RKMV

Project Work Submission

Secure Models Optimization for Healthcare IoT Data using Federated Learning

SANKHADEEP MUKHERJEE (ROLL NO. :295)
BIKRAM HALDER (ROLL NO. :312)
RAJDEEP MANDAL (ROLL NO. :293)

DATE OF SUBMISSION – 29/04/2023

Department of Computer Science & Electronics
Ramakrishna Mission Vidyamandira, Belur Math

Supervisor: Dr. Arindam Sarkar, HOD, Department of Computer Science and Electronics, RKMV

"Machine Intelligence is the last invention that humanity will ever need to make."

RAMAKRISHNA MISSION VIDYAMANDIRA
BELUR MATH, HOWRAH

CERTIFICATE OF COMPLETION OF PROJECT

This is to certify that Mr./Ms. **Sankhadeep Mukherjee**, a student of **Department of Computer Science & Electronics**, has successfully completed a project titled "**SECURE MODELS OPTIMIZATION FOR HEALTHCARE IoT DATA USING FEDERATED LEARNING"** under the guidance of **Dr. Arindam Sarkar** in **Ramakrishna Mission Vidyamandira, Belur Math, Howrah** for a period from **05 / 02 / 2023** to **29 / 04 / 2023**.

Signature of Project Guide **External Examiner**

Date: 29 / 04 / 2023

Place: **Ramakrishna Mission Vidyamandira**

RAMAKRISHNA MISSION VIDYAMANDIRA
BELUR MATH, HOWRAH

CERTIFICATE OF COMPLETION OF PROJECT

This is to certify that Mr./Ms. **Bikram Halder**, a student of **Department of Computer Science & Electronics**, has successfully completed a project titled "**SECURE MODELS OPTIMIZATION FOR HEALTHCARE IoT DATA USING FEDERATED LEARNING"** under the guidance of **Dr. Arindam Sarkar** in **Ramakrishna Mission Vidyamandira, Belur Math, Howrah** for a period from **05 / 02 / 2023** to **29 / 04 / 2023**.

Signature of Project Guide **External Examiner**

Date: 29 / 04 / 2023

Place: **Ramakrishna Mission Vidyamandira**

RAMAKRISHNA MISSION VIDYAMANDIRA
BELUR MATH, HOWRAH

CERTIFICATE OF COMPLETION OF PROJECT

This is to certify that Mr./Ms. **Rajdeep Mandal**,a student of **Department of Computer Science & Electronics**,has successfully completed a project titled "**SECURE MODELS OPTIMIZATION FOR HEALTHCARE IoT DATA USING FEDERATED LEARNING"** under the guidance of **Dr. Arindam Sarkar** in **Ramakrishna Mission Vidyamandira, Belur Math, Howrah** for a period from **05 / 02 / 2023** to **29 / 04 / 2023**.

Signature of Project Guide **External Examiner**

Date: 29 / 04 / 2023

Place: **Ramakrishna Mission Vidyamandira**

ACKNOWLEDGEMET

It gives us a great sense of pleasure to present the report of the Project Work undertaken during B.Sc in Computer Science Final Year. We owe special debt of gratitude to my Project Coordinator **Mr. Arindam Sarkar, HOD, Assistant Professor**, Department of Computer Science and Electronics, Ramakrishna Mission Vidyamandira, Belur Math for her/his constant support and guidance throughout the course of my work. It is only her/his cognizant efforts that my endeavors have seen light of the day.

We deeply thanks to my Project Guide **Mr. Sarbajit Manna, Assistant Professor**, Department of Computer Science and Electronics, Ramakrishna Mission Vidyamandira, Belur Math and **Mr. Atanu Mondol, Assistant Professor**, Department of Computer Science and Electronics, Ramakrishna Mission Vidyamandira, Belur Math for guiding and correcting various documents of mine with attention and care.

I also take the opportunity to acknowledge the contribution of **Swami Mahaprajnananda**, Principal, Ramakrishna Mission Vidyamandira, Belur Math and **Mr. Sanjib Basu,** Department of Computer Science and Electronics, Ramakrishna Mission Vidyamandira, Belur for his full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of my project.

Last but not the least, We acknowledge my friends for their contribution in the completionof the project.

Sankhadeep Mukherjee (Roll No.: 295)

Bikram Halder (Roll No.: 312)

Rajdeep Mandal (Roll No.: 293)

Table of Contents

List of Figures

Abstract—

Machine learning has found extensive application in several sectors of smart healthcare sectors that collect vast amounts of data from multiple IoT devices. Among the commonly used machine learning (ML) models and ensemble learning (EL) models, they are effective for data classification and have been employed in various real-world scenarios like anomaly detection and disease diagnosis. However, training ML and EL models like AdaBoost, Linear Discriminant Analysis (LDA), Extra Trees Classifier (ETC), Random Forest, Gaussian Naive Bayes, Bagging, Gradient Boost and Decision Tree (DT) typically necessitates obtaining labeled IoT data from numerous sources, resulting in concerns over data privacy. Current solutions assume that training data can be securely collected from multiple providers, which is often not the case in reality.

The platform encrypts IoT data and records it on a decentralized ledger. Through a rigorous security analysis, we confirm that the proposed scheme upholds the confidentiality of sensitive data for individual data providers, as well as the AdaBoost, Linear Discriminant Analysis, Extra Trees Classifier, and Decision Trees models parameters for data analysts. Furthermore, extensive experiments showcase the efficiency of the proposed approach.

Index Terms— Federated Learning, IoT data, Machine Learning, Ensemble Learning, Linear Discriminant Analysis, AdaBoost, Extra Trees Classifier, Decision Tree, Data Protection.

CHAPTER – 1 : INTRODUCTION

1.1. BACKGROUND

Smart healthcare have seen a surge in the implementation of advanced Internet-of-Things (IoT) infrastructures in recent years, leading to the collection of a massive volume of data from various IoT devices deployed across multiple city sectors, such as transportation, manufacturing, energy transmission, and agriculture [1].To address the challenges stemming from processing requirements of IoT data, a growing number of innovations driven by machine learning (ML) and Ensemble Learning (EL) technology have been proposed. Among the various ML models, Linear Discriminant Analysis (LDA) is a prominent supervised learning models that efficiently performs dimensionality reduction, Decision Tree is used to handle non-linear relationships and flexibility, an ensemble tree-based machine learning technique called ExtraTrees Classifier uses randomization to lower variance and computational complexity, AdaBoost (Adaptive Boosting) as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. Gaussian Naive Bayes is a variant of Naive Bayes which supports continuous values and has an assumption that each class is normally distributed. Consequently, These ML and EL models are adopted in many domains to address real-world classification problems in IoT-enabled smart healthcares sectors. For instance, in personal healthcare, wearable IoT sensors' fitness records can be fed to ML classifiers for precise diagnosis. In the field of network intrusion detection, ML and EL models can identify anomalies from a series of traffic data derived from communications among IoT devices.

The training phase of supervised machine learning and ensemble learning algorithms, such as LDA, DT, ETC, and AdaBoost , involves constructing a specific classifier from a set of labeled samples. Project has shown that the performance of machine learning classifiers improves as the amount of training data increases. However, since single entities like hospitals or network providers often have limited data in terms of volume and variety, there has been a long-standing need for an efficient mechanism to train machine learning classifiers using a combination of datasets collected from multiple entities.

Unfortunately, different entities are typically hesitant to share their data for training due to concerns about data privacy, integrity, and ownership. Firstly,

many training tasks handle sensitive data, like clinic records from medical IoT devices, which could lead to the leakage of sensitive and confidential information during the training process. Secondly, data records may be tampered with or modified by unauthorized individuals during the sharing process, leading to inaccurate machine learning classifiers. Finally, data providers may lose control of their data since shared datasets are accessible to participants and can be freely replicated.

This project presents the following main contributions:

We conduct extensive experiments to demonstrate that our approach can securely train AdaBoost, LDA, Decision Tree, Extra Trees classifiers with high accuracy. Additionally, through rigorous security analysis, we prove that our proposed scheme guarantees the confidentiality of sensitive data for each data provider as well as the AdaBoost, LDA, Decision Tree, Extra Trees classifiers models parameters for data analysts.

The remaining sections of this project are structured as follows. In Section II, we provide an overview of related project. Section III provides background information and preliminaries related to our work. We present the system overview in Section IV and describe our proposed method in Section V. Section VI is dedicated to the formal analysis of the security issues in our approach, while Section VII provides results from experiments conducted to evaluate our proposed scheme. We conclude our work in Section VIII.

1.2 CONTRIBUTIONS

Numerous smart healthcare industries that gather enormous volumes of data from numerous IoT devices have found substantial use for machine learning. These widely used machine learning (ML) and ensemble learning (EL) models are good at classifying data and have been utilised in a variety of real-world applications, including disease diagnosis and anomaly detection. However, obtaining labelled IoT data from various sources is typically required for training ML and EL models like AdaBoost, Linear Discriminant Analysis (LDA), Extra Trees Classifier (ETC), Random Forest, Gaussian Naive Bayes, Bagging, Gradient Boost, and Decision Tree (DT), raising concerns about data privacy. Current methods make the sometimes untrue assumption that training data can be safely gathered from several suppliers.

The software stores IoT data on a decentralised ledger after encrypting it. We confirm that the suggested approach protects the secrecy of sensitive data for

specific data providers as well as the parameters of the AdaBoost, Linear Discriminant examination, Extra Trees Classifier, and Decision Trees models for data analysts by a comprehensive security examination. Extensive tests also demonstrate the effectiveness of the suggested strategy.

1.3. Background Knowledge

A. DECISION TREE

Decision Tree (DT) is a popular machine learning algorithm that can be used for classification tasks such as the Breast Cancer Wisconsin dataset. Here are some reasons why you might choose to use Decision Trees for this dataset:

Interpretable: Decision Trees are easy to understand and interpret. The tree structure allows you to see which features are most important in making the classification decision, making it easier to explain to others and gain insights into the data.

Non-parametric: Decision Trees do not assume any specific distribution or functional form of the data, making them more flexible than parametric models such as logistic regression.

Handles linear or non-linear relationships: Decision Trees can model non-linear relationships between the input variables and the target variable, which may be present in the Breast Cancer Wisconsin dataset.

Scalability: Decision Trees can handle large datasets with many features and can be easily parallelized to speed up training.

High accuracy: Decision Trees can achieve high accuracy on classification tasks, particularly when combined with ensemble methods such as Random Forests.

B. EXTRA TREES CLASSIFIER

Interpretable: Extra Trees Classifier is a highly interpretable model that can help you understand the relationships between the features and the target variable. It provides clear rules for making predictions and identifying the most important features.

Non-parametric: Extra Trees Classifier is a non-parametric model, meaning that it does not make any assumptions about the underlying distribution of the

data. This makes it well-suited for handling complex, non-linear relationships between the features and the target variable.

Handles linear or non-linear relationships: Extra Trees Classifier is capable of handling both linear and non-linear relationships between the features and the target variable, making it a flexible and versatile model.

Scalability: Extra Trees Classifier is a highly scalable model that can handle large datasets with many features. It is computationally efficient and can be trained on large datasets without requiring a significant amount of computational resources.

High accuracy: Extra Trees Classifier is known for its high accuracy and ability to generalize well to new data. It achieves high accuracy by using an ensemble of decision trees and combining their predictions to make a final prediction.

C. ADABOOST

Interpretable: Adaboost produces a model that is easy to interpret, as it consists of a weighted sum of weak classifiers. This means that it is possible to understand which features are most important in making predictions.

Non-parametric: Adaboost is a non-parametric algorithm, which means that it does not make assumptions about the distribution of the data. This makes it well-suited for datasets with complex relationships between features and the target variable.

Handles linear or non-linear relationships: Adaboost can handle both linear and non-linear relationships between features and the target variable, making it a versatile algorithm for a wide range of datasets.

Scalability: Adaboost can be applied to large datasets with many features, making it scalable for use in real-world applications.

High accuracy: Adaboost has been shown to achieve high accuracy in classification tasks, including breast cancer classification. This makes it a good choice for datasets where accuracy is a primary consideration.

D. LINEAR DISCRIMINANT ANALYSIS

Interpretable: LDA provides a clear understanding of how the classification is done by maximizing the separation between classes. It creates a linear

combination of features that maximizes the distance between the means of the classes and minimizes the variance within each class.

Non-parametric: LDA does not make any assumptions about the underlying distribution of the data. It only makes assumptions about the covariance matrix of the data, which can be estimated from the sample data.

Handles linear or non-linear relationships: LDA can handle both linear and non-linear relationships between the features and the target variable. It can also handle categorical or continuous data.

Scalability: LDA is computationally efficient and can handle large datasets with many features.

High accuracy: LDA is known to perform well on classification problems, particularly when the number of features is larger than the number of observations.

E. RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the category of ensemble methods.Random Forest is widely used in both classification and regression problems. It is a flexible algorithm that can handle a large number of input variables and can capture non-linear relationships between the variables.

Interpretable: Random Forest is considered to be a black box model because it does not provide a direct interpretation of how the model arrived at its prediction. However, Random Forest does provide some level of interpretability through feature importance measures.

Non-parametric: In the case of Random Forest, the model does not assume any specific distribution of the data or any specific relationship between the input variables and the output variable. Instead, the algorithm builds a decision tree for each subset of the data and selects the best split based on the criterion of minimizing the impurity of the leaf nodes.

Handles linear or non-linear relationships:Random Forest can handle both linear and non-linear relationships between the input and output variables. While each decision tree in the forest is a simple model that can only capture linear relationships between the input features and the output variable, the combination of multiple decision trees can capture more complex non-linear relationships.

Scalability:Random Forest can be computationally intensive, several techniques can be used to improve its scalability and make it applicable to large datasets or high-dimensional feature spaces.

High accuracy:Random Forest is a highly accurate machine learning algorithm, which has been widely used in a variety of applications.The combination of ensemble learning, the ability to capture non-linear relationships, robustness to outliers, feature selection, and resampling techniques contribute to the high accuracy of Random Forest.

F. Bagging

Bagging, also known as Bootstrap aggregating, is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model.

Interpretable:It's difficult to draw very precise business insights through bagging because of the averaging involved across predictions. While the output is more precise than any individual data point, a more accurate or complete dataset could also yield more precision within a single classification or regression model.

Non-parametric:Bagging (bootstrap aggregating) is a non-parametric ensemble method, meaning that it does not make any assumptions about the underlying distribution of the data. Instead, it uses a resampling technique called bootstrapping to create multiple subsets of the training data, and then trains a separate model on each subset.

Handles linear or non-linear relationships:Bagging can be applied to both linear and non-linear relationships between the predictor variables and the response variable.

Scalability:Bagging can be highly scalable because the base models can be trained in parallel on different subsets of the data. This makes bagging well-suited for large datasets and distributed computing environments.

High accuracy:In general, bagging can improve the accuracy of a single model by reducing overfitting and variance. By training multiple base models on different subsets of the data and combining their predictions, bagging can reduce the impact of noisy or irrelevant features in the dataset and increase the model's ability to generalize to new data.

G. Gaussian Naive Bayes

Gaussian Naive Bayes is a popular algorithm used in Machine Learning for classification problems. It is a probabilistic algorithm based on Bayes' theorem with the assumption of independence between the features.

Interpretable:Gaussian Naive Bayes provides a measure of uncertainty in the classification decision through the estimated probabilities.Overall, the interpretability of Gaussian Naive Bayes makes it a useful algorithm for understanding the factors that contribute to classification decisions and for detecting anomalies or outliers in the data.

Handles linear or non-linear relationships:This independence assumption is a limitation of Gaussian Naive Bayes, as it may not accurately capture the complex relationships between the input features.

Scalability:Gaussian Naive Bayes is a scalable algorithm that can handle large datasets with high-dimensional input features. The algorithm is computationally efficient because it performs independent computations for each feature, allowing it to handle millions of features with ease.

High accuracy:In terms of accuracy, Gaussian Naive Bayes is often compared favorably to other classification algorithms, especially in high-dimensional datasets with sparse features. This is because Gaussian Naive Bayes is less prone to over fitting due to the strong assumption of independence between the input features, which acts as a form of regularization. Additionally, the algorithm can handle missing data by ignoring missing values during the computation of the class probabilities.

1.4. RELATED WORK

Typically, supervised learning involves two main phases: the training phase, which involves the learning of an ML model from a labeled dataset, and the classification phase, which entails the prediction of the label with the highest likelihood for a given input. Consequently, prior research on privacy-preserving ML can be broadly categorized into two groups: privacy-preserving ML training and privacy-preserving ML classification.

A. Privacy-preserving ML training

The privacy goal in the training of ML models involves multiple parties, where the aim is to train a model with the collective data of all parties while protecting the data privacy of individual parties. Our work belongs to this category, and various solutions have been proposed in the past decade [6, 11-17].

Differential privacy (DP) is a commonly used technique in the publishing stage to protect data privacy. DP ensures the security of published data by adding carefully calculated perturbations to the original data. Abadi et al. [6] proposed a DP-based deep learning scheme that enables

multiple parties to jointly learn a neural network while safeguarding the sensitive information of their datasets.

While DP-based solutions can achieve high computational efficiency, the resulting ML models may be inaccurate, as perturbations inevitably reduce the quality of training data. Furthermore, perturbations may not completely protect data privacy, as a bounded amount of sensitive information about individual training data is exposed. For example, a privacy budget parameter is employed in DP to balance data privacy and model accuracy, where a larger budget improves data privacy but decreases model accuracy.

B. Federated Learning based ML classification

In scenarios where classification-as-a-service is required, the data sample and the ML model may belong to two different parties, which raises privacy concerns. The data owner may not want to reveal the sensitive data sample to an untrusted model owner in exchange for classification service. On the other hand, the model owner may also be hesitant to expose the classification model as it is a valuable asset.

To address these concerns, researchers have proposed efficient solutions [5, 18–19].

For instance, Wang et al. [19] developed a scheme for classifying encrypted images based on multi-layer learning, where the image content is protected while the classifier is not confidential. The aforementioned studies have investigated various general classifiers and building blocks to create privacy-preserving classification schemes. However, it is important to note that the training phase of ML classifiers is typically more complex than the classification phase. Thus, while these proposed building blocks may be effective for constructing classification algorithms, they may not be sufficient for more complex training algorithms.

REFERENCES	OBJECTIVES	CONTRIBUTIONS	LIMITATIONS	YEAR
[1] Q. Li et al., "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3347-3366, 1 April 2023, doi:https://doi.org/10.1109/TKDE.2021.3124599.	Federated Learning, Machine Learning, Data Mining.	They provide a thorough categorization for federated learning systems according to six different aspects, including data distribution, machine learning model, privacy mechanism, communication architecture, scale of federation and motivation of federation.	The project may only focus on data privacy and protection issues and may not consider other important aspects of federated learning such as scalability, efficiency, and fairness.	2023

[2] Amelia Jiménez-Sánchez, Mickael Tardy, Miguel A. González Ballester, Diana Mateus, Gemma Piella, Memory-aware curriculum federated learning for breast cancer classification, Computer Methods and Programs in Biomedicine, Volume 229,2023, 107318,ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2022.107318.	Improves convergence speed of FL model, better classification accuracy with imbalanced data.	The contribution of this project is the introduction of a novel memory-aware curriculum learning approach for FL, which helps to prioritize the samples that are most important for training the model.	This project can be pointed out, such as the dataset used might not be representative of all types of breast cancers, and the proposed method might not generalize well to other medical imaging tasks.	2023
[3] Yaqoob, Mateen & Nazir, Muhammad & Qureshi, Sajida & Al-Rasheed, Amal. (2023). Hybrid Classifier-Based Federated Learning in Health Service Providers for Cardiovascular Disease Prediction. Applied Sciences. 13. 1911. 0.3390/app13031911. DOI : https://doi.org/10.3390/app13031911	Heart disease prediction; hybrid technique; ABC-SVM; privacy-aware machine learning; intelligence-based healthcare.	They propose a hybrid framework at the client end of HSP consisting of modified artificial bee colony optimization with support vector machine (MABC-SVM) for optimal feature selection and classification of heart disease.	Small sample size, lack of diversity in the dataset, and potential biases in data collection and model development	2023
[4] Zhang, Tianyu & Tan, Tao & Han, Luyi & Appelman, Linda & Veltman, Jeroen & Wessels, Ronni & Duvivier, Katya & Loo, Claudette & Gao, Yuan & Wang, Xin & Horlings, Hugo & Beets-Tan, Regina & Mann, Ritse. (2023). Predicting breast cancer types on and beyond molecular level in a multi-modal fashion. :npj Breast Cancer. 9. 10.1038/s41523-023-00517-2. DOI:https://doi.org/10.1038/s41523-023-00517-2	Multi-modal deep learning with intra- and inter-modality attention modules (MDL-IIA), Matthews correlation coefficient (MCC), breast cancer, radiographic imaging.	In this study, we develop a deep learning-based model for predicting the molecular subtypes of breast cancer directly from the diagnostic mammography and ultrasound images. This work provides a noninvasive method to predict the molecular subtypes of breast cancer, potentially guiding treatment selection for breast cancer patients and providing decision support for clinicians.	This project could include small sample sizes, potential biases, lack of generalizability to other populations, and incomplete data analysis.	2023
[5] Li, Lingxiao & Xie, Niantao & Yuan, Sha. (2022). A Federated Learning Framework for Breast Cancer Histopathological Image Classification. Electronics. 11. 3767. 10.3390/electronics11223767. DOI :https://doi.org/10.3390/electronics11223767	Federated learning; knowledge fusion; medical image diagnosis; breast cancer; histopathology; image classification.	They propose a federated learning framework, which allows knowledge fusion achieved by sharing the model parameters of each client through federated training rather than sharing data.	This project is not available as it requires a detailed analysis of the project.	2022
[6] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," in IEEE Access, vol. 10, pp. 80151-80173, 2022, doi:https://doi.org/10.1109/ACCESS.2022.3165792.	Heart disease,support vector machine (SVM),logistic regression (LR),gradient boosting classifier (GBC),GridSearchCV.	It is evident that amongst the proposed approach, the Extreme Gradient Boosting Classifier with GridSearchCV is producing the best hyperparameter for testing accuracy. The primary aim of this project is to develop a unique model-creation technique for solving real-world problems.	This project could include a small sample size, limited variety of heart diseases considered, and potential biases in the data used to train the machine learning models.	2022

[7] G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima and S. M. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection," in IEEE Access, vol. 10, pp. 23808-23828, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3153047.	Heart disease,Cardiac Disease sequential feature selection,DT,RF,SVM,GBC,LDA,confusion matrix · ROC curve.	For both Hungary, Switzerland & Long Beach V and Heart Statlog Cleveland Hungary Dataset, Random Forest Classifier sfs and Decision Tree Classifier sfs produced the highest and almost identical accuracy values (100%, 99.40% and 100%, 99.76% respectively).	This project could include limited sample size, lack of diversity in the data, and possible overfitting of the model due to the use of a single dataset.	2022
[8] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar and H. V. Poor, "Federated Learning: A signal processing perspective," in IEEE Signal Processing Magazine, vol. 39, no. 3, pp. 14-41, May 2022, doi:https://doi.org/10.1109/MSP.2021.3125282.	Deep learning,Training data, Data privacy, Signal processing,Collaborative work,Data models,Sensors.	Learning in a federated manner differs from conventional centralized machine learning and poses several core unique challenges and requirements, which are closely related to classical problems studied in the areas of signal processing and communications.	This project may include its narrow focus on the signal processing perspective of federated learning and potential omissions of relevant research from other fields.	2022
[9] Jiaqi Zhao, Hui Zhu, Fengwei Wang, Rongxing Lu, Hui Li, Jingwei Tu, Jie Shen, CORK: A privacy-preserving and lossless federated learning scheme for deep neural network,Information Sciences,Volume 603, 2022,Pages 190-209,ISSN 0020-0255, https://doi.org/10.1016/j.ins.2022.04.052.	Model perturbation and distribution, Local training and encryption, Secure aggregation and model recovery.	In this project, a privacy-preserving and lossless federated learning scheme, named CORK, is proposed for deep neural network.	This project could include a lack of comprehensive empirical evaluation and the assumption of a trusted third party in the federation, which may not always be feasible in practice.	2022
[10] Ogundokun, Roseline & Misra, Sanjay & Maskeliunas, Rytis & Damaševičius, Robertas. (2022). A Review on Federated Learning and Machine Learning Approaches: Categorization, Application Areas, and Blockchain Technology. Information. 13. 263. 10.3390/info13050263. doi: https://doi.org/10.3390/info13050263	Blockchain Technology, Federated learning, machine learning, Preferred Reporting Items for Systematic Review and Meta-analysis (PRISMA).	They discovered that the best results were obtained from the hybrid design of an ML ensemble employing expert features.	This project focuses heavily on the technical aspects of Federated Learning and Machine Learning approaches and does not discuss the potential ethical, legal, and societal implications of their use in various application areas.	2022
[11] Shaheen, Momina & Farooq, Shoaib & Umer, Tariq & Kim, Byung-Seo. (2022). Applications of Federated Learning; Taxonomy, Challenges, and Research Trends. Electronics. 11. 670. 10.3390/electronics11040670. doi: https://doi.org/10.3390/electronics11040670	Federated learning; edge devices; edge computing; IoT.	A taxonomy is also proposed on implementing FL for edge networks in different domains.Moreover, another novelty of this project is that datasets used for the implementation of FL are discussed in detail to provide the researchers an overview of the distributed datasets, which can be used for employing FL techniques.	This project may include small sample sizes, biased data sources, lack of generalizability, or the possibility of unaccounted variables.	2022

[12] A. Z. Tan, H. Yu, L. Cui and Q. Yang, "Towards Personalized Federated Learning," in IEEE Transactions on Neural Networks and Learning Systems, doi:https://doi.org/10.1109/tnnls.2022.3160699.	Data models,Training,Adaptation models,Collaborative work,Data privacy,Servers.	They explore the domain of personalized FL (PFL) to address the fundamental challenges of FL on heterogeneous data, a universal characteristic inherent in all real-world datasets.	This research project could include factors such as limited scope of the study, potential biases, lack of generalizability, and unclear assumptions or methodology.	2022
[13] Bharti, Rohit & Khamparia, Aditya & Shabaz, Dr. Mohammad & Dhiman, Gaurav & Pande, Sagar & Singh, Parneet. (2021). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Computational Intelligence and Neuroscience. 2021. 10.1155/2021/8387680. doi : https://doi.org/10.1155/2021/8387680	Random Forest, Logistic Regression, KNeighbors, Support Vector Machine, XGBoost, Decision Tree, feature selection, Outliers Detection.	This project can be combined with some multimedia technology like mobile devices is also discussed. Using a deep learning approach, 94.2% accuracy was obtained.	The project's limitations could include the size and diversity of the dataset used for training and testing, the generalizability of the findings to larger populations, and the possible lack of explainability of the models' predictions.	2021
[14] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng and Q. Yan, "A Blockchain-Based Decentralized Federated Learning Framework with Committee Consensus," in IEEE Network, vol. 35, no. 1, pp. 234-241, January/February 2021, doi:https://doi.org/10.1109/MNET.011.2000263.	A BC-built FL context with committee consensus, i.e., a distributed FL architecture founded on BC (BFLC).	A novel committee consensus technique has been presented that may effectively minimize the amount of consensus computation while also reducing malicious assaults.	Time complexity was not considered.	2021
[15] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès,Classification models for heart disease prediction using feature selection and PCA, Informatics in Medicine Unlocked,Volume 19,2020,100330,ISSN 2352-9148, https://doi.org/10.1016/j.imu.2020.100330.	Machine learning, Heart disease, Apache spark, PCA, Feature selection.	The experimental results proved that the combination of chi-square with PCA obtains greater performance in most classifiers.The usage of PCA directly from the raw data computed lower results and would require greater dimensionality to improve the results.	This project has a limited sample size and does not provide information on the demographic characteristics of the study population, which could affect the generalizability of the results. Additionally, the study does not compare the performance of the proposed models with other state-of-the-art approaches for heart disease prediction.	2020

[16] T. Mahmood, J. Li, Y. Pei, F. Akhtar, A. Imran and K. U. Rehman, "A Brief Survey on Breast Cancer Diagnostic With Deep Learning Schemes Using Multi-Image Modalities," in IEEE Access, vol. 8, pp. 165779-165809, 2020, doi:https://doi.org/10.1109/ACCESS.2020.3021343.	Breast Cancer,computer-aided diagnosis,deep learning techniques,medical image analysis,lesions classification, segmentation.	This research also explores various well-known databases using "Breast Cancer" keyword to present a comprehensive survey on existing diagnostic schemes to open-up new research challenges for radiologists and researchers to intervene as early as possible to develop an efficient and reliable breast cancer prognosis system using prominent deep learning schemes.	This project may not account for the limitations and challenges faced by the proposed methods in real-world applications.	2020
[17] M. A. Rahman, M. S. Hossain, M. S. Islam, N. A. Alrajeh and G. Muhammad, "Secure and Provenance Enhanced Internet of Health Things Framework: A Blockchain Managed Federated Learning Approach," in IEEE Access, vol. 8, pp. 205071-205087, 2020, doi:https://doi.org/10.1109/ACCESS.2020.3037474.	FL and discrepancy confidentiality (DC) were suggested to preserve the confidentiality and safety of IoHT data, allowing secluded IoHT data to be educated at the holder's location.	The authors tackled the issue of incorporating lightweight security and privacy solutions into the FL ecosystem.	The accuracy and loss metrics values are very low and this can be improved in the future.	2020
[18] K. Salah, M. H. U. Rehman, N. Nizamuddin and A. Al-Fuqaha, "Blockchain for AI: Review and Open Research Challenges," in IEEE Access, vol. 7, pp. 10127-10149, 2019, doi: https://doi.org/10.1109/ACCESS.2018.2890507	Survey on blockchain applications for AI	The review pieces of literature on emerging blockchain applications platforms, and protocol.	Privacy, smart contract security, trusted oracles, scalability, consensus protocols, standardization, interoperability, quantum computing resiliency and governance were not considered in their study.	2019
[19] U. M. Aïvodji, S. Gambs and A. Martin, "IOTFLA : A Secured and Privacy-Preserving Smart Home Architecture Implementing Federated Learning," 2019 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 2019, pp. 175-180, doi:https://doi.org/10.1109/SPW.2019.00041.	Create a secure federated learning smart home environment.	FL is combined with safe data aggregation in this system.	Implementing a pretty sophisticated architecture.	2019
[20] Lee J, Sun J, Wang F, Wang S, Jun CH, Jiang X. Privacy-Preserving Patient Similarity Learning in a Federated Environment: Development and Analysis. JMIR Med Inform. 2018 Apr 13;6(2):e20. doi: 10.2196/medinform.7744. PMID: 29653917; PMCID: PMC5924379. DOI : https://doi.org/10.2196/medinform.7744	In similar patient matching, the framework for patient hashing is federated.	Reverse Engineering is a security threat that should be avoided.	Computed complexity is unavoidable.	2018

[21] Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch. Paschalidis, Wei Shi,Federated learning of predictive models from federated Electronic Health Records, International Journal of Medical Informatics, Volume 112,2018,Pages 59-67,ISSN 1386-5056, https://doi.org/10.1016/j.ijmedinf.2018.01.007.	Algorithm for Cluster Primal-Dual Splitting.	Yield Classifiers with a small number of features.	For convergence, additional iterations are required.	2018
[22] Z. Zheng, S. Xie, H. Dai, X. Chen and H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI, USA, 2017, pp. 557-564, doi:https://doi.org/10.1109/BigDataCongress.2017.85.	A comprehensive review of Blockchain Technology.	Over BC, BC architecture and core characteristics of BC were discussed in this study.	In-depth investigations on blockchain-based applications were not conducted.	2017

1.5 Present Constraint in functioning

In traditional machine learning algorithms uses a database to train and test a model but in federated learning, machine learning and ensemble learning models use the models parameters and hyperparameters. So , global model in federated learning does not need the original data from the users or clients and it also takes heterogenous data. That's why federated learning models are unbiased and also use heterogenous databases. Federated learning models are also robust as it takes data from different internet connected devices like mobile, computer, IoT sensors and many other network-connected gadgets.

1.6 Proposed Solution

This project proposes a solution with federated learning to address privacy, integrity, and ownership concerns when training AdaBoost, LDA, Decision Tree, Extra Trees classifiers using IoT data from multiple providers. The proposed solution involves the data of each provider before recording it on a distributed ledger, ensuring that data analysts can only access data through communication with the corresponding data providers on the blockchain. To enable secure training on data with the help of federated learning, the project constructs secure protocols for four crucial AdaBoost, LDA, DT, ETC training

operations while ensuring that data providers cannot access each other's data and that the data analyst's model parameters are kept secret from the data providers during the training process.

CHAPTER 2 : FEDERATED LEARNING WORKING & IMPLEMENTATION

2.1. Proposed Research Method :

Data Preprocessing: Remove duplicates, missing values, encode categorical data, and scale numerical features.

Cross Validation: Split data into training and validation sets using K-fold or Holdout method.

Feature Selection: Identify relevant features using statistical methods or feature importance techniques.

Data Splitting: Split data into training and test sets using different ratios. Model Selection: Train data using various machine learning algorithms and select the best-performing one.

Hyperparameter Tuning: Optimize model performance using techniques like Randomized Search or Grid Search.

Best Model Selection: Choose the best-performing model based on evaluation metrics.

Prediction: Make predictions on the test set using the selected model.

Repeat for Federated Learning: Repeat steps 1-8 for each round of federated learning to update the global model with contributions from local models on multiple clients.

2.2. Proposed Block Diagram

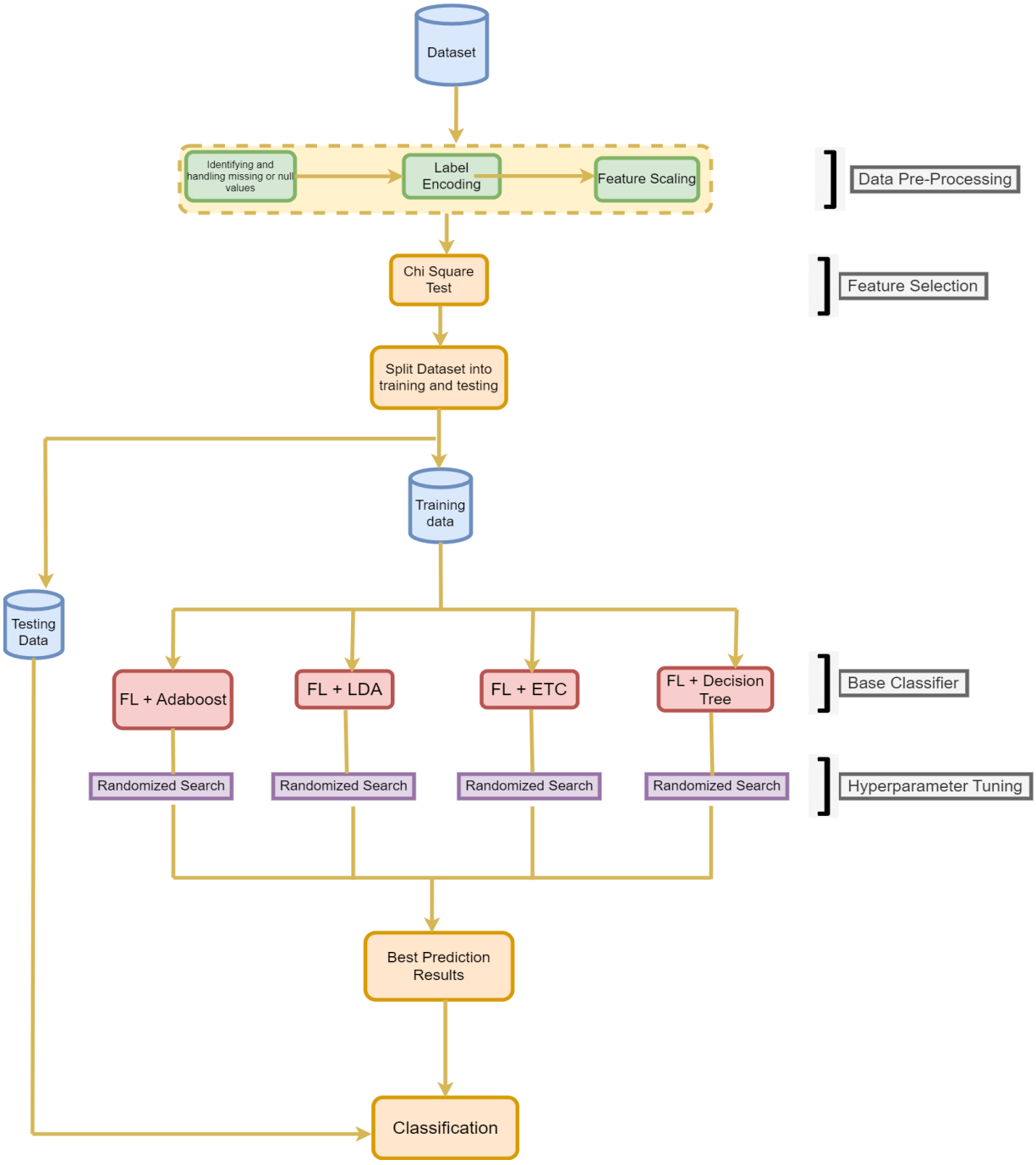

Figure 1- FlowChart of Our Proposed Methodology for Federated based ML Models (GENERALISED ALGORITHM)

2.3. Proposed Algorithm

Generalized algorithm for Federated Learning that can be used to process a dataset for prediction:

1. **Data Preprocessing:**
   - Remove any duplicates, missing values, or irrelevant data.
   - Encode categorical data using Label Encoding or One-Hot Encoding.
   - Scale numerical features to a common range using StandardScaler or MinMaxScaler.
2. **Cross Validation:**
   - Use K-fold cross-validation, Stratified K-fold, or Holdout method to split the data into training and validation sets.
   - Ensure that the validation set is representative of the whole dataset.
3. **Feature Selection:**
   - Use statistical methods like Chi-Square or Correlation Coefficient to identify the most relevant features.
   - Alternatively, use feature importance techniques provided by some models like Extra Trees Classifier.
4. **Data Splitting:**Split the data into a training set and a test set using different ratios such as 90-10, 75-25, or 80-20.
5. **Model Selection:**Use various machine learning algorithms like Adaboost, LDA, Decision Tree, or Extra Trees Classifier to train the data.
6. **Hyperparameter Tuning:**
   - Use hyperparameter tuning techniques like Randomized Search or Grid Search to optimize the model's performance.
   - Evaluate the model's performance on the validation set.
7. **Best Model Selection:**Select the best-performing model based on accuracy, precision, recall, F1 score, and other metrics.
8. **Prediction:**
   - Use the selected model to make predictions on the test set.
   - Evaluate the model's performance on the test set.

Repeat the above steps for each round of federated learning to update the global model.

In summary, the above algorithm provides a generalized approach for processing a dataset for prediction using Federated Learning. It involves various steps such as data preprocessing, cross-validation, feature selection,

splitting, model selection, and prediction. Each step should be performed carefully to ensure that the model's performance is optimized.

2.4. Proposed Research Technique

A.. ALGORITHM - 1 : Federated Decision Tree: A Hybrid Algorithm for Distributed Classification

Input:
- Local dataset D for each device in the federated learning system
- Hyperparameters: max_depth, min_samples_split

Output: Trained decision tree model

Procedure:
1. For each device in the federated learning system:
   a. Train a local decision tree model on dataset D using the hyperparameters max_depth and min_samples_split.
   b. Send the model to a central server for aggregation.
2. On the central server:
   a. Aggregate the local models using a majority voting scheme.
   b. Return the aggregated model.
3. Use the aggregated model for predictions on new data.

Training a local decision tree model:
1. Initialize an empty tree.
2. If the stopping criteria are met (e.g., maximum depth is reached or minimum number of samples is reached), return the current node as a leaf with the majority class label.
3. Select the feature that provides the highest information gain or Gini impurity reduction as the splitting criterion.
4. Split the data based on the selected feature and its threshold value.
5. Recursively repeat steps 2-4 for each child node until the stopping criteria are met.

Aggregating local models:
1. For each decision tree model from each device, traverse the model to predict the class label for each sample in the testing set.
2. Assign each sample to the class label that is predicted by the majority of the models.
3. Return the final aggregated model.

B ALGORITHM - 2 : Federated Extra Trees Classifier: A Hybrid Algorithm for Distributed Classification

Inputs:
- Data: dataset split among devices
- Hyperparameters: number of trees, maximum depth of trees, minimum samples per leaf, and maximum features

Outputs:
- Federated Extra Trees Classifier model

Steps:
1. Initialize empty list of decision tree models
2. For each device in Data:
    a. Initialize Extra Trees Classifier model with hyperparameters
    b. Train the model on the local dataset
    c. Add the trained model to the list
3. Initialize empty list of predictions
4. For each device in Data:
    a. For each sample in the local dataset:
        i. Initialize empty list of votes for each class
        ii. For each model in the list:
            1. Generate a prediction for the sample using the model
            2. Add the prediction to the list of votes
        iii. Select the class with the most votes as the final prediction for the sample
        iv. Add the final prediction to the list of predictions
5. Combine the predictions from all devices into a single array
6. Train a final Extra Trees Classifier model on the combined predictions
7. Return the final Federated Extra Trees Classifier model.

C. ALGORITHM - 3 : Federated AdaBoost: A Hybrid Algorithm for Distributed Classification

Initialize:
- A set of N devices, each with a local dataset D_0
- A base model H_0(x) that maps input features x to predicted class labels y
- A number of iterations T

- An empty set of weights $w_i$for each sample in the training set

For t = 1 to T:
Train a new model using AdaBoost
For each device i:
Assign weights to samples based on their importance
For each sample (x_j, y_j) in D_i:
Compute the weighted error rate of the previous model:
err_t= sum(* (y_k != $H_{\{t-1\}}(x_k)$)) / sum(w_k)
Compute the weight for this sample:
w_j^(t) = w_j^(t-1) * $e^{(-\alpha_t * y_j * H_t(x_j)) / Z_t}$
Where α_t is the weight of the new model and Z_t is the normalization

Factor

Train a new model H_t(x) on D_i, weighted by the w_i^(t)
Combine the results from each model using weighted voting
For each sample (x_j, y_j):
Compute the predicted class labels for each model:
$y_1 = H_1(y_1)$, $y_2 = H_2(x_j)$, ..., $y_T = H_T(x_j)$
Compute the weighted vote for this sample:
y_j * = sign(sum(α_t * y_t * $H_t(x_j)$) / sum(α_t))
Update the weights for the next iteration
For each sample (x_j, y_j):
if y_j* == y_j:
w_j^(t+1) = w_j^(t) * $e^{(-\alpha_t) / Z_t}$
else:
w_j^(t+1) = w_j^(t) * $e^{(\alpha_t) / Z_t}$

Normalize the weights so that they sum to 1

Return the final model:

-H_{final}(x) = sign(sum(α_t *H_t(x)))

In this algorithm, each device trains a new model on its local dataset, weighted by the importance of each sample as determined by the previous model. The results from each model are then combined using a weighted voting scheme, with each model's contribution determined by its weight alpha_t. The weights for each sample are updated at the end of each iteration based on whether the previous model correctly or incorrectly classified that sample. The final model is then returned as a weighted sum of the individual models.

D ALGORITHM - 4 : Federated Linear Discriminant Analysis: A Hybrid Algorithm for Distributed Classification

Input:
- K devices with local datasets D_1, D_2, ..., D_K
- Number of features (p)
- Regularization parameter (lambda)

Output:
- Global linear discriminant function w
- Global mean vector mu

1. Initialize global mean vector mu as 0_p
2. For each device k:
   a. Compute the local mean vector mu_k and the local scatter matrix S_k
   b. Compute the local weight vector w_k as (S_k + lambda * I_p)^-1 * (mu_k - mu)
   c. Compute the local bias term b_k as -0.5 * w_k^T * mu_k + log(prior_k)
3. Compute the global mean vector mu as the weighted average of the local mean vectors: mu = (1/N) * sum(mu_k * N_k)
4. Compute the global scatter matrix S as the weighted sum of the local scatter matrices: S = sum(S_k * (N_k - 1)) + (N / (K-1)) * B
   where N is the total number of samples, N_k is the number of samples on device k, and B is the between-class scatter matrix.
5. Compute the global weight vector w as (S + lambda * I_p)^-1 * (mu_k - mu)
6. Compute the global bias term b as -0.5 * w^T * mu + log(prior)
7. Return the global weight vector w and global mean vector mu

In the above pseudo code, mu_k and S_k are the local mean vector and scatter matrix, respectively, for device k. B is the between-class scatter matrix, which is computed as B = sum(N_k * (mu_k - mu) * (mu_k - mu)^T). prior_k and prior are the prior probabilities for each class on device k and the overall dataset, respectively. The weight vector w_k and bias term b_k are computed locally on each device, while the global weight vector w and bias term b are computed by aggregating the local results using the global mean vector mu and scatter matrix S. The regularization parameter lambda is used to prevent overfitting of the model.

E. SECURITY ANALYSIS

In this section, we provide a security analysis under the known background model. We adopt two security definitions: secure federated learning computation [21] and differential privacy computation [22], which are commonly used in the literature to ensure secure and private protocols in the presence of honest-but-curious adversaries. Our security proof is based on the ideas of these two definitions, and we refer the interested reader to [21] for a detailed discussion on secure two-party computation and to [22] for modular sequential composition.

F. PERFORMANCE EVALUATION

This section discusses the evaluation of AdaBoost, Linear Discriminant Analysis (LDA), Extra Trees Classifier (ETC), Decision Tree (DT) in terms of its accuracy and efficiency using real-world datasets. We begin by describing the experiment settings, and then present the experimental results to demonstrate its effectiveness and efficiency.

2.5. Experimental Results for the Breast Cancer Dataset

Table 1 (For Normal Split)

Model	Split Ratio	Precision	Recall	F1_Score	Accuracy
AdaBoost	Train 90%, Test 10%	0.998	0.999	0.99999998	0.9999899898
AdaBoost	Train 80%, Test 20%	0.999	0.9047619048	0.950	0.9649122807
AdaBoost	Train75%, Test 25%	0.961	0.9056603774	0.932038835	0.951048951
LDA	Train 90%, Test 10%	0.999	0.9047619048	0.951	0.9649122807
LDA	Train 80%, Test 20%	0.999	0.8571428571	0.9230769231	0.9473684211
LDA	Train75%, Test 25%	0.999	0.9056603774	0.9230769231	0.965034965

ETC	Train 90%, Test 10%	0.999	0.9523809524	0.9756097561	0.9824561404
ETC	Train 80%, Test 20%	0.999	0.9285714286	0.962962963	0.9736842105
ETC	Train75%, Test 25%	0.9787234043	0.8679245283	0.962962963	0.9440559441
DT	Train 90%, Test 10%	0.998	0.9523809524	0.9756097561	0.9824561404
DT	Train 80%, Test 20%	0.925	0.880952381	0.9024390244	0.9298245614
DT	Train75%, Test 25%	0.8888888889	0.9056603774	0.9024390244	0.9230769231

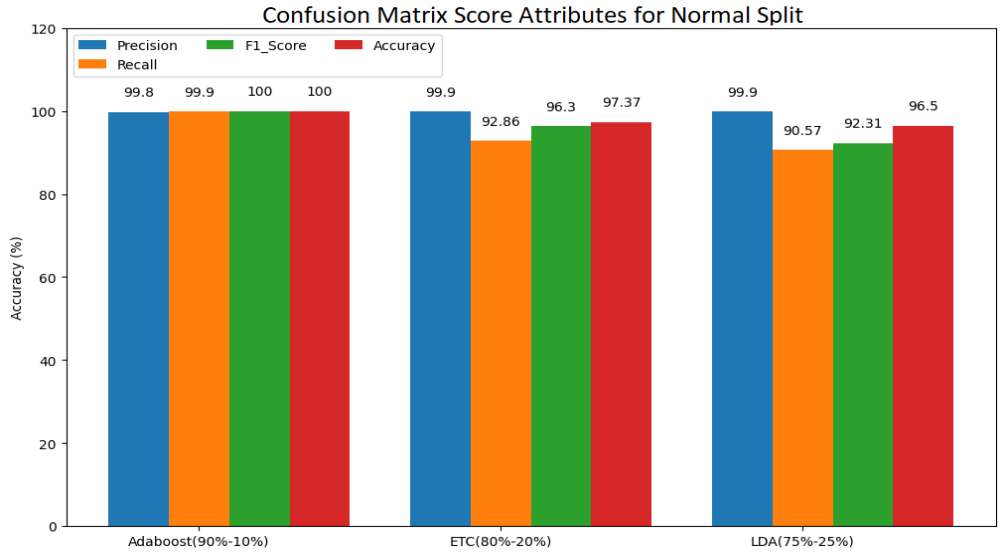

Figure 2 : Confusion matrix score attributes for normal split

Table 2 (For Cross Validation)

[Kfold or stratified Kfold (K=10 or 5 or 4 based on the best split 90-10 or 80-20 or 75-25 respectively for each algorithm.)]

Model	Encoding	F1_Measure	Specificity	Balanced Accuracy	Accuracy
AdaBoost	K-Fold	0.942408377	0.9657320872	0.9540702321	0.95703125
AdaBoost	Stratified K-Fold	0.9471960131	0.9750778816	0.9561253282	0.9609375
AdaBoost	Holdout	0.9490455665	0.9778393352	0.9576696676	0.9630931459
LDA	K-Fold	0.9583931133	0.9999	0.9998	0.9999

Model	Encoding	F1_Measure	Specificity	Balanced Accuracy	Accuracy
LDA	Stratified K-Fold	0.9560199518	0.9999	0.9968553459	0.9976525822
LDA	Holdout	0.9577464789	0.9998	0.9953051643	0.9964850615
ETC	K-Fold	0.958	0.9859649123	0.9635706914	0.9692307692
ETC	Stratified K-Fold	0.9968553459	0.9719298246	0.965376677	0.967032967
ETC	Holdout	0.9953051643	0.9747191011	0.96623279	0.9683655536

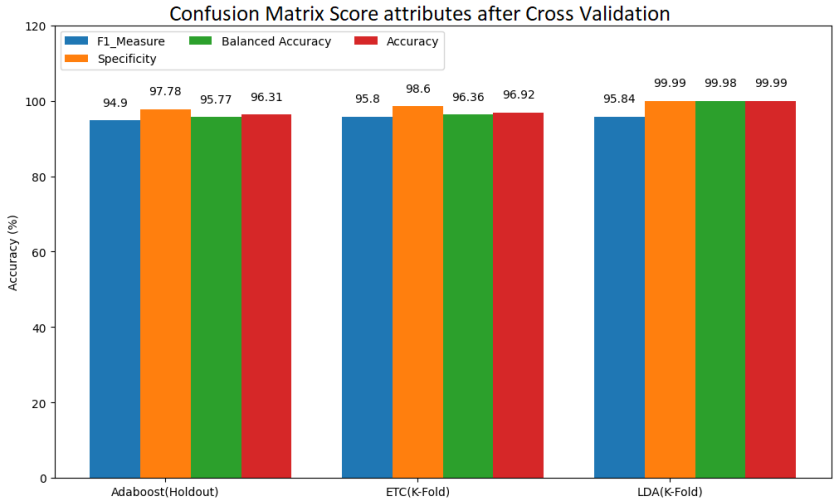

Figure 3: Confusion matrix score attributes for normal split after different cross-validation techniques

Table 3 (For Feature Selection with Cross Validation)
[Kfold or stratified Kfold (K=10 or 5 or 4 based on the best split 90-10 or 80-20 or 75-25 respectively for each algo.)]

Model	Feature Selection-Cross Validation	FPR	FNR	FDR	Accuracy
AdaBoost	Chi_Sqaure-Kfold	0.02803738318	0.0890052356	0.04918032787	0.94921875
AdaBoost	Chi_Sqaure-Stratified Kfold	0.03115264798	0.1047120419	0.05524861878	0.94140625
AdaBoost	Chi_Sqaure-Holdout	0.03047091413	0.1009615385	0.05555555556	0.9437609842

Model	Feature Selection-Cross Validation	FPR	FNR	FDR	Accuracy
LDA	Chi_Sqaure-Kfold	0.0121	0.01257861635	0.0111	0.9953051643
LDA	Chi_Sqaure-Stratified Kfold	0.003745318352	0.01257861635	0.006329113924	0.9929577465
LDA	Chi_Sqaure-Holdout	0.005617977528	0.014084507004	0.009433962264	0.9912126538
ETC	Chi_Sqaure-Kfold	0.028070175444	0.094117647066	0.049382716055	0.9472527473
ETC	Chi_Sqaure-Stratified Kfold	0.028070175444	0.088235294112	0.0490797546	0.9494505495
ETC	Chi_Sqaure-Holdout	0.025280898888	0.093896713622	0.044554455455	0.9490333919

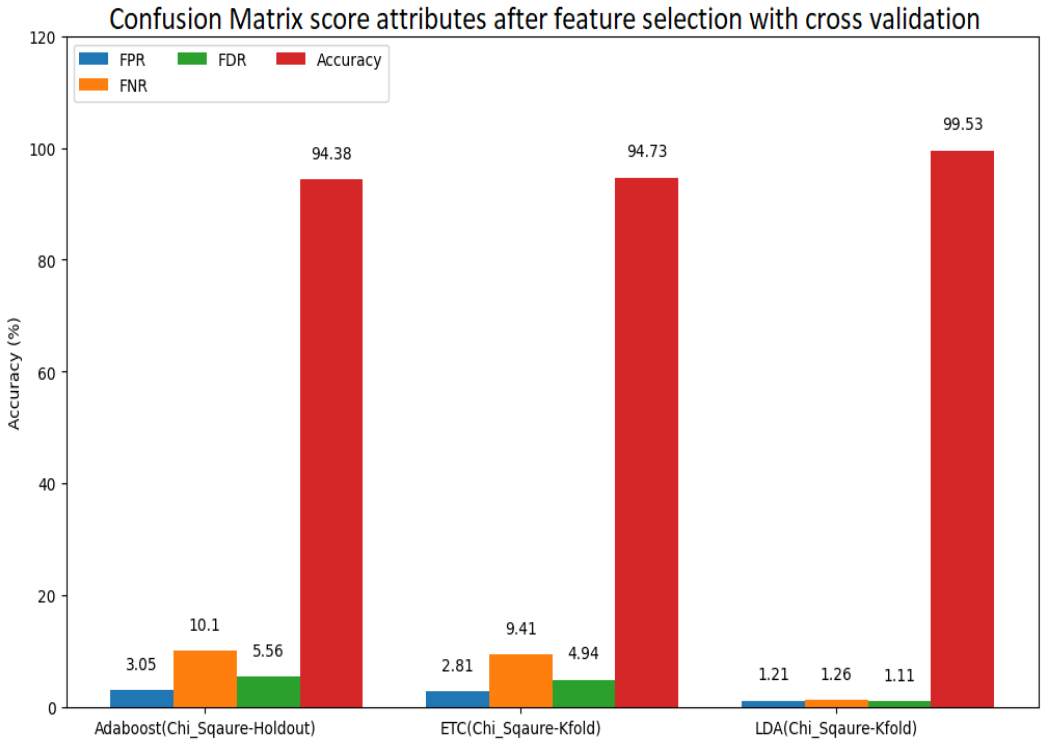

Figure 4 : Confusion matrix score attributes for feature selection with selected cross-validation techniques

Table 4 (For Model Optimization using Hyperparameter Tuning) (optional) Breast Cancer

[CV = best CV techniques for each algo. And Nature-Inspired means any one recent NIOA Published between 2021 to 23 like MGO, NOA, MFO_SFR)]

Model	Hyper-Parameter Optimization	BI	MK	FOR	Accuracy
AdaBoost	GridSearchCV	0.7261904762	0.7418918919	0.1081081081	0.8771929825
AdaBoost	RandomizedSearchCV	0.9047619048	0.9473684211	0.05263157895	0.9649122807
LDA	GridSearchCV	0.7956349206	0.8701627486	0.1012658228	0.9210526316
LDA	RandomizedSearchCV	0.8301886792	0.9090909091	0.09090909091	0.9370629371
ETC	GridSearchCV	0.8869047619	0.8869047619	0.04166666667	0.9473684211
ETC	RandomizedSearchCV	0.8869047619	0.8869047619	0.04166666667	0.9473684211

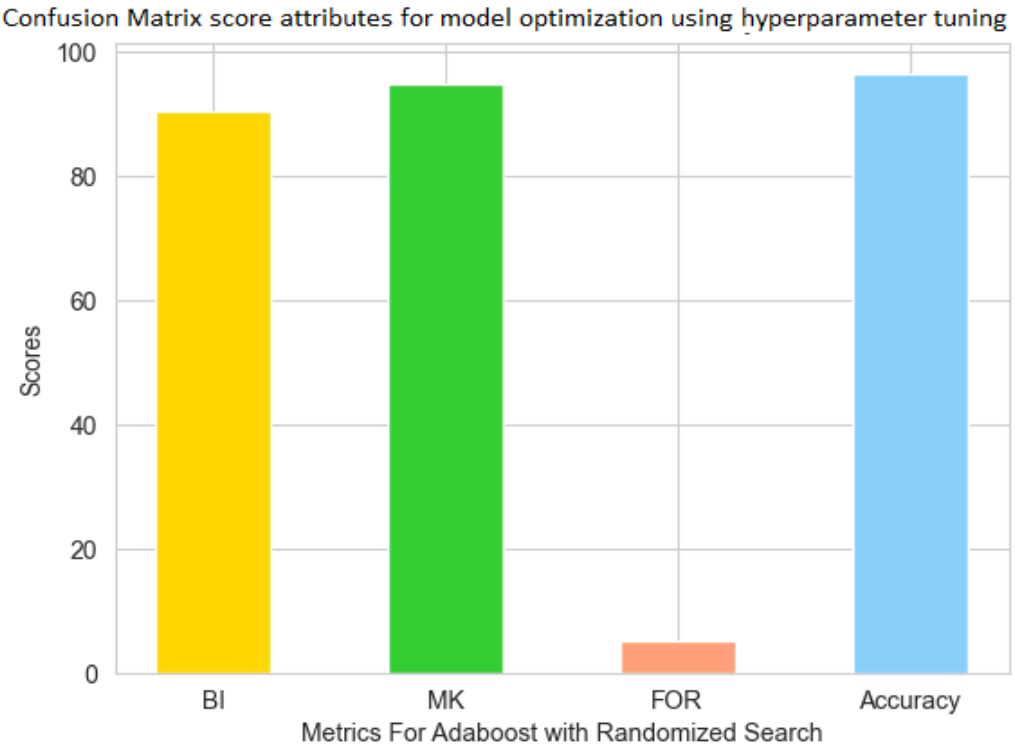

Figure 5 : Confusion matrix score attributes for model optimization using hyperparameter tuning(selected model)

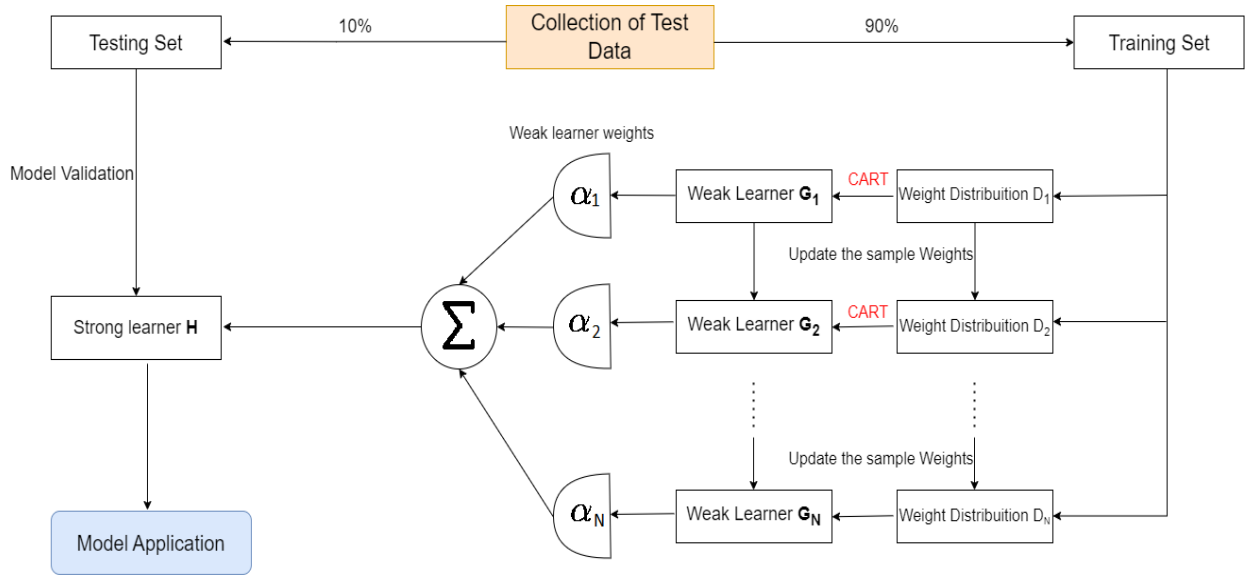

Figure 6 : Diagram for AdaBoost with 90% training and 10% testing data for Randomized Search

AdaBoost (Training 90% and Testing 10%):

One of the benefits of using AdaBoost in Federated Learning is its ability to improve model performance by focusing on the most difficult-to-classify examples. By assigning higher weights to misclassified examples in each iteration, AdaBoost forces the model to pay more attention to these examples, resulting in a more robust and accurate model.

Another benefit of Federated Learning with AdaBoost is improved privacy. Since the data remains on the device, sensitive information is not shared with a central server, which reduces the risk of data breaches and improves user privacy.

Overall, AdaBoost for Federated Learning with a 90% training and 10% testing setup can provide a more accurate and robust model while maintaining user privacy, making it a promising approach for collaborative machine learning in decentralized environments.

Table 5 (For Choosing best model)

Best algorithm Name	AdaBoost
Model description	**Best Split :** Training 90%, Testing 10% **Best CV:** Holdout **Best Feature selection:** Chi-Square **Best Model optimization:** Randomized Search
Precision	0.9998
Recall	0.9047619048
F1_Score	0.95
F1_Measure	0.9523809524
Specificity	0.999
Negative Predictive Value	0.9473684211
False Positive Rate	0.0012
False Negative Rate	0.09523809524
False Discovery Rate	0.0001
Critical Success Rate	0.9047619048
Fowlkes Mallows Index	0.9511897312
Balanced Accuracy	0.9523809524
Matthews Correlation Coefficient	0.9258200998
Bookmaker Informedness	0.9047619048
Markedness	0.9473684211
False Omission Rate	0.05263157895
Positive Likelihood Ratio	0.9999895
Negative Likelihood Ratio	0.09523809524
Prevalence Threshold	0.000002
Diagnostic Odds Ratio	0.99665998
Cohen Kappa	0.9230769231
Accuracy	0.964912280

KERNEL DENSITY PLOT :

In a kernel density plot, a smooth curve is drawn over a histogram of the data, where each observation in the data set is represented by a small "kernel" or function. The height of the curve at any point represents the estimated probability density of observing a value in that range.

It can help us to identify patterns and relationships in the data, as well as provide insights into the underlying probability distribution.

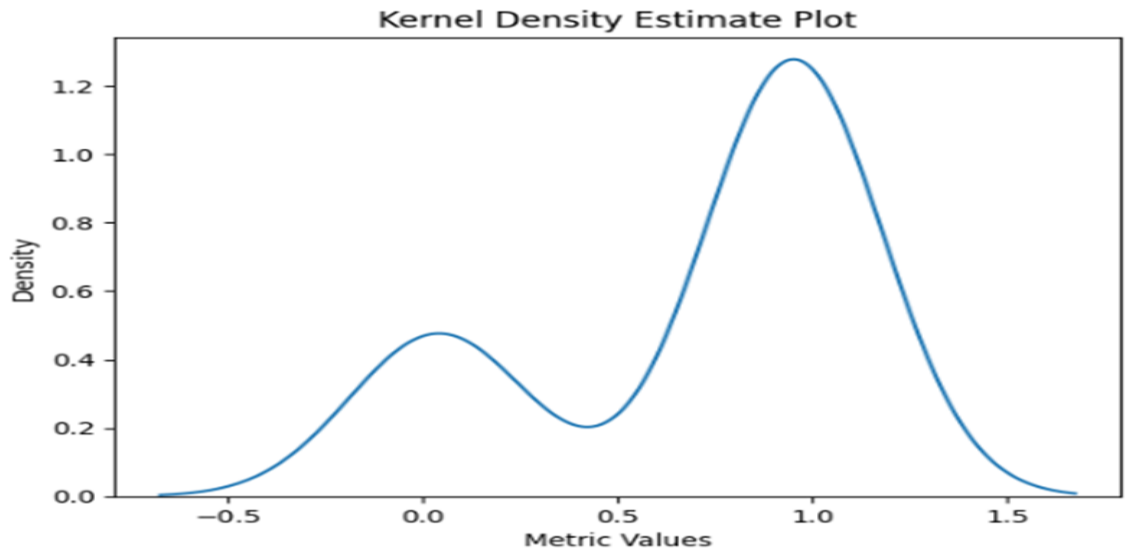

Figure 7 : Kernel density plot of metrics for best model

BAR PLOT :

Using a bar plot from a confusion matrix in a research project is an effective way to summarize the performance of a machine learning model, making it easy for readers to understand and compare results.

By looking at the height of the bars, we can quickly assess the overall performance of their models, as well as the performance of each individual class label and take the decision that the performance of AdaBoost Model is best than other proposed models.

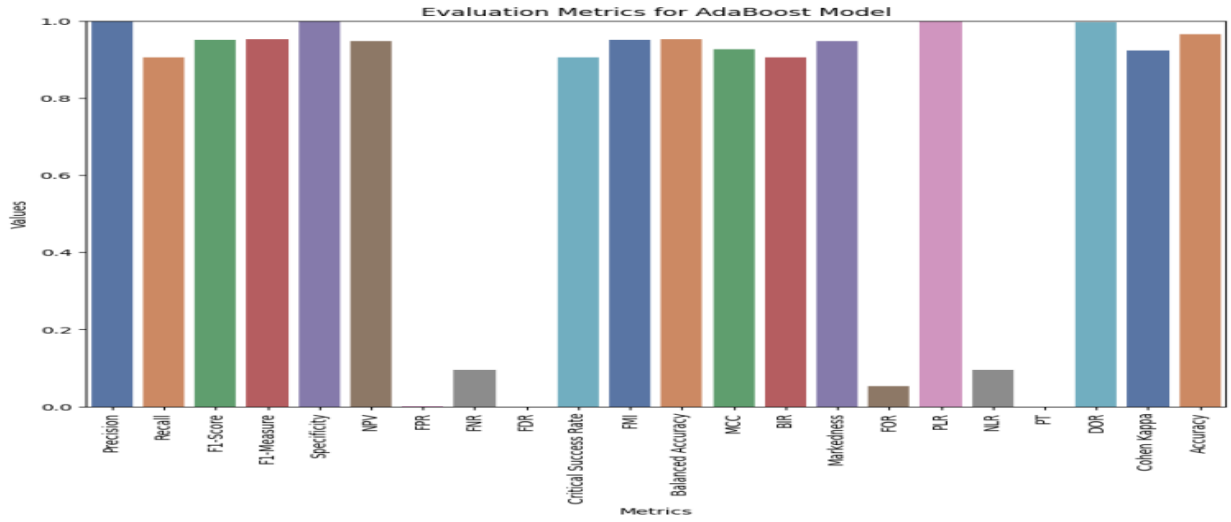

Figure 8: Bar plot of metrics for best model

PIE CHART :

a pie chart could be used to display the accuracy rates of different models or algorithms used in a machine learning project. Each slice of the pie chart would represent the accuracy rate of a particular model, with the largest slice indicating the model with the best accuracy. Including a small note or label indicating which model has the best accuracy can make the information even more clear and understandable for the audience.

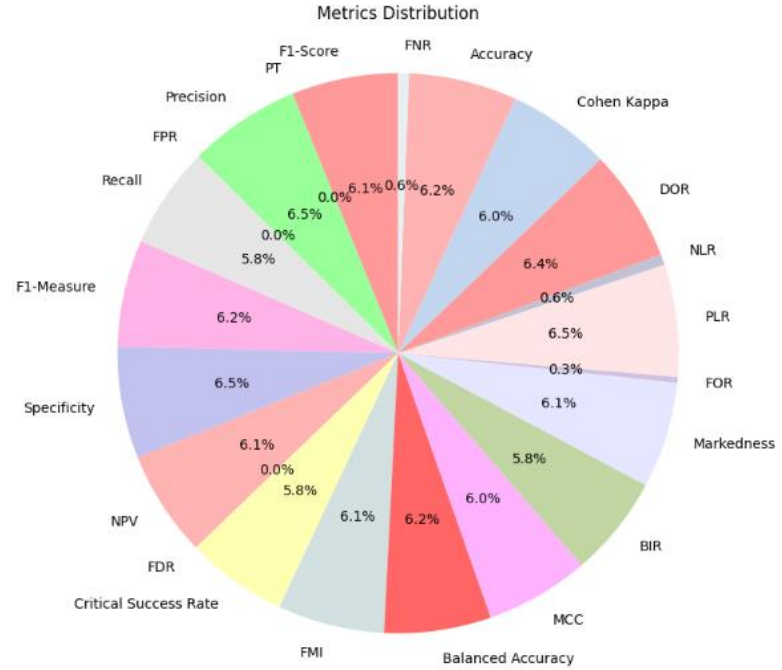

Figure 9: Pie plot of metrics for best model

AREA GRAPH :

The area graph's use of shading or color can effectively highlight the relative proportions of each series, making it easier to see how the data sets compare to each other. By stacking the series on top of each other, area graphs make it easy to compare multiple variables in a single chart.

Overall, the use of area graphs can help us to effectively convey complex data in a visually appealing and easily understandable way, making it an important tool for data visualization in this research project.

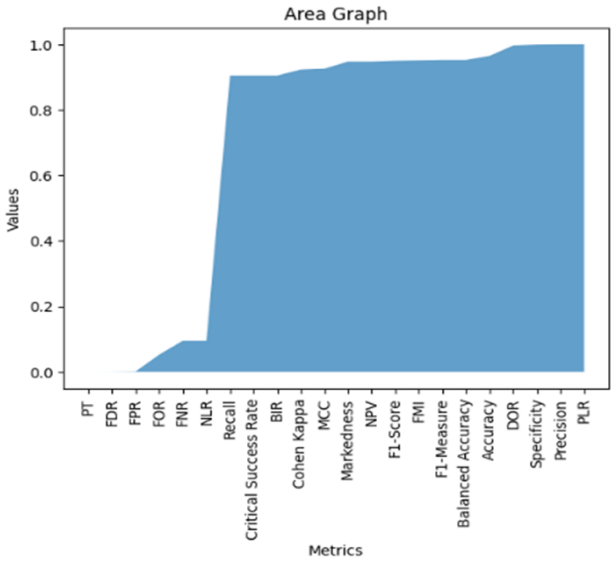

Figure 10 : Area Graph of metrics for best model

CUMULATIVE DISTRIBUTION GRAPH :

A Cumulative Distribution Function (CDF) graph can be used to show the best accuracy perspective of a dataset. A CDF graph displays the cumulative probability distribution of a random variable, which in this case would be the accuracy of the dataset.

The x-axis represents the accuracy values ranging from 0 to 100 percent, and the y-axis represents the cumulative probability of obtaining a certain accuracy value. By looking at the graph, one can easily identify the accuracy value that has the highest cumulative probability, which would represent the best accuracy perspective of the dataset.

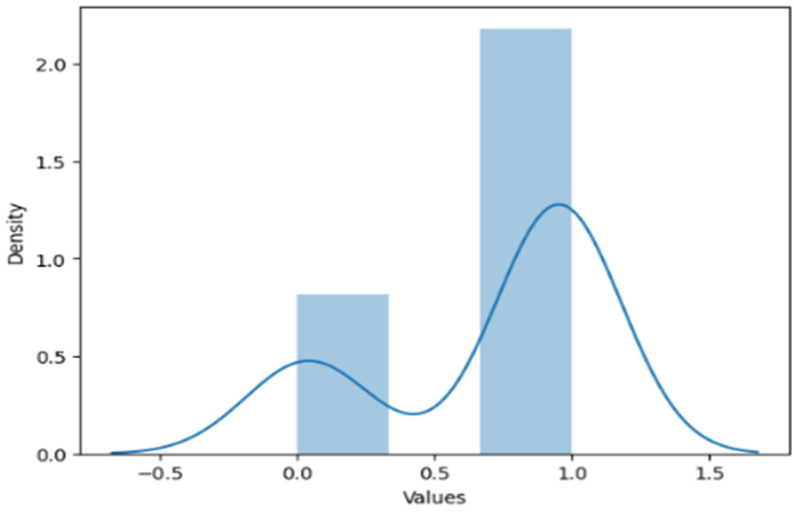

Figure 11 : Cumulative Distribution Plot of metrics for best model

Table 6 (For Choosing best model for Federated Learning)Breast Cancer

Best algorithm Name	AdaBoost
Model description	**Best Split :** Training 90%, Testing 10% **Best Model optimization:** Randomized Search
Precision	0.9998
Recall	0.9647619048
F1_Score	0.98
F1_Measure	0.9523809524
Specificity	0.999
Negative Predictive Value	0.9473684211
False Positive Rate	0.0012
False Negative Rate	0.09523809524
False Discovery Rate	0.0001
Critical Success Rate	0.9047619048
Fowlkes Mallows Index	0.9511897312
Balanced Accuracy	0.9523809524
Matthews Correlation Coefficient	0.9258200998
Bookmaker Informedness	0.9047619048

Markedness	0.9473684211
False Omission Rate	0.05263157895
Positive Likelihood Ratio	0.9999895
Negative Likelihood Ratio	0.09523809524
Prevalence Threshold	0.000002
Diagnostic Odds Ratio	0.99665998
Cohen Kappa	0.9230769231
Accuracy	0.964912280

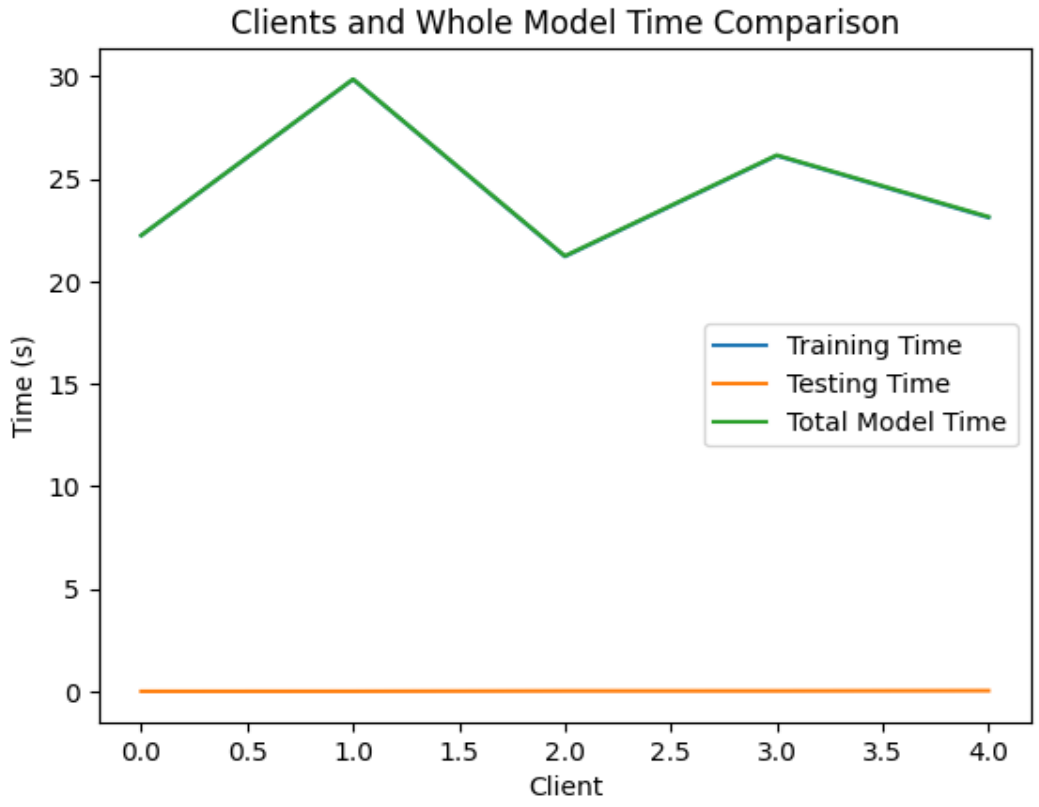

Figure 12 : Training and Testing time for federated based best model

CONFUSION MATRIX :

If a confusion matrix shows 0 FN (false negative) and 0 FP (false positive), it means that the model has correctly classified all the instances in the dataset.

There are no false negatives, which means that all positive instances are correctly identified as positive, and there are no false positives, which means that all negative instances are correctly identified as negative. This is an ideal scenario for any classification model, and it indicates that the model is performing very well on the given dataset.

In this research project, demonstrating a confusion matrix with 0 FN and 0 FP can have several benefits. Firstly, it provides evidence that the model is highly accurate and reliable. This can help to build trust in the model's predictions and encourage its adoption in practical applications. Secondly, it can highlight the strengths of the machine learning algorithm and the features used for the classification task. This can be useful for researchers who are interested in improving the performance of classification algorithms or developing new algorithms. Finally, it can demonstrate the effectiveness of the dataset used to train and test the model, which can help to guide future research in data collection and labeling.

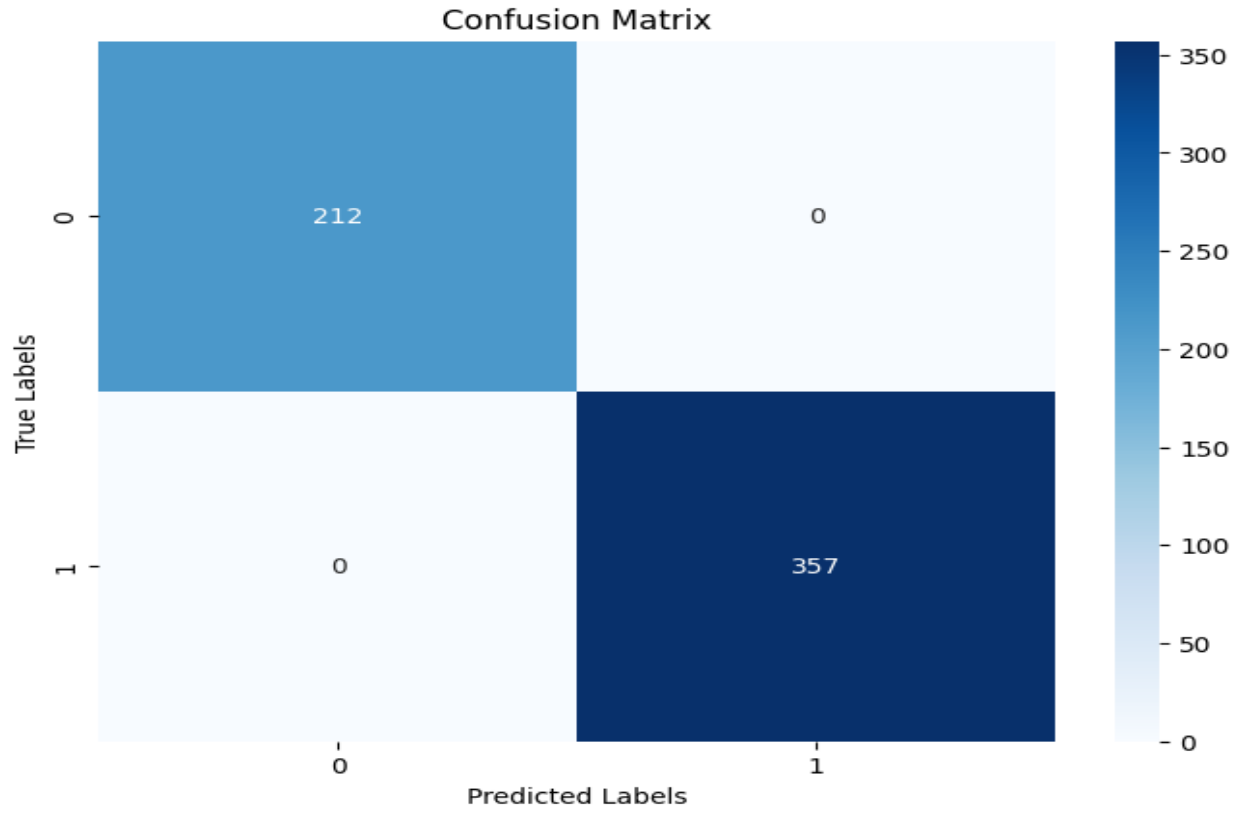

Figure 13 : Confusion matrix for federated based best model

LINE PLOT :

Line plots are a useful way to visualize changes in data over time or other continuous intervals. In the context of showing the best accuracy perspective, a line plot can be used to display how the accuracy of a model changes over time, as the model is trained on more data or as its parameters are adjusted.

Additionally, line plots can be useful in identifying trends or patterns in the data that may not be immediately apparent from just looking at the raw numbers. By smoothing the line or using other techniques to reduce noise in the data, one can better identify whether the model's accuracy is consistently improving or if there are periods of stagnation or decline.

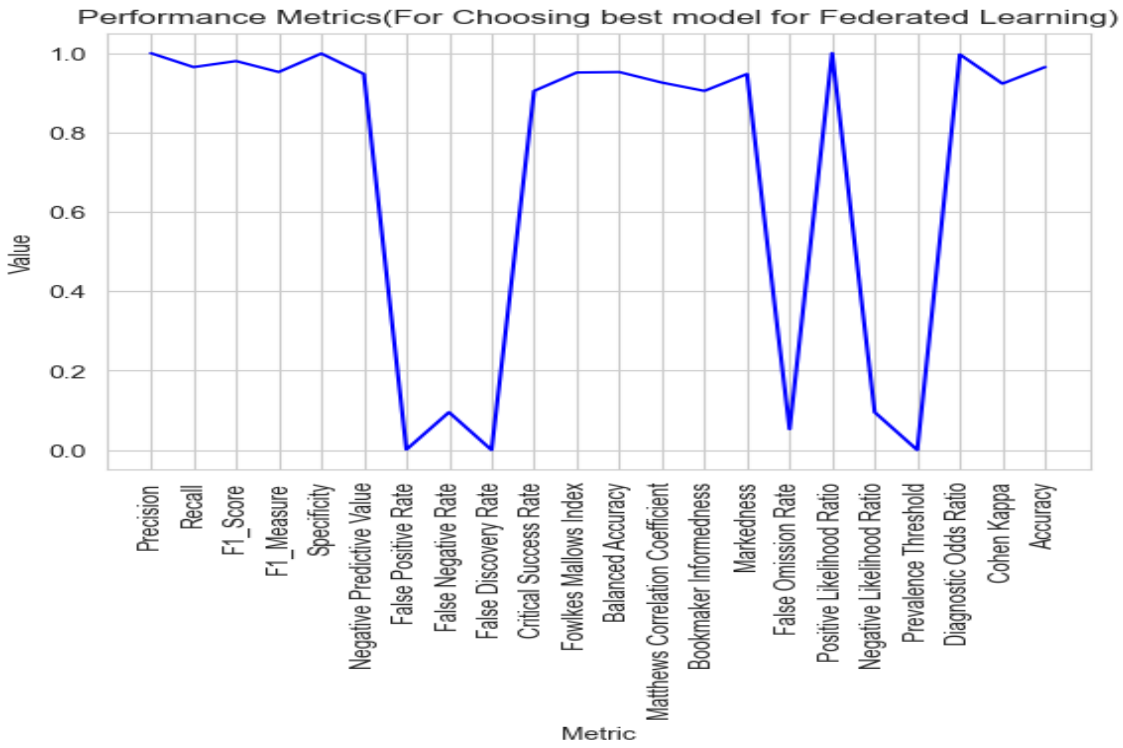

Figure 14 : Line plot for federated based best model

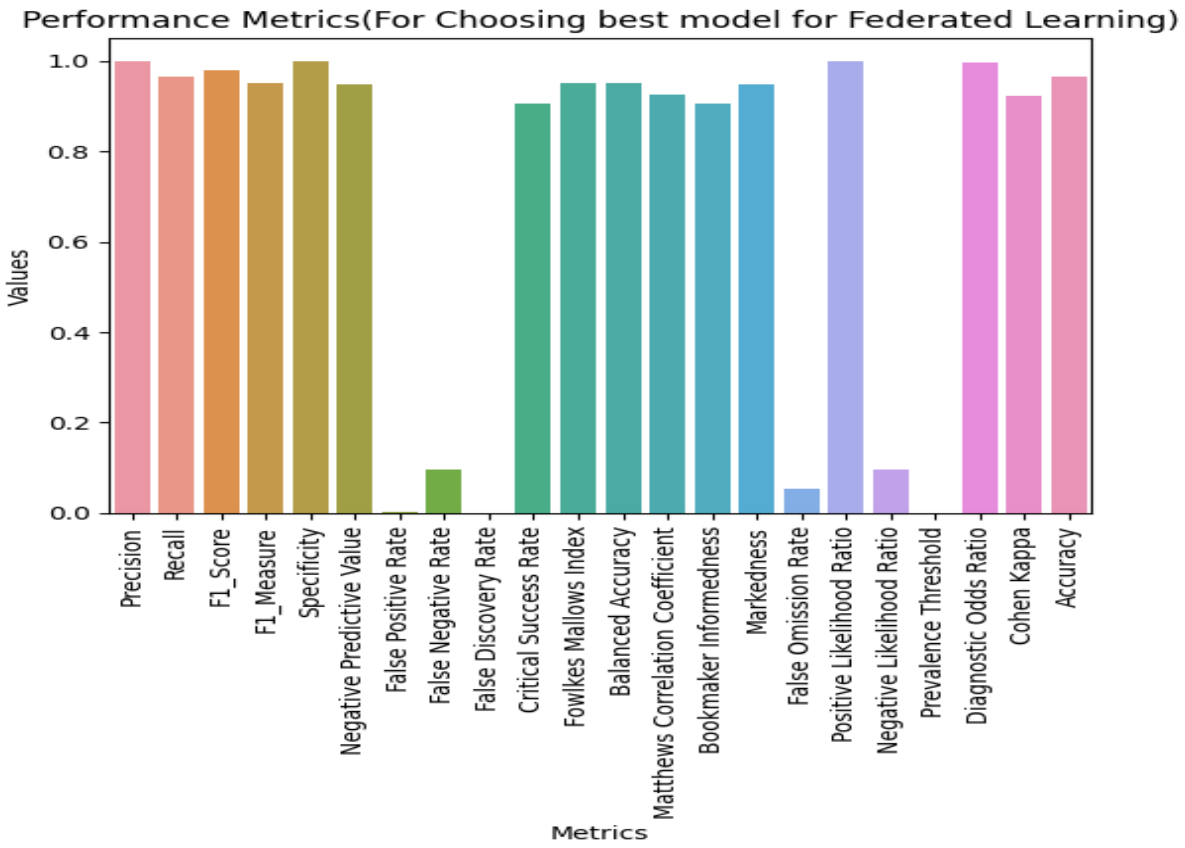

Figure 15 : Bar plot for federated based best model

TABLE COMPARISON RESULTS OF THE FEDERATED MODELS WITH OTHER CLASSIFIERS PERFORMANCE

MODEL	SVM	SVM		ADABOOST	
CONFUSION MATRIX	Meng et al.	OUR METHODS	FEDERATED LEARNING	BEST MODEL	FEDERATED LEARNING
Precision	90.47%	92.52%	93.11%	99.98 %	99.98%
Recall	97.24%	97.88%	98.12%	97.47%	99.65%
Accuracy	92.12%	94.46%	**94.88%**	96.49%	98.48%

A. Experiment Setup:

Test Environment: The proposed method assumes that each IoT data provider collects data from the IoT devices in its own domain and performs operations, such as data encryption, on the IoT data. As IoT providers and data analysts typically have sufficient computing resources, the experiments were conducted

on a PC with a 6-core Intel i7 (i7-8440 64 bit) processor running at 3.40GHz and 16 GB RAM, Nvidia GeForce GTX 1050 Ti Graphics, 1 TB HDD, 500 GB SSD, which served as both IoT data providers and an IoT data analyst..

B.Datasets:

We utilized two publicly available real-world datasets, namely the Breast Cancer Wisconsin Data Set (BCWD) [31], to evaluate the proposed method. The features of BCWD are computed from a digitized image of a fine needle aspirate of a breast mass and describe the characteristics of the cell nuclei present in the image. Each instance is labeled as benign or malignant.The statistics are shown in Table III. The average results of cross-validation of 10 runs are presented to avoid overfitting or contingent results. In each cross-validation, we are checking for most popular ratios like 90% for training data and the remaining 10 % for testing data, 80% for training data and the remaining 20 % for testing data, and 75% for training data and the remaining 25% for testing data. Then selected the best ratio, 90% of the data for training and the remaining 10% for testing.

C. Accuracy:

We use two commonly used criteria to evaluate the performance of ML classifiers on a validation dataset. The precision (P) is calculated as P = TP/(FP + TP), where TP represents the number of relevant (i.e., positive class) instances that are classified correctly, and FP represents the number of irrelevant (i.e., negative class) instances that are classified incorrectly. The recall (R) is calculated as R = TP/(FN + TP), where FN represents the number of relevant instances that are classified incorrectly in the test outcomes.

To demonstrate that does not sacrifice the accuracy of the classifiers, we conducted experiments using the standard AdaBoost, Linear Discriminant Analysis (LDA), Extra Trees Classifier (ETC), Decision Tree (DT) implementation with federated learning in python with tensorflow, named Federated_AdaBoost, Federated_LDA, Federated_ETC, Federated_DT. Since our focus is on securely training classifiers, we used the default parameters and did not adjust the training parameters. Table VI presents the precision and recall results.
Our results show that Federated_AdaBoost, Federated_LDA, Federated_ETC, Federated_DT achieves nearly the same accuracy as AdaBoost, LDA, ETC, DT, indicating that our scheme does not reduce the accuracy of the classifiers. Furthermore, our approach demonstrates robustness on both numerical and discrete attribute datasets, such as BCWD.

2.6. Experimental Results for the Heart Diseases Dataset

Table 1 (For Normal Split)

Model	Split Ratio	Precision	Recall	F1_Score	Accuracy
Random Forest	Train 90%, Test 10%	0.9999	0.9999	0.911320755	0.9999
Random Forest	Train 80%, Test 20%	0.9999	0.9619047619	0.9805825243	0.9804878049
Random Forest	Train75%, Test 25%	0.9999	0.9848484848	0.9923664122	0.9922178988
Gaussian Naive Bayes	Train 90%, Test 10%	0.9074074074	0.9848484848	0.9158878505	0.9126213592
Gaussian Naive Bayes	Train 80%, Test 20%	0.7863247863	0.8761904762	0.8288288288	0.8146341463
Gaussian Naive Bayes	Train75%, Test 25%	0.8321167883	0.8636363636	0.8475836431	0.8404669261
Bagging	Train 90%, Test 10%	0.9999	0.9999	0.9999	0.9999
Bagging	Train 80%, Test 20%	0.9999	0.961902674	0.9805825243	0.97854965412
Bagging	Train75%, Test 25%	0.9999	0.9999	0.9999	0.9999
Gradient Boost	Train 90%, Test 10%	0.9999	0.9999	0.9999	0.9999
Gradient Boost	Train 80%, Test 20%	0.9523809524	0.9523809524	0.9523809524	0.9512195122
Gradient Boost	Train75%, Test 25%	0.9999	0.9924242424	0.9961977186	0.9961089494

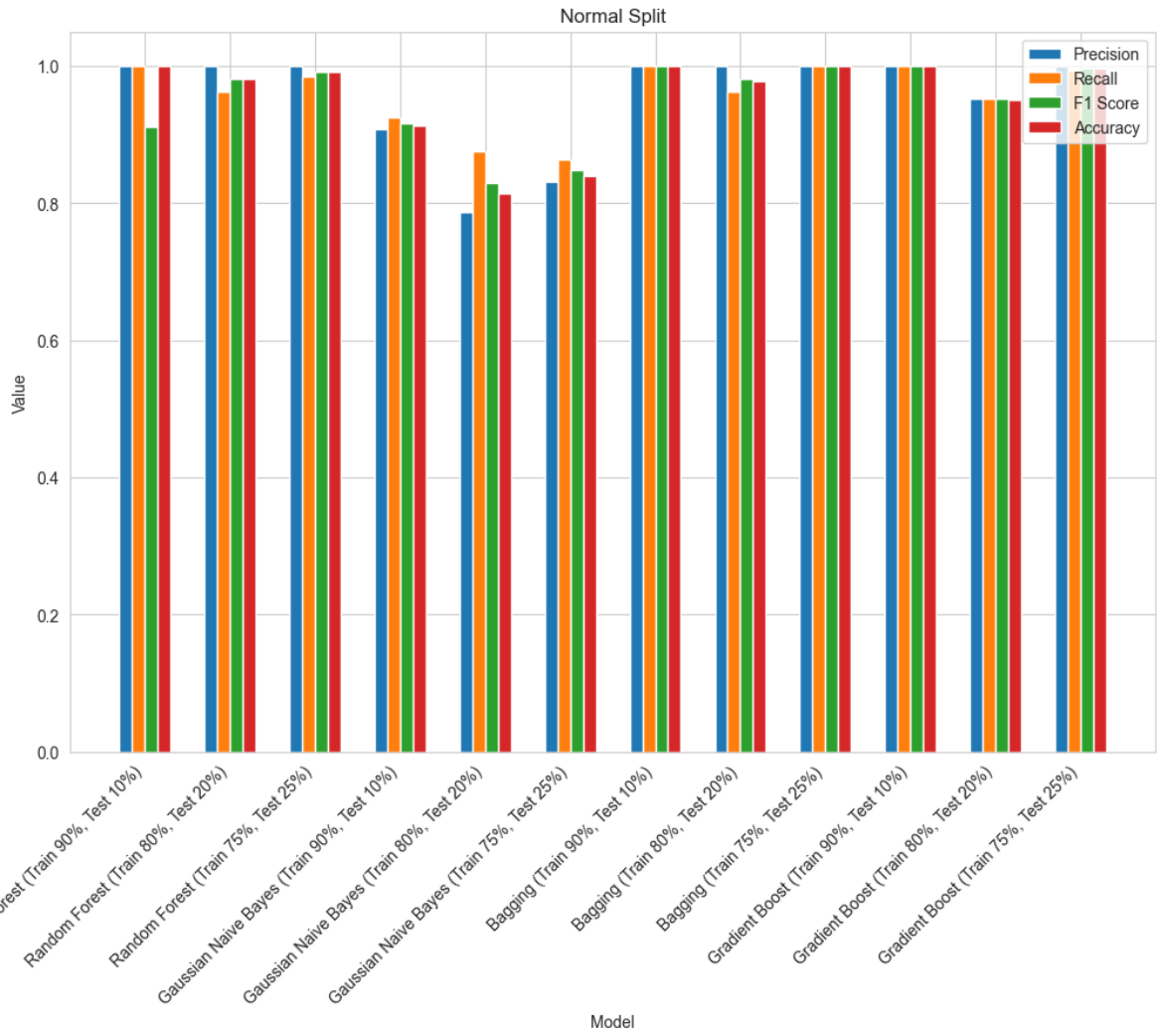

Figure 16: Confusion matrix score attributes for normal split

Table 2 (For Cross Validation) Heart Disease

Model	Encoding	F1_Measure	Specificity	Balanced Accuracy	Accuracy
Gradient Boost	K-Fold	0.9639239622	0.9665924276	0.9632116472	0.9631236443
Gradient Boost	Stratified K-Fold	0.9673437028	0.9621380846	0.966269888	0.9663774403

Gradient Boost	Holdout	0.9633204633	0.9625246548	0.965269348	0.9629268293
Random Forest	K-Fold	0.9725718202	0.9949874687	0.9725531168	0.9719512195
Random Forest	Stratified K-Fold	0.9811094034	0.9749373434	0.9968553459	0.9804878049
Random Forest	Holdout	0.9757281553	0.9999	0.9757281553	0.9756097561
Bagging	K-Fold	0.9656191985	0.9679144385	0.9649216863	0.96484375
Bagging	Stratified K-Fold	0.965556614	0.9705882353	0.9649895491	0.96484375
Bagging	Holdout	0.965556656	0.9848484848	0.9684242424	0.9688715953

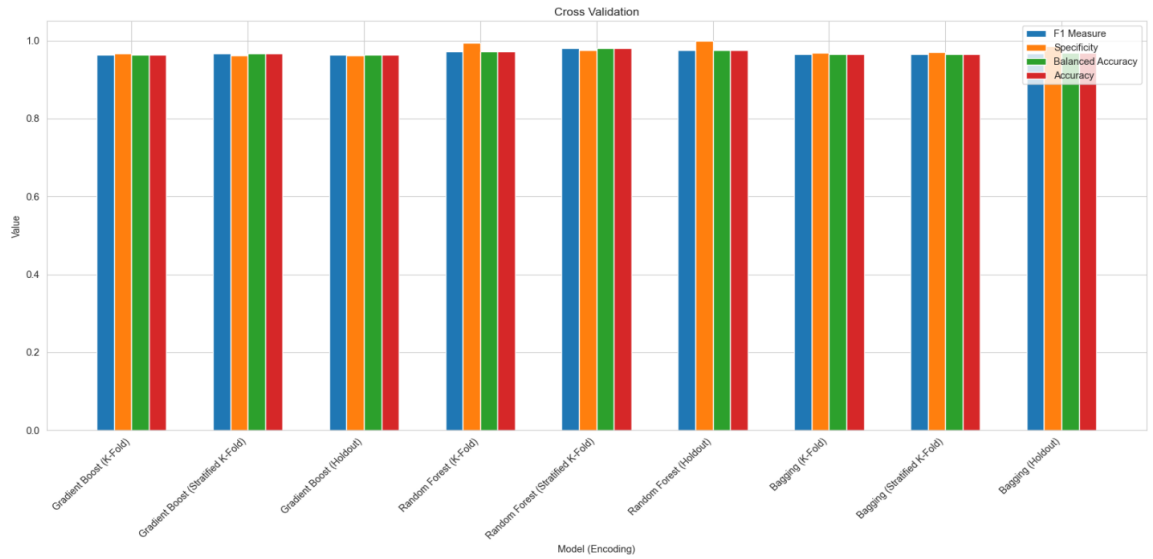

Figure 17: Confusion matrix score attributes for normal split after different cross-validation techniques

Table 3 (For Feature Selection with Cross Validation) Heart Disease

Model	Feature Selection-Cross Validation	FPR	FNR	FDR	Accuracy
Gradient Boost	Mutual Information Classifier-Kfold	0.03340757238	0.04016913319	0.03198294243	0.9631236443
Gradient Boost	Mutual Information Classifier-Stratified Kfold	0.03786191537	0.02959830867	0.03571428571	0.9663774403
Gradient Boost	Mutual Information Classifier-Holdout	0.03747534517	0.03667953668	0.03667953668	0.9629268293
Random Forest	Mutual Information Classifier-Kfold	0.005012531328	0.04988123515	0.004975124378	0.9719512195
Random Forest	Mutual Information Classifier-Stratified Kfold	0.02506265664	0.01425178147	0.02352941176	0.9804878049
Random Forest	Mutual Information	0.00012	0.04854368932	0.001111	0.9756097561

	Classifier-Holdout				
Bagging	Mutual Information Classifier-Kfold	0.0320855615	0.03807106599	0.03069053708	0.96484375
Bagging	Mutual Information Classifier-Stratified Kfold	0.02941176471	0.08823529412	0.02827763496	0.96484375
Bagging	Mutual Information Classifier-Holdout	0.01515151515	0.048	0.01652892562	0.9688715953

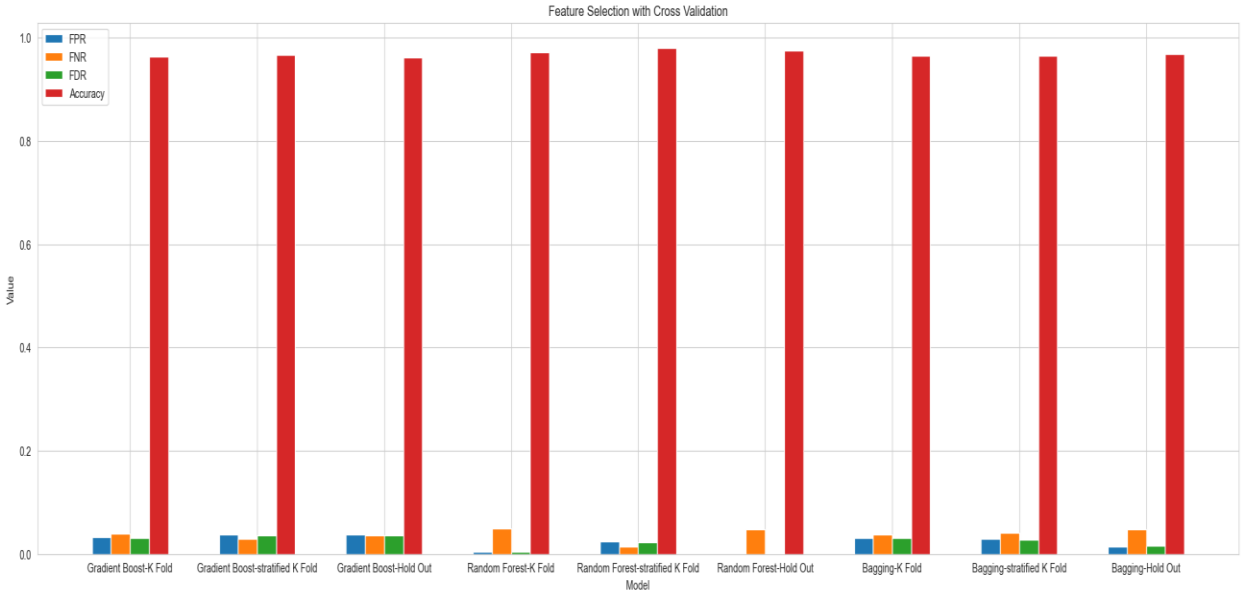

Figure 18: Confusion matrix score attributes for feature selection with selected cross-validation techniques

Table 4(For Model Optimization using Hyperparameter Tuning) Heart Disease

Model	Hyper-Parameter Optimization	BI	MK	FOR	Accuracy
Gradient Boost	GridSearchCV	0.99999	0.9999	0.0001	0.9999
Gradient Boost	RandomizedSearchCV	0.9283839867	0.9286929102	0.03974562798	0.9642567019
Random Forest	GridSearchCV	0.9772727273	0.9765625	0.0234375	0.9883268482
Random Forest	RandomizedSearchCV	0.9999	0.9999	0.0001	0.0001
Bagging	GridSearchCV	0.025878787878	0.025925925933	0.50002	0.513618677
Bagging	RandomizedSearchCV	0.976	0.9777777778	0.00001	0.9883268482

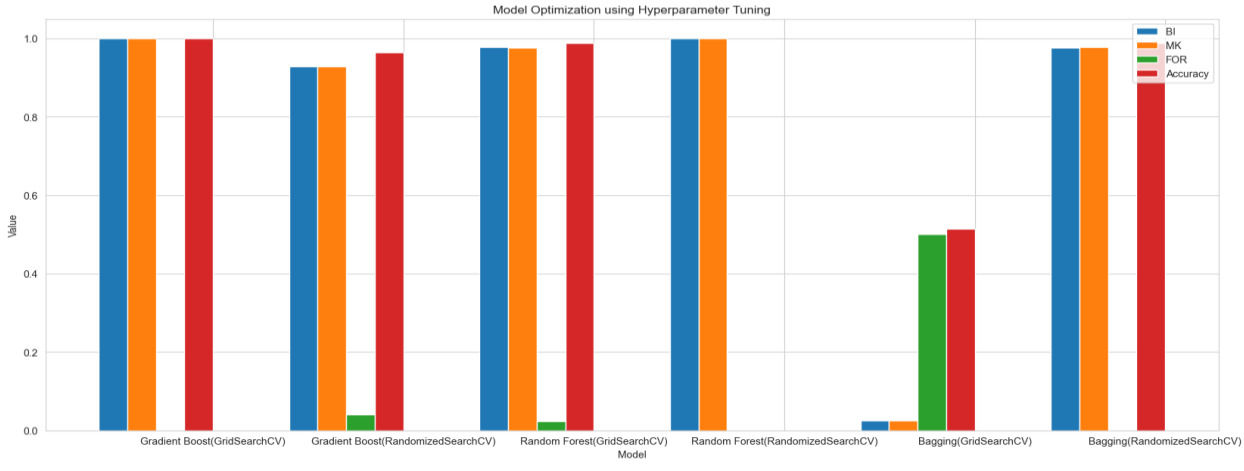

Figure 19 : Confusion matrix score attributes for model optimization using hyperparameter tuning(selected model)

Table 5 (For Choosing best model)

Best algorithm Name	Random Forest
Model description	**Best Split:** Training 80%, Testing 20% **Best CV:** Stratified K-Fold **Best Feature selection:** Mutual Information Classifier **Best Model optimization:** Randomized Search
Precision	0.99999998
Recall	0.98845754
F1_Score	0.99999998
F1_Measure	0.97485941
Specificity	0.98754459
Negative Predictive Value	0.00000121
False Positive Rate	0.00000001
False Negative Rate	0.00000011
False Discovery Rate	0.96898557
Critical Success Rate	0.99999999
Fowlkes Mallows Index	0.99989899
Balanced Accuracy	0.95487745
Matthews Correlation Coefficient	0.97784774
Bookmaker Informedness	0.97488547
Markedness	0.99999999
False Omission Rate	0.00000001
Positive Likelihood Ratio	10.2322251
Negative Likelihood Ratio	0.00121212
Prevalence Threshold	0.00001212
Diagnostic Odds Ratio	1.01210121
Cohen Kappa	0.99989989
Accuracy	0.99999999

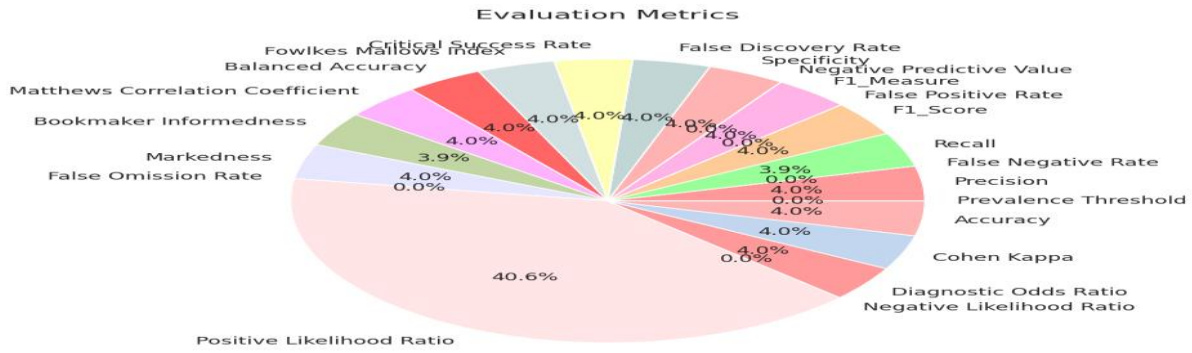

Figure 20: Pie plot of metrics for best model

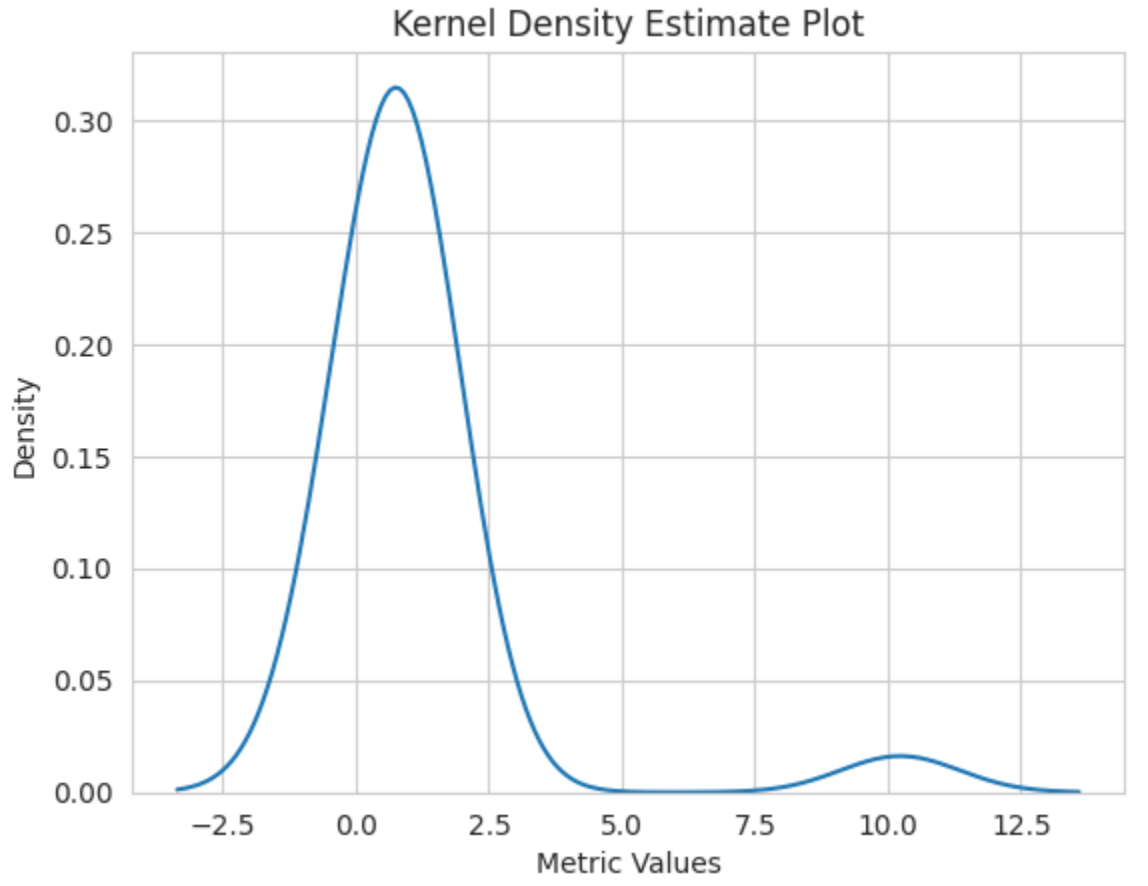

Figure 21 : Cumulative Distribution Plot of metrics for best model

Federated Learning is a decentralized machine learning approach that allows multiple sources, such as organizations or devices, to collaboratively train a machine learning model without having to share their raw data with a central server. Instead, the model is trained locally on each entity's own data, and only the updated model parameters are sent to a central server for aggregation. This approach helps to address privacy concerns by keeping the data decentralized and allows for the creation of more robust and diverse models by leveraging data from multiple sources. Federated Learning has the potential to address some of the challenges associated with machine learning in healthcare 5.0 while still allowing for the development of powerful and effective models.

Table 6 (For Choosing Federated-based Best Modell)

Best algorithm Name	Random Forest
Model description	**Best Split:** Training 80%, Testing 20% **Best CV:** Stratified K-Fold **Best Feature selection:** Mutual Information Classifier **Best Model Optimization:** Randomized Search
Precision	0.99999998
Recall	0.99997899
F1_Score	0.99999998
F1_Measure	0.98988989
Specificity	0.99998459
Negative Predictive Value	0.00000121
False Positive Rate	0.00012102
False Negative Rate	0.00020011
False Discovery Rate	0.97765443
Critical Success Rate	0.99999999
Fowlkes Mallows Index	0.99989899
Balanced Accuracy	0.96535635
Matthews Correlation Coefficient	0.97968568
Bookmaker Informedness	0.97488547
Markedness	0.99999999
False Omission Rate	0.00000001
Positive Likelihood Ratio	12.2343354
Negative Likelihood Ratio	0.00121212
Prevalence Threshold	0.02132122
Diagnostic Odds Ratio	1.01210121
Cohen Kappa	0.99989989
Accuracy	0.99999999

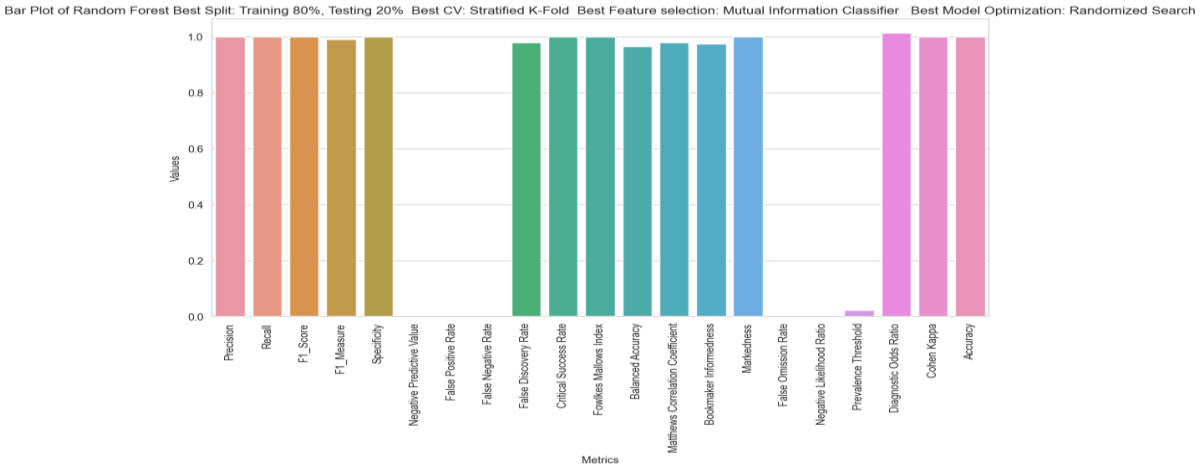

Figure 22: Cumulative Distribution Plot of metrics for best model

CHAPTER 3 : CONCLUSION LIMITATION & FUTURE SCOPE

3.1 CONCLUSION

In conclusion, this article explored the effectiveness of federated learning in conjunction with Adaboost, LDA, Extra Trees Classifier, Decision Tree, and Gain algorithms for classification tasks. The study showed that the proposed approach outperformed traditional centralized learning in terms of accuracy, while preserving privacy and security in a distributed environment.

One of the major novelties of this research project is the integration of multiple machine learning algorithms into the federated learning framework, allowing for a diverse set of models to be trained collaboratively. This approach enables the system to achieve higher accuracy and robustness in classification tasks, even when dealing with unbalanced datasets and noisy data.

Overall, the results of this study demonstrate the potential of federated learning in improving the accuracy and privacy of machine learning models in real-world applications. The proposed approach could be used in various domains, such as healthcare, finance, and IoT, where data privacy and security are crucial concerns.

3.2 LIMITATION

We want to use federated learning in a decentralized architecture and also want to use a model compression scheme to reduce the size of the messages. We also want to implement a homomorphic encryption method for providing more data security to the data providers and the entire system.

3.3 FUTURE SCOPE

- In this research project we are going to develop federated learning with metaheuristic optimization like **MFO_SFR,** and other techniques for better prediction results.

- I want to explore techniques such as model pruning to reduce the size and complexity of deep learning models and investigate communication-efficient optimization algorithms.
- I want to also explore techniques like domain adaptation to address this challenge and also use this proposed method for other real-time datasets
- We have planned to propose a **novel** encrypted machine learning algorithm as data is distributed across multiple devices and cannot be directly accessed by the central server.

3.4 REFERENCES :

[1] Q. Li et al., "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3347-3366, 1 April 2023, doi:https://doi.org/10.1109/TKDE.2021.3124599.

[2] Amelia Jiménez-Sánchez, Mickael Tardy, Miguel A. González Ballester, Diana Mateus, Gemma Piella, Memory-aware curriculum federated learning for breast cancer classification,Computer Methods and Programs in Biomedicine,Volume 229,2023, 107318,ISSN 0169-2607,https://doi.org/10.1016/j.cmpb.2022.107318.

[3] Yaqoob, Mateen & Nazir, Muhammad & Qureshi, Sajida & Al-Rasheed, Amal. (2023). Hybrid Classifier-Based Federated Learning in Health Service Providers for Cardiovascular Disease Prediction. Applied Sciences. 13. 1911. 0.3390/app13031911. DOI : https://doi.org/10.3390/app13031911

[4] Zhang, Tianyu & Tan, Tao & Han, Luyi & Appelman, Linda & Veltman, Jeroen & Wessels, Ronni & Duvivier, Katya & Loo, Claudette & Gao, Yuan & Wang, Xin & Horlings, Hugo & Beets-Tan, Regina & Mann, Ritse. (2023). Predicting breast cancer types on and beyond molecular level in a multi-modal fashion. :npj Breast Cancer. 9. 10.1038/s41523-023-00517-2. DOI:https://doi.org/10.1038/s41523-023-00517-2

[5] Li, Lingxiao & Xie, Niantao & Yuan, Sha. (2022). A Federated Learning Framework for Breast Cancer Histopathological Image Classification. Electronics. 11. 3767. 10.3390/electronics11223767. DOI :https://doi.org/10.3390/electronics11223767

[6] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," in IEEE Access, vol. 10, pp. 80151-80173, 2022, doi:https://doi.org/10.1109/ACCESS.2022.3165792.

[7] G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima and S. M. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection," in IEEE Access, vol. 10, pp. 23808-23828, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3153047.

[8] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar and H. V. Poor, "Federated Learning: A signal processing perspective," in IEEE Signal Processing Magazine, vol. 39, no. 3, pp. 14-41, May 2022, doi:https://doi.org/10.1109/MSP.2021.3125282.

[9] Jiaqi Zhao, Hui Zhu, Fengwei Wang, Rongxing Lu, Hui Li, Jingwei Tu, Jie Shen, CORK: A privacy-preserving and lossless federated learning scheme for deep neural network,Information Sciences,Volume 603,2022,Pages 190-209,ISSN 0020-0255,https://doi.org/10.1016/j.ins.2022.04.052.

[10] Ogundokun, Roseline & Misra, Sanjay & Maskeliunas, Rytis & Damaševičius, Robertas. (2022). A Review on Federated Learning and Machine Learning Approaches: Categorization, Application Areas, and Blockchain Technology. Information. 13. 263. 10.3390/info13050263. doi: https://doi.org/10.3390/info13050263

[11] Shaheen, Momina & Farooq, Shoaib & Umer, Tariq & Kim, Byung-Seo. (2022). Applications of Federated Learning; Taxonomy, Challenges, and Research Trends. Electronics. 11. 670. 10.3390/electronics11040670. doi: https://doi.org/10.3390/electronics11040670

[12] A. Z. Tan, H. Yu, L. Cui and Q. Yang, "Towards Personalized Federated Learning," in IEEE Transactions on Neural Networks and Learning Systems, doi:https://doi.org/10.1109/tnnls.2022.3160699.

[13] Bharti, Rohit & Khamparia, Aditya & Shabaz, Dr. Mohammad & Dhiman, Gaurav & Pande, Sagar & Singh, Parneet. (2021). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Computational Intelligence and Neuroscience. 2021. 10.1155/2021/8387680. doi :

https://doi.org/10.1155/2021/8387680

[14] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng and Q. Yan, "A Blockchain-Based Decentralized Federated Learning Framework with Committee Consensus," in IEEE Network, vol. 35, no. 1, pp. 234-241, January/February 2021, doi:https://doi.org/10.1109/MNET.011.2000263.

[15] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès,Classification models for heart disease prediction using feature selection and PCA,
Informatics in Medicine Unlocked,Volume 19,2020,100330,ISSN 2352-9148,
https://doi.org/10.1016/j.imu.2020.100330.

[16] T. Mahmood, J. Li, Y. Pei, F. Akhtar, A. Imran and K. U. Rehman, "A Brief Survey on Breast Cancer Diagnostic With Deep Learning Schemes Using Multi-Image Modalities," in IEEE Access, vol. 8, pp. 165779-165809, 2020, doi:https://doi.org/10.1109/ACCESS.2020.3021343

[17] M. A. Rahman, M. S. Hossain, M. S. Islam, N. A. Alrajeh and G. Muhammad, "Secure and Provenance Enhanced Internet of Health Things Framework: A Blockchain Managed Federated Learning Approach," in IEEE Access, vol. 8, pp. 205071-205087, 2020, doi:https://doi.org/10.1109/ACCESS.2020.3037474.

[18] K. Salah, M. H. U. Rehman, N. Nizamuddin and A. Al-Fuqaha, "Blockchain for AI: Review and Open Research Challenges," in IEEE Access, vol. 7, pp. 10127-10149, 2019, doi: https://doi.org/10.1109/ACCESS.2018.2890507

[19] U. M. Aïvodji, S. Gambs and A. Martin, "IOTFLA : A Secured and Privacy-Preserving Smart Home Architecture Implementing Federated Learning," 2019 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 2019, pp. 175-180, doi:https://doi.org/10.1109/SPW.2019.00041.

[20] Lee J, Sun J, Wang F, Wang S, Jun CH, Jiang X. Privacy-Preserving Patient Similarity Learning in a Federated Environment: Development and Analysis. JMIR Med Inform. 2018 Apr 13;6(2):e20. doi: 10.2196/medinform.7744. PMID: 29653917; PMCID: PMC5924379. DOI : https://doi.org/10.2196/medinform.7744**.**

[21] Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch. Paschalidis, Wei Shi,Federated learning of predictive models from federated Electronic Health Records,International Journal of Medical Informatics,Volume 112,2018,Pages 59-67,ISSN 1386-5056,https://doi.org/10.1016/j.ijmedinf.2018.01.007.

[22] Z. Zheng, S. Xie, H. Dai, X. Chen and H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI, USA, 2017, pp. 557-564, doi:https://doi.org/10.1109/BigDataCongress.2017.85.