Transactions on Machine Intelligence 5(4) (2022) 277–285

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI

Improving Accuracy in Breast Cancer Diagnosis Using Data Mining Techniques

F. Fatahi[1,*]

[1] Department of Computer Engineering, Faculty of Technical and Engineering, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran.

ARTICLE INFO

Article History:
Received 4 August 2022
Received in revised form 19 November 2022
Accepted 29 December 2022
Available online 30 December 2022

Keywords:
Data Mining, Disease Diagnosis, Breast Cancer, Principal Component Analysis, Regression and Classification Trees, Multilayer Perceptron.

ABSTRACT

This study introduces a novel, integrated approach for breast cancer diagnosis, addressing one of the most critical challenges in medical sciences: the lack of timely and precise detection. Breast cancer remains a leading cause of mortality worldwide, and early diagnosis plays a pivotal role in improving survival rates. Currently, diagnostic practices heavily rely on physicians' expertise, supported by complex and time-consuming laboratory tests, which are prone to human error and often lead to delays in treatment. To overcome these limitations, this research proposes a comprehensive methodology that combines principal component analysis (PCA) for dimensionality reduction, decision trees for feature selection, and artificial neural networks (ANNs) for classification and prediction. By integrating these techniques, the proposed system optimizes the use of database features, offering an adaptable, efficient, and accurate solution for breast cancer detection. The results demonstrate that this method achieves superior diagnostic accuracy compared to conventional techniques and existing artificial intelligence-based methods referenced in related studies. Furthermore, the system significantly reduces diagnostic costs and time without compromising performance. This research highlights the potential of combining machine learning and data mining techniques to enhance diagnostic precision, providing researchers and clinicians with an effective tool for improving early detection, treatment planning, and patient outcomes.

1. INTRODUCTION

In modern medical science, the collection of vast amounts of data on various diseases has become increasingly important. Medical centers gather these datasets for diverse purposes, including clinical research and patient management. Analyzing these datasets and extracting meaningful patterns and insights about diseases is one of the key objectives of medical data utilization. However, the overwhelming volume of medical data often leads to difficulties in identifying valuable patterns, hindering the extraction of significant results. To address this challenge,

* Corresponding Author: fatahi.fereshteh.f@gmail.com
Department of Computer Engineering, Faculty of Technical and Engineering, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran

http://dx.doi.org/10.47176/TMI.2022.277

data mining techniques are employed to uncover useful relationships between risk factors and disease prevalence, particularly for conditions that contribute significantly to human mortality [1].

Breast cancer is a type of malignancy that originates in breast tissue [2]. It is recognized as one of the leading causes of cancer-related mortality among women. Early diagnosis and timely treatment are crucial in reducing mortality rates and improving patient survival. Mammography plays a pivotal role in the early detection of breast cancer [3]. In cases where cancer has metastasized to other parts of the body, treatments are primarily aimed at enhancing the patient's quality of life and providing palliative care [4]. The survival rate for breast cancer patients in developed countries is relatively high, with approximately 80% to 90% of patients in the United Kingdom and the United States surviving for at least five years post-diagnosis [5,6].

According to medical experts, approximately 8,000 new cases of breast cancer are diagnosed annually in Iran, with an incidence rate of about 30 to 35 cases per 100,000 women [7]. The dataset used in this study is obtained from the Wisconsin Breast Cancer Database, which includes 699 samples and 10 key features [8]. These features encompass various clinical and histopathological attributes that play a significant role in breast cancer diagnosis and classification.

Table 1 provides an overview of the features present in the Wisconsin Breast Cancer Database, which is widely used for breast cancer classification. The dataset consists of 699 samples and includes 10 key attributes, as outlined below:

Table 1. Database Description

No.	Description	Feature Name
1	Clump Thickness	Clump Thickness
2	Uniformity of Cell Size	Uniformity of Cell Size
3	Uniformity of Cell Shape	Uniformity of Cell Shape
4	Marginal Adhesion	Marginal Adhesion
5	Single Epithelial Cell Size	Single Epithelial Cell Size
6	Bare Nuclei	Bare Nuclei
7	Bland Chromatin	Bland Chromatin
8	Normal Nucleoli	Normal Nucleoli
9	Mitoses	Mitoses
10	Class	Class

1.1. Data Mining Techniques for Breast Cancer Diagnosis

Clementine software is one of the leading tools in data mining, providing comprehensive support for all stages of the data mining process. In this study, Clementine 12.0 was used for analyzing and evaluating data mining techniques applied to the Wisconsin Breast Cancer dataset [9].

2. LITERATURE REVIEW

Numerous studies have been conducted to improve breast cancer detection and enhance diagnostic accuracy using data mining techniques. Some notable studies include:

Sayahi and Ashir employed the Wisconsin Hospital standard database, which contains 286 samples and 10 features, to evaluate various classification methods, including PART, Bagging, Radial Basis Function (RBF) networks, and Logistic Regression. Their study demonstrated improved accuracy through pruning techniques [10].

Nazarian et al. utilized the Wisconsin Breast Cancer dataset (699 samples and 10 features) and proposed a breast cancer diagnosis model inspired by bee colony optimization. Their method achieved remarkable results in classification accuracy [11].

Qasem Ahmed et al. worked with 547 records containing 22 attributes related to breast cancer patients from the Jihad University dataset. They applied decision trees (C5), support vector machines (SVM), and artificial neural

networks (ANN) for breast cancer classification. Their findings indicated that boosting and pruning techniques significantly enhanced classification accuracy [12].

Kiani and Atashi analyzed data from 995 breast cancer patients, incorporating 18 predictive factors per patient. They used SPSS.V20 software and developed a breast cancer recurrence prediction model based on the J48 decision tree algorithm [13].

Delen et al. leveraged large-scale datasets and employed artificial neural networks (ANNs), decision trees, and logistic regression models to predict breast cancer occurrence [14].

These studies demonstrate the effectiveness of various data mining techniques in improving breast cancer diagnosis, highlighting the potential for integrating machine learning approaches into clinical decision-making.

One of the innovative approaches in breast cancer detection involves the use of thermal imaging and advanced algorithms for segmenting cancerous regions. Ghayoumi Zadeh et al. (2017) proposed a fully automated method based on "fuzzy active contours" designed with fuzzy logic. This method utilized thermography images to segment the edges and central core of cancerous areas in 60 patients with an average age of 44.9 years. The results indicated that the Hausdorff distance between the manual and automated methods for the thermal core and edge was 0.4719 ± 0.4389 mm and 0.3171 ± 0.1056 mm, respectively, with an overall system accuracy of 91.98% and a sensitivity of 85%. This study highlights the significant potential of thermography as a non-invasive and rapid diagnostic tool, though it emphasizes the need for improved accuracy in more complex scenarios [15].

In the realm of mammography, the most common screening method, Rahmani Seryasat and Haddadnia (2017) introduced a novel framework based on ensemble learning for classifying benign and malignant masses. Their approach involved noise reduction, segmentation using a deformable model, extraction of features such as shape, edge, and fractal dimension, and the selection of an optimal feature subset using a genetic algorithm. The proposed architecture identified easy and difficult samples, training them with different classifiers. Tested on the mini-MIAS and DDSM databases, the system demonstrated competitive accuracy with state-of-the-art methods, although specific numerical results were not provided in the abstract. This study underscores the importance of feature selection and classifier combination, but the lack of quantitative details hinders direct comparison with other methods [16].

Similarly, Rahmani-Seryasat et al. (n.d.) proposed a new method for classifying breast cancer tumors using a neural network and the "growth area" technique. This approach employed a fuzzy adaptive threshold based on entropy to extract statistical features and spatial dependencies, separating tumors from normal tissue. Using the MIAS database with 238 images, the method achieved an accuracy of 86.66% for detecting abnormal masses and 38.05% for normal masses. The low accuracy for normal masses suggests potential weaknesses in data balance or method sensitivity, warranting further analysis [17].

Finally, Rahmani Seryasat et al. (n.d.) introduced another CAD system for mammography, utilizing an adaptive region-growing algorithm, noise reduction, and a combination of weak and strong classifiers. Extracted features included edge and texture properties, leveraging empirical mode functions. Tested on the mini-MIAS and DDSM databases, this system also showed competitive accuracy. The innovation lies in its segmentation algorithm and classifier architecture, but, similar to the previous study, the absence of specific numerical results limits precise evaluation [18].

Overall, these studies demonstrate progress in applying artificial intelligence to breast cancer detection. Thermography and mammography each offer unique advantages: thermography is non-invasive, while mammography provides high screening accuracy. However, challenges remain, such as low accuracy in specific cases (e.g., normal masses), insufficient methodological details, and the need for more comprehensive comparisons with existing methods. This literature review highlights the necessity of developing hybrid approaches and enhancing data analysis to improve the precision and reliability of CAD systems.

2.1. Proposed Methodology

This study aims to introduce an innovative data-driven approach to enhance the accuracy of breast cancer diagnosis. Compared to existing research, the proposed solution must achieve higher predictive accuracy while also identifying key factors influencing the diagnosis of this disease.

2.2. Proposed Method

The proposed algorithm consists of three main stages. In the first stage, the dataset is imported for preprocessing and feature selection.

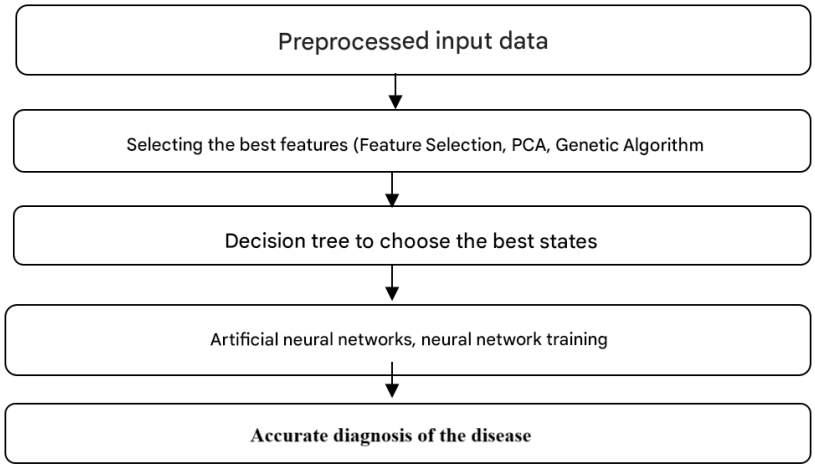

Fig. 1. Execution Process of the Proposed System

The data preprocessing phase is the most crucial and time-consuming stage in data mining projects. Since the input data serve as the foundation of the project, the accuracy of the output heavily depends on the precision of the input.

2.3. Data Normalization

Normalization is a process of rescaling data, which can be performed using various methods. It is particularly useful for classification algorithms such as neural networks and distance-based methods like k-nearest neighbors (KNN) and clustering. Normalization ensures that large-scale data values do not disproportionately influence the final results. In this study, Z-score normalization is applied to standardize the dataset.

2.4. Classification and Regression Trees (CART)

CART is a decision tree implementation used to classify and predict future observations. This method minimizes impurity within each category. A node is considered completely pure when all its subgroup elements belong to a single target class. Predictor fields and target fields can be either numerical or categorical. Additionally, all splits are binary, meaning each node is divided into only two subgroups.

2.5. Multi-Layer Perceptron (MLP) Neural Network

The neural network model employed in this study is a multi-layer perceptron (MLP). The backpropagation algorithm is used for training the network. One of the major challenges in neural networks is determining the optimal network structure based on the available data. To address this issue, a rapid training strategy is adopted. This approach ensures that the network topology is appropriately configured by the software while reducing the number

of hidden layers to facilitate faster training and better generalization. The network weights are updated using the backpropagation training algorithm.

2.6. Principal Component Analysis (PCA) for Feature Reduction

Since PCA only retains the most significant features, applying this technique is essential for dimensionality reduction. However, PCA must be implemented with caution to prevent excessive loss of data accuracy. The results obtained from applying PCA should be comparable to those achieved without it, ensuring that data precision remains intact.

2.7. Final Model Implementation in Clementine Software

The final model is developed using Clementine software. Confusion matrices are included for evaluating the CART decision tree and the MLP neural network. Additionally, an analysis node is incorporated to assess the predictive capabilities of the models, ensuring accurate predictions. The final model diagram is presented in Figure 2.

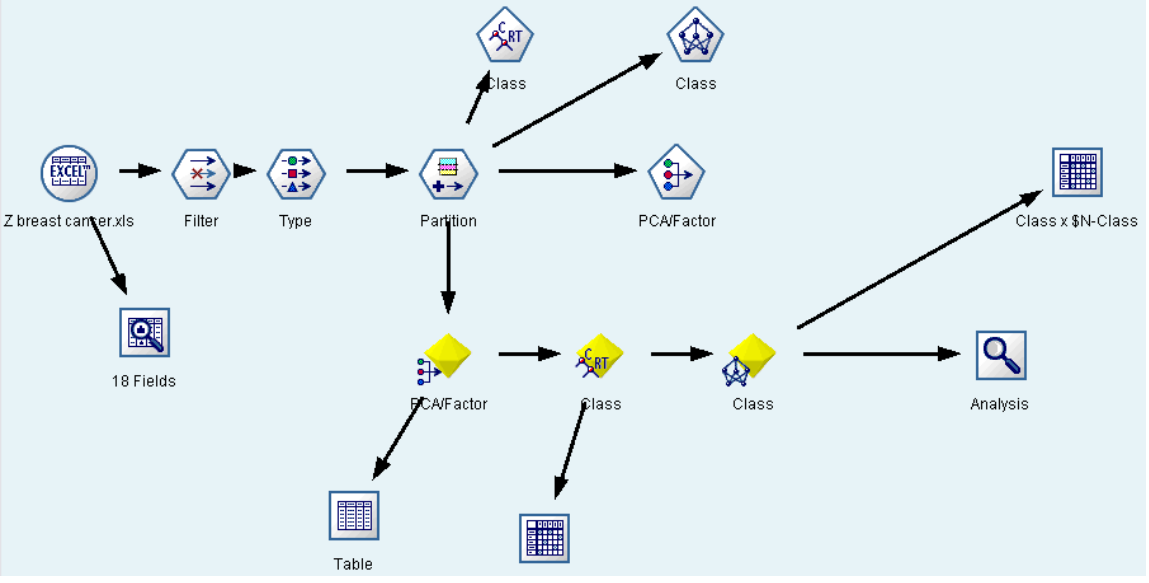

Fig. 2. Final Diagram of the Proposed Method

Additionally, the feature importance levels in the generated C&R decision tree are illustrated in Figure 3.

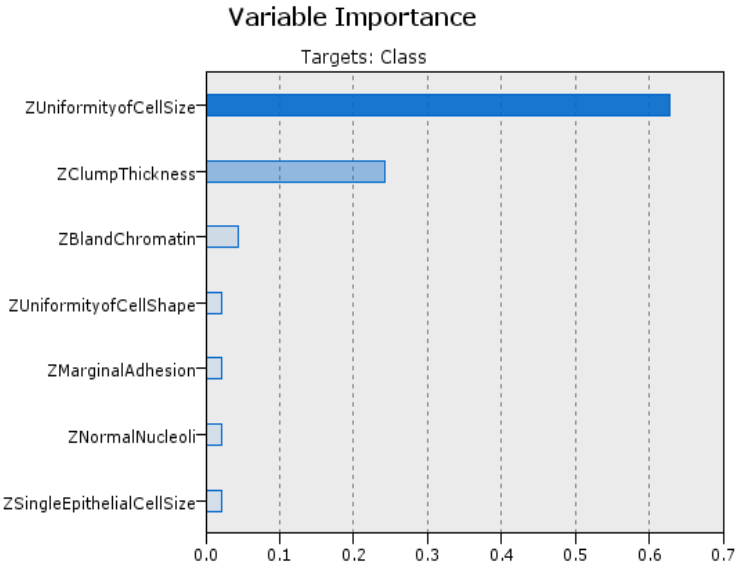

Fig. 3. Feature Importance Levels

3. ANALYSIS AND EVALUATION OF RESULTS

A confusion matrix, also known as an occurrence matrix, is a visual tool used to illustrate classification accuracy. It displays the relationship between actual samples and predicted samples.

Table 2. Confusion Matrix

Predicted Class	Benign (Benign)	Malignant (Malignant)
Actual Class		
Benign	FN	TP
Malignant	TN	FP

- **TP (True Positives):** The number of benign samples correctly classified as benign.
- **TN (True Negatives):** The number of malignant samples correctly classified as malignant.
- **FP (False Positives):** The number of malignant samples incorrectly classified as benign.
- **FN (False Negatives):** The number of benign samples incorrectly classified as malignant.

3.1. Performance Metrics

- **Accuracy:** The ratio of correctly classified samples to the total number of samples.
- **Sensitivity (Recall):** The ratio of correctly identified benign cases to the total actual benign cases.
- **Specificity:** The ratio of correctly identified malignant cases to the total actual malignant cases.
- **Error Rate:** The ratio of misclassified samples to the total number of samples.

These performance metrics are calculated using the following formulas:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Sensivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

$$Error = \frac{FN+FP}{TP+TN+FP+FN} \quad (4)$$

Figure 4 presents a comparison of the evaluation criteria results obtained from the combination of the Decision Tree algorithm and Principal Component Analysis (PCA), as well as the combination of the Neural Network algorithm and PCA, against previous methods.

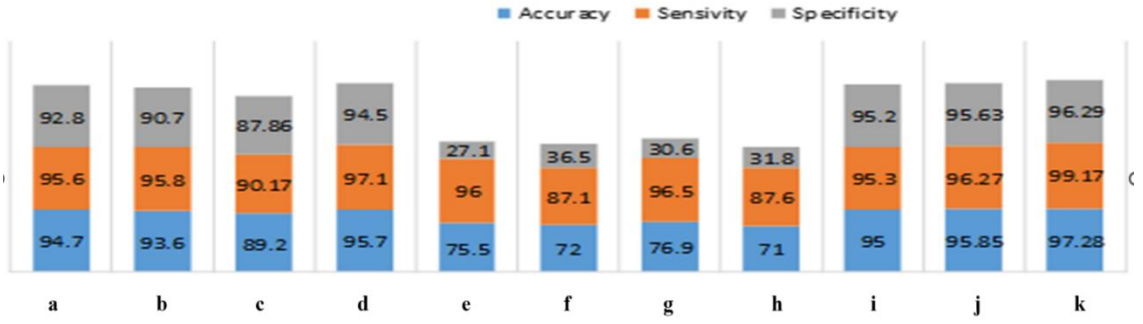

Figure 4. Comparison of Evaluation Metrics Between Proposed Models and Previous Methods.
a: MLP, b:C5, c: Logistic Regression, d: SVM, e: J48, f: PART, g: BAGGING, h: RBFN, i: ANN, j: PCA+KART, k: PCA+MLP

In Table 3, the proposed method is compared with all the methods discussed in this paper in terms of accuracy. As can be observed, the performance of the proposed algorithm is superior in comparison to other methods, indicating that to reduce feature dimensions in disease diagnosis, it is preferable not to use elimination-based algorithms, or rather, to benefit from a combination of all functions in a suitable manner.

Table 3. Performance of the proposed algorithm compared to other algorithms

Breast Cancer Prediction Accuracy (%)	Data Mining Technique
95.7	SVM
75.5	J48
72	PART
76.9	Bagging
95.1	Artificial Neural Network
71	RBFN
89.20	Logistic Regression
94.7	MLP
93.62	C5
96.53	Proposed Method

4. CONCLUSION

The aim of this paper is to examine the role and domains of predictive data mining applications in medical sciences and propose a framework for constructing, evaluating, and utilizing data mining models in this field. This paper demonstrates that data mining predictions provide essential tools for researchers and physicians to improve disease prevention, diagnostic methods, and treatment programs. According to the findings of this paper, the combination of the mentioned methods was successful in achieving high identification accuracy by relying on the characteristics of the database in the form of a combination and interaction, which, in comparison with conventional methods on one hand and artificial methods on the other hand, is suitable in its own right in the referenced sources.

Declaration

We acknowledge that we used ChatGPT to enhance the academic writing of our manuscript while ensuring the originality and integrity of our work.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] World Health Organization. (2014). World cancer report 2014 (Chapter 5.2). Geneva: World Health Organization. ISBN: 92-832-0429-8.

[2] National Cancer Institute (NCI). (2014). Breast cancer. Retrieved June 29, 2014, from https://www.cancer.gov

[3] Yeh, J.-Y., Chan, S.-W., & Wu, T.-H. (2016). Mining breast cancer classification rules from mammograms. Journal of Intelligent Systems, 25(1), 19–36. https://doi.org/10.1515/jisys-2014-0122

[4] National Cancer Institute (NCI). (2014, June 26). Breast cancer treatment (PDQ®). Retrieved June 29, 2014, from https://www.cancer.gov

[5] Office for National Statistics. (2013, October 29). Cancer survival in England: Patients diagnosed 2007–2011 and followed up to 2012. Retrieved June 29, 2014, from https://www.ons.gov.uk

[6] National Cancer Institute (NCI). (2014). SEER stat fact sheets: Breast cancer. Retrieved June 18, 2014, from https://seer.cancer.gov/statfacts/html/breast.html

[7] National Cancer Institute (NCI). (2014). Male breast cancer treatment. Retrieved June 29, 2014, from https://www.cancer.gov

[8] University of California, Irvine. (n.d.). Breast cancer Wisconsin (original) dataset. Retrieved from http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin(Original)

[9] Alizadeh, S., & Malek Mohammadi, S. (2014). Data mining and knowledge discovery step by step with Clementine software (3rd ed.). Tehran: Khajeh Nasir Toosi University of Technology Press.

[10] Siahi, M., & Ashir, A. (2015). Breast cancer recurrence prediction using data mining. National Electronic Conference on Recent Advances in Engineering and Basic Sciences, Islamic Azad University, Dezful Branch.

[11] Nazarian, M., Abbasi Dezfuli, M., & Haroon Abadi, A. (2013). A method for breast cancer diagnosis using data mining techniques and virtual bee colony algorithm. 5th National Conference on Electrical and Electronic Engineering of Iran, Islamic Azad University, Gonabad Branch.

[12] Ahmed, L. Q. (2013). Using data mining techniques for prediction of breast cancer recurrence. Iranian Journal of Breast Disease, 5(4), 23–34.

[13] Kiani, B., & Atashi, A. (2014). A prognostic model based on data mining techniques to predict breast cancer recurrence. Journal of Health and Biomedical Informatics, 1(1), 26–31.

[14] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 4(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002

[15] Zadeh, H. G., Haddadnia, J., Seryasat, O. R., & Isfahani, S. M. M. (2016). Segmenting breast cancerous regions in thermal images using fuzzy active contours. EXCLI Journal, 15, 532.

[16] Seryasat, O. R., & Haddadnia, J. (2018). Evaluation of a new ensemble learning framework for mass classification in mammograms. Clinical Breast Cancer, 18(3), e407–e420. https://doi.org/10.1016/j.clbc.2017.05.009

[17] Rahmani-Seryasat, O., Haddadnia, J., & Ghayoumi-Zadeh, H. (2015). A new method to classify breast cancer tumors and their fractionation. Ciência e Natura, 37(4), 51–57. https://doi.org/10.5902/2179460X19428

[18] Seryasat, O. R., & Haddadnia, J. (2017). Assessment of a novel computer-aided mass diagnosis system in mammograms. Biomedical Research, 28(7), 3129–3135.