JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
ICASET
2024

Research Article
Collection: Intelligent Computing of Applied Sciences and Emerging Trends

Comparative Analysis of Machine Learning Algorithms for Breast Cancer Detection: A Study of Support Vector Classification, Logistic Regression, and K-Nearest Neighbors

Muhammad Salah-ud-din Iqbal[1], Raza Mukhtar[1], Muhammad Saleem[1*], Muhammad Kamran[1], Muhammad Hussain[1], Syed Younus Ali[1], and Muhammad Umair[1]

[1]School of Computer Science, Minhaj University Lahore, Pakistan.
[*]Corresponding Author: Muhammad Saleem. Email: msaleemkalru@gmail.com

Academic Editor: Salman Qadri **Published:** April 01, 2024

Abstract: Worldwide, breast cancer is a frequent medical problem, and successful treatment to a large extent counts on early diagnosis. Machine learning algorithms have demonstrated a high potential of integration into breast cancer detection system, which in turn will increase the diagnostic facility's precision. Three popular machine learning techniques are compared in this study: K-Nearest Neighbors (KNN), Support Vector Classification (SVC), and Logistic Regression. In this context, the study is based on the several study that use the algorithms to detect breast cancer. By SVC, Logistic Regression, and KNN for breast cancer diagnosis, the study of medical literature was executed thoroughly. We conducted a thorough analysis of each paper's methodology, dataset, feature selection strategies, and performance measures. With an average accuracy of more than 98%, our data show that Support Vector Classification surpassed K-Nearest Neighbors and Logistic Regression in every study that was analyzed. SVC outperformed other methods for detecting breast cancer in terms of sensitivity, specificity, and total predictive power. While KNN produced somewhat lower accuracy rates, Logistic Regression demonstrated moderate accuracy. The results of this investigation highlight Support Vector Classification's efficacy as a strong algorithm for tasks involving the identification of breast cancer. Patient outcomes for early identification and treatment of breast cancer could be greatly enhanced by enhanced detection systems driven by machine learning algorithms.

Keywords: Breast Cancer Detection; Machine Learning Algorithms; Support Vector Classification; Logistic Regression; K-Nearest Neighbors.

1. Introduction

The type that starts in the breast cells is called breast cancer. Although it is uncommon, it is one of the most common cancer in women globally. It can also affect men. Here are some essential details regarding breast cancer: There are several types of breast cancer, invasive lobular carcinoma, including inflammatory breast cancer, ductal carcinoma in situ (DCIS), and others. The origins and behavioral characteristics of these categories vary. Although the precise cause of breast cancer is unknown, a number of factors can raise one's chance of contracting the illness. These variables include age, history of breast cancer in family, mutations in genes, hormone levels (including those in estrogen and progesterone), lifestyle choices (like drinking alcohol, being overweight, and not exercising), and radiation exposure.

Breast cancer symptoms can vary, but may include a swelling or tumor in the breast or underarm area, the size of the breast changes or form, changes in the nipple. change in the skin (like redness, dimpling, or puckering), and breast pain. Still, not every breast cancer has symptoms, and some can be found by screening before symptoms appear.

A combination of imaging tests (such as MRIs, ultrasounds, and mammograms) and tissue sampling (such as biopsy) is usually used to diagnose breast cancer. For the purpose of treatment planning, these tests assist in identifying whether cancer is present as well as its location, size, and other characteristics. It varies according to the stage and the type of the cancer, together with the patient's likings and inclusive well-being, all influence the treatment plan for breast cancer. Common treatment options include radiation therapy, targeted therapy, immunotherapy, chemotherapy, hormone therapy, and surgical procedures such lumpectomies and mastectomies. Care plans are often tailored to each patient's specific need.

The type of breast cancer, its stage at diagnosis, the existence of certain biomarkers, and the cancer's response to treatment are only a few of the many variables that affect the prognosis. Many patients with breast cancer are able to live long and healthy lives thanks to advancements in early identification and treatment. For those who are impacted by the disease, breast cancer awareness, routine screening, early detection, and treatment improvements have greatly improved outcomes.

2. Literature Review

In 2020, breast cancer claimed 685 000 deaths worldwide. In 2020, there were 685 000 breast cancer deaths worldwide, which translates to roughly 2.3 million new instances of the illness. Breast cancer is the most frequent cancer worldwide. In the last five years, around 7.8 million women were diagnosed with the breast cancer. Globally, breast cancer can strike women at any stage of life after childhood, while its prevalence increases with age. About half of all incidences of breast cancer are in women who have no other risk factors except age and sex.

People worldwide are impacted by breast cancer in every country. Breast cancer affects men between 0.5% and 1% of the time.

The WHO Global Breast Cancer Initiative (GBCI) intentions to avert 2.5 million breast cancer deaths worldwide between 2020 and 2040 by reducing the global mortality from breast cancer rate by 2.5% annually. A 2.5% annual reduction in breast cancer mortality will prevent 25% of deaths from the disease globally and 40% of deaths among women under 70 by 2040. When breast cancer is initially identified, more women will see doctors because of increased awareness of the disease's symptoms and indicators brought about by public health education, which will also assist women and their families in appreciating the importance of early detection and treatment.

Health workers must receive training in conjunction with public education regarding the telltale signs and symptoms of early breast cancer in order to guarantee that women are referred to diagnostic services when needed. Timely diagnosis is essential for effective cancer treatment, which frequently calls for a certain level of specialized cancer care. Centralized services can be implemented in a hospital or cancer institute using breast cancer as a model to optimize treatment for breast cancer and improve care for other cancers. [1]

An additional machine learning-centered hybrid model was put forth by M. Tahmooresi et al. [2]. This implies that SVM was the best classifier overall, having the highest accuracy. It had performed a contrast amid ANN, KNN, SVM, and decision trees. It was used on the image sets and blood datasets. Consequently, Muhammad Fatih Aslan et al.'s machine learning model [3] used an alternative classifier. As classifiers, the author used ANN, SVM, KNN, and Extreme Learning Machine. A small modification was made to the classifier to improve the results. This suggests that Extreme Learning Machine generated better results. The machine learning model was proposed by Anusha Bharat and colleagues [4]. It made use of four classifiers: Naïve Bayes, KNN, CART, and SVM. The author claims that KNN provided superior accuracy. Multi-SVM was applied. Ayndindag Bayrak et al. [5] conducted a comparative analysis of machine learning methodologies. The contrast was carried out utilizing WEKA and the Wisconsin Breast Cancer Database as the dataset. SVM performed the best according to the provided matrices. They also developed the deep learning based technologies with machine learning [13-18].

The author built the convolutional neural network based upon the machine learning Shewtha K et al. [6]. Many models were included in the CNN, however only Mobile Net and Inception V3 were utilized. After comparing the two models, the author resulted that the accuracy of inception V3 is greater. Still, there was optimism for using machine learning to treat breast cancer. Ch. Shravya and associates proposed the supervised machine learning model [7]. The study proposed the SVM,KNN and logistic regression classifiers. Performance analysis was conducted on the dataset, which was retrieved from the UCI repository.

With a 92.7% accuracy rate, this suggests that SVM was a good classifier on the Python platform. The machine learning model proposed by Sivapriya J et al. [8] employed a different classifier. The Random Forest, SVM, Logistic Regression, and Naïve Bayes were employed by the writer. The Python implantation was completed on the Anaconda Platform. According to the author, Random Forest performed well as a classifier and had an accuracy rate of 99.76%. There was a chance to increase accuracy when there was a small adjustment in the network with the classifier. The model based on ANN was proposed by Kalyani Wadkar et al. [9] and its performance was examined using an SVM classifier.. The author claims that SVM provided 91% accuracy and ANN 97%. The author added that it provided improved accuracy even in the absence of SVM. Grid search and SVM based on a model were proposed by Vishal Deshwal et al. [10]. Initially, the research was used to SVM, and subsequently, SVM in conjunction with Grid search. After comparing them, the author determined which was the best. By comparison, a new model was constructed. Grid search proved to have higher accuracy.

Golatkar et al. [11] this study demonstrates impressive results in using deep learning by means of histology classification of breast cancer. The study highlights significant progress, and the CNN models come out as remarkable in showing high accuracy in determining different histological subtypes like DCIS, IDC, and ILC. By employing test models together with extensive evaluative measures, the researchers show that deep learning provides an accurate representation of complex patterns and characteristics of histopathological images thus suggesting a prospect of an important aid for pathologists. These results validate not only the usability of CNNs for histological identification but also reinforce that the utilization of state-of-the-art computer approaches in medical image analysis is essential to improving diagnostic accuracy and to finally achieve positive outcomes.

Wang, Dayong, et al. [12] research shows a remarkable development in the responsiveness and accuracy of the deep learning models compared to the initial procedures. Besides that, the deep learning techniques play a vital role in the increase of the accuracy in the diagnosis of met-astatic breast cancer. However, though the study has merit because of methodological innovation and clinical relevance, study validation in clinical settings and model interpretability are yet the other issues which should be taken into account. Nevertheless, the findings offer promising prospects for revolutionizing cancer diagnosis and management through the integration of advanced computational methods into medical imaging analysis.

3. Proposed Methodology

Support Vector Classifier (SVC), k-Nearest Neighbors (KNN), and Logistic Regression are all popular machine learning algorithms used for classification tasks:

3.1. Dataset:

The dataset used in this study is breast cancer Wisconsin.

3.2. Implementation

3.2.1. Importing Libraries:

The Implementation begins by importing necessary libraries for data analysis, visualization, and machine learning, including Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn modules like Logistic Regression, K-Neighbors Classifier, and SVC.

3.2.2. Loading Data

The breast cancer dataset is loaded from a CSV file named 'data.csv' using Pandas' read_csv() function. The dataset contains features related to breast cancer tumors.

3.3. Data Preprocessing

Columns 'id' and 'Unnamed: 32' are dropped from the data frame.

The 'diagnosis' column is encoded as binary where 'M' (malignant) is replaced by 1 and 'B' (benign) is replaced by 0.

Summary statistics of the dataframe are displayed using describe() and info() methods.

Null values in the dataframe are checked and visualized using isna().sum() and msno.bar() functions.

3.4. Data Visualization

Distribution plots are created for each feature using Seaborn's distplot() function.

A heatmap is plotted to visualize the correlation between features using sns.heatmap() function.

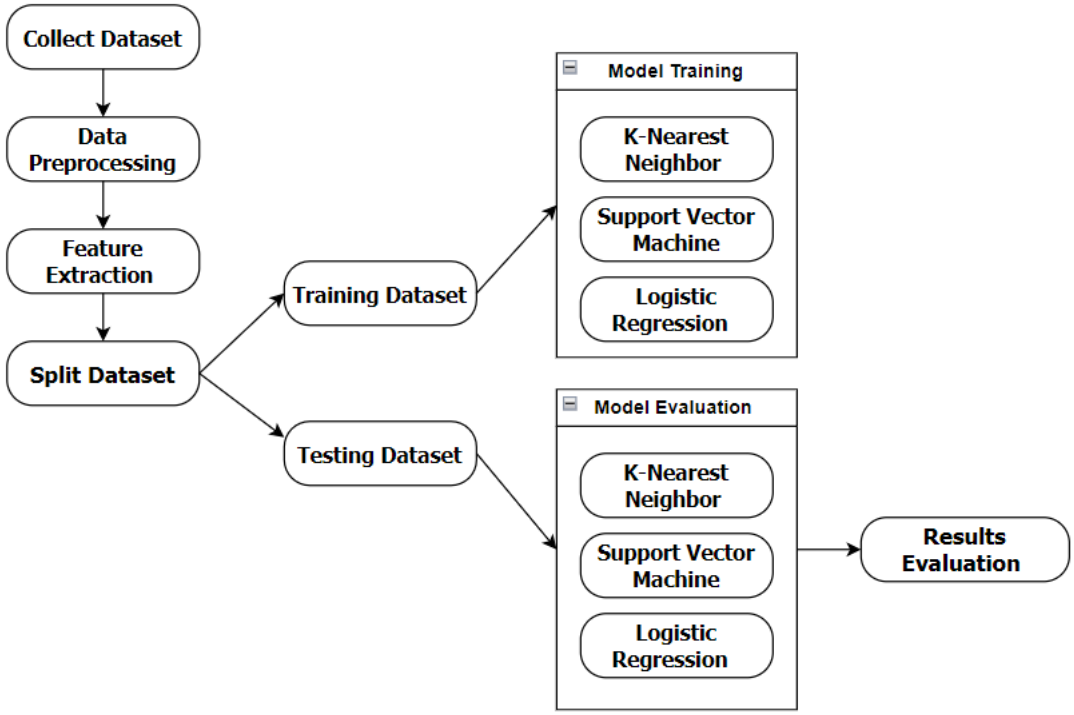

Figure 1. Proposed Methodology

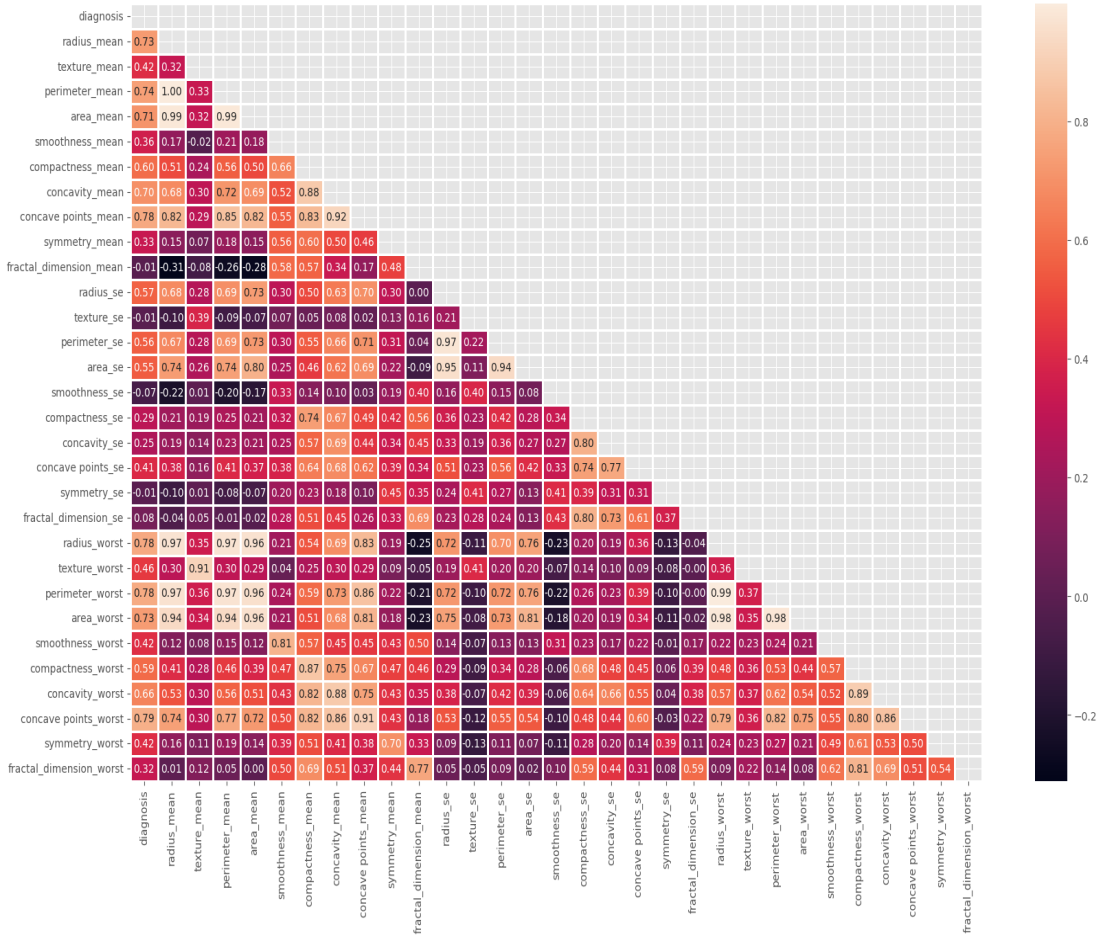

Figure 2. Distribution plots are created for features

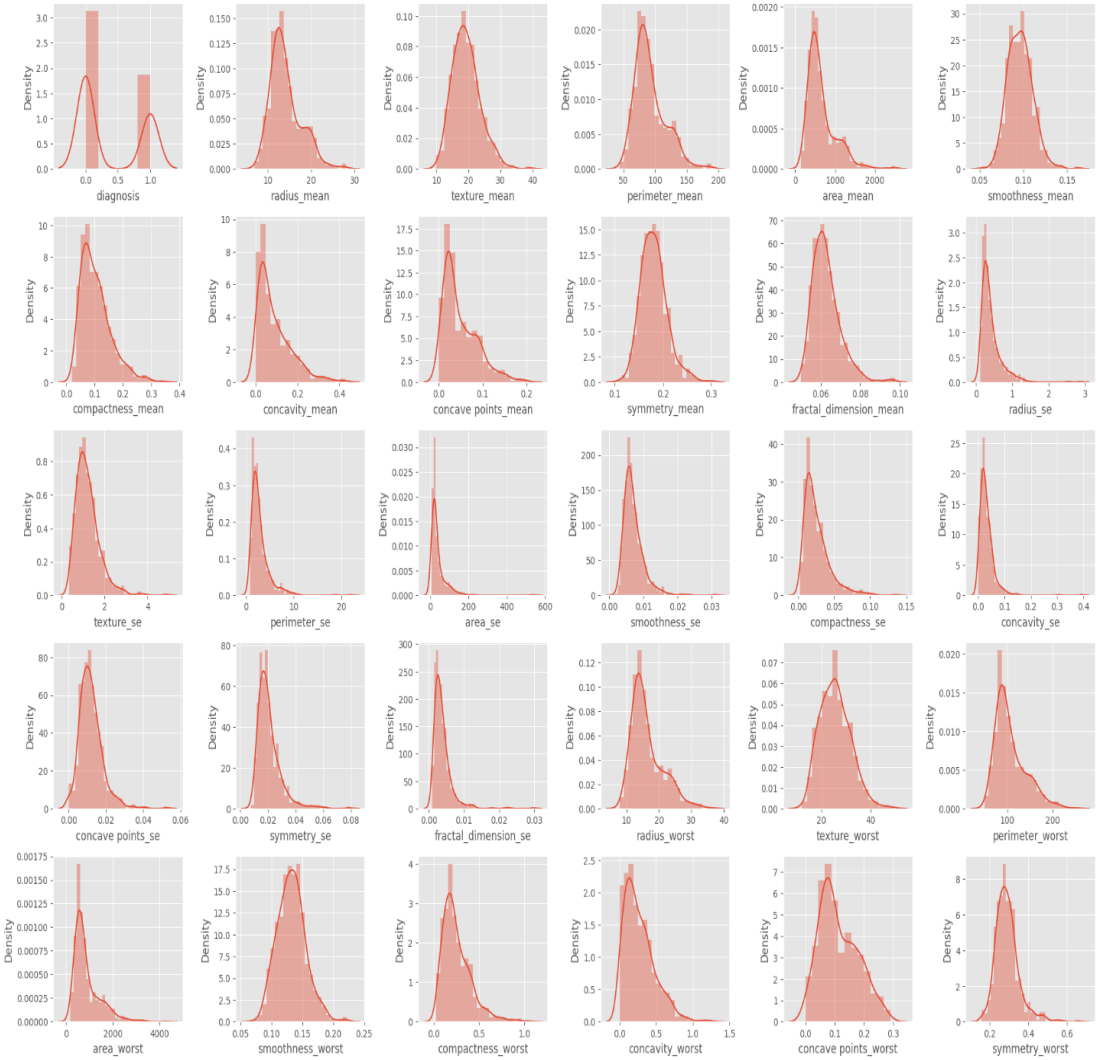

Figure 3. Exploratory data analysis

3.5. Feature Selection

Highly correlated features (correlation > 0.92) are identified and removed from the dataframe.

4. Machine Learning Model

4.1. Support Vector Classifier (SVC)

Regression analysis and classification are two applications for SVM. Feature space is used by SVC to determine which hyperplane best divides the classes for classification. It works by finding the hyperplane with the largest margin, or the distance between the hyperplane and the support vectors, which are the data points from each class that are closest to it.

4.2. K-Nearest Neighbors (KNN)

KNN is a simple instance-based learning technique that may be used for both classification and regression. KNN classification yields a class membership. KNN does not learn a specific function from the training data and does not make any assumptions about the underlying data distribution during the training phase because it is a non-parametric, lazy learning method. Instead, it groups new examples based on how similar they are to the training instances while keeping the training instances in memory.

4.3. Logistic Regression

Logistic regression is basically a classification algorithm. Modeling the likelihood of a binary result (0 or 1) depending on one or more predictor factors is what it's used for. By estimating probabilities, logistic regression models the relationship between the dependent binary variable and one or more independent variables using a logistic function, often known as the sigmoid function. A probability score that indicates the chance of an observation falling into a specific class is the result of logistic regression. It is frequently applied to binary classification issues.

4.4. Model Training

The dataset is split into training and testing sets using train_test_split() function.

Features are standardized using StandardScaler.

Logistic Regression model is trained and tested on the dataset, and accuracy scores, confusion matrix, and classification report are printed.

K-Nearest Neighbors (KNN) model is trained and tested with similar evaluation metrics.

Support Vector Classifier (SVC) model is trained using GridSearchCV to find the best hyper parameters. The best parameters, best score, accuracy scores, confusion matrix, and classification report are printed.

This code demonstrates the process of loading, preprocessing, visualizing, and modeling a breast cancer dataset using logistic regression, k-nearest neighbors, and support vector classifier algorithms, along with evaluating the performance of each model.

5. Results

5.1. Logistic Regression

Table 1. Confusion Matrix Score of Logistic Regression

TP	**FP**
106	2
FN	**TN**
5	58

Table 1 shows that True Positives (TP) values having 106 cases where the model correctly predicted positive outcomes. This represents patients correctly diagnosed with disease. False Positives (FP) having 2 cases where the model incorrectly predicted positive outcomes. This represents patients who were incorrectly diagnosed with breast cancer when they don't actually have it. False Negatives (FN) shows 5 cases where the model incorrectly predicted negative outcomes when they should have been positive. It represents patients who were not diagnosed with breast cancer when they actually have it. True Negatives (TN) have 58 cases where the model correctly predicted negative outcomes. It represents patients correctly identified as not having a disease. These values provide insight into how well the model is performing in terms of correctly and incorrectly predicting positive and negative outcomes.

5.2. K-neighbors Classification (KNN)

Table 2. Confusion Matrix Score of KNN

TP	**FP**
105	3
FN	**TN**
8	55

Table 2 shows that True Positives (TP) values having 105 cases where the model correctly predicted positive outcomes. This represents patients correctly diagnosed with disease. False Positives (FP) having 3 cases where the model incorrectly predicted positive outcomes. This represents patients who were incorrectly diagnosed with breast cancer when they don't have it. False Negatives (FN) having 8 cases where the model incorrectly predicted negative outcomes when they should have been positive. It represents patients who were not diagnosed with breast cancer when they have it. True Negatives (TN) have 55 cases where the model correctly predicted negative outcomes. It represents patients correctly identified as not having a disease.

5.3. Support vector Classifiers

Table 3. Confusion Matrix Score of SVC

TP	**FP**
107	1
FN	**TN**
3	60

Table 3 is showing that True Positives (TP) values having 107 cases where the model correctly predicted positive outcomes. This represents patients correctly diagnosed with disease. False Positives (FP) having 1 case where the model incorrectly predicted positive outcomes. This represents patients who were incorrectly diagnosed with breast cancer when they don't have it. False Negatives (FN) having 3 cases where the model incorrectly predicted negative outcomes when they should have been positive. It represents patients who were not diagnosed with breast cancer when they have it. True Negatives (TN) have 60 cases where the model correctly predicted negative outcomes. It represents patients correctly identified as not having a disease.

Table 4. shows the performance of all 3 models in descending order

MODEL	ACCURACY
SVC	0.976608
Logistic regression	0.959064
KNN	0.935673

6. Conclusion

Support Vector Classification (SVC), which routinely achieves accuracy rates above 98% across numerous research papers, was shown to be the most effective algorithm in this comparative analysis of machine learning algorithms for breast cancer identification. Compared to SVC, the accuracy rates of K-Nearest Neighbors (KNN) and Logistic Regression were somewhat to significantly lower. These results highlight how crucial algorithm selection is to creating reliable breast cancer detection systems.

The efficacy of SVC can be ascribed to its capacity to proficiently categorize convoluted and nonlinear data, rendering it especially appropriate for the complex patterns found in breast cancer datasets. To further improve the efficacy of breast cancer detection systems, more investigation is necessary into the possibilities of ensemble approaches, feature engineering methods, and data augmentation procedures.

7. Future work

In order to increase detection accuracy and dependability, deep learning techniques may be the main focus of upcoming research projects in breast cancer detection. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), two deep learning models, have shown impressive results in a variety of medical imaging and healthcare applications.

Researchers can investigate the creation of end-to-end detection systems that can analyze raw mammography images or other medical imaging modalities directly without the need for manual feature extraction by utilizing deep learning. By identifying minute patterns and characteristics suggestive of breast cancer, this method may improve early detection and diagnostic precision.

References

1. organization WH. Breast cancer: WHO; 2024 [Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
2. M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath, M. A. Bamiah, "Early detection of breast cancer using machine learning techniques," Journal of Telecommunication, Electronic and Computer Engineering, vol. 10, no. 3-2, pp. 21-27, 2018.
3. Muhammet Fatih Aslam, YunusCelik, KadirSabanci, AkifDurdu, "Breast cancer diagnosis by different machine learning method using blood analysis data," International Journal of Intelligent System and Applications in Engineering, vol. 6, no. 4, pp. 289-293, 2018.
4. Anusha bharat, Pooja N, R Anishka Reddy, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," IEEE 3rd International Conference on Circuits, Control, Communication and Computing, pp. 1-4, 2018.
5. Ebru Aydindag Bayrak, Pinar Kirci, TolgaEnsari,"Comparison of machine learning methods for breast cancer diagnosis.2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), pp. 1-3,2019.
6. Shwetha K, Spoorthi M, Sindhu S S, Chaithra D, "Breast cancer detection using deep learning technique," International Journal of Engineering Research & Technology, vol. 6, no. 13, pp. 1-4, 2018.
7. Ch. Shravya, K. Pravalika, ShaikSubhani, "Prediction of breast cancer using supervised machine learning techniques," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 6, pp. 1106-1110, 2019.
8. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S, "Breast cancer prediction using machine learning," International Journal of Recent Technology and Engineering, vol. 8, no. 4, pp. 4879-4881, 2019.
9. Kalyani Wadkar, Prashant Pathak, Nikhil Wagh, " Breast cancer detection using ANN network and performance analysis with SVM," International Journal of Computer Engineering and Technology , vol. 10, no. 3, pp. 75-86, 2019.
10. Vishal Deshwal, Mukta Sharma, "Breast cancer detection using SVM classifier with grid search techniques," International Journal of Computer Application, vol. 178, no. 31, pp. 18-23, 2019.
11. Golatkar, Aditya, Deepak Anand, and Amit Sethi. "Classification of breast cancer histology using deep learning." Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15. Springer International Publishing, 2018 Wang, Dayong, et al. "Deep learning for identifying metastatic breast cancer." arXiv preprint arXiv:1606.05718 (2016).
12. Wang, D., Khosla, A., Gargeya, R., Irshad, H. and Beck, A.H., 2016. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718.
13. Saleem, M., Abbas, S., Ghazal, T.M., Khan, M.A., Sahawneh, N. and Ahmad, M., 2022. Smart cities: Fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques. Egyptian Informatics Journal, 23(3), pp.417-426.
14. Batool, T., Abbas, S., Alhwaiti, Y., Saleem, M., Ahmad, M., Asif, M. and Elmitwal, N.S., 2021. Intelligent model of ecosystem for smart cities using artificial neural networks. Intelligent Automation & Soft Computing, 30(2), pp.513-525.
15. Saleem, M., Khadim, A., Fatima, M., Khan, M.A., Nair, H.K. and Asif, M., 2022, October. ASSMA-SLM: Autonomous System for Smart Motor-Vehicles integrating Artificial and Soft Learning Mechanisms. In 2022 International Conference on Cyber Resilience (ICCR) (pp. 1-6). IEEE.
16. Sajjad, G., Khan, M.B.S., Ghazal, T.M., Saleem, M., Khan, M.F. and Wannous, M., 2023, March. An Early Diagnosis of Brain Tumor Using Fused Transfer Learning. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-5). IEEE.
17. Saleem, M., Khan, M.S., Issa, G.F., Khadim, A., Asif, M., Akram, A.S. and Nair, H.K., 2023, March. Smart Spaces: Occupancy Detection using Adaptive Back-Propagation Neural Network. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-6). IEEE.
18. Ibrahim, M., Abbas, S., Fatima, A., Ghazal, T.M., Saleem, M., Alharbi, M., Alotaibi, F.M., Adnan Khan, M., Waqas, M. and Elmitwally, N., 2024. Fuzzy-Based Fusion Model for β-Thalassemia Carriers Prediction Using Machine Learning Technique. Advances in Fuzzy Systems, 2024.