trends and the future opportunities: A scoping review. American Journal of Orthodontics and Dentofacial Orthopedics [Internet]. 2021;160(2):170-192.e4.

[10] Chen P-HC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nat Mater [Internet]. 2019;18(5):410–414.

[11] He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. Nat Med [Internet]. 2019;25(1):30–36.

[12] Chen Y, Liu Q, Guo D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. J Med Virol [Internet]. 2020;92(4):418–423.

[13] Reddy RR, Mamatha C, Reddy RG. A Review on Machine Learning Trends, Application and Challenges in Internet of Things. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) [Internet]. IEEE; 2018. p. 2389–2397. Available from: https://ieeexplore.ieee.org/document/8554800/.

[14] Tătaru OS, Vartolomei MD, Rassweiler JJ, et al. Artificial Intelligence and Machine Learning in Prostate Cancer Patient Management—Current Trends and Future Perspectives. Diagnostics [Internet]. 2021;11(2):354.

[15] Rossa-Roccor V, Karim ME. Are US adults with low-exposure to methylmercury at increased risk for depression? A study based on 2011–2016 National Health and Nutrition Examination Surveys (NHANES). Int Arch Occup Environ Health [Internet]. 2021;94(3):419–431.

[16] Fung BCM, Wang K, Chen R, et al. Privacy-preserving data publishing. ACM Comput Surv [Internet]. 2010;42(4):1–53.

[17] 19. T. Craig and M. Ludloff, Privacy & Big Data. O'Reilly Media Aug 2011. No Title.

[18] Zaman ANK, Obimbo C, Dara RA. A Novel Differential Privacy Approach that Enhances Classification Accuracy. Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering - C3S2E '16 [Internet]. New York, New York, USA: ACM Press; 2016. p. 79–84. Available from: http://dl.acm.org/citation.cfm?doid=2948992.2949027.

[19] Aujla GS, Jindal A, Chaudhary R, et al. DLRS: Deep Learning-Based Recommender System for Smart Healthcare Ecosystem. ICC 2019 - 2019 IEEE International Conference on Communications (ICC) [Internet]. IEEE; 2019. p. 1–6. Available from: https://ieeexplore.ieee.org/document/8761416/.

[20] Aledhari M, Razzak R, Parizi RM, et al. Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications. IEEE Access [Internet]. 2020;8:140699–140725.

[21] Backes M, Berrang P, Bieg M, et al. Identifying Personal DNA Methylation Profiles by Genotype Inference. 2017 IEEE Symposium on Security and Privacy (SP) [Internet]. IEEE; 2017. p. 957–976. Available from: http://ieeexplore.ieee.org/document/7958619/.

[22] and K. Lauter DJWTFMN. "Privately evaluating decision trees and random forests," in Proc. Privacy Enhancing Technol., vol. 2016, no. 4, pp. 335–355, Oct. 2016.

[23] Damgard I, Escudero D, Frederiksen T, et al. New Primitives for Actively-Secure MPC over Rings with Applications to Private Machine Learning. 2019 IEEE Symposium on Security and Privacy (SP) [Internet]. IEEE; 2019. p. 1102–1120. Available from: https://ieeexplore.ieee.org/document/8835310/.

[24] Zhan J. Using Homomorphic Encryption For Privacy-Preserving Collaborative Decision Tree Classification. 2007 IEEE Symposium on Computational Intelligence and Data Mining [Internet]. IEEE; 2007. p. 637–645. Available from: http://ieeexplore.ieee.org/document/4221360/.

[25] Vaidya J, Clifton C, Kantarcioglu M, et al. Privacy-preserving decision trees over vertically partitioned data. ACM Trans Knowl Discov Data [Internet]. 2008;2(3):1–27.

[26] Zhan WD and Z. , "Building decision tree classifier on private data,." Computer Society, Inc. 2002;in Proc. I(Australia, Australian):1–8.

[27] Shokri R, Shmatikov V. Privacy-Preserving Deep Learning. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security [Internet]. New York, NY, USA: ACM; 2015. p. 1310–1321. Available from: https://dl.acm.org/doi/10.1145/2810103.2813687.

[28] B. McMahan, E. Moore, D. Ramage, S. Hampson and BA y. A. "Communication-efficient learning of deep networks from decentralized data," . ,. Conf Artif Intell Statist. 2017;Florida, U(in Proc. 20th Int):1273–1282.

[29] Yin X, Zhu Y, Hu J. A Comprehensive Survey of Privacy-preserving Federated Learning. ACM Comput Surv [Internet]. 2022;54(6):1–36.

[30] Kulkarni V, Kulkarni M, Pant A. Survey of Personalization Techniques for Federated Learning. 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) [Internet]. IEEE; 2020. p. 794–797. Available from: https://ieeexplore.ieee.org/document/9210355/.

[31] Nguyen DC, Pham Q-V, Pathirana PN, et al. Federated Learning for Smart Healthcare: A Survey. ACM Comput Surv [Internet]. 2023;55(3):1–37.

[32] Zhang H, Bosch J, Olsson HH, et al. AF-DNDF: Asynchronous Federated Learning of Deep Neural Decision Forests. 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) [Internet]. IEEE; 2021. p. 308–315. Available from: https://ieeexplore.ieee.org/document/9582575/.

[33] Truex S, Baracaldo N, Anwar A, et al. A Hybrid Approach to Privacy-Preserving Federated Learning. Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security [Internet]. New York, NY, USA: ACM; 2019. p. 1–11. Available from: https://dl.acm.org/doi/10.1145/3338501.3357370.

[34] de Hoogh S, Schoenmakers B, Chen P, et al. Practical Secure Decision Tree Learning in a Teletreatment Application. 2014. p. 179–194. Available from: https://link.springer.com/10.1007/978-3-662-45472-5_12.

[35] Kim H-J, Kim H-I, Chang J-W. A Privacy-Preserving kNN Classification Algorithm Using Yao's Garbled Circuit on Cloud Computing. 2017 IEEE 10th International Conference on Cloud Computing (CLOUD) [Internet]. IEEE; 2017. p. 766–769. Available from: https://ieeexplore.ieee.org/document/8030668/.

[36] Riazi MS, Javaheripi M, Hussain SU, et al. MPCircuits: Optimized Circuit Generation for Secure Multi-Party Computation. 2019 IEEE International Symposium on Hardware Oriented Security and Trust (HOST) [Internet]. IEEE; 2019. p. 198–207. Available from: https://ieeexplore.ieee.org/document/8740831/.

[37] Dwork C, McSherry F, Nissim K, et al. Calibrating Noise to Sensitivity in Private Data Analysis. Journal of Privacy and Confidentiality [Internet]. 2017;7(3):17–51.

[38] Zhang Y, Hao Z, Wang S. A Differential Privacy Support Vector Machine Classifier Based on Dual Variable Perturbation. IEEE Access [Internet]. 2019;7:98238–98251.

[39] Qi X, Ma X, Bai X, et al. Differential Privacy Preserving Data Publishing Based on Bayesian Network. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) [Internet]. IEEE; 2020. p. 1718–1726. Available from: https://ieeexplore.ieee.org/document/9343115/.

[40] Li D, Wang J, Tan Z, et al. Differential Privacy Preservation in Interpretable Feedforward-Designed Convolutional Neural Networks. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) [Internet]. IEEE; 2020. p. 631–638. Available from: https://ieeexplore.ieee.org/document/9343157/.

[41] Blum A, Dwork C, McSherry F, et al. Practical privacy. Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems [Internet]. New York, NY, USA: ACM; 2005. p. 128–138. Available from: https://dl.acm.org/doi/10.1145/1065167.1065184.

[42] Zhao L, Ni L, Hu S, et al. InPrivate Digging: Enabling Tree-based Distributed Data Mining with Differential Privacy. IEEE INFOCOM 2018 - IEEE Conference on Computer Communications [Internet]. IEEE; 2018. p. 2087–2095. Available from: https://ieeexplore.ieee.org/document/8486352/.

[43] Yamamoto F, Wang L, Ozawa S. New Approaches to Federated XGBoost Learning for Privacy-Preserving Data Analysis. 2020. p. 558–569. Available from: https://link.springer.com/10.1007/978-3-030-63833-7_47.

[44] Li Q, Wu Z, Wen Z, et al. Privacy-Preserving Gradient Boosting Decision Trees. Proceedings of the AAAI Conference on Artificial Intelligence [Internet]. 2020;34(01):784–791.

[45] Gao S, Li X, Shi Z, et al. Towards Fair and Decentralized Federated Learning System for Gradient Boosting Decision Trees. Michienzi A, editor. Security and Communication Networks [Internet]. 2022;2022:1–18.

[46] Abspoel M, Escudero D, Volgushev N. Secure training of decision trees with continuous attributes. Proceedings on Privacy Enhancing Technologies [Internet]. 2021;2021(1):167–187.

[47] Li J, Tian Y, Zhu Y, et al. A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. Artif Intell Med [Internet]. 2020;103:101814.

[48] Patil A, Singh S. Differential private random forest. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI) [Internet]. IEEE; 2014. p. 2623–2630. Available from: https://ieeexplore.ieee.org/document/6968348.

[49] Rana S, Gupta SK, Venkatesh S. Differentially Private Random Forest with High Utility. 2015 IEEE International Conference on Data Mining [Internet]. IEEE; 2015. p. 955–960. Available from: http://ieeexplore.ieee.org/document/7373418/.

[50] Fletcher S, Islam MZ. Differentially private random decision forests using smooth sensitivity. Expert Syst Appl [Internet]. 2017;78:16–31.

[51] Liu Y, Ma Z, Yang Y, et al. RevFRF: Enabling Cross-Domain Random Forest Training With Revocable Federated Learning. IEEE Trans Dependable Secure Comput [Internet]. 2022;19(6):3671–3685.

[52] Feng Z, Xiong H, Song C, et al. SecureGBM: Secure Multi-Party Gradient Boosting. 2019 IEEE International Conference on Big Data (Big Data) [Internet]. IEEE; 2019. p. 1312–1321. Available from: https://ieeexplore.ieee.org/document/9006000/.

[53] Li Q, Wen Z, He B. Practical Federated Gradient Boosting Decision Trees. Proceedings of the AAAI Conference on Artificial Intelligence [Internet]. 2020;34(04):4642–4649.

[54] Bai J, Song X, Cui S, et al. Scalable Private Decision Tree Evaluation with Sublinear Communication. Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security [Internet]. New York, NY, USA: ACM; 2022. p. 843–857. Available from: https://dl.acm.org/doi/10.1145/3488932.3517413.

[55] Zheng Y, Duan H, Wang C, et al. Securely and Efficiently Outsourcing Decision Tree Inference. IEEE Trans Dependable Secure Comput [Internet]. 2022;19(3):1841–1855.

[56] Qu Y, Uddin MP, Gan C, et al. Blockchain-enabled Federated Learning: A Survey. ACM Comput Surv [Internet]. 2023;55(4):1–35.

[57] Li F, Zhang X, Zhang X, et al. Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets. Inf Sci (N Y) [Internet]. 2018;422:242–256.

[58] Ohta S, Kurebayashi R, Kobayashi K. Minimizing False Positives of a Decision Tree Classifier for Intrusion Detection on the Internet. Journal of Network and Systems Management [Internet]. 2008;16(4):399–419.

[59] Maimon, O. Z., & Rokach L. Data mining with decision trees: theory and applications. World scientific; 2014.

[60] Bakhshinategh B, Zaiane OR, ElAtia S, et al. Educational data mining applications and tasks: A survey of the last 10 years. Educ Inf Technol (Dordr) [Internet]. 2018;23(1):537–553.

[61] Al Hamad M, Zeki AM. Accuracy vs. Cost in Decision Trees: A Survey. 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) [Internet]. IEEE; 2018. p. 1–4. Available from: https://ieeexplore.ieee.org/document/8855780/.

[62] Pitchayadejanant K, Nakpathom P. Data mining approach for arranging and clustering the agro-tourism activities in orchard. Kasetsart Journal of Social Sciences [Internet]. 2018;39(3):407–413.

[63] Patel H, Patel A, Patel A. A Comprehensive Analysis of Privacy-Preserving Techniques in Machine Learning. 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N) [Internet]. IEEE; 2024. p. 1836–1841. Available from: https://ieeexplore.ieee.org/document/10895292/.

[64] Nguyen TT, Huynh TT, Ren Z, et al. A Survey of Privacy-Preserving Model Explanations: Privacy Risks, Attacks, and Countermeasures. 2024;

[65] Pulido-Gaytan LB, Tchernykh A, Cortés-Mendoza JM, et al. A Survey on Privacy-Preserving Machine Learning with Fully Homomorphic Encryption. 2021. p. 115–129. Available from: https://link.springer.com/10.1007/978-3-030-68035-0_9.

[66] Park J, Lim H. Privacy-Preserving Federated Learning Using Homomorphic Encryption. Applied Sciences [Internet]. 2022;12(2):734.

[67] Sav S, Pyrgelis A, Troncoso-Pastoriza JR, et al. POSEIDON: Privacy-Preserving Federated Neural Network Learning. 2021;

[68] Alghamdi W, Salama R, Sirija M, et al. Secure Multi-Party Computation for Collaborative Data Analysis. Vijayan V, Senthil Kumar TS, editors. E3S Web of Conferences [Internet]. 2023;399:04034.

[69] Abadi M, Chu A, Goodfellow I, et al. Deep Learning with Differential Privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security [Internet]. New York, NY, USA: ACM; 2016. p. 308–318. Available from: https://dl.acm.org/doi/10.1145/2976749.2978318.

[70] Choudhury O, Gkoulalas-Divanis A, Salonidis T, et al. Differential Privacy-enabled Federated Learning for Sensitive Health Data. 2020;

[71] Liu T, Di B, Song L. Privacy-Preserving Federated Edge Learning: Modeling and Optimization. IEEE Communications Letters [Internet]. 2022;26(7):1489–1493.

[72] He Z, Wang L, Cai Z. Clustered Federated Learning With Adaptive Local Differential Privacy on Heterogeneous IoT Data. IEEE Internet Things J [Internet]. 2024;11(1):137–146.

[73] Zhang L, Zhu T, Xiong P, et al. A Game-Theoretic Federated Learning Framework for Data Quality Improvement. IEEE Trans Knowl Data Eng [Internet]. 2023;35(11):10952–10966.

[74] Jia H, Xing Z, Song W. A New Hybrid Seagull Optimization Algorithm for Feature Selection. IEEE Access [Internet]. 2019;7:49614–49631.

[75] Arora S, Singh H, Sharma M, et al. A New Hybrid Algorithm Based on Grey Wolf Optimization and Crow Search Algorithm for Unconstrained Function Optimization and Feature Selection. IEEE Access [Internet]. 2019;7:26343–26361.

[76] Jia H, Li J, Song W, et al. Spotted Hyena Optimization Algorithm With Simulated Annealing for Feature Selection. IEEE Access [Internet]. 2019;7:71943–71962.

[77] Sheng Y, Zeng L, Cao S, et al. FedBoost: Bayesian Estimation Based Client Selection for Federated Learning. IEEE Access [Internet]. 2024;12:52255–52266.

[78] Wu J, Ji H, Yi J, et al. Optimizing Client Selection in Federated Learning Base on Genetic Algorithm. Cluster Comput [Internet]. 2025;28(6):400.

[79] Wu Z, Zhou M, Lin Z, et al. Improved Genetic Algorithm and XGBoost Classifier for Power Transformer Fault Diagnosis. Front Energy Res [Internet]. 2021;9.