International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue:12 | Dec 2024 www.irjet.net p-ISSN: 2395-0072

Machine Learning for the Identification and Prevention of Breast Cancer: A Literature Review

Tanushree T[1], Shreya Kanti M[2], Rashmitha K[3],Prathibha H B[4]

[1234]*Bachelor of Engineering, Information Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, affiliated to VTU Belagavi, Karnataka, India.*

Abstract - : *In developing countries, breast cancer is one of the main causes of mortality for women, and early detection and treatment are essential. Invasive ductal carcinoma (IDC) and ductal carcinoma in situ (DCIS) are the two primary subtypes of the illness. More precise diagnostic models are the result of developments in artificial intelligence (AI) and machine learning (ML). Magnetic resonance imaging (MRI) and convolutional neural networks (CNNs) are important diagnostic and prophylactic techniques for breast cancer. CNN-based models, such as CNNI-BCC, help classify subtypes of the disease. These models, however, demand a large amount of processing power. Using three feature selection techniques—low-variance elimination, recursive feature elimination, and univariate feature selection—this study suggests an efficient deep learning model for detecting breast cancer from digital mammograms with different densities. 3002 photos from 1501 patients who had digital mammograms between February 2007 and May 2015 are included in the dataset. The following six classification techniques were assessed: logistic regression (LR), support vector classifier (SVC), k-nearest neighbors (KNN), decision trees (DT), random forests (RF), and linear support vector classifier (linear SVC). The findings demonstrate that the suggested approach provides excellent accuracy and efficiency while using less processing resources.*

Key Words: Breast Cancer ; Invasive Ductal Carcinoma (IDC);Ductal Carcinoma In Situ (DCIS);Artificial Intelligence (AI); Convolutional Neural Networks (CNNs);Magnetic Resonance Imaging (MRI);Digital Mammograms.

1. INTRODUCTION

Cancer is a global health issue that affects people of all ages and backgrounds. Among various cancers, breast cancer is particularly prevalent in women and is the second wleading cause of cancer-related deaths in women, after lung cancer [1]. Early detection and diagnosis are critical for improving survival rates, and machine learning (ML) techniques can significantly enhance this process. Breast cancer develops in the breast cells and can present in several forms, including Ductal Carcinoma In Situ (DCIS), which is non-invasive, and Invasive Ductal Carcinoma (IDC), the most common form, accounting for 80% of cases [2]. Early symptoms may include unexplained swelling in the breast or armpit, changes in breast size or appearance, pain or discomfort, nipple discharge, or skin changes. For early detection, mammography remains a vital tool. However, mammography can be less effective in women with dense breast tissue, where diagnostic ultrasonography, thermography, or radiography might offer more precise results. Magnetic Resonance Imaging (MRI), particularly Dynamic Contrast-Enhanced MRI (DCE-MRI) [3], is highly sensitive for identifying and characterizing breast cancer lesions. Recent advances in artificial intelligence (AI) have further improved diagnostic accuracy by enabling AI-based systems to analyze medical images, such as mammograms, and identify cancerous lesions. In this context, deep learning models can be used to analyze computerized mammograms of varying breast tissue densities. This study proposes an effective deep learning approach to enhance breast cancer detection across different breast densities, ultimately improving the accuracy of early diagnosis.

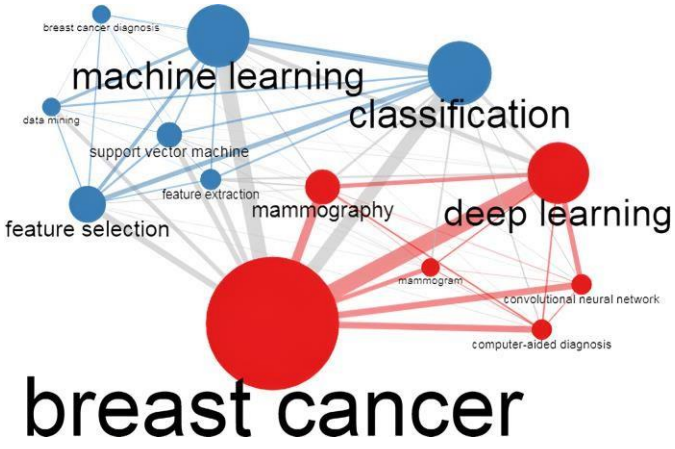

Fig-1: An analysis of author keywords using a co-occurrence network

The image highlights the relationship between advanced computational methods and breast cancer diagnosis. At its core, "breast cancer" connects to technologies like machine learning and deep learning. Machine learning is linked to terms such as "classification," "support vector machine," and "feature selection," emphasizing its role in predictive modeling , Deep learning focuses on applications like "convolutional neural networks," "mammography," and

"computer-aided diagnosis," enhancing imaging and automation. The diagram shows how these technologies work together to advance breast cancer detection and classification.

2. Literature review

In order to identify BC, a number of research have employed machine learning (ML) approaches in the healthcare industry in recent years. Other scientists have applied the algorithms to difficult problems because they yield satisfactory results [4]. In breast cancer photos, invasive ductal carcinoma was predicted and diagnosed using a CNN algorithm, and it succeeded in an approximate 88% accuracy rate [5,6]. Furthermore, it is frequently employed in the medical industry to diagnose and foresee abnormal events in order to gain a better knowledge of diseases that cannot be cured, like cancer [7]. Imaging and genetic-based breast cancer screening methods have been the subject of numerous investigations. Additionally, as far as we are aware, no research has been done that combines the two of these methods. . The authors of [8] provided an overview of the various methods for histological image analysis (HIA) in the detection of breast cancer. These methods are based on several kinds of convolutional neural networks (CNN) [9]. The authors categorized their work according to the type of dataset they used. Everything was arranged in reverse chronological order, starting with the most recent event. According to the findings of this study, ANNs were initially applied in the field of HIA somewhere around the middle of 2012. ANNs and PNNs were the most widely used algorithm types [10]. The suggested approach was created to assist medical professionals in determining whether to do a breast biopsy on a worrisome lesion following a review of mammography results. Several distinct classification algorithms are used in the random forest (RF) classifier, an ensemble approach. Using a decision tree, each of them can be implemented. Using multiple decision trees improves classification accuracy (Dai, Chen, et al., 2018) [11]. To put it simply, the RF is an ensemble classifier composed of several decision trees that cooperate to increase output and prediction precision. In Nguyen, Wang, et al. (2013) [12], scientists developed a classifier based on RF. The results of training the algorithm they employed on two datasets are promising: It performed well and with great accuracy.

3. Proposed Methodology

This study explores the use of six machine learning models for breast cancer diagnosis: Random Forest (RF), Decision Tree (DT), k-Nearest Neighbors (KNN), Logistic Regression (LR),Support Vector Classifier (SVC), and Linear SVC. Breast Cancer Wisconsin (Diagnostic) data was used to train and assess the models. To guarantee consistent and dependable results, the data underwent preprocessing procedures such as normalization, low-variance feature removal, and data balancing. To preserve the most informative qualities for classification, important feature selection strategies such recursive feature elimination and univariate selection were used. The study emphasizes the value of preprocessing, which involved separating the data into training (70%) and testing (30%) sets and eliminating duplicate values. In order to improve interpretability, decision trees were displayed and feature significance analysis was performed for tree-based classifiers such as RF and DT. To increase the models' resilience and dependability, ensemble approaches including bagging in RF and stacking methods were applied. To assess the effectiveness of the model, performance indicators such as accuracy, precision, recall, sensitivity, and F1-score were utilized. Additionally, deep learning architectures such as convolutional neural networks (CNNs) were used to investigate sophisticated picture categorization approaches. For deep feature extraction, pre-trained networks like VGG and ResNet were employed in addition to more conventional techniques like Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG). By utilizing the advantages of several feature extraction and classification techniques, this multimodal approach showed a notable improvement in the identification of benign and malignant tumors. Genetic Programming (GP) was also used in the study to optimize machine learning pipelines. The investigation of various pipeline topologies was made easier by GP's evolutionary methodology, which included mutation, crossover, and selection operators. Because of its flexibility, the system was able to improve classification accuracy, optimize hyperparameters, and refine preprocessing procedures. The pipelines that performed the best were chosen for deployment thanks to the iterative approach. Overall, the work shows that feature selection, preprocessing, ensemble learning, and evolutionary algorithms can be used to diagnose breast cancer. The findings highlight how well machine learning algorithms can classify tumors, opening the door for more advancements in clinical diagnostics.

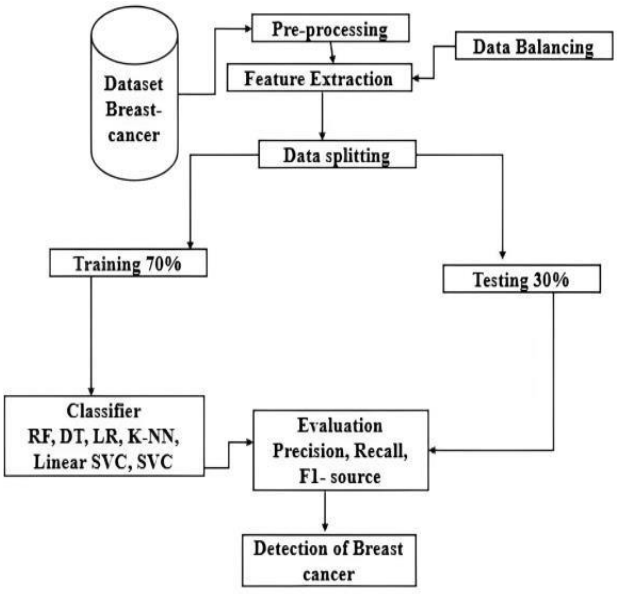

Fig-2: Proposed methodology.

4. Results and Discussion

Using the assessment measures EDA dataset, we have presented a systematic approach in this study to compare the state-of-the-art. We initially gathered the material from kaggle.com, which we accessed on April 15, 2023. For the simulation, we utilized Python Jupiter Notebook 6.4.12. The initial findings of explanatory data analyses employing many models. Python extensions like Pandas, Seaborn, Plotly, and Bokeh are examples of EDA resources. Pandas is a powerful data analysis library that provides Data Frames and Series. Plotly is a versatile framework for making interactive visualizations including dashboards and 3D plots, whereas Seaborn is a high-level interface for making statistical visualizations. A library called Bokeh is used to create interactive dashboards that can be used online. The predictions are applied to the dataset's subsets in this graph. Thirty percent of the test data and seventy percent of the training data make up each batch. The table provided a summary and comparison of various metrics, including radius mean and id.

Table-1: Exploratory data analysis

Smoothne ss Mean	Compact ness Mean	Concavity Mean	Concave Points Mean	Radiu s Worst
0.11840	0.27760	0.3001	0.14710	25.38
0.08474	0.07864	0.0869	0.07017	24.99
0.10960	0.15990	0.1974	0.12790	23.57
0.14250	0.28390	0.2414	0.10520	14.91
0.10030	0.13280	0.1980	0.10430	22.54
Smoothne ss Mean	Compact ness Mean	Concavity Mean	Concave Points Mean	Radiu s Worst

id	Diagnosis	Radius Mean	Texture Mean	Perimeter Mean	Area Mean
1	842302 M	17.99	10.38	122.80	1001.0
2	842517 M	20.57	17.77	132.90	1326.0
3	84300903M	19.69	21.25	130.00	1203.0
4	84348301M	11.42	20.38	77.58	386.1
5	84358402M	20.29	14.34	135.10	1297.0

Texture Worst	Perimeter Worst	Area Worst	Smoothness Worst	Compactne ss Worst
17.33	184.60	2019.0	0.1622	0.6656
23.41	158.80	1956.0	0.1238	0.1866
25.53	152.50	1709.0	0.1444	0.4245
26.50	98.87	567.7	0.2098	0.8663
16.67	152.20	1575.0	0.1374	0.2050

Concavity Worst	Concave Points Worst	Symmetry Worst	Fractal Dimension Worst
0.7119	0.2654	0.4601	0.11890
0.2416	0.1860	0.2750	0.08902
0.4504	0.2430	0.3613	0.08758
0.6869	0.2575	0.6638	0.17300
0.4000	0.1625	0.2364	0.07678

Without the exploratory data analysis (EDA) phase, the data analysis process described above would not be complete. In order to identify patterns, draw inferences, identify outliers, and formulate hypotheses, data analysis involves displaying the data. Before advancing to more intricate statistical and machine learning techniques, EDA is helpful for understanding the structure and characteristics of the dataset. Recognizing the dataset's characteristics, finding issues with data quality, and spotting early trends that could inspire additional modeling and analysis EDA is an essential step in the data analysis process since it is used to make all of the decisions. It is also essential to the reliability of the data-driven outcomes.

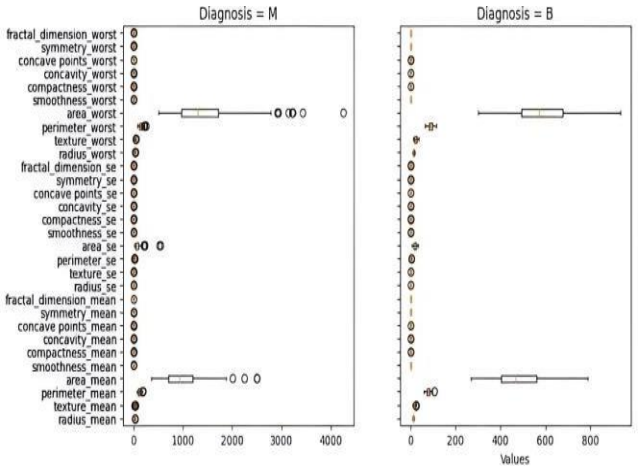

Fig-3: Boxplots of breast cancer dataset.

The boxplots of the dataset on breast cancer are shown in Figure 3. The distribution of breast cancer detection scores is seen in these boxplots in this figure. Every graph that we obtained had an abnormality. The median of the data will decrease if they are eliminated, which could make it more difficult to correctly diagnose tumors, especially if a malignant tumor is mistakenly classified as benign. However, in order to avoid interfering with the identification of serious cancers, we will only eliminate abnormalities in benign tumors. The maximum value of the "worst" area after processing is 932.7, compared to 1210.0 prior to processing. Prior to processing, the maximum area of the mean value was 992.1; now, it is 788.5. This could result in false positives, wherein benign tumors are incorrectly classified as malignant. On the other hand, this approach is better than the other. By identifying tumors at an early stage, when treatment is most effective, this diagnosis may save lives. In order to take an image, a mammographer will place the patient's breast on the plate of the mammography machine and gently compress it with a second plate to spread out the breast tissue. According to standard procedure, each breast is X-rayed twice: once from the side (mediolateral oblique view) and once from the top down (cranio-caudal view). There is the opportunity to add more views if necessary.

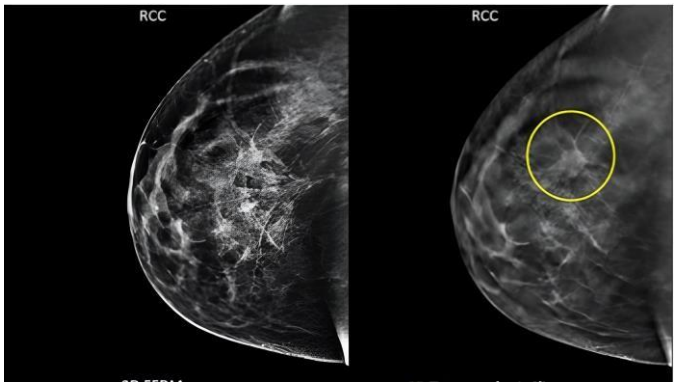

Fig-4: Breast cancer look on mammography.

Figure 4 illustrates how the location, size, and stage of breast cancer might affect its appearance on mammograms. To find breast cancer, radiologists look for anomalies and trends. Masses, microcalcifications, and architectural deformities are typical findings.Nodules, spiculated borders, and asymmetries. Microcalcifications are microscopic calcium deposits in breast tissue, whereas masses, microcalcifications, and architectural abnormalities are early indicators of breast cancer.K-fold cross-validation is a breast cancer screening method that divides a dataset into K equal-sized subgroups. The other K-1 subsets are used for training, and one of the K subsets is used as the test set. K evaluations and trainings are performed on the model. Performance metrics are calculated on the test set to assess how well the model generalizes to new data. After all K iterations, the performance metrics from each fold are averaged to provide a comprehensive overview.

Breast Cancer detection using machine learning: K-fold cross-validation is necessary for robustness, overfitting detection, optimal hyperparameter tuning, bias and variance assessment, and realistic performance estimations. It guarantees that the evaluation is more trustworthy and less dependent on a specific data split, allowing for more accurate and exact predictions of the model's performance on fresh, untested data.

The top 10 correlation findings across all variables are shown in e 6. Diagnostic, radius_mean, testure_mean, perimeter_mean, area_mean, smoothness_mean, and other parameters are displayed on the y-axis. Whereas a positive or negative value denotes a strong relationship between the two variables, a correlation coefficient between -1 and 1 implies no association at all. It is crucial to remember that correlation can only quantify the linear relationship between variables. Each of these indicators has a correlation with the patient's prognosis of at least 69%.Diagnostics 2023, 13, 3113 16 of 22 that the correlation can be used to make measurements. Each of these indicators has a correlation with the patient's prognosis of at least 69%.

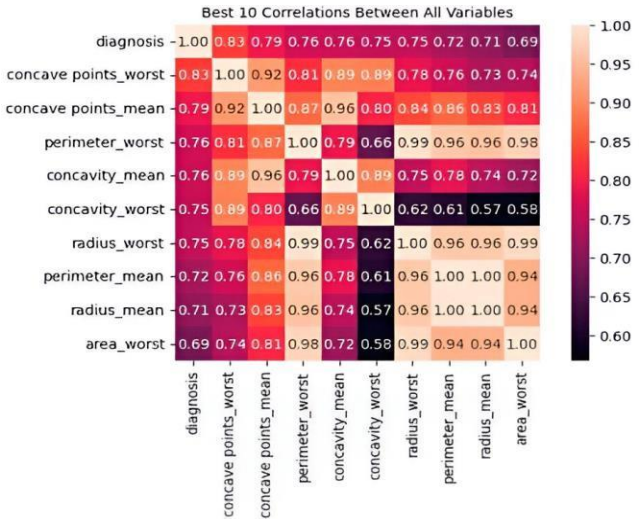

Fig-5: Correlation between the ten best variables.

	Model	Scores
0	Random Forest Classifier	96.491228
3	Decision Tree	93.859649
1	Logistic Regression	92.982456
2	KNeighbour Clasifier	92.105263
5	Linear SVC	89.473684
4	SVC	87.719298

Fig-6: Accuracy result for classifiers.

The accurate outcomes of every classifier we employed are displayed in Figure 6. With an accuracy of 96.49%, the random forest is the best result in this figure. The explanation of others are listed below.

* When compared to other models, random forest models perform better at predicting the target variable, as seen by their greatest accuracy scores.

* KNN and decision trees are comparable to other models and have greater values than LG and SVC, suggesting that they might not be the best models for predicting the target variable.

* SVC and logistic regression both perform well in predicting the target variable, as evidenced by their comparable values.

Because of its adaptability, simplicity in interpretation, and capacity to identify which characteristics are nearly essential to the categorization decision-making process, random forest may be a good substitute for other methods when it comes to cancer detection.

3. CONCLUSIONS

Using the Breast Cancer Wisconsin (diagnostic) dataset, this study assesses six machine learning models for the categorization of breast cancer. The Standard Scaler module was used for data preparation, and scikit-learn in Python was used for feature selection. Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Support Vector Classifier (SVC) are among the models that were put to the test. Confusion matrices were used to evaluate performance parameters such accuracy, precision, recall, sensitivity, F1-score, and AUC. Correlations between characteristics and patient outcomes were emphasized by exploratory data analysis, which showed that RF had the highest accuracy, followed by DT and KNN. SVC and logistic regression performed similarly. RF is notable for its adaptability, readability, and capacity to recognize important characteristics. Given that breast cancer can be classed as either invasive ductal carcinoma (IDC) or ductal carcinoma in situ (DCIS), the study highlights the importance of early identification. Diagnostic imaging methods such as thermography, ultrasonography, and mammography are essential. Advances in artificial intelligence, namely in the areas of deep learning and convolutional neural networks, are boosting the accuracy of mammography analysis. An extremely sensitive imaging technique is still breast MRI. To further enhance breast cancer detection and treatment, future studies should investigate the incorporation of cutting-edge AI approaches and promote cooperation between data scientists, medical experts, and researchers. This could revolutionize early diagnosis and patient outcomes.

REFERENCES

[1] Bayrak, E.A.; Kırcı, P.; Ensari, T. Comparison of machine learning methods for breast cancer diagnosis. In Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 24–26 April 2019.

[2] Loomans-Kropp, H.A.; Umar, A. Increasing Incidence of Colorectal Cancer in Young Adults. *J. Cancer Epidemiol.* **2019**, *2019*, 9841295. [CrossRef] [PubMed].

[3] Din, N.M.U.; Dar, R.A.; Rasool, M.; Assad, A. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. *Comput. Biol. Med.* **2022**, *149*, 106073. [CrossRef] [PubMed].

[4] Pang, T.; Wong, J.H.D.; Ng, W.L.; Chan, C.S. Deep learning radiomics in breast cancer with different modalities: Overview and future. *Expert Syst. Appl.* **2020**, *158*, 113501. [CrossRef]

[5] Zhou, X.; Li, C.; Rahaman, M.M.; Yao, Y.; Ai, S.; Sun, C.; Wang, Q.; Zhang, Y.; Li, M.; Li, X.; et al., A comprehensive review for breast histopathology image analysis using classical and deep neural networks. *IEEE Access* **2020**, *8*, 90931–90956. [CrossRef]

[6] Wisesty, U.N.; Mengko, T.R.; Purwarianti, A. Gene mutation detection for breast cancer disease: A review. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *830*, 032051. [CrossRef]

[7] Mahmood, T.; Li, J.; Pei, Y.; Akhtar, F.; Imran, A.; Rehman, K.U. A Brief Survey on Breast Cancer Diagnostic with Deep Learning Schemes Using Multi-Image Modalities. *IEEE Access* **2020**, *8*, 165779–165809. [CrossRef]

[8] Sutanto, D.H.; Ghani, M.K.A. A benchmark of classification framework for non-communicable disease prediction: A review. *ARPN J. Eng. Appl. Sci.* **2015**, *10*, 9941–9955.

[9] Gautam, R.; Kaur, P.; Sharma, M. A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings. *Prog. Artif. Intell.* **2019**, *8*, 401–424. [CrossRef]

[10] Mahmood, M.; Al-Khateeb, B.; Alwash, W.M. A review on neural networks approach on classifying cancers. *IAES Int. J. ArtifIntell.* **2020**, *9*, 317–326. [CrossRef]

[11] Dai, B.; Chen, R.-C.; Zhu, S.-Z.; Zhang, W.-W. Using Random Forest Algorithm for Breast Cancer Diagnosis. In Proceedings of the 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 6–8 December 2018; pp. 449–452. [CrossRef].

[12] Nguyen, C.; Wang, Y.; Nguyen, H.N. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. Biomed. Sci. Eng.* **2013**, *06*, 551–560. [CrossRef].