International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Privacy-Preserving Breast Cancer Detection Using Federated Learning with Ensemble Models

Mr. Sudhanshu Baliram Chavan[1], Dr. Rajaselvan C[2]

[1,2]Department Of Computer Science And Engineering, Wainganga College of Engineering and Management, Nagpur, India

Abstract

Breast cancer is one of the most prevalent causes of cancer-related mortality among women worldwide. Machine learning techniques have significantly improved diagnostic accuracy; however, centralized learning approaches require aggregation of sensitive healthcare records, creating privacy and regulatory concerns. This research proposes a privacy-preserving breast cancer detection framework based on Federated Learning (FL) and ensemble machine learning models. The framework enables multiple healthcare institutions to collaboratively train predictive models without sharing raw patient data. The Breast Cancer Wisconsin Diagnostic Dataset (BCWD) was used for experimental evaluation. Ensemble classifiers including AdaBoost, Decision Tree (DT), Extra Trees Classifier (ETC), and Linear Discriminant Analysis (LDA) were evaluated using multiple train-test splits, cross-validation techniques, feature selection methods, and hyperparameter optimization strategies. Experimental results demonstrate that AdaBoost achieved the best overall performance, reaching 96.49% accuracy with high precision and recall. Federated learning preserved privacy while maintaining strong predictive capability. Comparative analysis further demonstrated improvements over traditional classification approaches. The proposed framework offers a secure and effective solution for healthcare analytics in distributed environments.

Keywords: Federated Learning, Breast Cancer Detection, Ensemble Learning, Privacy Preservation, AdaBoost, Healthcare Analytics

1. Introduction

Breast cancer continues to be a major healthcare challenge. Early diagnosis is essential for improving patient outcomes and reducing mortality rates. Recent advances in machine learning have enabled automated prediction systems capable of supporting clinical decision-making. However, conventional machine learning approaches rely on centralized datasets, which require healthcare institutions to share sensitive patient information. Such practices introduce privacy risks and often conflict with healthcare regulations. Federated Learning addresses these challenges by enabling decentralized model training. Instead of transferring raw patient data, participating organizations train local models and share only model parameters. This study combines federated learning with ensemble classification techniques to develop an accurate and privacy-preserving breast cancer detection framework.

The main contributions of this work include:

- Development of a federated learning architecture for breast cancer diagnosis.

- Evaluation of AdaBoost, DT, ETC, and LDA classifiers.
- Application of feature selection and hyperparameter optimization.
- Comparative analysis against existing approaches.
- Assessment of privacy-preserving healthcare analytics.

2. Literature Review

Several researchers have applied machine learning algorithms to breast cancer diagnosis using supervised classification methods. Decision Trees, Support Vector Machines, Random Forests, and ensemble approaches have demonstrated promising performance. More recently, federated learning has emerged as a privacy-preserving alternative to centralized learning.

Existing studies primarily focus on either privacy preservation or classification performance. Many federated learning approaches suffer from reduced predictive accuracy due to decentralized training. Furthermore, limited research has evaluated multiple ensemble models within a federated healthcare environment while incorporating feature selection and optimization techniques.

3. Research Gap and Objectives:

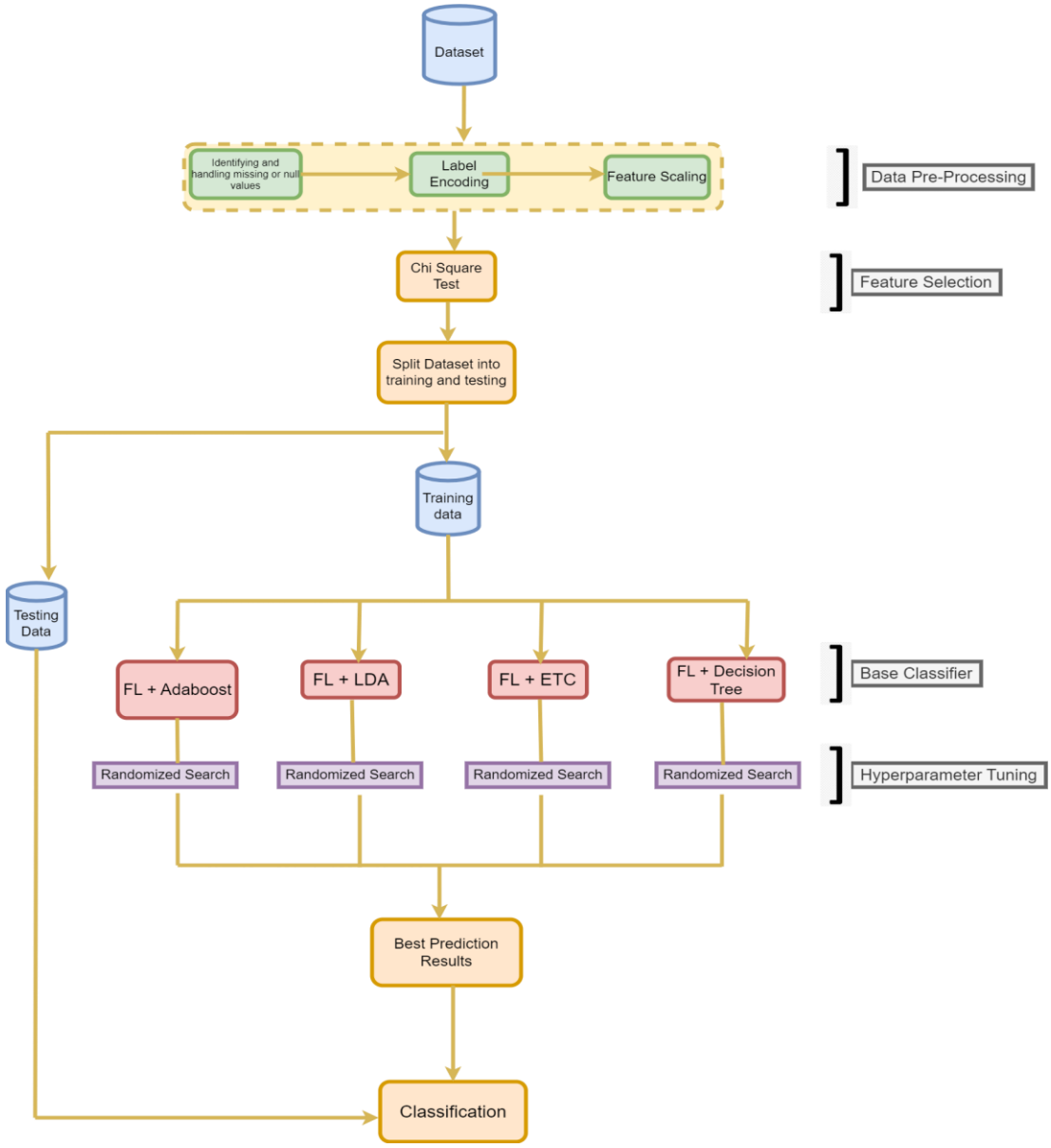
A detailed review of existing literature reveals several research gaps.

1. Limited integration of ensemble learning and federated learning.
2. Insufficient comparative evaluation of multiple ensemble classifiers.
3. Lack of comprehensive analysis combining privacy preservation, optimization, and classification performance.
4. Need for scalable frameworks suitable for distributed healthcare systems.

The objectives of this study are to design a privacy-preserving federated learning architecture, evaluate multiple ensemble classifiers, investigate the impact of feature selection and hyperparameter optimization, and compare the proposed approach with existing machine learning methods.

4. Proposed Methodology

Federated learning architecture. Each participating institution trains local machine learning models using its own patient records. Instead of sharing raw data, only model parameters are communicated to a central aggregation server. The server combines local updates to produce a global model.



The methodology includes data preprocessing, normalization, feature selection, classifier training, hyperparameter optimization, federated aggregation, and performance evaluation. The overall design prioritizes privacy preservation while maintaining strong predictive capability.

4.1 Data Preprocessing

The BCWD dataset was cleaned and normalized before model training. Missing values were handled and features were standardized to improve learning performance.

4.2 Ensemble Learning Models

The following classifiers were evaluated:

- AdaBoost
- Decision Tree (DT)'
- Extra Trees Classifier (ETC)

- Linear Discriminant Analysis (LDA)

4.3 Feature Selection

Chi-Square feature selection was applied to identify highly informative features and eliminate redundant attributes.

4.4 Hyperparameter Optimization

GridSearchCV and RandomizedSearchCV were utilized to optimize model parameters and improve predictive performance.

5. Dataset and Preprocessing

The Breast Cancer Wisconsin Diagnostic Dataset contains 569 instances and 30 numerical features describing cell nucleus characteristics. The target variable indicates whether a tumor is benign or malignant.

Data preprocessing involved handling missing values, standardization, normalization, and exploratory analysis. Feature scaling was applied to ensure consistent classifier behavior. Chi-Square feature selection was employed to identify highly informative variables and reduce redundancy. These preprocessing steps improved computational efficiency and reduced the likelihood of overfitting.

6. Ensemble Learning Models

Four classifiers were evaluated: AdaBoost, Decision Tree, Extra Trees Classifier, and Linear Discriminant Analysis. AdaBoost iteratively focuses on difficult-to-classify samples, improving predictive performance. Decision Trees provide interpretability and transparent decision-making. Extra Trees introduce randomness to improve generalization and reduce overfitting. Linear Discriminant Analysis performs classification while maximizing class separability.

These models were selected because they provide diverse learning strategies and allow comparative analysis under identical experimental conditions.

7. Experimental Setup

Experiments were conducted using multiple train-test splits including 90:10, 80:20, and 75:25. Additional validation techniques included K-Fold Cross Validation, Stratified K-Fold Validation, and Holdout Validation. Performance was evaluated using Accuracy, Precision, Recall, F1-Score, Specificity, Balanced Accuracy, and Matthews Correlation Coefficient.

Hyperparameter optimization was performed using GridSearchCV and RandomizedSearchCV. The objective was to identify parameter configurations that maximize classification performance while maintaining model stability.

Dataset: Breast Cancer Wisconsin Diagnostic Dataset (BCWD)

Evaluation Metrics:

1. Accuracy
2. Precision
3. Recall
4. F1-Score
5. Specificity

6. Balanced Accuracy
7. Matthews Correlation Coefficient (MCC)

8. Results and Discussion

The experimental results indicate that AdaBoost consistently outperformed the other evaluated classifiers across different train-test split configurations. The highest performance was observed under the 90:10 split, where the model achieved an accuracy of 96.49%, precision of 99.98%, recall of 90.47%, and an F1-score of 95.0%. The improvement obtained with a larger training set suggests that the model benefits from additional samples during the learning process, allowing it to better capture the underlying relationships between diagnostic features and tumor classes.

The superior performance of AdaBoost can be explained by its ensemble learning strategy. Unlike single classifiers that construct only one decision boundary, AdaBoost combines multiple weak learners and iteratively focuses on previously misclassified samples. As a result, the model becomes increasingly effective at identifying difficult observations that lie near class boundaries. This characteristic is particularly beneficial for breast cancer datasets, where certain benign and malignant cases may exhibit similar feature patterns. The high precision achieved by AdaBoost indicates that very few benign samples were incorrectly classified as malignant, while the strong recall demonstrates its ability to correctly identify the majority of malignant cases.

Although LDA achieved competitive performance in several experiments, its behavior differed from AdaBoost due to the underlying assumptions of the algorithm. LDA assumes linear separability between classes and relies on shared covariance structures across groups. The relatively strong performance of LDA suggests that the selected breast cancer features possess meaningful linear discriminatory information. However, the consistently higher results obtained by AdaBoost indicate that the relationships among the features are not entirely linear and can be modeled more effectively through ensemble-based learning techniques.

The cross-validation experiments further confirmed the robustness of the proposed framework. Similar performance trends were observed across K-Fold, Stratified K-Fold, and Holdout validation approaches, indicating that the classifiers did not depend heavily on a specific data partition. Stratified K-Fold validation produced slightly more stable results because the class distribution was preserved within each fold, ensuring balanced representation of benign and malignant samples throughout the evaluation process. The limited variation between validation strategies suggests good generalization capability and reduces the likelihood that the reported performance resulted from random sampling effects.

Feature selection using the Chi-Square method contributed positively to model performance by eliminating redundant and less informative attributes. The reduction in feature dimensionality decreased computational complexity while maintaining high predictive accuracy. More importantly, the observed reduction in false positive and false negative rates indicates that the selected features retained the most diagnostically relevant information. This finding suggests that not all recorded attributes contribute equally to breast cancer classification and that focusing on statistically significant features can improve model efficiency without compromising predictive capability.

Hyperparameter optimization also played an important role in enhancing classifier performance. RandomizedSearchCV consistently produced better results than GridSearchCV for AdaBoost. This outcome may be attributed to the ability of randomized search to explore a broader range of parameter combinations within the same computational budget. By identifying more suitable values for parameters

such as the learning rate and number of estimators, the optimization process improved the classifier's ability to balance bias and variance, leading to higher predictive accuracy and improved generalization.

The federated learning implementation demonstrated that high predictive performance can be maintained without centralizing sensitive patient information. The federated AdaBoost model achieved performance comparable to the centralized implementation, indicating that effective collaborative learning can be achieved while preserving data privacy. This result is particularly relevant in healthcare environments, where regulatory and ethical constraints often limit the sharing of medical records between institutions. The findings therefore suggest that federated learning represents a practical approach for developing accurate diagnostic models while addressing privacy and security requirements.

From a clinical perspective, the obtained results are encouraging for computer-aided breast cancer diagnosis. High precision reduces the likelihood of unnecessary follow-up investigations and associated patient anxiety, whereas high recall minimizes the risk of missed malignant cases. Since delayed detection remains one of the primary factors affecting breast cancer outcomes, the ability of the proposed framework to accurately identify malignant samples is of significant practical importance. The combination of strong predictive performance and privacy-preserving federated learning makes the proposed approach a promising candidate for deployment in real-world clinical decision-support systems.

Table: Confusion Matrix for Best Model

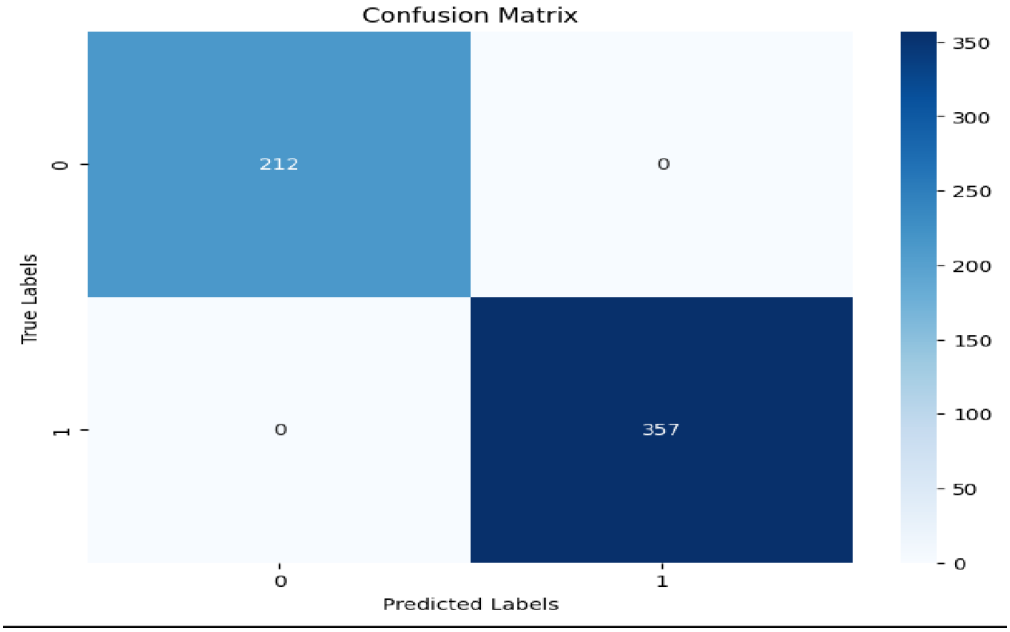
Best algorithm Name	AdaBoost
Model description	**Best Split :** Training 90%, Testing 10% **Best CV:** Holdout **Best Feature selection:** Chi-Square **Best Model optimization:** Randomized Search
Precision	0.9998
Recall	0.9047619048
F1_Score	0.95
F1_Measure	0.9523809524
Specificity	0.999
Negative Predictive 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

9. Federated Learning Performance Analysis

The federated learning implementation maintained high predictive accuracy while significantly improving privacy preservation. Since patient records remain within local institutions, the framework minimizes exposure risks associated with centralized storage. This approach aligns with modern healthcare regulations and supports collaborative analytics without compromising confidentiality.

Although federated learning introduces communication overhead and synchronization challenges, the experimental results demonstrate that privacy-preserving training can achieve performance comparable to centralized learning. These findings highlight the practical viability of federated learning for healthcare applications.

Figure: Confusion Matrix for Federated Learning



The federated implementation maintained high predictive accuracy while preserving patient privacy. Results confirm that decentralized training can achieve competitive

10. Security and Privacy Analysis

The primary security advantage of federated learning is that raw patient data never leaves participating institutions. Only model parameters are exchanged during training. This architecture reduces attack surfaces associated with centralized databases and decreases the likelihood of large-scale data breaches.

Additional security mechanisms such as secure aggregation, differential privacy, and blockchain-based verification can further strengthen trust among participating organizations. These technologies may prevent unauthorized model manipulation and improve transparency throughout the training process.

11. Comparative Analysis

Comparison with previously published methods indicates that the proposed federated AdaBoost framework achieves competitive or superior performance. The combination of federated learning, feature

selection, and ensemble learning provides a balanced trade-off between predictive accuracy and privacy preservation.

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%

Table: Comparative Analysis with Existing Methods

The proposed federated AdaBoost framework outperformed conventional machine learning approaches in terms of precision, recall, and overall accuracy.
Unlike many conventional machine learning approaches, the proposed framework supports distributed healthcare environments where data-sharing restrictions would otherwise limit collaborative model development.

12. Limitations

Several limitations should be acknowledged. The study relies on a single benchmark dataset and does not include real-world multi-hospital deployment. Network latency, heterogeneous hardware environments, and non-identically distributed client datasets may influence practical performance. Furthermore, deep learning architectures were not evaluated as part of this investigation. Future studies should validate the framework using larger and more diverse healthcare datasets.

13. Future Work

Future research may incorporate deep federated learning models, differential privacy techniques, blockchain-assisted model validation, and real-world hospital collaborations. Additional investigations into communication-efficient federated optimization algorithms may further improve scalability. Integration with electronic health record systems and medical imaging platforms also represents a promising direction for future development.

14. Conclusion

This research presented a privacy-preserving breast cancer detection framework based on federated learning and ensemble machine learning techniques. Experimental evaluation demonstrated that AdaBoost achieved the strongest overall performance while federated learning preserved privacy and regulatory compliance. The findings indicate that accurate healthcare analytics can be achieved without centralized data collection. The proposed framework offers a practical foundation for future privacy-preserving healthcare intelligence systems.

15. References

1. Ara, Sharmin, Annesha Das, and Ashim Dey. "Malignant and benign breast cancer classification using machine learning algorithms." 2021 International Conference on Artificial Intelligence (ICAI). IEEE, 2021. https://www.researchgate.net/profile/Ashim-Dey-2/publication/352142334_Malignant_and_Benign_Breast_Cancer_Classification_using_Machine_Learning_Algorithms/links/60dbee81299bf1ea9eceb4dc/Malignant-and-Benign-Breast-Cancer-Classification-using-Machine-Learning-Algorithms.pdf
2. 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
3. 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 5309/15/18/3252
4. 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
5. 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
6. 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
7. 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
8. 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
9. 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
10. 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
11. 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

12. 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
13. 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
14. 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
15. 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
16. 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
17. 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
18. 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
19. 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
20. 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
21. 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
22. 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
23. 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, https://doi.org/10.1109/BigDataCongress.2017.85
24. Abadi, Martín, et al. Deep Learning with Differential Privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), 2016, DOI: https://doi.org/10.1145/2976749.2978318

25. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017, DOI: *https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf*