Breast Cancer Prediction from Multimodal Datasets Using Deep Learning Techniques

R. Sathya[1], P.Rohini [2]
[1] Student, Department of CSE, Chendhuran College of Engineering&Technology, Pudukkottai, India
[2]Assistant Professor, Department of CSE, Chendhuran College of Engineering&Technology, Pudukkottai, India
Email id : sathyaravikumar1295@gmail.com[1], rohini.ccet@gmail.com[2]

Article Received: 28 April 2025 *Article Accepted: 29 April 2025* *Article Published: 30 April 2025*

Citation
R. Sathya, P.Rohini, "Breast Cancer Prediction from Multimodal Datasets Using Deep Learning Techniques", Journal of Next Generation Technology (ISSN: 2583-021X), vol. 5, no. 2, pp. 113-121. April 2025. DOI: 10.5281/zenodo.15633280

ABSTRACT: Using deep learning techniques to predict breast cancer from histopathology datasets is a major breakthrough in medical diagnosis. Usually, the procedure starts with the acquisition of histological pictures of samples of breast tissue, frequently taken from biopsy slides. The malignant (cancerous) or benign (non-cancerous) status of each sample is then carefully noted on these photos. Preprocessing techniques including scaling, normalization, and augmentation are used to improve dataset variety and reduce overfitting problems in order to guarantee optimal performance. Convolutional Neural Networks (CNNs) are the recommended option because of their proficiency in extracting hierarchical information from images. Next, an appropriate deep learning architecture is chosen. The dataset must be separated into training, validation, and test sets in order to train the model. The training set is used to train the model, and its performance is used to adjust the hyperparameters. thus avoiding overfitting on the validation set. The model's performance is thoroughly assessed on the test set after training using a variety of metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, to determine how well it generalizes to new data. In order to provide insights into the model's behavior, methods such as the sequential framework are also used to illustrate the predictions and decision-making process. Once the model performs satisfactorily, it can be used in clinical settings as an independent program or incorporated into pre-existing medical systems, guaranteeing adherence to ethical and regulatory requirements.

Keywords: Convolutional Neural Networks, Deep Learning, Breast Cancer, Multi model Dataset.

I. INTRODUCTION

A subset of machine learning known as "deep learning" entails teaching multi-layer artificial neural networks to identify patterns in data. Deep learning algorithms can be applied to a variety of tasks, including natural language processing, picture and audio recognition, and even playing games like chess and go. Deep learning's primary benefit over conventional machine learning techniques is its capacity to automatically extract features from unprocessed data without the requirement for feature engineering. This is achieved by stacking several layers of neurons, each of which transforms the incoming data in a nonlinear way.The network can progressively learn more intricate representations of the input data by using the output of one layer as the input for the subsequent layer.

Algorithms include Generative Adversarial Networks (GANs) for producing lifelike images and videos, Convolutional Neural Networks (CNNs) for processing images and videos, and Recurrent Neural Networks (RNNs) for processing sequential data, including natural language processing. Large volumes of labeled data and substantial processing power are needed to train deep learning models. Artificial neural networks, the foundation of deep learning algorithms, draw inspiration from the composition and operations of the human brain.TheNetworks process information hierarchically and are made up of layers of connected nodes, or neurons. The network's initial layer receives the input data and uses it to extract fundamental features. This layer's output is then sent to the layer after it, which uses the output from the layer before it to extract increasingly intricate features, and so on. In order to reduce the discrepancy between the expected and actual output, a deep learning model is trained by modifying the weights and biases of the network's neurons. Stochastic gradient descent is the most often used optimization technique. 4. A primary benefit of deep learning is its capacity to manageunstructureddata,including as text, pictures, and videos. While recurrent neural networks(RNNs) are better suited for sequential data processing, including natural language processing, convolutional neural networks (CNNs) are especially good at processing pictures and video. Numerous industries, including healthcare, banking, and transportation, have been significantly impacted by deep learning[1]. Deep learning algorithms, for instance, are used in finance to identify fraudulent transactions, in medical imaging to aid in the diagnosis of diseases like cancer, and in transportation to enhance the performance of self-driving automobiles.

Deep learning is not without its difficulties, though. To effectively train the models, a significant amount of labeled data is required, which is one of the main obstacles. This might be very difficult for applications when gathering data is costly or sparse.

Furthermore, deep learning models are frequently "black boxes," which makes it difficult to understand how the model makes its predictions. Applications like healthcare and banking, where interpretability is crucial, may find this troublesome.

II. LITERATURE SURVEY

Md. Milon Islam, among others,[2] offered a comparison of five machine learning methods for breast cancerprediction,includingsupportvector machine, logistic regression, random forests, K-nearest neighbors, and artificial neural networks. Each of the five machine learning techniques' fundamental characteristics and operation were demonstrated. The ANNs achieve the highest accuracy of 98.57%, while the RFs and LR yield the lowest accuracy of 95.7%. In the medical field, the diagnosing process is both time-consuming and highly costly. According to the system, machine learning techniques can be used as a clinical aid for breast cancer diagnosis and will be highly beneficial for new medical professionals in the event that a diagnosis is made incorrectly. We can infer from the findings that machine learning methods can accurately and automatically identify theillness. One effective supervised classification method is the random forest classifier. One ensemble technique that can be analyzed as a variation of the nearest neighbor predictor is the RF classification. Ensemble learning is the process of intentionally developing and integrating statistical techniques, such as classifiers or experts, to address a particular computational intelligence issue. Instead of producing a single classification tree from a given dataset, RF creates a forest of classification trees.

Amik Rawal, et al.[3] compares the effectiveness of four classifiers, including SVM, Random Forest, Logistic Regression, and kNN, which are among the,most important algorithms for data mining. It can be identified by a portable cancer diagnostic instrument or by a mammogram during a screening check. Cancer staging can be directly connected to the changes in cancerous breast tissues as the disease progresses. The degree to which a patient's breast cancer has spread is indicated by their stage (I–IV). Stages are determined by statistical indications such tumor size, distant metastases, lymph node metastases, and so forth. Patients must endure breast cancer surgery, chemotherapy, radiation therapy, and endocrine therapy to stop the cancer from spreading. The objective of the The goal of the research is to distinguish between malignant and benign patients, as well as to determine how to best parameterize our classification methods to attain high accuracy. Genetic alterations and mutations are among the causes of breast cancer. There are numerous varieties of breast cancer, but two prevalent ones are invasive carcinoma and ductal carcinoma in situ (DCIS). Others, such as angiosarcoma and phyllodes 10 tumors, are less frequent. Numerous algorithms are available for classifying the outcomes of breast cancer. Breast cancer side symptoms include pain, headaches, exhaustion, and osteoporosis, bone loss, and peripheral neuropathy, or numbness. Numerous algorithms are available for categorizing and forecasting the course of breast cancer.

Reza Rabiei, among others,[4] said machine learning techniques could forecast breast cancer since early identification of the condition could help reduce the progression of the illness and lower the death rate by using the proper therapeutic measures at the right moment. Using various machine learning techniques, having access to larger datasets from various institutions (multi-center study), and taking into account important characteristics from numerous pertinent data sources may enhance modeling effectiveness. The following are regarded as limitations in the current study: modeling based on information from a single database and the absence of genetic data access that might have an impact on the study's conclusions.

Nevertheless, various machine learning techniques were applied while taking laboratory, demographic, and mammography characteristics into account, leading to evaluating the effectiveness of various methods for breast cancer prediction. The Multi-Layer Perceptron (MLP) is a deep artificial neural network that consists of an input layer for signal reception, an output layer for prediction, and various hidden layers that serve as the computing engine in between. A backpropagation algorithm, which is a component of supervised networks, trains the MLP. Data is driven from input nodes to output nodes in this network. In order to adjust the weights, any errors in the output must be somehow returned from the output to the input. The post-diffusion algorithm is the most often used technique for this.

Benlahmard, Habib, et al.[5] compares the effectiveness of five classifiers: Random Forest, Logistic, Support Vector Machine (SVM), and The research community ranks regression, decision trees (C4.5), and K-Nearest Neighbors (KNN Network) as three of the most important data mining techniques and among the top ten. According to data published by the International Agency for Research on Cancer (IARC) in December 2020, breast cancer has surpassed lung cancer as the most prevalent cancer diagnosed in women

globally. The total number of cancer diagnoses has almost doubled over the last 20 years, from an expected 10 million in 2000 to 19.3 million in 2020.

One in five people on the planet will have cancer at some point in their lives. According to projections, the number of cancer diagnoses will continue to rise. more in the upcoming years, and it will surpass 2020 by about 50% by 2040. Additionally, there have been more cancerrelated deaths—6.2 million in 2000 compared to 10 million in 2020.

Puja Gupta, et al.uses the hyperparameters of six machine learning models to demonstrate how they operate for the Wisconsin Breast Cancer dataset. Deep learning and various machine learning techniques have been used to classify cells as either benign or cancerous under supervision. Because Adam Gradient Learning combines the advantages of RMSProp and AdaGrad, it finds the maximum accuracy. RMSProp performs well with nonstationary signals, while AdaGrad is ideally suited to computer vision issues. By using the rectified linear unit (ReLU) function, the model was able to train more quickly and perform better without experiencing a vanishing gradient issue. integrates root mean square propagation (RMSProp) with the adaptive gradient technique (AdaGrad).

III. PROPOSED SYSTEM

The deployment of a cancer prediction diagnostic system, specifically Using deep learning-based neural network algorithms to identify breast cancer from histopathological pictures is a crucial first step in early identification and prevention, which is in line with the main objective of lowering cancer-related mortality. Finding both genetic and environmental factors is essential for creating efficient diagnostic and preventive methods, as cancer is frequently caused by mutations in genes encoding essential cell regulating proteins. The emphasis of this suggested approach is on using deep learning methods—specifically, Convolutional Neural Networks, or CNNs—for precise histopathological image categorization in order to ascertain the existence and severity of cancer.

A CNN model may be trained to distinguish between benign and malignant tumors using a dataset of annotated histopathology pictures. facilitating early intervention and detection.

The goal of the CNN-based diagnostic system is to increase the accuracy of forecasts and boost the dependability of cancer diagnosis. The model may learn complex patterns and traits suggestive of malignant growth through the iterative process of training the neural network on labeled data, allowing it to make well-informed predictions on data that hasn't been seen before. The system may outperform conventional diagnostic techniques by employing a deep learning technique, providing more accurate andeffective detection of cancer.

Additionally, the suggested system aims to differentiate between benign and malignant tumors and identify the degree of cancer. This degree of classification granularity is essential for directing prognosis evaluations and therapy choices, guaranteeing that patients receive the right care according to the severity of their ailment. Fig.1 shows the work flow of the proposed system.

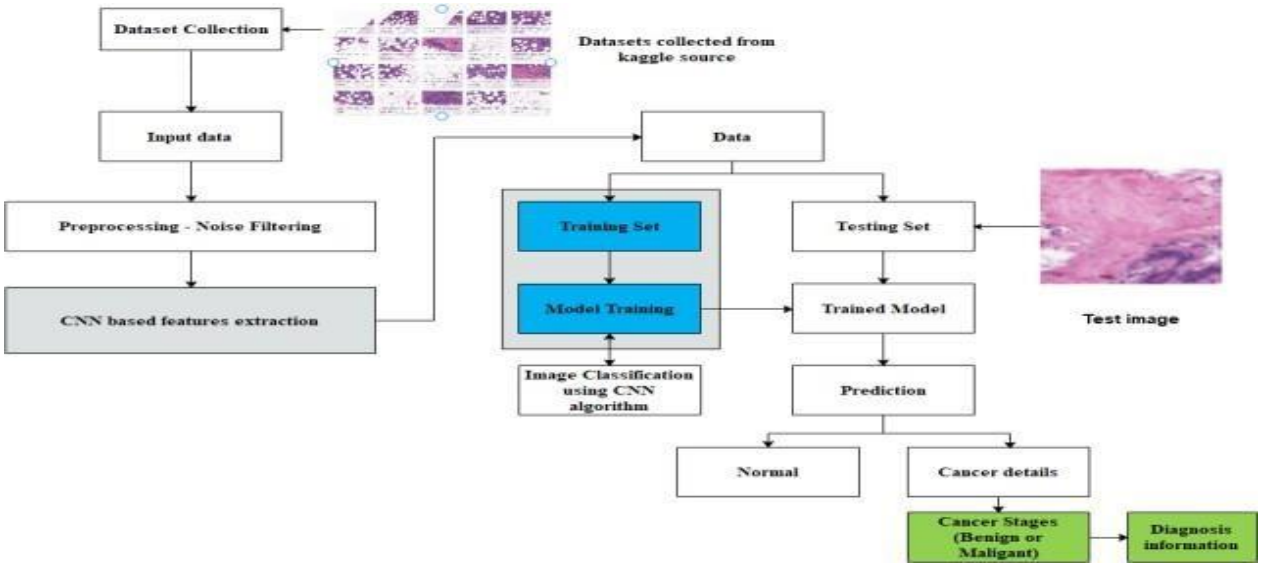


Fig. 1. Work flow of the proposed system

The Convolutional Neural Network (CNN) algorithm, which is frequently used for image classification tasks such as identifying cancer from histopathological images, generally takes the following steps: 18 Preprocessing: Preprocessing steps are applied to the input images to improve their quality and get them ready for analysis.

Common preprocessing methods include resizing the images to a uniform size, normalizing pixel values to a common scale, and applying data augmentation to increase the diversity of the training dataset. The algorithm begins with a set of input images representing histopathological slides of breast tissue samples, which serve as the raw data for the classification task.

Training: By passing input images through the network, calculating the loss, and using backpropagation to update the parameters, the CNN is trained using a labeled dataset (training set). Until the model achieves a sufficient level of performance, training is carried out overanumberofepochs.

Validation: A different validation dataset is used to track the CNN's performance during the training phase. In order to avoid overfitting and guarantee generalization to new data, this enables early halting andhyperparameteradjustment.

Testing: To determine whether the CNN can generalize to new, unseen images, its performance is assessed using a held-out test dataset once training is finished. This gives an approximation of the model's performance and accuracy in practical settings.

Plotting a confusion metric can be used to assess disease classification. We can determine the link between the expected and actual values for each target attribute with the aid of a confusion matrix. When contrasting the imbalanced data with the balance dataset'sconfusionmatrix We can see that the CNN model can recognize all of the labels. By analyzing tissue images and predicting whether a particular condition will be present or absent, a deep learning-based lung disease prediction system can aid in the diagnosisprocess.

lung cancer. We can distinguish between the three categories of sickness in this module. Additionally, give prescriptions for illnesses that are impacted

IV. SYSTEM IMPLEMENTATION

LIST OF MODULES

• Image acquisition

• Preprocessing

• Model training

• Classification

• Diagnosis details

IMAGE ACQUISITION: A critical stage in the analysis of tissue samples for research and diagnostic purposes is the acquisition of histopathological images. With the use of sophisticated algorithms and computeraided image analysis tools, pathologists or researchers examine these pictures to detect cellular structures, tissue shape, and any anomalies. This module allows us to enter college image collections from the KAGGLE website. It has labels with the numbers 0 and 1, which stand for cancer and normal.

PREPROCESSING :Operations involving images at the lowest level of abstraction, when both input and output are intensity images, are commonly referred to as pre-processing. This module allows you to resize the image using predetermined dimensions Additionally, use the median filtering procedure to eliminate image noise.

MODEL TRANING: A Convolutional Neural Network (CNN) goes through numerous crucial phases of training. Initially, a labeled dataset is put together, making sure it accurately depicts the intended task. as picture classification, and is thereafter separated into test, validation, and training sets.

Data pretreatment procedures are used to standardize pixel values and enhance the training data for better model generalization after dataset preparation. The design of the CNN architecture is the next important stage, during which an appropriate model is selected or developed, detailing the architecture's layers, filter sizes, and activation works.

Following the definition of the architecture, the model is assembled, which includes choosing optimization algorithms, loss functions, and assessment metrics. After the model has been created, training begins with the training dataset, and through a process known as backpropagation, the model gains the ability to identify patterns in the input data.

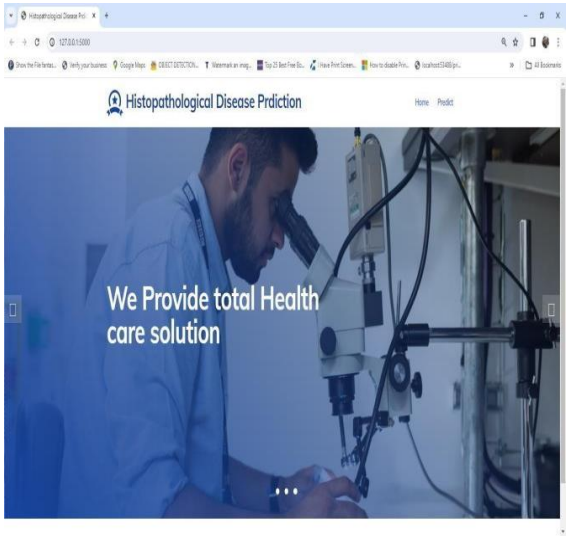
CLASSIFICATION: To ensure peak performance, the training procedure is iteratively adjusted in response to the validation set's feedback. Once the performancewassufficient,

The model's generalization abilities are assessed using the test set. This module allows the user to enter an image and use model files produced by the CNN algorithm to carry out the classification procedure.

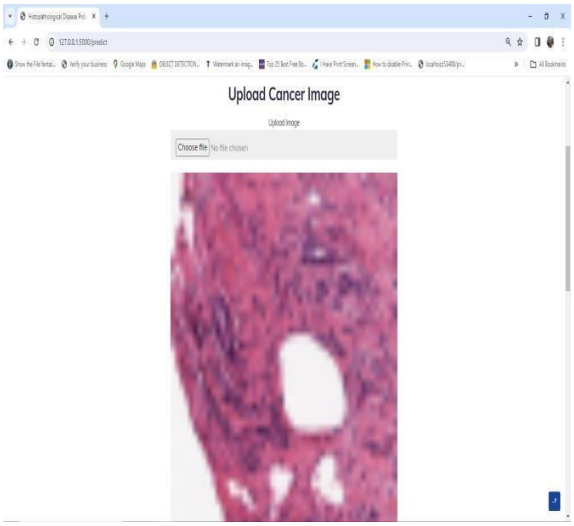
DIAGNOSIS DETAILS :This module involves using the trained CNN in a practical setting, such a computer-aided diagnostic (CAD) system, to help healthcare providers identify and diagnosis of mouth cancer as shown in Fig.2 .

By entering the photograph, we can anticipate the types of cancer and provide information on the disease's precautions; a caption or response is generated. Fig. 3. Shows the image selection and corresponding result.

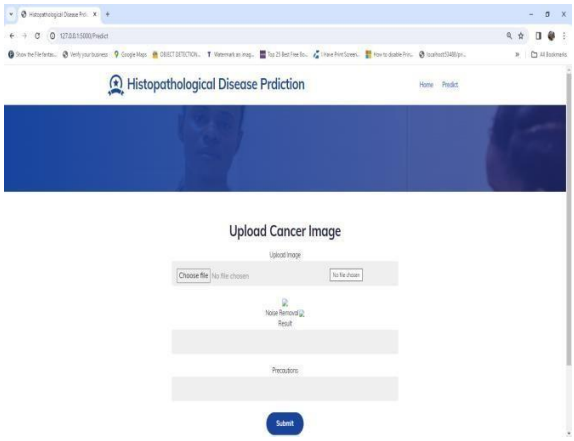
DISEASE PREDICTION



DISEASE PREDICTION



UPLOAD TISSUE IMAGE



CLASSIFICATION

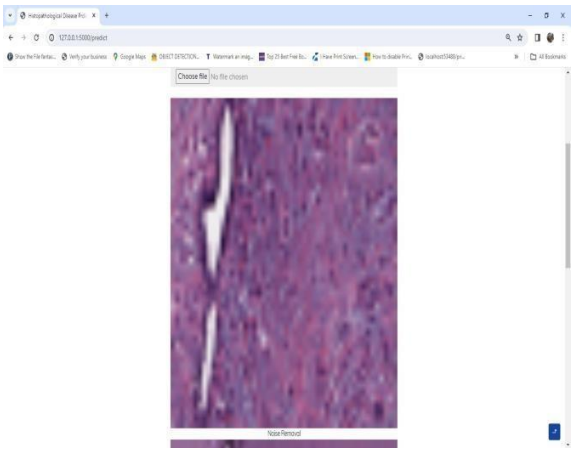


Fig. 2. Identification of Mouth cancer

IMAGE SELECTION **RESULTS**

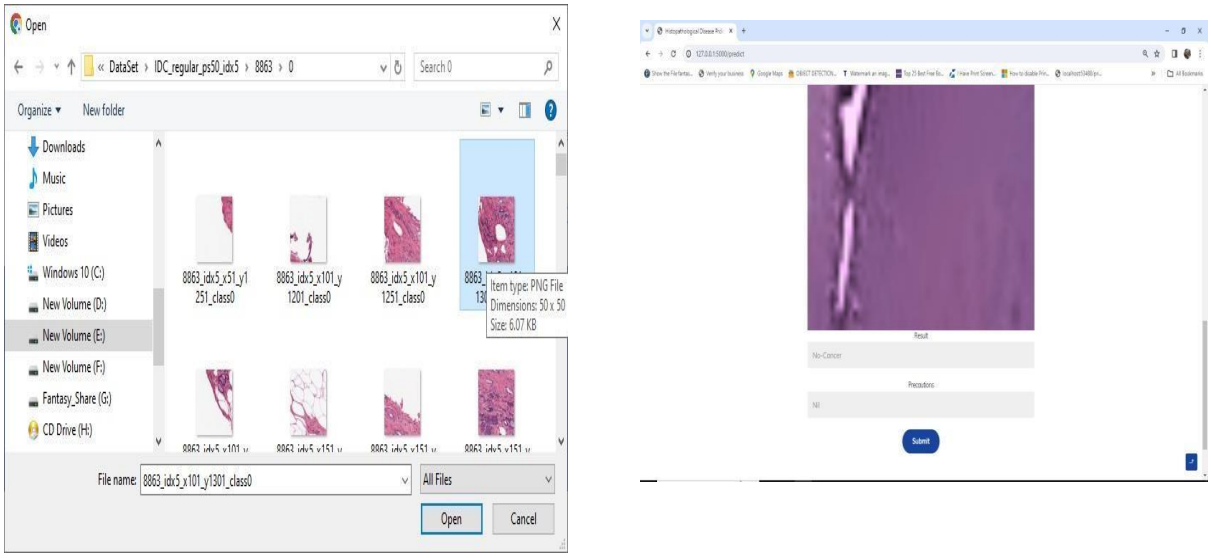


Fig. 3. Image selection and the result

V. CONCLUSION

To sum up, the development of a dataset of breast tissue histopathology pictures tagged with details regarding the existence or In order to create efficient machine learning models for the identification and categorization of lung cancer, the lack of malignant tissue is an essential first step.

Researchers can create an extensive dataset that represents the heterogeneity present in breast cancer pathology by obtaining a variety of breast tissue samples and carefully annotating regions of interest that contain malignant tissue. Along with making it easier to create machine learning models, the production of datasets of tagged histopathology images for Research on breast cancer also encourages cooperation and information exchange among scientists.

These datasets allow researchers to compare outcomes and pinpoint best practices by acting as uniform standards for assessing how well various algorithms and methodologies work. Furthermore, multidisciplinary cooperation between pathologists, radiologists, computer scientists, and other specialists is promoted by the availability of annotated datasets, leading to novel methods for the diagnosis and treatment of lung cancer.

VI. FUTURE ENHANCEMENTS

In order to diagnose and predict breast cancer, future research in histopathological image analysis should aim to improve data collection, Examine cutting-edge modeling approaches, include multiomics data, conduct clinical validation, and deal with ethical and interpretability issues.

By tackling these obstacles and possibilities, scientists might hasten the development of more precise, effective, and customized methods for diagnosing and treating breast cancer with tumor characteristics.

References

[1]. R. S. Sirisati, A. Eenaja, N. Sreeja, et al., "Human Computer Interaction Gesture recognition Using Deep Learning Long Short Term Memory (LSTM) Neural Networks," Journal of Next Generation Technology (ISSN: 2583-021X), vol. 4, no. 2, 2024.

[2]. Islam, Md Milon, et al. "A comparative study using machine learning techniques for breast cancer prediction." SN Computer Science I (2020: 1–14).

[3]. Rawal, Ramik. "Predicting breast cancer through machine learning.", Journal of Emerging Technologies and Innovative Research (JETIR), 13.24 (2020):7.

[4]. “Prediction of breast cancer using machine learning approaches" by Rabiei, Reza, et al. Biomedical Physics and Engineering Journal 12.3 (2022): 297.

[5]. Mohammed Amine, Naji, and colleagues, "Machine learning algorithms for breast cancer diagnosis and prediction." 487–492 in Procedia Computer Science 191 (2021).

[6]. "Breast cancer prediction using varying parameters of machine learning models" by Gupta, Puja, and Shruti Garg. 593-601 in Procedia Computer Science 171 (2020).