Review Article

ACCENTS Transactions on Image Processing and Computer Vision, Vol 10(26)
ISSN (Online): 2455-4707
http://dx.doi.org/ 10.19101/ TIPCV.2023.924003

Machine learning and data mining for breast cancer detection: a comprehensive review

Manish Singh[*] and Animesh Kumar Dubey
Department of Computer Science and Engineering, Patel College of Science and Technology, Bhopal, Madhya Pradesh, India

Received: 20-September-2023; Revised: 18-December-2023; Accepted: 20-December-2023

Abstract

Breast cancer remains a pervasive global health concern, contributing significantly to cancer-related morbidity and mortality among women. Traditional diagnostic methods, such as mammography and clinical breast exams, though valuable, possess limitations, including sensitivity issues and the risk of false positives. In response to these challenges, the emergence of data mining and machine learning technologies has opened new avenues for breast cancer detection. This review examines the application of data mining and machine learning approaches in breast cancer detection and analysis, emphasizing recent advancements and critical findings. The review included an analysis and discussion of the effectiveness of these technologies in improving diagnostic accuracy, the examination of commonly algorithms, and the identification of research gaps. The review highlights the transformative potential of data-driven medical diagnostics, offering valuable insights for researchers, clinicians, and policymakers.

Keywords

Breast cancer detection, Data mining and machine learning, Algorithm analysis, Medical diagnostic innovation.

1.Introduction

Breast cancer remains one of the most prevalent and devastating forms of cancer affecting women worldwide [1]. It accounts for a significant proportion of cancer-related morbidity and mortality, making it a critical public health concern [2–5]. Despite advancements in healthcare, the early detection and accurate diagnosis of breast cancer continue to pose challenges. Traditional methods, such as mammography and clinical breast exams, have been instrumental in screening efforts but are not without limitations [6, 7]. These include issues of sensitivity, particularly in dense breast tissue, and the risk of false positives leading to unnecessary interventions [6–10]. The advent of data mining and machine learning (ML) technologies offers a transformative potential in breast cancer detection [11, 12]. These advanced computational approaches, such as linear regression, logistic regression, support vector machine (SVM), Naïve Bayes (NB), decision tree (DT), k-nearest neighbors (KNN), clustering (k-Means, fuzzy c-Means), random forest (RF), and Apriori, can analyze complex and voluminous datasets. They extract meaningful patterns and make predictions with a level of accuracy and efficiency often surpassing human capabilities [13–17]. This shift towards a more data-driven approach in medical diagnostics is motivated by the need for higher accuracy, objectivity, and the capability to leverage large-scale health data for personalized medicine [18–20].

This review is conducted with the methodological rigor to scrutinize and synthesize the burgeoning body of literature on the application of data mining and machine learning in breast cancer detection. The selection criteria for the literature included peer-reviewed articles published within a specified timeframe, focusing on a range of data mining and machine learning techniques applied to breast cancer detection. The review process involves a critical evaluation of the methodologies, outcomes, and effectiveness of these studies, providing a comprehensive and unbiased overview of the current landscape.

The primary objective of this review is to collate and analyze existing research findings on the application of data mining and machine learning in breast cancer detection. This involves assessing the effectiveness of

[*]Author for correspondence

these technologies in improving diagnostic accuracy, understanding the types of data and algorithms most commonly used, and identifying any gaps or inconsistencies in the current research. This approach aims to provide a coherent and comprehensive picture of the current state of the art, offering valuable insights for researchers, clinicians, and policymakers.

The contributions of this review include providing a consolidated reference for the latest advancements in breast cancer detection using data mining and machine learning, aiding the research community in staying informed about current developments. It also identifies promising methodologies and technologies that could pave the way for future innovations in cancer diagnostics. Highlighting the gaps and challenges in the current research, this review sets the stage for future studies, potentially guiding the direction of subsequent investigations and technological advancements.

This systematic review seeks to bridge the gap between the rapidly evolving field of data mining and machine learning and the critical domain of breast cancer detection. By providing a thorough analysis of existing research and pointing out potential future directions, this study aims to contribute significantly to the enhancement of breast cancer diagnostic methods, ultimately aiding in the fight against this global health challenge. *Figure 1* illustrates the major areas where machine learning and data mining are applied in the context of breast cancer detection and analysis.

The structure of this paper is organized as follows: Section 2 presents a thorough literature review. Section 3 provides a discussion based on the related work, and Section 4 concludes the paper.

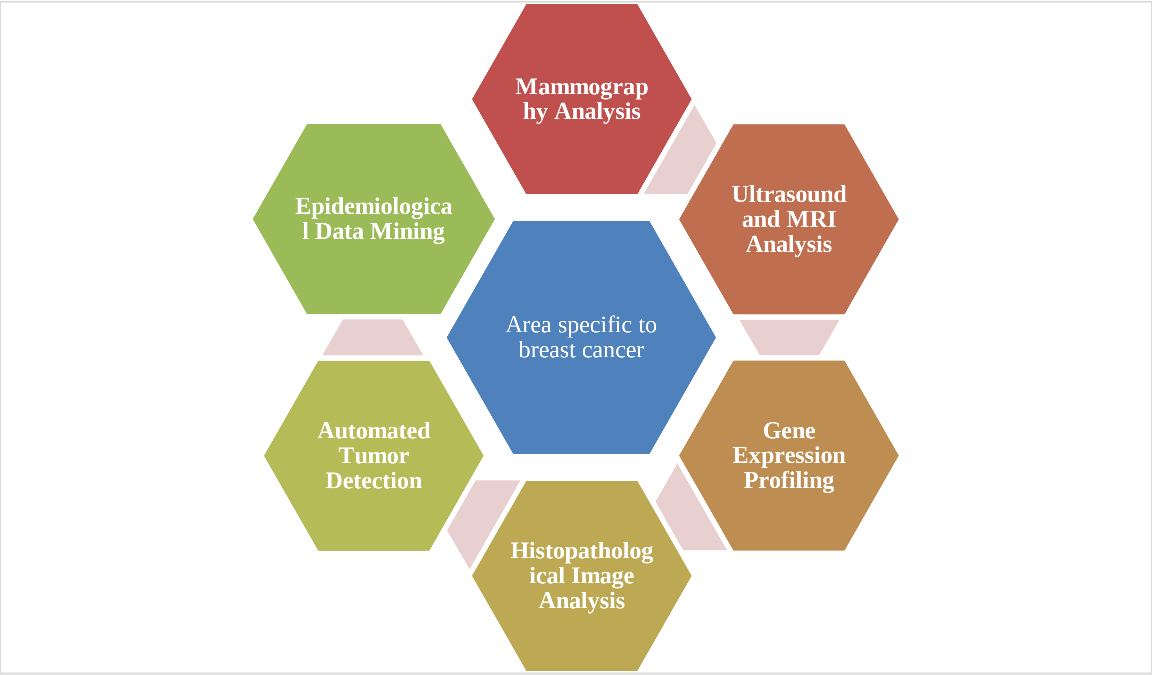
Mammograp hy Analysis
Ultrasound and MRI Analysis
Epidemiologica l Data Mining
Area specific to breast cancer
Automated Tumor Detection
Gene Expression Profiling
Histopatholog ical Image Analysis

Figure 1 Major areas where machine learning and data mining are applied in the context of breast cancer detection and analysis

2.Literature review

The 2022 study by Rovshenov and Peker [21] utilized artificial intelligence for early breast cancer detection, using artificial neural network (ANN), SVM, and RF algorithms on the Wisconsin dataset. The ANN achieved 99% accuracy. Advantages include efficient early detection and wider accessibility. Limitations involve dataset specificity and the need for long-term validation of results.

In 2022, Jiang et al. [22] designed the Photoacoustic Pen (PAPen) for breast cancer surgery, enhancing sentinel lymph node detection. Using photoacoustic sensing for deep penetration, the PAPen combines an ultrasound transducer and optical fiber, accurately locating lymph nodes by capturing photoacoustic signals from contrast agents, with a detection depth of up to 50 mm.

In 2022, Basha and Sindhu [23] evaluated breast cancer prediction using the anisotropic diffusion algorithm (ADA) against the variational partial differential equation (PDE) Method. On a health dataset, ADA (p=0.001) outperformed variational PDE, achieving 94% accuracy compared to 93%. The study highlights ADA's superior precision and accuracy in predicting breast cancer.

Nelli [24] highlighted the need for improved early breast cancer detection, citing limitations of current methods like mammography. Their review focused on gold nanoparticles (AuNPs) and biosensors in nanotechnology, offering advanced, sensitive diagnostic solutions. AuNPs are gaining traction for their potential in revolutionizing cancer diagnosis, monitoring, and treatment.

In 2022, Dubey et al. [25] explored eight machine learning techniques for breast cancer detection using the Wisconsin Diagnostic dataset. Focusing on differentiation between benign and malignant tumors, the study found SVM and ANN to be the most effective, achieving an accuracy of 98.08% in their performance assessment.

In 2023, Botlagunta et al. [26] developed a non-invasive machine learning system for diagnosing metastatic breast cancer (MBC) using blood profile data from EMRs. Utilizing Python and machine learning techniques, they achieved a Decision Tree classifier accuracy of 83% with an AUC of 0.87. This system could help physicians in early MBC detection.

In 2023, Nemade and Fegade [27] explored machine learning algorithms for breast cancer prediction, using classifiers like Naïve Bayes, Logistic Regression, SVM, KNN, DT, and ensembles like RF, Adaboost, and XGBoost. Their study found the Decision Tree and XGBoost to have the highest accuracy (97%) with XGBoost achieving an AUC of 0.999.

In 2023, Sugimoto et al. [28] conducted a review on the advancements and applications of machine learning ML in breast cancer. The review discusses the impact of machine learning on image processing, especially deep and convolutional neural networks for tumor and lymph node detection. It also covers the use of quantitative omics techniques for molecular pathology understanding and the development of new biomarkers. Additionally, the review highlights the role of machine learning in analyzing clinical-pathological features and predicting treatment outcomes, ultimately underscoring machine learning’s contribution to personalized breast cancer medicine.

Elsadig et al. [29] highlighted the global challenge of breast cancer in women, emphasizing the role of socioeconomic barriers in late diagnoses. The review stresses the importance of collaborative healthcare efforts and explores various machine learning algorithms for improving accuracy and early detection of breast cancer, aiming to enhance women's health outcomes.

In their research, Manikandan et al. [30] aimed to classify the survival status of breast cancer patients using the SEER dataset. The study utilized machine learning and deep learning for data preprocessing and analysis. Employing Variance Threshold and Principal Component Analysis for feature selection, they applied techniques like Ada Boosting, XG Boosting, and DT. The DT algorithm excelled with 98% accuracy, outperforming others in both train-test split and k-fold cross-validation approaches.

Ebrahim et al. [31] compared various machine learning algorithms for breast cancer prediction using a 1.7 million record dataset from the National Cancer Institute. The study included classical methods like DT, Logistic regression, and ensemble techniques, as well as deep learning approaches like deep neural networks. DT and ensemble techniques emerged as the most accurate, each achieving 98.7% accuracy.

3.Discussion and analysis

The comprehensive review of machine learning and data mining applications in breast cancer detection indicates a paradigm shift in medical diagnostics. Traditional methods like mammography and clinical breast exams, while foundational, present limitations such as sensitivity issues and false positives. In contrast, machine learning and data mining techniques demonstrate transformative potential, largely attributed to their ability to process complex and voluminous datasets, uncover patterns, and predict outcomes with remarkable accuracy.

The literature indicates a growing trend towards leveraging these advanced computational methods for enhancing diagnostic precision in breast cancer. This approach aligns with the broader objective of personalized medicine, aiming to tailor healthcare based on individual characteristics. Studies reviewed encompass a diverse range of algorithms and applications, from early detection to the prediction of metastatic progression, highlighting the versatility and depth of ML in this field.

Advantages of Machine Learning and Data Mining in Breast Cancer Detection

Enhanced Diagnostic Accuracy: machine learning algorithms, due to their computational power, have shown higher accuracy in detecting breast cancer compared to traditional methods. This accuracy is pivotal in early detection, which is crucial for effective treatment.

Efficiency in Handling Large Datasets: The ability of machine learning to analyze large-scale health data enables the extraction of meaningful insights that might be overlooked by human analysis.

Predictive Analytics for Personalized Medicine: The predictive capabilities of these algorithms facilitate personalized treatment plans by analyzing patient data and predicting disease progression or treatment outcomes.

Versatility in Applications: The diverse range of algorithms, from SVMs to deep learning models, allows for varied applications in different aspects of breast cancer detection and treatment.

Reduced Human Error: Automation in diagnostics reduces the risk of human error, leading to more reliable and consistent outcomes.

Limitations and Challenges

Data Dependency and Quality Issues: The effectiveness of machine learning models is heavily reliant on the quantity and quality of data. Poor data quality or biased datasets can lead to inaccurate predictions.

Complexity and Interpretability: Some machine learning models, especially deep learning algorithms, are often seen as 'black boxes' due to their complexity, making it difficult to interpret how decisions are made.

Generalizability of Models: Models trained on specific datasets may not perform well when applied to different populations or data types, limiting their generalizability.

Resource Intensity: Developing, training, and maintaining machine learning models require significant computational resources and expertise, which may not be readily available in all healthcare settings.

Ethical and Privacy Concerns: The use of patient data in machine learning raises concerns regarding privacy and ethical use of information, necessitating robust governance frameworks.

4.Conclusion

This comprehensive review highlights the transformative impact of data mining and machine learning in breast cancer detection and analysis. These advanced computational approaches offer higher accuracy, objectivity, and the ability to leverage large-scale health data for personalized medicine. The discussed studies showcase the versatility of machine learning algorithms and their potential to improve breast cancer diagnostics. While challenges and limitations exist, the findings indicate the importance of ongoing research and collaboration in advancing breast cancer detection methods. This review aims to contribute to the ongoing efforts to combat breast cancer, ultimately leading to improved patient outcomes and a reduction in the global burden of this devastating disease.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Ranjbarzadeh R, Dorosti S, Ghoushchi SJ, Caputo A, Tirkolaee EB, Ali SS, et al. Breast tumor localization and segmentation using machine learning techniques: overview of datasets, findings, and methods. Computers in Biology and Medicine. 2023; 152:106443.

[2] Babichev S, Yasinska-Damri L, Liakh I. A hybrid model of cancer diseases diagnosis based on gene expression data with joint use of data mining methods and machine learning techniques. Applied Sciences. 2023; 13(10):6022.

[3] Dubey AK, Gupta U, Jain S. Breast cancer statistics and prediction methodology: a systematic review and analysis. Asian Pacific Journal of Cancer Prevention. 2015; 16(10):4237-45.

[4] Liza FT, Das MC, Pandit PP, Farjana A, Islam AM, Tabassum F. Machine learning-based relative performance analysis for breast cancer prediction. In world AI IoT congress (AIIoT) 2023 (pp. 0007-0012). IEEE.

[5] Kiliç AE, Karakoyun M. Breast cancer detection using machine learning algorithms. International Journal of Advanced Natural Sciences and Engineering Researches. 2023; 7(3):91-5.

[6] Akhtar N, Pant H, Dwivedi A, Jain V, Perwej Y. A breast cancer diagnosis framework based on machine learning. International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET). 2023:2395-1990.

[7] Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. International Journal of Computer Assisted Radiology and Surgery. 2016; 11:2033-47.

[8] Nemade V, Pathak S, Dubey AK, Barhate D. A review and computational analysis of breast cancer using different machine learning techniques. International Journal of Emerging Technology and Advanced Engineering. 2022; 12(3):111-8.

[9] Kuruba C, Pushpalatha N, Ramu G, Suneetha I, Kumar MR, Harish P. Data mining and deep learning-based hybrid health care application. Applied Nanoscience. 2023; 13(3):2431-7.

[10] Al-Dmour NA, Said RA, Alzoubi HM, Alshurideh M, Ali L. Breast cancer prediction using machine learning and image processing optimization. In the effect of information technology on business and marketing intelligence systems 2023 (pp. 2067-79). Cham: Springer International Publishing.

[11] Ramakrishna MT, Venkatesan VK, Izonin I, Havryliuk M, Bhat CR. Homogeneous adaboost ensemble machine learning algorithms with reduced entropy on balanced data. Entropy. 2023; 25(2):245.

[12] Wu R, Luo J, Wan H, Zhang H, Yuan Y, Hu H, et al. Evaluation of machine learning algorithms for the prognosis of breast cancer from the Surveillance, epidemiology, and end results database. Plos one. 2023; 18(1):e0280340.

[13] Nemade V, Pathak S, Dubey AK. Deep learning-based ensemble model for classification of breast cancer. Microsystem Technologies. 2023:1-5.

[14] Sharma A, Hooda N, Gupta NR, Sharma R. Efficient RIEV: a novel framework for the prediction of breast cancer cases using ensemble machine learning. Network Modeling Analysis in Health Informatics and Bioinformatics. 2023; 12(1):29.

[15] Dubey A, Gupta U, Jain S. Medical data clustering and classification using TLBO and machine learning algorithms. Computers, Materials and Continua. 2021; 70(3):4523-43.

[16] Avcı H, Karakaya J. A novel medical image enhancement algorithm for breast cancer detection on mammography images using machine learning. Diagnostics. 2023; 13(3):348.

[17] Liu J, Lei J, Ou Y, Zhao Y, Tuo X, Zhang B, et al. Mammography diagnosis of breast cancer screening through machine learning: a systematic review and meta-analysis. Clinical and Experimental Medicine. 2023; 23(6):2341-56.

[18] Sun X, Qourbani A. Combining ensemble classification and integrated filter-evolutionary search for breast cancer diagnosis. Journal of Cancer Research and Clinical Oncology. 2023:1-7.

[19] Rashed AE, Elmorsy AM, Atwa AE. Comparative evaluation of automated machine learning techniques for breast cancer diagnosis. Biomedical Signal Processing and Control. 2023; 86:105016.

[20] Prajapati JB, Paliwal H, Prajapati BG, Saikia S, Pandey R. Quantum machine learning in prediction of breast cancer. In quantum computing: a shift from bits to qubits 2023 (pp. 351-82). Singapore: Springer Nature Singapore.

[21] Rovshenov A, Peker S. Performance comparison of different machine learning techniques for early prediction of breast cancer using Wisconsin breast cancer dataset. In 3rd international informatics and software engineering conference (IISEC) 2022 (pp. 1-6). IEEE.

[22] Jiang D, Zhao J, Zhang Y, Cong B, Shen Y, Gao F, et al. Integrated photoacoustic pen for breast cancer sentinel lymph node detection. In international ultrasonics symposium (IUS) 2022 (pp. 1-3). IEEE.

[23] Basha HM, Sindhu G. Improved accuracy of early stage breast cancer detection using anisotropic diffusion algorithm and Variational partial differential equation method. In international conference on sustainable computing and data communication systems (ICSCDS) 2022 (pp. 1683-9). IEEE.

[24] Nelli S. Prediction of Early Stage Breast Cancer by Injection of Gold Nano Particles and Analyzing Images using Data Analytics. In 2nd international conference on mobile networks and wireless communications (ICMNWC) 2022 (pp. 1-5). IEEE.

[25] Dubey C, Shukla N, Kumar D, Singh AK, Dwivedi VK. Breast cancer modeling and prediction combining machine learning and artificial neural network approaches. In international conference on computing, communication, and intelligent systems (ICCCIS) 2022 (pp. 119-24). IEEE.

[26] Botlagunta M, Botlagunta MD, Myneni MB, Lakshmi D, Nayyar A, Gullapalli JS, et al. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Scientific Reports. 2023; 13(1):485.

[27] Nemade V, Fegade V. Machine learning techniques for breast cancer prediction. Procedia Computer Science. 2023; 218:1314-20.

[28] Sugimoto M, Hikichi S, Takada M, Toi M. Machine learning techniques for breast cancer diagnosis and treatment: a narrative review. Annals of Breast Surgery. 2023; 7.

[29] Elsadig MA, Altigani A, Elshoush HT. Breast cancer detection using machine learning approaches: a comparative study. International Journal of Electrical & Computer Engineering (2088-8708). 2023; 13(1).

[30] Manikandan P, Durga U, Ponnuraja C. An integrative machine learning framework for classifying SEER breast cancer. Scientific Reports. 2023; 13(1):5362.

[31] Ebrahim M, Sedky AA, Mesbah S. Accuracy assessment of machine learning algorithms used to predict breast cancer. Data. 2023; 8(2):35.

Manish Singh is currently pursuing an M.Tech in Computer Science at Patel College of Science and Technology, RGPV, Bhopal, Madhya Pradesh. He completed his Bachelor of Engineering (B.E.) in Computer Science Engineering at the same institution, PCST, RGPV University, in Bhopal, MP. His areas of interest include Python Development, AWS (EC2, Lambda Functions), Machine Learning, and Web Applications Development.
Email: msingh21497@gmail.com

Animesh Kumar Dubey is currently serving as an Assistant Professor in the Department of Computer Science and Engineering at Patel College of Science and Technology, RGPV, Bhopal, Madhya Pradesh, India. He holds a Bachelor of Engineering (B.E.) and an M.Tech. degree in Computer Science Engineering from Rajiv Gandhi Technical University, Bhopal, Madhya Pradesh. With over 15 publications in reputable, peer-reviewed national and international journals and conferences, his expertise extends across a range of subjects. His primary research interests include Data Mining, Optimization, Machine Learning, Cloud Computing, and Artificial Intelligence.
Email: animeshdubey123@gmail.com