Chapter 8

Book Chapter

Identifying Breast Cancer Using Machine Learning Algorithms

Rajendra M. Jotawar *[1], Sonal Patange †[2], and Yashaswini M ‡[3]

[1]Assistant Professor, Acharya Institute of Technology
[2]Assistant Professor, Department of MCA, Jain College of Engineering
[3]Department of MCA, Acharya Institute of Technology

Abstract

Breast cancer is a leading cause of death for women in underdeveloped nations, where early detection and treatment are crucial. This study explores the effectiveness of various machine-learning techniques in breast cancer detection through image processing, including CNNs, transfer learning models (AlexNet, Inception V3), SVMs, and traditional algorithms like Extreme Gradient Boosting and Naive Bayesian classifiers. Optimization techniques such as Particle Swarm Optimization (PSO) are integrated to enhance performance. A comprehensive literature survey highlights existing methodologies and achievements, providing insights for future research in this critical domain.

Keywords: Breast cancer. Machine Learning. Convolutional Neural Networks. Transfer Learning. Optimization Techniques.

*Email: rajendra2842@acharya.ac.in Corresponding Author
†Email: sonalpatange@jainbgm.in
‡Email: m.yashaswin@gmail.com

1 Introduction

Early detection of breast cancer is crucial for improving patient outcomes, as it continues to be a major public health issue. Advanced algorithms that can greatly help in the early diagnosis of breast cancer have been made possible by the developments in machine learning, especially in the field of medical imaging systems. To detect breast cancer through image processing, this study investigates the effectiveness of several machine-learning techniques. Convolutional Neural Networks (CNNs) are a class of machine learning algorithms that have attracted a lot of attention because of their automatic learning of hierarchical features from unprocessed image data. Because of this capability, CNNs are especially well-suited for image classification tasks, such as medical imaging for cancer detection. CNN architectures that use transfer learning, such as Inception V3 and AlexNet, have shown impressive gains in performance, particularly when dealing with situations where there is a shortage of labeled data. Support Vector Machines (SVM) are another powerful tool in the arsenal of machine learning techniques. Known for their effectiveness in handling high-dimensional data and their ability to delineate complex decision boundaries, SVMs provide a robust method for breast cancer detection. Additionally, traditional machine learning algorithms such as Extreme Gradient Boosting (XGBoost) and Naive Bayesian classifiers offer a baseline for comparison, showcasing the trade-offs between model complexity and performance.

Furthermore, optimization techniques like Particle Swarm Optimization (PSO) can be integrated to enhance the performance of these algorithms, providing a comprehensive approach to the detection process. This research attempts to clarify these different algorithms' advantages and disadvantages in breast cancer detection by performing a comparative analysis of them. Understanding the difficulties in detecting tumors in breast tissue and investigating various image-processing techniques related to early detection are the main goals of this research. Our goal in doing this research is to use cutting-edge machine learning techniques to improve breast cancer diagnosis, which is an ongoing effort. Figure 1 shows Benign and Malignant Cells.

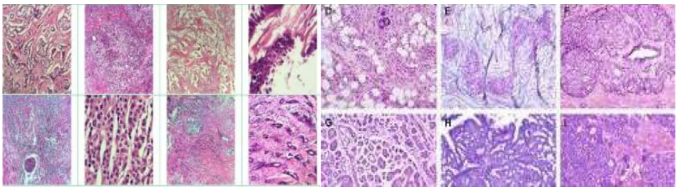

Figure 1. Benign and Malignant Cells

By leveraging the unique capabilities of each algorithm, this study seeks to provide a detailed comparative analysis, offering insights into their applicability and efficacy in medical imaging for breast cancer detection. The findings from this research will be instrumental in guiding the selection of appropriate machine-learning models for enhanced diagnostic accuracy and early intervention.

2 Literature Survey

Khalid et al.'s (2023) approach, which is hybrid and combines several machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF), and Convolutional Neural Networks (CNNs), yielded an accuracy of 98.7%. Their approach greatly increases diagnostic accuracy, combining clinical data and image processing methods. Utilizing each algorithm's unique strengths, the suggested framework surpassed separate algorithms, offering a more dependable and resilient solution for the identification of breast cancer. Zuo et al.'s (2023) conducted a comparative study on deep learning models for the early detection of breast cancer, published in IEEE Transactions on Medical Imaging. The study achieved an accuracy of 97.5% by evaluating the performance of various models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transfer Learning techniques. Their research highlights the particular effectiveness of transfer learning in enhancing model performance when dealing with limited datasets, demonstrating its potential to significantly improve the early detection and diagnosis of breast cancer. Machine Learning Algorithms have transformed medical imaging analysis by automatically extracting important data from digital mammograms. CNNs, a type of deep learning model, have received a lot of interest due to their ability to detect detailed patterns in picture data. For "Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features" got accuracy 76.25% (Wang et al., 2019)

Habibi's (2020) have used SVMs, XGBoost, and Naïve Bayes algorithms have showed potential in breast cancer classification. SVMs are ideal for binary classification tasks, correctly discriminating between benign and malignant tumors. XGBoost, a gradient boosting method, and Naïve Bayes, a probabilistic classifier, provide accurate and efficient algorithms for breast cancer classification. The Rizki Habibi for "Svm Performance Optimization Using PSO for Breast Cancer Classification" got accuracy 78.91%. In order to preprocess mammography images, Badriya Al Maqbali's (2021) used Contrast Limited Adaptive Histogram Equalization (CLAHE) in conjunction with a median filtering algorithm. Following preprocessing, the preprocessed data is passed to the region-growing algorithm for segmentation. Subsequently, feature extraction was used to extract features, including textures, gradient features, and geometric features. Hybrid Wolf Pack Algorithm and Particle Swarm Optimization, or hybrid WPA-PSO (Wolf Pack Algorithm

DOI:10.48001/978-81-966500-0-1-8 | ISBN: 978-81-966500-0-1 | Copyright ©2024 QTanalytics®

– Particle Swarm Optimization), has been used for feature selection. Finally, they used NN classifiers for classification. They have achieved 83.83% accuracy.

Mahesh et al.'s (2022) For XG Boost and Naïve Bayesian "Performance Analysis of XG Boost Ensemble Methods for Survivability with the Classification of Breast Cancer" got accuracy 81%. Leow et al.'s (2023) used Inception V3 and AlexNet for breast cancer classification with histopathological images, developing a CNN-based model to differentiate benign from malignant cases. They also tested five pre-trained CNN architectures, including ResNet-50, VGG-19, Inception V3, and AlexNet. ResNet-50 was used as a feature extractor for random forest and k-nearest neighbors classifiers. The accuracies achieved were 90% for Inception V3 and 81% for AlexNet.

Using FT-IR (Fourier-transform infrared spectroscopy) technology, Badriya Al Maqbali's (2021) classified sample data obtained from individuals with cervical cancer, CIN (cervical intraepithelial neoplasia) I, CIN II, CIN III, and hysteromyoma. They used the PSO-CNN model to do classification, and the accuracy achieved was 87.2%. Particle swarm optimization has been used for classification by Papasani et al.'s (2022).Decision tree learning (DTL), logistic regression (LR), Naive Bayes (N-Bayes), K-nearest neighbor (KNN), and other machine learning classifiers were used for the classification, and the outcomes were compared.

3 Methodology

Data Collection:

- Dataset Selection: The study utilizes publicly available datasets containing medical images of breast tissue, such as the Mammographic Image Analysis Society (MIAS) dataset and the Digital Database for Screening Mammography (DDSM).
- Data Preprocessing: The collected images undergo preprocessing techniques, including resizing, normalization, and augmentation, to ensure uniformity and enhanced model performance.

Algorithm Selection:

- Convolutional Neural Networks (CNNs): CNN architectures like AlexNet, Inception V3, and ResNet-50 are employed for their ability to automatically learn hierarchical features from raw image data. Transfer learning is utilized to leverage pre-trained models and improve performance, particularly in scenarios with limited labeled data.
- Support Vector Machines (SVMs): SVMs are chosen for their effectiveness in handling high-dimensional data and delineating complex decision boundaries, making them suitable for breast cancer detection tasks.
- Extreme Gradient Boosting (XGBoost): XGBoost, a gradient boosting method, is utilized for its robustness and efficiency in classification tasks.
- Naive Bayesian Classifiers: Naive Bayesian classifiers provide a probabilistic approach

to classification, serving as a baseline for comparison with more complex algorithms.

- Particle Swarm Optimization (PSO): PSO is integrated to improve classification accuracy and optimize certain algorithms' parameters, thereby enhancing their performance.

4 System Process

The collection consists of 2,77,524 RGB 50X50 pixel digitized picture patches that were taken from 162 H&E-stained breast histopathology samples. These are minuscule patches extracted from computer images of breast tissue samples.Benign (non-cancerous) and malignant (cancerous) cells are denoted by the numbers ”0” and ”1” in patches of cells, respectively. Lobular carcinoma, papillary carcinoma, mucinous carcinoma, and ductal carcinoma are the four types of malignant tumors. Benign tumors include Phyllodes tumor, Tubular adenoma, Fibroadenoma, and adenosis.

- Data Preprocessing: Preparing and dividing the breast cancer dataset into training and testing sets are the tasks performed by the Data Preprocessing Module. The functions include loading the dataset, preprocessing it using techniques like feature scaling and normalization, and splitting it into training and testing sets.
- Feature Extraction: It extracts relevant features from the breast cancer dataset, which are essential for training the machine learning models. Functions are Extracting features from raw data, Feature selection or dimensionality reduction techniques.
- Model Training: It trains machine learning models using various algorithms such as CNN, SVM, PSO, XGBoost, Naïve Bayesian, Inception V3, and AlexNet. Functions are Train CNN model, Train SVM model, Train PSO model, Train XGBoost model, Train Naïve Bayesian model, Train Inception V3 model, Train AlexNet model.
- The Model Evaluation: It assesses the efficacy of trained models by employing metrics such as sensitivity, specificity, F1-score, recall, accuracy, and precision. The functions include generating a confusion matrix, calculating sensitivity and specificity, and assessing model performance (accuracy, precision, recall, and F1-score).
  Accuracy = (FN+TP)/(TP+FP+TN+FN)
  Precision = TP/(TP+FP)
  Recall = TP/(TP+FN)
  F1 Score = 2* (Precision*Recall)/(Precision+Recall)
  where TP= True Positive,
  TN= True Negative, FP= False Positive,
  and FN= False Negative.
  Result Analysis: It analyzes and compares the results obtained from different machine learning algorithms to identify the most effective approach for breast cancer detection. Functions are Compare performance metrics across models, visualize results.
  The flow of the process can be better understood in figure 2

DOI:10.48001/978-81-966500-0-1-8 | ISBN: 978-81-966500-0-1 | Copyright ©2024 QTanalytics®

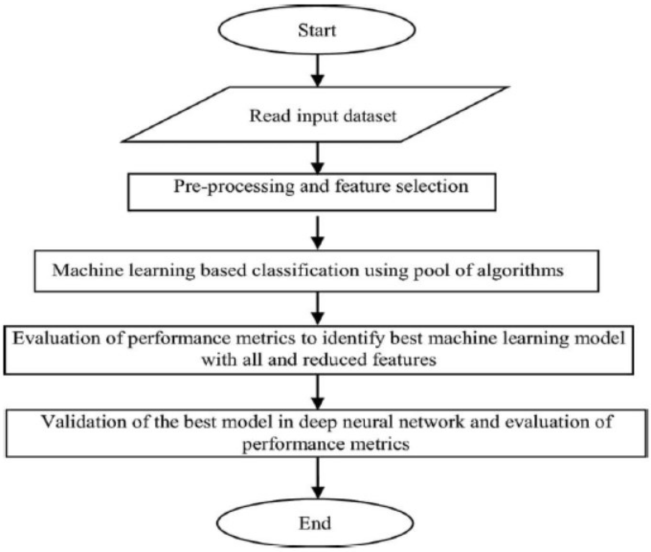
Start
Read input dataset
Pre-processing and feature selection
Machine learning based classification using pool of algorithms
Evaluation of performance metrics to identify best machine learning model with all and reduced features
Validation of the best model in deep neural network and evaluation of performance metrics
End

Figure 2. Flow of the Process

5 Conclusion

The comparative study on breast cancer detection using machine learning algorithms provide valuable insights into the effectiveness of various approaches for this critical task. Deep learning models like Convolutional Neural Networks (CNNs) and sophisticated techniques like Extreme Gradient Boosting (XGBoost) demonstrate high accuracy rates, albeit requiring substantial computational resources and data for effective training. Conversely, simpler models such as Naïve Bayesian offer respectable performance with reduced computational demands, making them suitable for resource-constrained environments or scenarios where interpretability is paramount. The potential of ensemble methods, exemplified by XGBoost, in integrating diverse models to enhance overall performance. By leveraging the strengths of individual algorithms, ensemble approaches can mitigate weaknesses and yield superior results. In practical applications, the choice of algorithm should align with the specific requirements of the healthcare setting, considering factors such as computational efficiency, interpretability, and scalability. For the field to advance and, eventually, improve patient outcomes, issues with data quality, model interpretability, and computational efficiency must be resolved. The healthcare industry can keep making important advancements in the early diagnosis and treatment of breast cancer by embracing innovation and teamwork.

References

Badriya Al Maqbali. (2021). Hybrid Wolf Pack Algorithm and Particle Swarm Optimization Algorithm for Breast Cancer Diagnosis. Multimedia Research, 4(3), 9–16. https://doi.org/10.46253/j.mr.v4i3.a2

Habibi, R. (2020). Svm Performance Optimization Using PSO for Breast Cancer Classification. Budapest International Research in Exact Sciences (BirEx) Journal, 3(1), 741–754. https://doi.org/10.33258/birex.v3i1.1499

Khalid, A., Mehmood, A., Alabrah, A., Alkhamees, B. F., Amin, F., AlSalman, H., & Choi, G. S. (2023). Breast Cancer Detection and Prevention Using Machine Learning. Diagnostics, 13(19). https://doi.org/10.3390/diagnostics13193113

Leow, J. R., Khoh, W. H., Pang, Y. H., & Yap, H. Y. (2023). Breast cancer classification with histopathological image based on machine learning. International Journal of Electrical and Computer Engineering, 13(5), 5885–5897. https://doi.org/10.11591/ijece.v13i5.pp5885-5897

Mahesh, T. R., Vinoth Kumar, V., Muthukumaran, V., Shashikala, H. K., Swapna, B., & Guluwadi, S. (2022). Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer. Journal of Sensors, 2022. https://doi.org/10.1155/2022/4649510

Papasani, A., Devarakonda, N., Polkowski, Z., Thotakura, M., & Bhagya Lakshmi, N. (2022). Feature Selection Using PSO Optimized-Framework with Machine Learning Classification System via Breast Cancer Survival Data, 513–531. https://doi.org/10.1007/978-981-16-9573-5_38

Wang, Z., Li, M., Wang, H., Jiang, H., Yao, Y., Zhang, H., & Xin, J. (2019). Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion with CNN Deep Features. IEEE Access, 7, 105146–105158. https://doi.org/10.1109/ACCESS.2019.2892795

Zuo, D., Yang, L., Jin, Y., Qi, H., Liu, Y., & Ren, L. (2023). Machine learning-based models for the prediction of breast cancer recurrence risk. BMC Medical Informatics and Decision Making, 23(1). https://doi.org/10.1186/s12911-023-02377-z

DOI:10.48001/978-81-966500-0-1-8 | ISBN: 978-81-966500-0-1 | Copyright ©2024 QTanalytics®