Application of federated learning in predicting breast cancer

Jiarui Chai

Intelligence Science and Technology, Sun Yat-sen University, 518000 Shenzhen, China

Abstract. The prediction and diagnosis of breast cancer relies on multimodal data, such as imaging, genetic information, and patient lifestyle habits. Federated learning provides a framework to protect data privacy, allowing multiple institutions to share model training without sharing the original data. This paper proposes a breast cancer prediction model combined with federated learning, where each participant trains the model locally using multimodal data such as imaging, genes, and treatment history. During the local training process, the data is normalized and feature extracted, initially classified using support vector machines (SVM) or penalized logistic regression and optimized using stochastic gradient descent (SGD). Subsequently, each participant then sends the updated model parameters to the central server, where the FedAvg algorithm combines them to produce a global model. The model achieves data protection, also accurately predicts the progression and recurrence risk of breast cancer. Although federated learning effectively solves the privacy protection problem, the issues of data heterogeneity and model interpretability still need to be addressed. In the future, interpretability technologies (such as SHAP and LIME) and transfer learning can be combined to improve the transparency and adaptability of the model.

1 Introduction

Breast cancer [1] is a malignant tumor occurring in breast tissue, usually originating from breast ducts or lobules. Breast cancer can be classified into numerous varieties based on the type and growth of cancer cells. These types include invasive ductal carcinoma, invasive lobular carcinoma, and ductal carcinoma in situ. Although it is one of the most prevalent malignancies in women worldwide, survival rates can be greatly increased with early detection and treatment. Heredity is one of the risk factors for breast cancer, age, hormone level, lifestyle, etc. Typical signs of breast cancer include breast lumps, breast pain, skin changes, nipple secretions, etc.

At the same time, medical information of breast cancer patients, including diagnosis, treatment plan, gene data, etc., are highly sensitive personal data. Once these data are leaked, they may not only raise ethical and legal issues, but also have profound negative impacts on patients' lives, such as causing psychological stress, social discrimination, and even affecting insurance eligibility or employment opportunities. Therefore, protecting patient privacy is

Corresponding author: chaijr@mail2.sysu.edu.cn

not only a moral responsibility, but also a necessary requirement to comply with laws and regulations. In short, privacy protection is not only related to individual rights and dignity, but also the core foundation for ensuring information security, improving treatment effect and promoting scientific research progress during breast cancer treatment.

Although some progress has been made in privacy protection technology in breast cancer treatment, there are still significant research gaps, which need to be solved and optimized urgently. First, genetic data privacy protection is a key challenge, as breast cancer treatment increasingly relies on genomics data, and existing regulations lack globally harmonized standards for long-term storage and cross-border sharing, increasing the risk of data misuse and breaches. Last but not least, due to patients' lack of control over data, transparency and traceability are limited, which weakens informed consent mechanisms and reduces patients' trust in data use. Therefore, this article introduces federated learning (FL) technology, it provides a workable answer to the privacy protection problem during breast cancer treatment.FL avoids the direct sharing of raw data by allowing institutions to store sensitive data locally and train models collaboratively. For example, researchers have developed a FL-based system that allows multiple hospitals to co-train breast cancer prediction models without sharing patient data, improving privacy and diagnostic accuracy. FL has been effectively used in breast cancer diagnosis distributed radiology and pathology image analysis, especially in scenarios that require cross-agency data, addressing data privacy and security issues. By combining differential privacy technology with FL, privacy protection can be maintained even for shared model parameters, reducing the risk of re-identification. However, the application of FL in breast cancer treatment still faces challenges such as "communication overhead", "model convergence" and data differences between institutions, but its potential to improve data sharing efficiency and privacy undoubtedly opens up a new direction for breast cancer treatment and is expected to promote the development of precision medicine. The content should be reduced. In addition, related works in federated learning should be provided

Therefore, thanks to the criticality of this field, many new federated learning methods [2, 3] have been proposed in recent years, providing new directions for solving these problems. In order to assess the potential of federated learning and currently available privacy-preserving technologies in the treatment of breast cancer, a thorough analysis of this topic is therefore required. Federated learning makes it possible for multiple institutions to train models together without exchanging raw data, which improves data interchange efficiency while simultaneously successfully safeguarding patient privacy. This paper intends to further raise the standard of privacy protection in breast cancer therapy by utilizing federated learning technology in conjunction with a number of techniques and ensure the development of precision medicine under the premise of ensuring data security. This technology will enable healthcare organizations to securely and efficiently share and analyze data globally, ultimately providing more personalized and precise treatment options for those suffering with breast cancer.

2 Methods

2.1 Introduction of federated learning

Federated learning [4] is a distributed machine learning method whose core principle is to achieve multi-party collaborative training of models without sharing raw data. Every participant update and processes the model locally. Then transmit the model parameters to a central server via encryption, and the server integrates the changes to produce a global model. The key idea is to use data from several sources while protecting data privacy to

improve model performance. Together, these elements—local model training, model update sharing, central aggregation, and privacy-preserving technologies—ensure data protection and satisfy regulatory requirements.

2.2 FedAvg

FedAvg [5] is a core algorithm in federated learning, which trains the model locally by each client and updates the global model after weighing the model parameters, so as to achieve distributed collaborative learning under data privacy protection.

2.2.1 CNN

Federated Averaging-Based Convolutional Neural Network (CNN) [6] is one of the commonly used models in the federated learning framework, especially for processing image data. In FedAvg, CNNs can extract image features locally on the client side and perform classification tasks. The architecture typically consists of a convolutional layer, a pooling layer, and a fully connected layer, where each client trains on local image data, extracts key features, and sends model weights to a central server.

In each hospital or medical facility, CNNs are used to analyze local breast cancer image data (such as X-rays or MRIs [7]). CNNs are able to extract tumor features from images and classify them. After several rounds of training, each institution's local CNN model updates model parameters (such as the weights of the convolutional layer) and sends these updates to a central server.

In the meantime, here's an innovative experience using CNN technology：Enhanced privacy protection and Predictive models combined with clinical data.

Although FedAvg protects privacy by not sharing data, the sensitivity of breast cancer data still requires additional privacy protections. This paper proposes to combine differential privacy and homomorphic encryption [8] to provide stronger protection for federated learning in breast cancer diagnosis. Differential privacy prevents attackers from extrapolating the original data from model parameters by adding noise to model updates. Homomorphic encryption allows the server to aggregate in a model-encrypted state, ensuring that data cannot be decrypted even if it is intercepted in transit.

In the prediction task of breast cancer, in addition to imaging data, multimodal data such as genetic information, lifestyle habits, and the patient's treatment history are also involved. Under the federated learning framework, each participant (such as a hospital) can retain its local data to ensure patient privacy. Each participant trains the model locally, combines imaging data with other clinical data, and forms a comprehensive breast cancer prediction model. This model can not only diagnose the current state of breast cancer, but also predict future disease progression, risk of recurrence, or response to treatment options. Each participant uses a secure federated learning mechanism to transmit the new model parameters to the central server after each training cycle. The central server then uses weighted aggregation to create a global model. This method effectively integrates information from different sources and improves the overall prediction ability of the model, while avoiding direct sharing of raw data and protecting data privacy. Through this multimodal federated learning method, the model can increase the precision and dependability of breast cancer diagnosis and prognosis by fully using the data advantages of all parties.

2.3 SGD

SGD [9] is a core algorithm for optimizing machine learning models. It gradually updates model parameters by using the gradient information of the loss function, setting an appropriate learning rate, and randomly selecting small batches of data for training. In the federated learning framework, SGD allows each participant to retain data locally and independently update model parameters, and then securely send the updated parameters to the central server for aggregation. Combined with techniques such as momentum and regularization, SGD improves training efficiency and helps the model gradually approach the optimal solution over multiple iterations, while protecting data privacy and enhancing breast cancer detection accuracy and prediction. This combination ensures that the model can share knowledge between different participants without leaking original data.

2.3.1 SVM

Support vector machine (SVM) [10] is a classical supervised learning algorithm that is widely used in classification, regression, and other tasks. SVM is frequently utilized in the investigation and diagnosis of breast cancer for the purposes of prognosis prediction, classification, and detection, especially in classification tasks based on breast cancer imaging and genetic data.

Under the framework of federated learning, the implementation details of SVM include the following aspects: First, each participant (such as a hospital) collects and preprocesses imaging data (such as X-rays, MRI) or genetic data locally, and the preprocessing steps include denoising [11], standardization and feature extraction. Secondly, the key features of the image (such as shape and texture) or important features in gene expression are obtained through feature extraction methods [12], and feature selection methods (such as recursive feature elimination) are combined for dimensionality reduction. Then, the participants train the SVM model locally, select appropriate kernel functions (such as RBF kernels [13]) and adjust hyperparameters (such as regularization parameters and kernel function parameters) to ensure efficient training of the model on local data. During the training process, the use of the SMO algorithm [14] can accelerate the calculation of large-scale data to ensure efficient convergence of the model. Lastly, each participant sends the updated model parameters to the central server for federated learning-based aggregation in order to validate the model's effectiveness.

The innovation is that SVM can not only process a single data type (such as imaging data), but also combine multimodal data for comprehensive analysis. Through the federated learning framework, data privacy protection is ensured, and the original data is not required to be shared by the participants. This combination of multi-source data can significantly improve the accuracy of tumor classification and recurrence prediction. Through feature fusion technology (such as feature-level fusion [15]), SVM can process heterogeneous data from multiple sources and improve the reliability of diagnosis. At the same time, combined with recursive feature elimination [16] (RFE) for feature selection, it can effectively screen out the most discriminative features in high-dimensional data, help reduce redundancy, and improve the computational efficiency and accuracy of the model. In this sense, the use of SVM in federated learning guarantees data confidentiality while simultaneously improving the model's performance.

2.3.2 Penalized logistic regression

Penalized logistic regression refers to a form of logistic regression where the coefficients are estimated by optimizing the log-likelihood function along with an added penalty term on the

magnitude of the coefficients. This penalization helps to control the size of the coefficients, improving the model's generalization ability.

In federated learning, each participant first collects local data and performs standardization, missing value processing, and encoding. L1 or L2 regularization is used for feature selection to ensure that the model only uses important features. Each participant trains a penalized logistic regression model on local data and uses SGD for parameter optimization. Each participant uses the federated learning framework to communicate the modified model parameters to the central server following each training round, where they are aggregated to generate a global model. In this process, SGD combined with federated learning can not only accelerate the training process, but also handle large-scale and heterogeneous data. Since the data distribution of each participant may be different, SGD allows each participant to optimize its model locally to better adapt to the data characteristics. An iterative training process is formed by sending the revised global model back to each participant so they can begin the next round of local training.

3 Challenges and future prospects

Combining multimodal data with machine learning approaches (such SVM, penalized logistic regression, and stochastic gradient descent) has generated some encouraging outcomes in the detection of breast cancer and prediction. However, these approaches continue to encounter multiple limitations and challenges, especially in terms of interpretability and applicability.

3.1 Interpretability

Machine learning models, especially in federated learning environments, often lack transparency. Although each participant is able to train the model locally and protect patient privacy, the lack of interpretability may lead to a decrease in medical staff's trust in the model output. Therefore, for the purpose of improving the interpretability of models in federated learning, it is imperative to use explainable artificial intelligence technologies, such as SHapley Additive exPlanations (SHAP) [3] and Local Interpretable Model-Agnostic Explanations (LIME) [17]. These technologies can analyze the factors that affect each prediction result of the model without leaking the original data, thereby helping doctors understand the model's decision-making procedure and improving the credibility of clinical applications.

3.2 Applicability and distribution

Differences in data distribution between different hospitals may affect the generalization ability of federated learning models. Each participant uses their own data for training locally, which may result in the model performing well in some hospitals and poorly in others. Therefore, future research can use transfer learning and domain adaptation techniques to enable the model to better adapt to disparate data distributions. Under the framework of federated learning, by fine-tuning the shared model, the applicability and robustness of the model can be effectively improved.

3.3 Data privacy and security

Although federated learning protects data privacy, it still faces security risks during the transmission of model parameters. Malicious attackers may cause data leakage or model

reverse engineering by attacking the transmitted model parameters. Therefore, ensuring the security of data during transmission and adopting methods such as encryption mechanisms and secure multi-party computing (SMPC [18]) will be an important direction for further studies to enhance federated learning security even more.

3.4 Real-time prediction and personalized medicine

With the advancement of technology, future models can achieve real-time prediction and provide doctors with instant decision support. Under the framework of federated learning, by combining the patient's historical data and characteristics, the model can provide personalized predictions and treatment recommendations for each patient. This real-time prediction capability can help doctors quickly assess the condition and develop personalized treatment plans to improve the quality of medical care.

3.5 Ethical and regulatory issues

With the widespread application of machine learning technology in the medical field, related ethical and regulatory issues have become increasingly prominent. In the federated learning framework, all parties should jointly formulate ethical standards and regulatory frameworks to ensure that data privacy and security are protected. At the same time, researchers and medical institutions need to pay attention to the fairness and transparency of the model to promote the healthy development of machine learning in breast cancer diagnosis.

4 Conclusion

The federated learning to breast cancer prediction is reviewed in this work, with a focus on how patient privacy protection might enhance diagnosis accuracy and prediction model generalization, reducing data transmission requirements, and combining multiple data sources (such as imaging, genetic data, etc.). Federated learning, which combines data from several sources without directly sharing data, successfully addresses the problem of data privacy in classic centralized learning. However, even if federated learning has demonstrated a lot of promise for the medical industry, it still faces some challenges in terms of model convergence, communication overhead, and heterogeneous data processing. In the future, related research can focus on optimizing the training efficiency of the model, exploring the combination of transfer learning and domain adaptation technology, and further introducing technologies to enhance the model's security and privacy protection features and provide stronger support for personalized prediction of complex diseases such as breast cancer.

References

1. X. Zhang, X. Dong, Y. Guan, M. Ren, D. Guo, Y. He, Research progress on epidemiological trend and risk factors of female breast cancer. Cancer Research on Prevention and Treatment, 48(1), 87-92 (2021).
2. L. Zhang, Analysis of breast cancer diagnosis based on a machine-learning algorithm. Operations Research and Blurring, 14(4), 397-405 (2024).
3. Y. Supriya, R. Chengoden, Breast cancer prediction using Shapely and Game theory in Federated Learning environment. IEEE Access, (2024).

4. J. Wen, Z. Zhang, Y. Lan, et al., A survey on federated learning: challenges and applications. International Journal of Machine Learning and Cybernetics, 14(2), 513-535 (2023).
5. Z. Li, V. Sharma, S. P. Mohanty, Preserving data privacy via federated learning: Challenges and solutions. IEEE Consumer Electronics Magazine, 9(3), 8-16 (2020).
6. M. Islam, M. T. Reza, M. Kaosar, et al., Effectiveness of federated learning and CNN ensemble architectures for identifying brain tumors using MRI images. Neural Processing Letters, 55(4), 3779-3809 (2023).
7. J. Ogier du Terrail, A. Leopold, C. Joly, et al., Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nature Medicine, 29(1), 135-146 (2023).
8. H. Fang, Q. Qian, Privacy preserving machine learning with homomorphic encryption and federated learning. Future Internet, 13(4), 94 (2021).
9. Y. N. Tan, V. P. Tinh, P. D. Lam, et al., A transfer learning approach to breast cancer classification in a federated learning framework. IEEE Access, 11, 27462-27476 (2023).
10. J. Peta, S. Koppu, Enhancing breast cancer classification in histopathological images through Federated Learning Framework. IEEE Access, 11, 61866-61880 (2023).
11. T. U. Islam, R. Ghasemi, N. Mohammed, Privacy-preserving federated learning model for healthcare data. In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 281-287 (2022).
12. A. Chowdhury, H. Kassem, N. Padoy, et al., A review of medical federated learning: Applications in oncology and cancer research. In International MICCAI Brainlesion Workshop, Cham: Springer International Publishing, 3-24 (2021).
13. F. Zerka, V. Urovi, F. Bottari, et al., Privacy preserving distributed learning classifiers–sequential learning with small sets of data. Computers in Biology and Medicine, 136, 104716 (2021).
14. S. A. Mohammed, S. Darrab, S. A. Noaman, et al., Analysis of breast cancer detection using different machine learning techniques. In Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, Proceedings 5. Springer Singapore, 108-117 (2020).
15. R. B. Eshun, Supervised Predictive and Federated Learning for Integrative Analysis of Medical Data. North Carolina Agricultural and Technical State University (2024).
16. A. Umar, M. S. Aliyu, J. Awwalu, An Ensemble Model Based on Recursive Feature Selection for Breast Cancer Prediction. nijocet.fud.edu.ng (2016).
17. A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, et al., Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. Cancer Innovation, 3(5), e136 (2024).
18. A. P. Kalapaaking, V. Stephanie, I. Khalil, et al., SMPc-based federated learning for 6G-enabled internet of medical things. IEEE Network, 36(4), 182-189 (2022).