Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 07 Issue 02
2024

Research Article
https://doi.org/10.56979/702/2024

Comparative Analysis of Machine Learning Algorithms for Breast Cancer Classification

Muhammad Shaharyar Ramzan[1*], Gohar Mumtaz[1], Nazish Rasheed[1], and Zeeshan Mubeen[2]

[1]Faculty of Computer Science and Information Technology, Superior University, Lahore, 54000, Pakistan.
[2]Riphah International University, Lahore, 54000, Pakistan.
[*]Corresponding Author: Muhammad Shaharyar Ramzan. Email: shaharyarramzan56@gmail.com

Received: July 21, 2024 **Accepted:** August 21, 2024 **Published:** September 01, 2024

Abstract: Breast cancer is one the most fatal diseases among women. Therefore, the need to develop reliable diagnostic tools to early detect breast cancer for treatment. Machine Learning is the powerful approach to breast cancer classification. This systematic review aims to provide a comprehensive comparison of various machine learning algorithms used to the classification of breast cancer. The key algorithms, including support vector machines (SVM), decision tree (DT), random forest (RF), K-nearest neighbor (KNN), Logistic Regression (LR) and Convolutional neural network (CNN), evaluate the performance of the metrics and the accuracy of the model. We analyze the numerous data from the various papers, and then find the strengths and the limitations of each algorithm in the different scenarios. In addition, we discuss the impact of the data preprocessing techniques, selection methods and the role of ensemble learning in the classification performance. We found that no single algorithm consistently outperforms others across all metrics, suggesting a hybrid approach may offer the most robust solution.

Keywords: Diseases; Women; Cancer; Algorithms; Classification.

1. Introduction

Breast cancer is the most common disease among women globally, with 2.26 million cases diagnosed in 2020[1], leading to mortality of the women. Breast cancer affects one out of every 3000 breastfeeding women or pregnant women. According to research, if a woman diagnoses breast cancer while pregnant, her chances of survival are very low in pregnant women [2]. According to statistics of the World Health organization, breast cancer is the most life-threatening disease for women with a rate of 12.5% in some developed countries. In addition, studies show that it is a commonly diagnosed type of cancer in the world [3]. In the comparison of the other types of cancer and the number of cases reported worldwide, which lead to the cases of women dying is breast cancer [4].

The machine learning algorithms revolutionized in various fields in recent years. Using machine learning developed the Computer Aided Diagnosis (CAD) system for analysis of breast ultrasound images. Several techniques have been introduced for the screening and diagnosis of breast cancer one of the most is a mammography, which detects the cancer initial stage.in general, X-ray does not detect breast cancer due to small size of cells seen from the outside. The machine learning algorithms have emerged as a tool for the predictive analyses in the content of breast cancer [6]. Machine learning and artificial intelligence introduce transformative approaches for diagnosis of disease. Machine learning is a subset of artificial intelligence. In the recently developed machine learning algorithms, it trains it to take the decision correctly. These algorithms revolutionized the domain of medicine [7].

The development of artificial intelligence technologies brings many solutions for breast cancer detection in the health domain. Machine learning and artificial neural networks, can categorize the patient data and develop a predictive model which helps to detect cancer. The predictive cancer using the various machine learning algorithms, K Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Naive

Bayes (NB) [1], support vector machines (SVM), artificial neural networks (ANN), and Logistic Regression (LR).

1.1. Objective of the Study

The objective of this systematic literature review compares various algorithms for the cancer classification. This comparison objective of this paper is to finds accurate machine learning techniques, and to understand their strengths and limitations. The main goal of this systematic literature review is to provide useful guidance for the future researcher for the diagnosis and treatment of breast cancer.

1.2. Breast Cancer Classification (BBC)

The objective of the Breast cancer classification is to diagnose breast cancer. The pathologists used nine characteristics to classify breast cancer, including various factors like clump thickness, cell size, cell shape, adhesion, epithelial cell size, nuclei characteristics, chromatin, nucleoli, and mitosis count. Machine learning techniques and algorithms used for Breast cancer Classification which helps to classify the best cancer characteristics for the treatment.

1.3. Machine Learning Approaches

Machine learning is the branch of artificial intelligence. Machine learning algorithms used for the diagnosis and treatment of breast cancer and other diseases. This systematic literature review provides a useful and valuable guide for the future researcher and compares the various algorithms of machine learning which diagnose and detect breast cancer.

1.3.1. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a classification algorithm which implements easy on the base of nearest neighbors. In the classification process there are several problems faced with imbalanced data. Using the synthetic Minority Oversampling techniques for the data imbalance. The Ensemble methods can be applied to improve the performance of the imbalanced data classification. Which is one adaptive Boosting. This study applies the K-Nearest Neighbors combined with adaptive Boosting and SMOTE can handle the imbalanced data classification [8].

1.3.2. Convolutional neural network (CNN)

Convolutional neural network (CNN) is the part of artificial intelligence and has been in the process of revaluation since the last decade. CNN was introduced in 1993 and unfortunately this project was remained dead for a long time, recently it achieves the improvement and performance due to the powerful graphics processing unit (GPU) [9]. Convolutional neural networks automatically classify mammograms into benign and malignant and control the overfitting and dataset imbalance to overcome issues.

1.3.3. Logistic Regression (LR)

Logistic Regression is the statistical model used for the binary classification. Logistic Regression has two classes, such as "M" for Malignant or "B" for benign in breast cancer diagnosis. Logistic Regression is the probability of the outcome on the perdition variable fettering a logistic regression function, which enables the perdition medical diagnosis and serves as the baseline model for the perdition of cancer. Logistic Regression is useful for the feature selection using the Recursive Feature Elimination (RFE), the most predictive feature for cancer diagnosis. While the interpretable logistic Regression can struggle with non-linear relationships [10].

1.3.4. Decision Tree (DT)

Decision Tree is the most famous algorithm which is used for classification and predicting data. Decision Tree algorithms are easy to understand and helpful to solve the problem. When we combine the random methods, the result of this is very accurate. The Decision Tree is helpful to diagnose breast cancer. Decision Tree has compared the other methods like Support Vector Machine (SVM) and Naive Bayes (NB), showing they can classify data well [11].

1.3.5. Support Vector Machines (SVM)

Support Vector Machine is the useful and strong tool for the machine learning tool, which helps to diagnose breast cancer, and useful for complex tasks. In the breast cancer diagnosis, the Support Vector Machines is the outstanding performance to identify the accurate the beings' cases. Support Vector Machine effective during the reduction of the data using the techniques, which show its adaptability. Support Vector Machine is effective which requires high accuracy [7].

1.3.6. Random Forest (RF)

The Random Forest is the algorithms that useful for classification and regression, and it create many decision trees form the random data and make strong perdition. The Random Forest methods used for handling the missing value and uncertain data well. Random Forest avoids the overfitting by the multiples trees, which make the accurate results. For the breast cancer diagnosis, the Random Forest performance is better than the other algorithms like decision tree and linear regression. Random Forest helps the early diagnosis the breast cancer and improve the treatment and reduce the cost of treatment [12].

2. Research Methodology

The objective of this systematic literature review summarizes and synthesizes the results of the existing primary studies that address the specific research questions. We try to find the accurate algorithms for diagnosing breast cancer, and provide a valuable guide to future researchers. We used various algorithms and evaluated their performance and accuracy with their results. The comparison of the algorithms helps us to find the best algorithm for the diagnosis of breast cancer. Early detection of breast cancer can save many lives of young women.

2.1. Research Questions

The systematic literature review objective to finds the various research questions which will help us to find the best algorithms for the breast cancer diagnosis. There are some research questions the review aims to address, such as

RQ 1- Which machine learning algorithms commonly used in classification of breast cancer.

RQ 2- Compare the algorithms which are accurate and efficient due to strong results.

RQ 3 -Identify each algorithm's strength and limitations.

2.2. Search approach

The objective of the primary studies for the related classification techniques. During the search of the research papers, we searched various databases, Google Scholar, PubMed and IEEE Xplore to collect the related papers of the topic. These databases helped in collecting the material and deeply explore the Topic. This method helps to write systematic literature reviews and addresses the various research gaps and limitations. This systematic literature review will provide a comprehensive guide for future research. The help of this LSR they find the various limitations and improve them accurately. The following flow diagram shows the finding of the related papers.

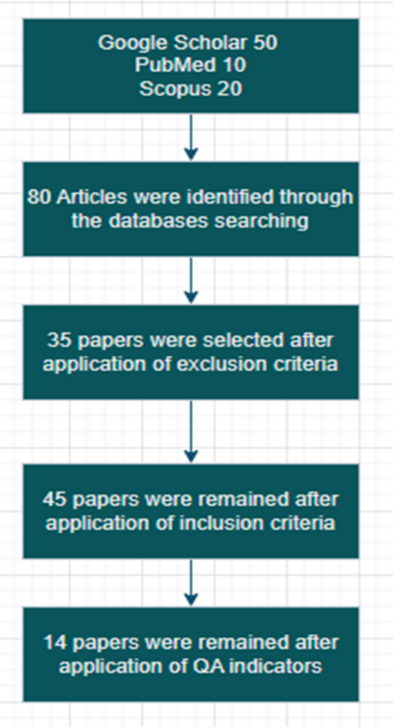
Google Scholar 50
PubMed 10
Scopus 20
80 Articles were identified through the databases searching
35 papers were selected after application of exclusion criteria
45 papers were remained after application of inclusion criteria
14 papers were remained after application of QA indicators

Figure 1. Selection Process of Related papers

2.3. Study selection

The study selection objective to identify the most relevant research papers of the research question. This process involved the various steps, which included establishing the inclusion and exclusion criteria and independent evaluations.

2.3.1. Inclusion Criteria

IC 1- The research paper is about the classification techniques in breast cancer.

IC 2- The research paper developed new classification techniques in breast cancer.

IC 3 - The paper's main objective is to evaluate existing classification techniques which are applied to breast cancer diagnosis.

IC 4 - Comparison of the various classification techniques in breast cancer diagnosis.

IC 5 - The criteria for selecting the paper is the recent and complete without the duplication.

2.3.2. Criteria of Exclusion:

EC 1- The objective of the breast cancer classification and dealing machine learning algorithms.

EC 2- The papers published earlier than 2020.

EC 3-The paper is not written in English.

2.4. Evaluation Process

There are paper classified into the three categories:

Included: The research paper includes one and none of the excluded criteria.

Exclusion: The article needs to satisfy one requirement for exclusion.

Uncertain: paper does not clearly fit into included or excluded based on the available information.

2.5. Quality Assessment

The selection of the studies for the comparison of machine learning algorithms for breast cancer classification to ensure the reliability and validation of the paper. We evaluate the quality assessment with the help of the various paper-based sets of predefined quality criteria. The process of the assessment was performed independently by the researchers and disagreements were resolved through discussion.

Quality Criteria

QC 1 - The study explores the research question related to breast cancer classification using the machine learning algorithms.

QC 2 - The methodological Rigor is the process of collection of data and analysis techniques for the replicable methodology.

QC 3 - The study provides a comprehensive evaluation of the machine learning algorithms including the performance metrics as accuracy, precision, recall, F 1 score, and ROC-AUC.

QC 4 - The study provides the transparency of the used datasets including the source, preprocessing and modification made.

QC 5 - The objective of the study is the comparison and analysis of the machine learning algorithms with appropriate benchmarking of the established methods.

QC 6 - The study provides the comprehension details and to allow replication of the experiments, algorithms parameters and implementation details.

2.6. Data Extraction

Data extraction is the process of systematically gathering the relevant data or information from the selected topic to facilitate a comparative analysis of machine learning algorithms for the breast cancer classification. The process involves extracting data related to research questions and the objective of the study. The data extraction process helps the research to get more relevant data and collect meaningful information.

2.7. Data Synthesis

Data synthesis is the process which involves analyzing and integrating the exact data from the selected study to draw meaningful information about the comparative performance of the various machine learning algorithms for breast cancer classification. The synthesis main objective is to identify the patterns, trends and insights that could be useful for the future researchers and clinical applications. There are various steps to analyzing the data synthesis.

DS 1 - Descriptive analysis is the process of summarizing the characteristics of the studies which include the number of studies, publication years and types of datasets which were used and geographical distribution of the research. Cataloged algorithms used for noting the frequency and the variations in their application.

DS 2 - Performance Metrics is the process of statistical analysis to compare the performance metrics of various algorithms. Which included the calculating mean, median, standard deviation, and conducting tests where applicable.

DS 3 - Methodological approach is used to analyze the data preprocessing methods used to identify the common practices and unique approaches. The comparison of the validation techniques used to assess the algorithms performance.

2.8. Risks to the Accuracy

Risks to the accuracy the process machine learning algorithms for breast cancer classification, which must be considered to ensure the reliability and robustness of the finding. There threats can be categorized into the internal, external, construct, and conclusion validity. Internal validity processes how accurately the comprehensive study measures the effects of the independent variables on the dependent variables. Internal validity is the process to identify which the study accurately measures the effect of the independent variable on the dependent variables. The primary threat to internal validity is selection bias, which may occur if the relevant study were missed or the inclusion and exclusion criteria were not rigorously applied. External validity processes the generalizability of study to finding the other settings of populations. The threat in this category is dataset representativeness, where datasets used in the included study do not represent the broader population of breast cancer patients. Conclusion validity is to relate the degree to which conclusions drawn from the study are credible and justified. One threat in the area of statistical power, which can be affected by the limited sample sizes and the variability in study.

3. Results and discussion

In this section discuss the results and their results with the comparative analysis of algorithms for the classification of cancer. We analyze various algorithms, including Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), K-Nearest Neighbors (K-NN) and Convolutional Neural Networks (CNN). The datasets used the performance metrics evaluated and the outcomes observed are systematically discussed.

RQ 1- Which breast cancer diagnosis techniques used for the classification and their results?

There is various methodology adopted to diagnose breast cancer. K- Nearest Neighbor (KNN) classification on the data based with some features like tumor size and shape. It used metrics such as Euclidean distance to evaluate the nearest neighbors. The study of the 75 patients (62 malignant 13 benign), SMOTE balanced the datasets to 40% benign and 60% malignant cases. The optimal K was found to be 9, achieving 99.6% accuracy [13]. The support vector machine achieved the accuracy of 98.25% and made effective algorithms for breast cancer diagnosis. Random Forest closely follows the support vector machine with the accuracy of 98.24% showing the strong performance in the classification of breast cancer. Decision Trees are known for the interpretability and ease of use with the specific accuracy rates not provided, but the part of the comparative analysis including the reasonable performance. Logistic Regression achieved a high accuracy rate of 93.37% showing its strong capability in identifying breast cancer correctly [7]. MatconvNet achieved 94.2% accuracy and YOLO achieved 93.3% with 100% sensitivity and 87% ensemble of CNN models achieved 95.7% accuracy. These results showing the strong performance of CNN in diagnosis the breast cancer [14].

RQ 2: Which are the benefits and drawbacks of the breast cancer classification systems?

K-Nearest Neighbor (KNN) is highly accurate, achieving 99.6% accuracy with an optimal K value of 9, particularly when using a balanced dataset as achieved through SMOTE (Synthetic Minority Over-sampling Technique). KNN is effective in small to moderately sized datasets, especially when features like tumor size and shape are well defined. However, KNN can be computationally intensive with larger datasets and its performance is sensitive to the choice of K and the scaling of features. Support Vector Machine is the powerful classification technique for boasting a high accuracy of 98.25%. SVM is the most effective for binary classification problems and it is robust against overfitting, marking it a reliable choice for breast cancer detection. Support vector machines require parameter tuning to achieve optimal performance and the resulting model can understand the decision process is crucial. Random Forest is the SVM with an accuracy of 98.24%. The method benefits from the robustness of the ensembles learning, which combines the various and multiple decision trees to reduce the risk of overfitting and handle large datasets effectively. The strong performance of the Random Forests can be slow to train the models generated are less interpretable than single decision trees, which can limit their use in clinical settings where transparency is important. Decision Trees used for their interpretability and the ease of use and making them a valuable tool in medical diagnostics. While the accuracy rates were not provided in the study. Decision trees are noted for their reasonable performance in the analyses. Logistic Regression achieved a high accuracy of 93.37% for correctly identifying breast cancer. It is very simple to implement and effective for binary classification problems. Logistic Regression is a linear relationship between the features and the outcome, which can be a limitation when dealing with the complex, non-linear relationships in the data. Convolutional Neural Networks are models like MatConvNet and YOLO which have ensemble models. CNN is particularly effective in extracting features from images data, which is used in medical imaging tasks.

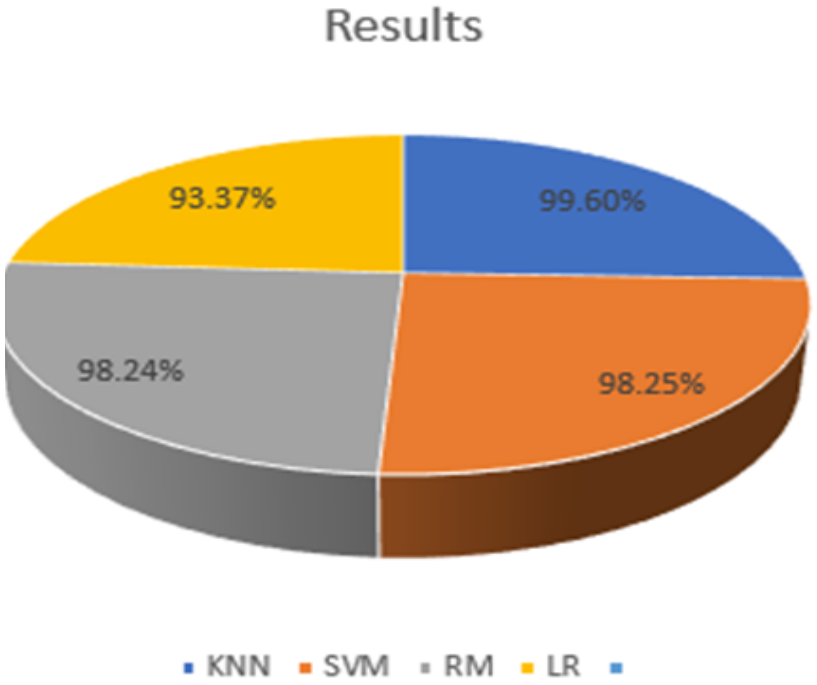

Figure 2. Show the results of various algorithms

RQ 3- How accurate are the classification methodologies utilized in the diagnosis of breast tumors in general?

Overall performance the breast cancer classification, K-Nearest Neighbor achieved 99.6% accuracy using the optimal K value of 9. Support Vector Machine and Random Tees showed reasonable performance, while the Logistic Regression achieved 93.37% accuracy. Convolutional Neural Network ensemble of CNN models reaching 95.7%. Overall, these techniques provide highly effective tools for breast cancer diagnosis.

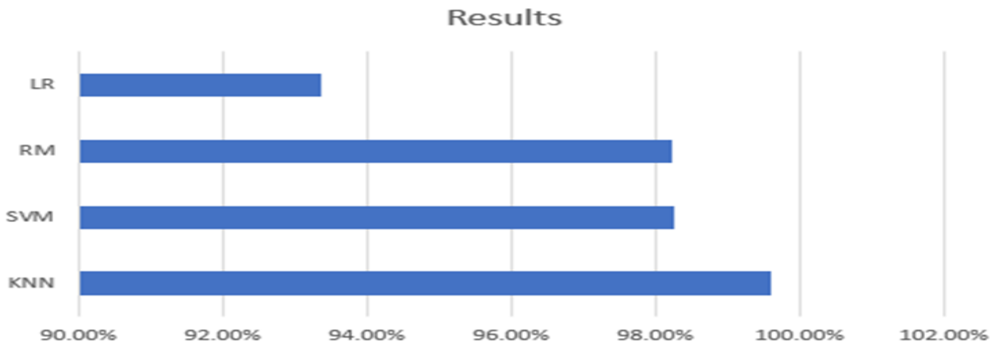

Figure 3. Show the Overall Performance of Classification Techniques

RQ 4- What are the classification techniques which are applied in breast cancer detection better than others?

In the process of the classification several methodologies have been applied to the breast cancer diagnosis with some demonstrating performance compared to others. One of the methods is K-Nearest Neighbor. By using the features like tumor size and shape and employing Euclidean distance to evaluate the nearest neighbors, KNN achieved the impressive accuracy of 99.6%. This high accuracy was attained after balancing the dataset using SMOTE to ensure a representative distribution of benign and malignant cases. The optimal K value was found to be 9 which contributed to its top performance.

Another effective technique is the support vector machine which reached an accuracy of 98.25%. SVM is suitable for the binary classification problems and a reliable algorithm for the breast cancer diagnosis. It creates effective decision boundaries even in complex datasets and makes it a strong performer in this field. Random Forest showed robust performance closely following SVM with the accuracy of 98.24%. This method benefits from its ensemble approach and improves the generalization. Logistic Regression achieved a high accuracy rate of 93.37% for the diagnosis of breast cancer. Overall KNN, SVM and Random Forest have the superior performance in the breast cancer diagnosis with the KNN leading the way. These techniques provide highly effective tools for early and accurate detection. Which is crucial for improving the patient treatment outcomes and reducing the mortality rates with breast cancer.

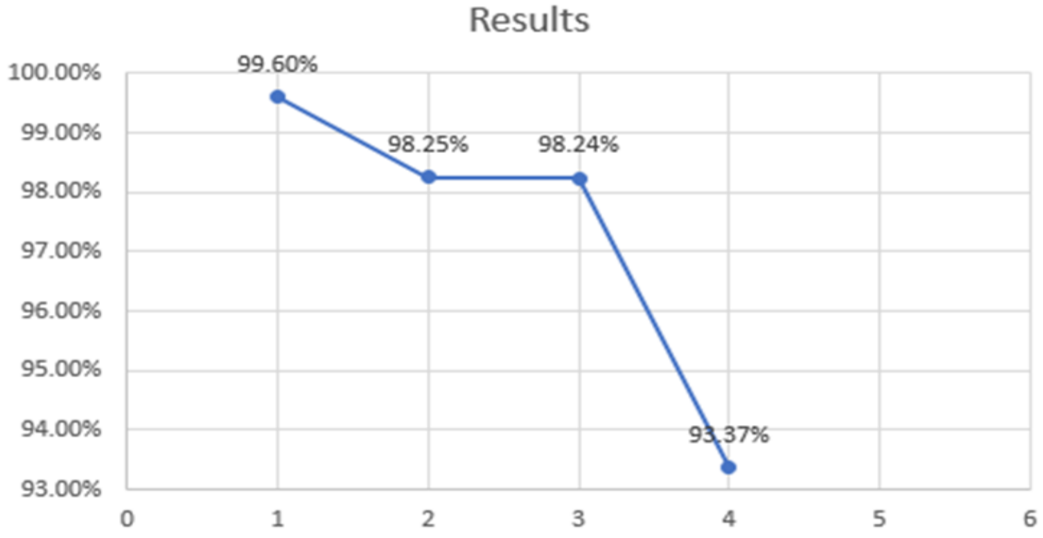

Figure 4. Show the Better Performance of the Algorithms

4. Conclusions

Breast cancer is one of the most fatal diseases among women. Early detection and treatment for breast cancer needs reliable and diagnostic tools. This systematic review provides a comparison of various machine learning algorithms used for the breast cancer classification. We study and analyze the numerous papers of each algorithm and its strengths and limitations in the different scenarios. KNN achieved the highest accuracy of 99.6% with an Optimal K value of 9 when using balanced datasets. SVM and Random Forest also achieved a strong performance with the accuracies of 98.25% and 98.24%. Decision Trees known for their interpretability and ease of use, achieved reasonable performance but lacked the accuracy rates. Logistic Regression achieved a notable accuracy of 93.37% with the effectiveness and correctly identifying breast cancer diagnosis. MatComNet and YOLO achieved accuracies of 94.2% and 93.3% with the ensemble of CNN models reaching 95.7% with performance of CNN in this domain. This hybrid approach may improve the early diagnosis and treatment outcomes and reduce the mortality rates associated with breast cancer.

References

1. Ö. Çağrı Yavuz, M. H. Calp, and H. Ceren Erkengel, 'Prediction of breast cancer using machine learning algorithms on different datasets', Ing. Solidar., vol. 19, no. 1, pp. 1–32, Jun. 2023, doi: 10.16925/2357-6014.2023.01.08.
2. M. A. Wajeed et al., 'A Breast Cancer Image Classification Algorithm with 2c Multiclass Support Vector Machine', J. Healthc. Eng., vol. 2023, pp. 1–12, Jul. 2023, doi: 10.1155/2023/3875525.
3. L. Sun and S. Li, 'A Study of Breast Cancer Classification Algorithms by Fusing Machine Learning and Deep Learning', Appl. Sci., vol. 13, no. 5, p. 3097, Feb. 2023, doi: 10.3390/app13053097.
4. M. A. Elsadig, A. Altigani, and H. T. Elshoush, 'Breast cancer detection using machine learning approaches: a comparative study', Int. J. Electr. Comput. Eng. IJECE, vol. 13, no. 1, p. 736, Feb. 2023, doi: 10.11591/ijece.v13i1.pp736-745.
5. M. Ebrahim, A. A. H. Sedky, and S. Mesbah, 'Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer', Data, vol. 8, no. 2, p. 35, Feb. 2023, doi: 10.3390/data8020035.
6. T. Sun, 'Breast cancer prediction based on multiple machine learning algorithms', vol. 92, 2024.
7. J. Li, 'Evaluative Comparison of Machine Learning Algorithms for Precision Diagnosis in Breast Cancer', Highlights Sci. Eng. Technol., vol. 85, pp. 354–362, Mar. 2024, doi: 10.54097/40fmfw48.
8. R. S. Yuliani, A. Rusgiyono, and R. Santoso, 'K-NEAREST NEIGHBOR DENGAN ADAPTIVE BOOSTING DAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE UNTUK KLASIFIKASI DATA TIDAK SEIMBANG', J. Gaussian, vol. 12, no. 2, pp. 231–241, Jul. 2023, doi: 10.14710/j.gauss.12.2.231-241.
9. S. Nadkarni and K. Noronha, 'Breast cancer detection using ensemble of convolutional neural networks', Int. J. Electr. Comput. Eng. IJECE, vol. 14, no. 1, p. 1041, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1041-1047.
10. B. Zhang, 'Using Logistic Regression and Support Vector Classification to Predict Cancer', Highlights Sci. Eng. Technol., vol. 92, pp. 288–294, Apr. 2024, doi: 10.54097/bkvnxg90.
11. T. Islam et al., 'Predictive Modeling for Breast Cancer Classification in the Context of Bangladeshi Patients: A Supervised Machine Learning Approach with Explainable AI'.
12. Md Zahidul Islam, Md Nasiruddin, Shuvo Dutta, Rajesh Sikder, Chowdhury Badrul Huda, and Md Rasibul Islam, 'A Comparative Assessment of Machine Learning Algorithms for Detecting and Diagnosing Breast Cancer', J. Comput. Sci. Technol. Stud., vol. 6, no. 2, pp. 121–135, Jun. 2024, doi: 10.32996/jcsts.2024.6.2.14.
13. R. S. Yuliani, A. Rusgiyono, and R. Santoso, 'K-NEAREST NEIGHBOR DENGAN ADAPTIVE BOOSTING DAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE UNTUK KLASIFIKASI DATA TIDAK SEIMBANG', J. Gaussian, vol. 12, no. 2, pp. 231–241, Jul. 2023, doi: 10.14710/j.gauss.12.2.231-241.
14. S. Nadkarni and K. Noronha, 'Breast cancer detection using ensemble of convolutional neural networks', Int. J. Electr. Comput. Eng. IJECE, vol. 14, no. 1, p. 1041, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1041-1047.
15. Baig, M. S., Imran, A., Yasin, A. U., Butt, A. H., & Khan, M. I. (2022). Natural language to SQL queries: A review. International Journal of Innovations in Science & Technology, 4, 147-162. 50sea.
16. Bilal, M., Ali, M. A., Nichol, J. E., Qiu, Z., Mhawish, A., & Khedher, K. M. (2023). Reduced major axis regression. In Encyclopedia of Mathematical Geosciences (pp. 1199-1203). Springer International Publishing Cham.