Application of Machine Learning in Prediction of Breast Cancer

Chuanqi Yu

Information Management and Information Systems, Guizhou University, Guiyang, 550000, China
wenwen@ldy.edu.rs

Abstract. Breast cancer is a malignant tumor that develops from the cells of breast tissue, typically from the ducts or glands. Artificial Intelligence (AI) technology has become a viable option for the automated detection of breast cancer. The study aims to provide a comprehensive review about the application of various machine learning models, including Logistic Regression, Support Vector Machine (SVM), and Artificial Neural Network (ANN) in predicting breast cancer risk. The discussion showed that deep learning neural networks outperformed other models, demonstrating promising potential in clinical applications. The findings suggest that deep learning neural networks offer superior accuracy in breast cancer prediction, highlighting their significance in healthcare advancements. Key challenges such as model interpretability, data distribution differences, and privacy are addressed, emphasizing the need for transparent models and secure data handling techniques like federated learning in the future studies. This paper can be considered as an effective reference for this field.

Keywords: Machine learning, breast cancer, deep learning

1 Introduction

Breast cancer is a malignant tumor originating from breast tissue, usually originating from breast ducts or glands. The causes of its formation are diverse, including genetic mutations, environmental factors, lifestyle, etc., but the specific pathogenesis is not yet fully understood. The harmfulness of breast cancer is mainly reflected in its high incidence, easy spread and impact on patients' physical and mental health. According to statistics, there are hundreds of thousands of new cases of breast cancer in China every year, while millions of women around the world may suffer from breast cancer every year [1], which has become one of the important challenges for women's health.

The traditional detection of breast cancer mainly depends on doctors' manual diagnosis, including clinical examination, imaging examination and histological examination. Hothis paperver, this manual diagnostic method has certain limitations and shortcomings, including low market efficiency, high misdiagnosis rate, high labor cost, and subjectivity in the diagnostic process. Therefore, the use of Artificial Intelligence (AI) technology for automated detection of breast cancer has become a

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_9

feasible solution. By combining AI technology, it is possible to identify breast cancer lesions more accurately and efficiently, so as to improve the early diagnosis rate, reduce the misdiagnosis rate, and the workload of doctors, and provide patients with more timely and accurate diagnosis and treatment programs.

Artificial intelligence has witnessed significant advancements in recent years, revolutionizing various domains. Historically, AI development dates back to the 1950s, with the concept gaining traction over subsequent decades. Notable milestones include the development of expert systems, decision trees, random forests, and neural networks. AI's applications span diverse fields, including biology, chemistry, civil engineering, transportation, and particularly healthcare. In healthcare, AI has made significant strides in predictive analytics for various diseases. For instance, algorithms have been employed in predicting heart diseases, pneumonia, and tumor growth. Growing interest has been seen in applying AI methods to predict breast cancer in recent years. Several studies have explored different algorithms for breast cancer prediction. For example, researchers have utilized Support Vector Machines (SVM), deep learning neural networks, and ensemble methods like gradient boosting and random forests. Noteworthy examples in Yolanda D Austria et al.'s paper, these methods were extensively and widely used, and their performance in comparing cancer aspects was compared [2]. Additionally, D Delen also developed prediction models for large datasets using artificial neural networks and decision trees [3].

To address this paper's research objective of enhancing breast cancer prediction, this paper embarked on a systematic review. In this paper study, this paper evaluated multiple machine learning models to ascertain the most effective approach for breast cancer prediction. These models included logistic regression, SVM, and deep learning neural networks. The discussion presented in this research showed how well the suggested methodology predicted breast cancer risk. This overview summarizes the key aspects of this paper's technical approach and highlights the effectiveness of this paper methodology in addressing the research objective.

2 Method

2.1 The Introduction of Machine Learning Algorithms

The process of machine learning generally includes data collection, data preprocessing, feature engineering model construction, and training and testing

Data collection is the process of abstracting business logic into problems that algorithms can handle, and determining the types of supervised, unsupervised, or semi supervised learning based on data and objectives. At the same time, relevant data needs to be collected.

Data preprocessing is the process of cleaning and organizing raw data, including deduplication, standardization, error correction, as well as handling empty values and garbled characters. At this stage, preliminary exploratory analysis of the data is also necessary to understand the distribution characteristics of the data and the correlation between independent and dependent variables.

Feature engineering is a crucial step in machine learning, involving the extraction, screening, construction, and transformation of features, with the aim of obtaining a set of features that can effectively represent samples and aid in model training. The quality of features directly determines the quality of the model.

Select appropriate algorithms for model training and validation, and use processed features and training datasets for model construction. During the model training process, it is usually necessary to evaluate and adjust model parameters through methods such as cross validation to achieve optimal performance.

Model evaluation uses an independent test dataset to evaluate the trained model and check whether its performance on new data meets the expected requirements. Common evaluation indicators include accuracy, recall, F1 score, etc.

Online service and application optimization can deploy evaluated and optimized models to provide services in actual production environments, and continuously optimize models and data processing processes based on feedback from actual applications to adapt to new needs and data changes

2.2 Machine Learning Models

Several machine learning models from the sklearn package were used in this work, including Support Vector Machine, Logistic Regression and neural network models.

Logistic Regression. A statistical technique called logistic regression is employed to describe the connection between one or more independent variables and a categorical dependent variable. It is frequently used to forecast the likelihood of an event occurring by fitting data to a logistic function in cases where the dependent variable is binary (having just two possible outcomes).

The estimation of the likelihood that a given input falls into a specific category is the fundamental concept of logistic regression. By using a logistic function on the linear combination of the independent variables, this is accomplished. The logistic function, also known as the sigmoid function, maps any real value into a range betthis paperen 0 and 1, representing the probability of the event happening.

The principle of logistic regression involves estimating the parameters of the model using a process called maximum likelihood estimation. The model calculates the likelihood of observing the actual outcome given the input variables and adjusts the parameters to maximize this likelihood. By iteratively updating the parameters based on the training data, logistic regression finds the coefficients that most accurately match the data and are useful for forecasting upcoming observations. In the study by Liu et al., the Logistic Regression algorithm from the Sklearn machine learning library was employed to classify breast cancer diagnosis datasets. The results indicate that selecting the maximum texture and maximum perimeter as features yields a classification accuracy of 96.5%, which represents an improvement over previous method [4].

SVM. Encouragement Vector Machines are supervised learning models that examine data for regression and classification. They are paired with learning algorithms. SVM

carries out classification in the context of determining the hyperplane that best differentiates betthis paperen different classes. The key idea is to maximize the margin betthis paperen the hyperplane and the nearest data points of any class, allowing for better generalization to new data. SVM can also handle non-linear decision boundaries through techniques like the kernel trick.

In this research conducted by Huang et al., the objective was to thoroughly evaluate the prediction performance of SVM and SVM ensembles on both small and large scale breast cancer datasets. The study compared the classification accuracy, ROC, F-measure, and computational training times of SVM and SVM ensembles. The findings indicate that for small scale datasets, SVM ensembles employing a linear kernel with the bagging method and those using an RBF kernel with the boosting method are superior, particularly when feature selection is applied during data preprocessing. Conversely, for large scale datasets, SVM ensembles that utilize an RBF kernel with the boosting method outperform other classifiers [5].

Neural Networks. Neural networks are a class of algorithms with pattern recognition capabilities, somewhat inspired by the structure of the human brain. They categorize or group unprocessed information in order to interpret sensory data using a form of machine perception. Layers of connected nodes, or "neurons," that process and send data make up a neural network. Neural networks' primary principle is to employ linked layers of nodes to identify intricate patterns in data. The number of layers and neurons can vary widely based on the specific architecture being used. In general, more complex problems may require deeper networks with more layers and neurons.

In this study by Saritas et al., an artificial neural network (ANN) was developed to diagnose breast cancer in patients. The ANN, combined with BI-RADS evaluation, utilizes patient age, mass shape, mass border, and mass density to determine not only the presence of cancer but also its type. The implemented ANN model achieved a disease prediction accuracy of 90.5% and a healthy prediction accuracy of 80.9% in the test data. These high predictive values demonstrate that the ANN model is fast, reliable, and risk-free, making it a valuable tool for assisting physicians [6].

3 Discussion

To effectively discuss the limitations and challenges in the field of AI in healthcare, particularly regarding interpretability, distribution differences, and privacy, it's essential to consider both current issues and future prospects.

Interpretability: Interpretability is the capacity to comprehend and elucidate the decision-making process used by AI algorithms [7, 8]. In healthcare, where decisions can have critical consequences, such as in breast cancer prediction, interpretability is crucial for gaining trust and acceptance from medical professionals. Present AI models, including deep neural networks, frequently act as "black boxes," making it difficult to communicate their judgments to medical professionals. Addressing this, techniques like expert systems, SHAP, Grad-CAM are being explored. Expert systems provide transparent decision-making processes, SHAP offers explanations of individual

predictions, and Grad-CAM highlights regions in images influencing predictions, aiding interpretability.

Distribution Differences: Distribution differences refer to variations in data distributions across different datasets or domains, which can degrade AI model performance when applied to new data [9]. This issue is particularly relevant in healthcare due to variations in patient demographics, medical protocols, and equipment. To mitigate this, methods such as transfer learning, domain adaptation, and domain generalization are employed. Transfer learning enables models trained on one dataset to be adapted to another related dataset, while domain adaptation adjusts models to perform well on target domains with different data distributions.

Privacy: Privacy concerns arise from the sensitive nature of medical data used in AI models, such as mammographic images and patient records. Maintaining patient privacy is critical to ensuring ethical AI deployment in healthcare [10]. Federated learning and differential privacy are emerging solutions. Federated learning allows training models across multiple decentralized institutions without exchanging raw data, thereby preserving privacy. In order to safeguard personal information while preserving the ability to conduct insightful analysis, differential privacy adds noise to data.

Future Prospects: Looking ahead, advancements in AI offer promising prospects. Improved interpretability through hybrid models combining deep learning with explainable AI techniques like SHAP and Grad-CAM could enhance trust and adoption in clinical settings. Addressing distribution differences will benefit from more robust domain adaptation techniques capable of handling complex healthcare data dynamics. Additionally, ensuring privacy through federated learning will enable collaborative model training across institutions while protecting patient confidentiality.

4 Conclusion

This review synthesizes the application of machine learning in breast cancer detection, detailing steps from data collection to model deployment. It highlights the effectiveness of logistic regression, Support Vector Machines, and neural networks in handling classification problems in healthcare. Key challenges such as model interpretability, data distribution differences, and privacy are addressed, emphasizing the need for transparent models and secure data handling techniques like federated learning. The article advocates for advancements in explainable AI and robust data adaptation methods to enhance the deployment of AI in clinical settings, ensuring both efficacy and ethical compliance.

References

1. Chen, W.Q., Zheng, R.S.: Incidence, Mortality, and Survival Status of Female Breast Cancer in China. Chinese Journal of Clinical Oncology 42(13), 668-674 (2015).
2. Austria, Y.D., Jay-ar, P.L., Maria Jr, L.B.S., et al.: Comparison of machine learning algorithms in breast cancer prediction using the Coimbra dataset. Cancer 7(10), 23-1 (2019).

3. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine 34(2), 113-127 (2005).
4. Liu, L.: Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In: 2018 International Conference on Robots & Intelligent System (ICRIS), pp. 157-160. IEEE, May 26, (2018).
5. Huang, M.W., Chen, C.W., Lin, W.C., Ke, S.W., Tsai, C.F.: SVM and SVM ensembles in breast cancer prediction. PloS One 12(1), e0161501 (2017).
6. Saritas, I.: Prediction of breast cancer using artificial neural networks. Journal of Medical Systems 36, 2901-2907 (2012).
7. Qiu, Y., Chen, H., Dong, X., Lin, Z., Liao, I.Y., Tistarelli, M., Jin, Z.: Ifvit: Interpretable fixed-length representation for fingerprint matching via vision transformer. arXiv preprint arXiv:2404.08237 (2024).
8. Zhang, Y., Tiňo, P., Leonardis, A., Tang, K.: A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence 5(5), 726-742 (2021).
9. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. Advances in Neural Information Processing Systems 29 (2016).
10. Sav, S., Pyrgelis, A., Troncoso-Pastoriza, J.R., Froelicher, D., Bossuat, J.P., Sousa, J.S., Hubaux, J.P.: POSEIDON: Privacy-preserving federated neural network learning. arXiv preprint arXiv:2009.00349 (2020).