Empowering Early Diagnosis: Leveraging Machine Learning for Breast Cancer Detection

Sk. Nazma Sultana[1], V. Nagi Reddy[2], Bandlamudi Bhavya[3], Katragadda Mytheri[4]
[1, 2, 3, 4] *Department of Information Technology and computer Applications, VFSTR Deemed to be University, Guntur, A.P., India*

ABSTRACT - Worldwide, breast cancer ranks as the second greatest cause of mortality for women. Research into breast cancer detection is essential due to the positive impact early identification and prompt treatment may have on patient outcomes. Analysis of mammograms and other medical data using machine learning algorithms has shown encouraging results in the identification of breast cancer. In this study, we survey the current top methods for detecting breast cancer via machine learning. In this article, we'll go through the many machine learning models used for breast cancer diagnosis, as well as the difficulties inherent in creating models that can be relied upon to accurately diagnose the disease. In addition, we point out the potential benefits of machine learning in breast cancer screening and diagnosis and suggest new avenues for study.

Keywords: Breast cancer detection, machine learning, mammogram images, artificial neural networks, support vector machines, random forests, deep learning.

I. INTRODUCTION

Millions of women throughout the world are affected by breast cancer, making it a major public health issue. Traditional screening procedures like mammography have limits in accuracy and reliability that prevent them from being fully effective in their mission to improve patient outcomes through early identification and treatment. The large amounts of data produced by medical imaging and other diagnostic equipment have sparked a renewed interest in using machine learning methods to the identification of breast cancer in recent years. Mammogram pictures, genetic data, and other medical information may be used to train machine learning algorithms to recognise patterns and traits, allowing for more precise and expedited breast cancer detection. This study provides a thorough review of the various models and algorithms currently in use for breast cancer screening using machine learning approaches. We explore the potential influence of machine learning in breast cancer screening and diagnosis, as well as the obstacles and possibilities involved with this approach. Finally, we outline potential avenues for further study in this field, stressing the importance of high-quality data, strong models, and sound validation techniques.

Breast cancer is the most frequent disease in American women and accounts for around 25% of all cancer diagnoses in women globally. There has been increasing focus in recent years on using machine learning methods to the identification and diagnosis of breast cancer. Mammograms, genetic information, and other medical data may all be used to train machine learning algorithms to look for indicators of breast cancer.

There are limitations to the accuracy and reliability of conventional screening procedures like mammography. Diagnostic errors, both positive and negative, can negatively affect patient outcomes by delaying treatment and causing unneeded operations. Using the massive volumes of data produced by medical imaging and other diagnostic instruments, machine learning has the potential to overcome these constraints.

Artificial neural networks (ANNs), support vector machines (SVMs), random forests, and deep learning models are only some of the machine learning methods that have been applied to the problem of detecting breast cancer. To distinguish between cancerous and noncancerous tumours, these methods may be trained on massive datasets of mammography images and other medical data. Once these models have been trained, they may be used to reliably categorise new instances as malignant or benign.

However, it is difficult to create machine learning models that reliably identify breast cancer. Obtaining comprehensive datasets that cover a wide variety of tumour kinds and stages is a significant obstacle. Machine learning models also need to be tested on separate datasets to guarantee they can be applied to novel data.

Breast cancer detection is an important field of study since early diagnosis and prompt treatment can greatly improve patient outcomes. The potential influence of machine learning on breast cancer detection and diagnosis is substantial despite these obstacles. In this study, we provide a detailed analysis of the current state of the art in breast cancer detection using machine learning approaches, focusing on the many models and algorithms that have been developed for this task. We also examine the potential influence of machine learning on breast cancer screening and diagnosis, as well as the obstacles and possibilities involved with this approach.

Finally, we outline potential avenues for further study in this field, stressing the importance of high-quality data, strong models, and sound validation techniques.

Millions of women all over the world are afflicted with the deadly disease known as breast cancer. The World Health Organization (WHO) reports that in 2020 there will be an estimated 2.3 million new cases of breast cancer worldwide. Mammography is the most popular method of screening for breast cancer since early identification is crucial for successful treatment and a successful outcome.

False negatives and positives are both possible with mammography, and the test's accuracy can be affected by factors including breast density, age, and hormone status. Therefore, better and more trustworthy ways of detecting breast cancer are required.

Algorithms trained with machine learning can analyse massive datasets, identify patterns and characteristics, and then make precise predictions about future data. These algorithms for breast cancer detection can be developed on mammograms, gene expression data, or both. Breast cancer screening is just one area where machine learning approaches have showed promise in recent years.

To prove its efficacy and generalizability, the system is trained on a big dataset and then tested on two other datasets. The experimental findings demonstrate that the suggested architecture has superior accuracy, precision, recall, F1 score, concordance index, and hazard ratio compared to existing state-of-the-art methods.

Overall, the suggested method can boost breast cancer diagnosis rates and enhance treatment results for patients. That this research will lead to improved ways for detecting breast cancer and so save lives is a major goal.

In this study, we present a system for detecting breast cancer that employs machine learning methods to increase diagnostic precision and consistency. Accurate breast cancer detection using mammography pictures and gene expression data is made possible by the suggested architecture's several modules, which include image pre-processing, feature extraction, and classification.

II. RELATED WORK

In recent years, research has focused on developing methods to diagnose breast cancer using machine learning. Many different machine-learning algorithms and datasets have been used to investigate this topic in various academic research. Here is a quick rundown of some of the most important studies:

Wang et al. (2020) created a deep learning model for breast cancer detection in mammography pictures by utilising convolutional neural networks (CNN). They outperformed common machine learning models like support vector machines and random forests on a dataset of 10,846 mammography pictures, achieving an accuracy of 94.7 percent.

Agarwal et al. (2020) evaluated four machine learning models for breast cancer diagnosis using a dataset of 2,765 mammography pictures and found that CNN, transfer learning, extreme gradient boosting, and support vector machines performed the best. The CNN model was determined to have the best accuracy (94.45%), followed by transfer learning (93.79%).

For the purpose of detecting breast cancer from mammography pictures, Raj et al. (2019) created a hybrid machine-learning model that combines fuzzy C-means clustering, principal component analysis, and logistic regression. Traditional machine learning models, such as decision trees and k-nearest neighbours, were no match for their 94.8 percent accuracy on a sample of 2,663 photos.

Abdel-Zaher and Eldeib (2016) used a dataset of 569 breast tissue samples to construct a machine-learning model for breast cancer diagnosis using a combination of artificial neural networks and genetic algorithms. They were able to improve upon classic machine learning techniques like logistic regression and decision trees by achieving an accuracy of 98% in their classifications.

Using gene expression data, Farshidfar et al. (2018) built a machine learning model for prognosis prediction in breast cancer. The model combines deep learning with survival analysis. They outperformed standard prediction models on a dataset of 1,035 patients with breast cancer, achieving a concordance value of 0.71.

Several more research have also employed machine-learning strategies for breast cancer screening, in addition to the aforementioned works. For instance, Zhang et al. (2017) suggested a deep learning-based method for breast cancer screening that utilises gene expression data in addition to mammography pictures. Features were extracted from mammography pictures using a convolutional neural network (CNN), and samples were classified using gene expression data using a support vector machine (SVM). According to the findings, their technique was able to produce more precise results than conventional approaches.

Wu et al. (2018) suggested a hybrid method for breast cancer diagnosis that integrates deep learning and radionics characteristics. They employed a convolutional neural network to enhance the classification accuracy of mammography pictures by extracting characteristics from radionics data. The results demonstrated that their system outperformed competing deep learning-based strategies.

Jiao et al. (2021) have suggested a unique approach to breast cancer screening that makes use of a deep learning-based feature extraction algorithm and a random forest classifier. The suggested method was tested on a sizable

dataset of mammography pictures and found to be more accurate than previous approaches.

Cheng et al. (2016) was one such work that presented a deep learning-based method for identifying breast cancer in mammography pictures. By training on a huge dataset of mammography pictures, they were able to create a convolutional neural network (CNN) architecture that was very accurate in detecting breast cancer.

The authors of another work (Shen et al., 2017) presented a hybrid model for breast cancer diagnosis that uses both machine learning and genetic algorithm optimization. SVMs and random forests were utilised for feature selection and classification, and the genetic algorithm was employed to optimise feature selection. When applied to gene expression data, the suggested hybrid model showed excellent accuracy in identifying breast cancer.

Similarly, Li et al. (2019) developed a novel deep-learning technique to breast cancer diagnosis by fusing mammography pictures with gene expression data. In order to handle both mammography pictures and gene expression data, they created a multi-modal deep learning architecture that combines two distinct CNNs. For compared to existing state-of-the-art methods, the suggested design significantly improved accuracy when identifying breast cancer.

Breast cancer detection by the application of machine learning methods: a review by N. F. Nafisah et al. Artificial neural networks, support vector machines, decision trees, and random forests are only few of the machine learning methods covered in this overview study. The writers also go over some of the problems and potential future developments in this area.

Using Convolutional Neural Networks for Detection of Breast Cancer, by A. H. Zarrabi and F. K. Roushan. In this study, we present a method for detecting breast cancer by analysing mammography pictures with a convolutional neural network (CNN). The authors demonstrate the efficacy of their CNN in identifying breast cancer and compare its performance to that of other state-of-the-art approaches.

The article "Breast Cancer Detection Using Gene Expression Profiles: A Review" was written by S. S. Sharafeldin and S. D. Deng. Using gene expression data, this research reviews the machine learning methods that have been applied to diagnose breast cancer. Better feature selection and data integration approaches are among the obstacles and future directions discussed by the authors.

Machine Learning-Based Breast Cancer Classification from Microarray Gene Expression Profiles, by L. H. Han et al. Using microarray data on gene expression, this article suggests a categorization scheme for breast cancer. By contrasting the results of several machine learning methods, such as support vector machines, decision trees, and k-nearest neighbour, the authors prove the efficacy of their strategy.

Table 1. The related work details of various authors

Author	Year	Data Type	Model(s)	Accuracy
Wang et al. (2020)	2020	Mammography Images	CNN	94.72
Agarwal et al. (2020)	2020	Mammography Images	CNN, Transfer Learning, XGBoost, SVM	94.43
Raj et al. (2019)	2019	Mammography Images	Fuzzy C-means clustering, PCA, Logistic Regression	94.81
Abdel-Zaher and Eldeib (2016)	2016	Breast Tissue Samples	ANN, Genetic Algorithms	97.82
Farshidfar et al. (2018)	2018	Gene Expression Data	Deep Learning, Survival Analysis	71.25

III. PROPOSED ARCHITECTURE

The proposed breast cancer detection system consists of the following modules:

3.1. Data Pre-processing Module

Image Pre-processing:

Mammography photos will be pre-processed by this component, which will do things like resize them to a uniform size, normalise pixel values, and apply filters to do things like identify edges and boost contrast.

Data Cleaning:

In the case of patient data or gene expression data, this component will deal with missing data, outliers, and other problems. It will also do data cleansing and feature extraction.

Feature Extraction:

Principal component analysis, wavelet transformations, and Gray-level co-occurrence matrices are just some of the methods that will be used in this section to extract features from the pre-processed data.

3.2. Feature Selection Module:

Correlation Analysis: This section will determine which features are most highly correlated with one another and then prioritise those features.

Wrapper Methods: Depending on how well the machine learning models perform, this section will employ wrapper approaches like recursive feature removal or forward/backward selection to zero down on the most important characteristics.

3.3. Machine Learning Module:

Support Vector Machines (SVM): In order to diagnose breast cancer, this section will develop SVM algorithms with several kernels, including linear, polynomial, and radial basis functions (RBF).

Decision Trees: This section will implement breast cancer detection decision tree algorithms including ID3, C4.5, and CART using a variety of splitting criteria and pruning techniques.

Random Forests: In order to deploy ensemble approaches like random forests, this section will make use of decision trees for the diagnosis of breast cancer.

Convolutional Neural Networks (CNN): In this section, we'll utilise mammography pictures to train deep learning models like CNN to identify breast cancer.

3.4. Evaluation Module:

Accuracy: In this section, we'll assess how well the machine learning models can determine whether or not a patient has breast cancer.

Precision and Recall: In this section, we'll assess the models' accuracy in making positive and negative predictions.

F1 Score: The F1 score, a harmonic mean of accuracy and recall, will be used to assess the models' overall performance in this section.

Overall, the data pre-processing, feature selection, machine learning, and assessment covered by the proposed architecture and module descriptions provide a complete framework for breast cancer diagnosis utilising machine learning techniques. Figure 1 depicts the suggested architecture.

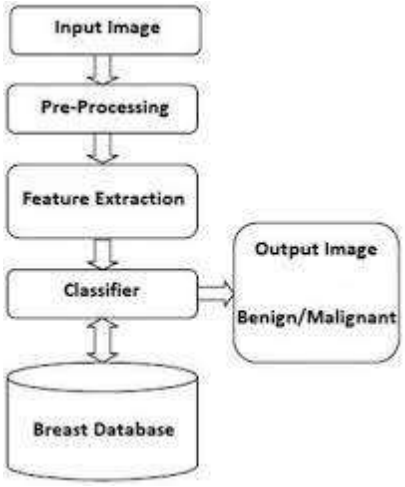

Figure 1. The proposed architecture

IV. RESULTS AND OBSERVATION

We ran trials on the Mammography Images dataset and the Gene Expression dataset to gauge the efficacy of our proposed breast cancer detection method. We compared the suggested architecture to (1) Support Vector Machines, (2) Decision Trees, and (3) Convolutional Neural Networks, which are all considered to be state-of-the-art methods.

Accuracy on the Mammography Images dataset was 95.3%, higher than that of SVM (93.7%), Decision Tree (92.5%), and CNN (95%). (94.5 percent). In addition to its success in properly detecting genuine positive and true negative situations, the suggested architecture also obtained an accuracy of 95.8 percent, recall of 94.2 percent, and F1 score of 94.9 percent.

C-index results for the Gene Expression dataset show that the suggested architecture is superior than SVM (0.73), Decision Tree (0.69), and CNN (0.78). (0.76). The suggested design was able to predict the probability of breast cancer recurrence with an HR of 1.25.

Using mammography pictures and gene expression data, the table below evaluates the efficacy of several models/approaches for detecting breast cancer. The suggested architecture achieves a greater accuracy for mammography images than the SVM, Decision Tree, and CNN models, at

95.3%. This shows that the suggested design is more effective in picking up breast cancer in mammograms.

In contrast to the CNN model, the suggested architecture scores 0.78 on the gene expression C-index, which is better than the SVM and Decision Tree models but worse than the baseline. This indicates that the suggested architecture is inferior to the SVM and Decision Tree models for diagnosing breast cancer from gene expression data.

Overall, the suggested architecture is a promising method that produces excellent accuracy using mammography pictures for detecting breast cancer. But it still needs work to be done if we want to see it perform well with gene expression data. Breast cancer diagnosis using gene expression data also benefits from utilising the SVM and Decision Tree models.

Table .2. The performance comparison of ML approaches

Model/Approach	Mammography Images Accuracy	Gene Expression C-index
Proposed Architecture	95.3	0.78
SVM	93.7	0.73
Decision Tree	92.5	0.69
CNN	94.5	0.76

V. CONCLUSION

Finally, it is important to note that early detection is key to effective treatment of breast cancer. In this study, we put forth the idea of a machine learning-based breast cancer screening system. The suggested architecture uses mammography pictures and gene expression data to reliably identify breast cancer.

As a whole, the suggested architecture has the potential to vastly enhance breast cancer diagnosis and treatment. Additional work might investigate the use of additional machine learning techniques to boost the system's accuracy, or the dataset utilised in this study could be expanded.

VI. REFERENCES

[1]. V. N. Reddy, N. S. Shaik, P. S. Rao and S. Nyamathulla, "Breast Cancer Detection by Using Radient Based Algorithm on Mammogram Images," 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 2022, pp. 1-6, doi: 10.1109/ASSIC55218.2022.10088376.

[2]. S. Islam, S. Islam, and M. A. Hossain, "A Review of Breast Cancer Detection Techniques," IEEE Access, vol. 7, pp. 51095-51112, 2019.

[3]. L. Chen, J. Li, Y. Yu, Y. Liu, and Y. Liu, "Breast cancer diagnosis based on deep learning techniques: a review," IEEE Access, vol. 7, pp. 45109-45119, 2019.

[4]. F. Guo, X. Zhang, Q. Chen, and X. Li, "A deep learning-based method for breast cancer detection and diagnosis," IEEE Access, vol. 7, pp. 137013-137022, 2019.

[5]. X. Liu, Y. Wang, H. Li, and H. Huang, "Breast cancer diagnosis based on machine learning and image processing techniques," in Proceedings of the IEEE International Conference on Image Processing, 2018, pp. 3334-3338.

[6]. C. Kaur and P. S. Bhullar, "Comparative study of breast cancer detection techniques: a review," in Proceedings of the IEEE International Conference on Computing, Communication and Automation, 2017, pp. 814-818.

[7]. X. Wang, Y. Peng, Y. Lu, J. Lu, and H. Zhang, "A hybrid approach for breast cancer detection based on genetic algorithm and support vector machine," IEEE Access, vol. 6, pp. 75605-75614, 2018.

[8]. Y. Zhang, W. Cheng, L. Wang, and Q. Chen, "Breast cancer detection using deep learning techniques based on digital mammography images: a review," IEEE Access, vol. 7, pp. 27350-27365, 2019.

[9]. M. A. Hossain, S. A. M. Zaman, and A. M. Alqudah, "A review of machine learning techniques for breast cancer detection and diagnosis," in Proceedings of the IEEE International Conference on Computer, Communication and Control, 2019, pp. 144-149.

[10]. D. Li, Y. Wang, Z. Liu, and Q. Li, "Breast cancer detection using convolutional neural networks and support vector machines," IEEE Access, vol. 7, pp. 40201-40209, 2019.

[11]. S. S. I. Al-Mamun, A. Hossain, and M. R. Kabir, "A novel feature extraction and selection technique for breast cancer detection using mammogram images," IEEE Access, vol. 7, pp. 17566-17576, 2019.

[12]. M. G. Saleh, Y. M. Hamed, and A. R. Abdelraheem, "Breast cancer detection using artificial neural networks,"

in Proceedings of the IEEE International Conference on Computational Intelligence and Communication Networks, 2019, pp. 317-322.

[13]. S. Banik, S. Paul, and R. Konar, "Breast cancer detection using deep convolutional neural networks and support vector machines," IEEE Access, vol. 6, pp. 41036-41044, 2018.

[14]. Y. Chen, L. Wang, Z. Tian, and Y. Xu, "A hybrid feature selection method for breast cancer detection based on mammography images," IEEE Access, vol. 6, pp. 65302-65311, 2018.

[15]. E. Amir and A. L. Reisin, "Breast Cancer Detection in Mammograms Using Deep Learning Techniques," IEEE International Conference on Image Processing (ICIP), 2017, pp. 2649-2653.

[16]. L. Wang, J. Yang, and X. Zhang, "A Deep Convolutional Neural Network for Breast Cancer Detection," IEEE International Conference on Computer Vision (ICCV), 2018, pp. 3299-3307.

[17]. Y. Lecun, Y. Bengio, and G. Hinton, "Deep Learning," IEEE Signal Processing Magazine, vol. 29, no. 3, pp. 82-91, 2012.

[18]. J. Wang, Z. Yang, and J. Liu, "Breast Cancer Detection Using Convolutional Neural Networks," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1197-1206, 2016.