2021 International Conference on
Artificial Intelligence (ICAI)
Islamabad, Pakistan, April 05-07, 2021

Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms

Sharmin Ara
Department of CSE, CUET
Chittagong-4349, Bangladesh
u1604044@student.cuet.ac.bd

Annesha Das
Department of CSE, CUET
Chittagong-4349, Bangladesh
annesha@cuet.ac.bd

Ashim Dey
Department of CSE, CUET
Chittagong-4349, Bangladesh
ashim@cuet.ac.bd

2021 International Conference on Artificial Intelligence (ICAI) | 978-1-6654-3293-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAI52203.2021.9445249

***Abstract*—At the moment, the most prevalent form of cancer diagnosed in women across the globe is breast cancer. It develops in the breast tissue and is one of the most frequent causes of women's death. This cancer can be cured if it is diagnosed at preliminary stage. Malignant and benign are two types of tumor found in case of breast cancer. Malignant tumors are deadly as their rate of growth is much higher than benign tumors. So, early identification of tumor type is pivotal for the appropriate treatment of a patient having breast cancer. In this work, Wisconsin Breast Cancer Dataset has been used which was collected from UCI repository. Our goal is to analyze the dataset and evaluate the performance of various machine learning algorithms for predicting breast cancer. Here, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Decision Tree, Naive Bayes and Random Forest classifiers have been implemented for classifying tumors into benign and malignant. The accuracy of each algorithm is calculated and compared to find the most suitable one. Based on the analysis, Random Forest and Support Vector Machine outperform other classifiers with accuracy of 96.5%. These classifiers can be used to build an automatic diagnostic system for preliminary diagnosis of breast cancer.**

***Keywords*—Analysis of WBCD Dataset, Malignant, Benign, Classification, Breast Cancer Diagnosis, Machine Learning Algorithms**

I. Introduction

A form of cancer which originates in breast is known as breast cancer (BC). Uncontrolled division or expansion of cells initiates cancer. Usually, breast cancer cells create a tumor and can be detected on an X-ray. Among women, breast cancer has become one of the most familiar illnesses resulting in death. As one of the most regular cancers in women, breast cancer always has shown extremely high occurrence and death rate affecting about 10% women at some stages of their lives. After lung cancer, it is the 2nd biggest reason behind female's death. Among all cancers, breast cancer accounts for 25% together with 12% of all new cases in women [1]. It is possible to detect breast cancer by classifying tumors. Malignant and benign are two separate types of tumor as found in breast cancer cases. Malignant tumors spread at a higher rate compared to benign. To differentiate between these tumors, doctors need a reliable diagnostic technique. But it is usually very difficult for the tumors to be distinguished even by the specialists. So, a reliable automatic diagnostic system is direly required for the diagnosis of tumor type.

To determine breast cancer, patients are frequently subjected to a barrage of examinations which include ultrasound, biopsy and mammography according to the varying nature of breast cancer symptoms. Of these methods, the most indicative is biopsy that involves the extraction of sample tissues or cells for investigation. The sample of cells is collected from a breast through Fine Needle Aspiration (FNA) procedure and then sent for analysis under a microscope to a pathology laboratory [2]. Numerical features such as radius, texture, perimeter and area can be calculated from microscopic images of cells and tissues. Data subsequently obtained from FNA are analyzed to predict the probability of the patient having malignant tumor in combination with various imaging data. The prophesy and probability of survival can be remarkably enhanced by early diagnosis of BC, as it allows patients to receive timely clinical treatment. The proper identification of BC and patient categorization into malignant or benign classes is an extremely significant avenue of research. Various methods for predicting breast cancer have been established in recent years. Classification techniques for instance, Random Forest (RF), Support Vector Machine (SVM), Adaboost Classifier, K-Nearest Neighbors (KNN) and XGboost classifier have been used in the recent literature [3].

In this work, Wisconsin Breast Cancer Dataset (WBCD) of the FNA biopsy system has been used and different machine learning (ML) classifiers have been implemented to determine the form of breast cancer in a suspected patient. Six classification models were used including Random Forest, Logistic Regression, Decision Tree, Naive Bayes, SVM and KNN. In order to discover the most suited model for predicting breast cancer, the results acquired are then evaluated to compare the algorithms. The main objective of our paper are:

- To analyze the WBCD dataset for finding relation between the features.
- To apply different established classifiers on the WBCD dataset for comparing them.
- To find most satisfactory approach supporting the dataset with good prediction accuracy.

The remainder of this paper is outlined as follows: literature review is presented in Section II. The overall methodology is illustrated in Section III. Section IV represents the obtained results. At last, Section V concludes the entire work.

978-1-6654-3293-1/21/$31.00 ©2021 IEEE

II. Literature Review

In the previous works, many models have been proposed which use different feature sets and methods of machine learning to diagnose breast cancer. The scarcity of large datasets and inequality between negative and positive classes are the main challenges in the research area of breast cancer prediction.

In [1], the prime goal of the analysis is to detect the algorithm that operates quicker, more reliably, and more effectively in breast cancer prediction. With a precision of 99.76%, Random Forest surpasses all other algorithms. In [2], [4], [5], authors have performed comparative analysis of the precise breast cancer prediction offered by existing ML algorithms. The dataset used in these papers is WBCD. In [6], authors have used Random Forest classifier for the identification and prediction of breast cancer to determine whether or not the person has breast cancer. This offers the highest identification accuracy since both classification and regression approaches are used for Random Forest algorithm. In [7], a study on breast cancer was provided by the authors to develop predictive models for breast cancer survival. In this paper, three breast cancer survivability prediction models were applied to two classes: benign and malignant cancer. In [8], authors highlighted all previous research on ML algorithms used for breast cancer determination. They suggested that the problem of limited available dataset can be solved by data augmentation techniques. In [9], authors presented a technique that can be used to detect and identify cell morphology in automated systems that carry out the classification using computer-aided mammogram image features. In [10], authors have compared various classification and clustering algorithms in the survey. The result shows that the algorithms for classification are better predictors than the clustering algorithms.

In [11], the method of automatic detection of anomalies in mammograms is discussed. Applying the fuzzy-C-means and thresholding strategy, suspect region-of-interest (ROI) was segmented. The proposed algorithm for the Mini-MIAS dataset was validated. They concluded that the performance of suspicious region detection in mammograms can be improved by subtracting preprocessed enhanced and preprocessed enhanced inverted images. In [12], for the identification of the malignant and benign state, an algorithm was proposed by the authors depending on a fuzzy inference system. Comparison of conventional performance criteria such as sensitivity, accuracy and specificity suggests that their introduced solution outperforms Artificial Neural Network (ANN) and SVM classification. In [13], different deep learning concepts related to breast mammogram analysis are reviewed and contributions to this area are summarized. This work summarizes the past of mammogram research, recent advances, and the current state of the art.

III. Methodology

We have configured a series of steps to come up with the most reliable results in order to determine whether stage of the tumor is malignant (cancerous) or benign (non-cancerous). Our overall methodology can be presented in following subsections-

A. Dataset Description
B. Dataset Analysis
C. Training and Testing

A. Dataset Description

Dr. William H. Wolberg of the University of Wisconsin Hospital in Madison, Wisconsin, USA has developed the WBCD dataset used for this paper which is publicly accessible. This dataset includes 357 and 212 cases of benign and malignant breast cancer respectively as shown in Fig. 1.

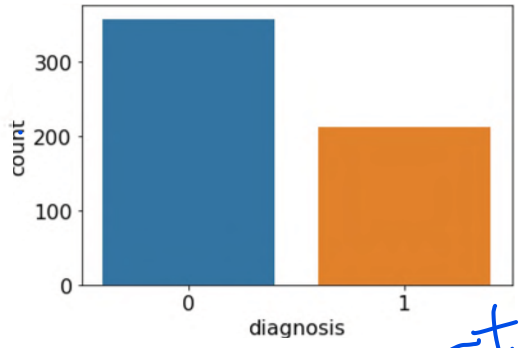


Fig. 1. Class distribution.

stupicast

The dataset comprises 32 columns, with the ID number being the first column and the diagnosis outcome (0-benign and 1-malignant) being the second column. The rest of the columns (3-32) contain three measurements (mean, standard deviation, and mean of worst) of ten features. These features represent the shape and size of the target cancer cell nucleus. The sample of cells is collected from a breast through Fine Needle Aspiration (FNA) procedure in biopsy test. For each cell nucleus, these features are determined by analyzing under a microscope in a pathology laboratory. All values of the features are stored up to four significant digits. There were no null entries in the dataset. The 10 real-value features are described in Table I.

B. Dataset Analysis

For the dataset analysis, the whole dataset has been considered. In Fig. 2, the mean radius feature of the dataset is counter-plotted. From the figure, it can be observed that suspected patients not bearing cancer have a mean radius of around 1, whereas suspected patients bearing cancer have a mean radius of more than 1. ?

Now, in Fig. 3, the correlation among the features of the WBCD dataset is shown in a heatmap. Correlation heatmap shows a 2D correlation matrix between two discrete dimensions where the first dimension value is considered as a row and the second dimension value as a column of the heatmap. In this heatmap, the colored cells in a monochromatic scale are used to show the resultant correlation between the features of the dataset. Increasing intensity of color represents increasing correlation. The value of the color of the cells is proportional

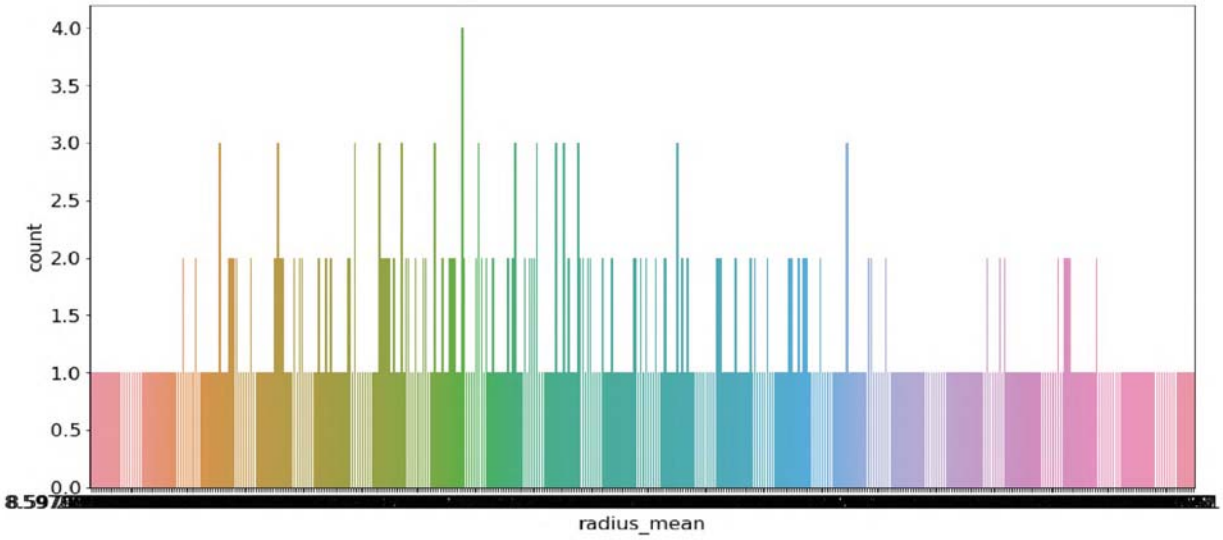


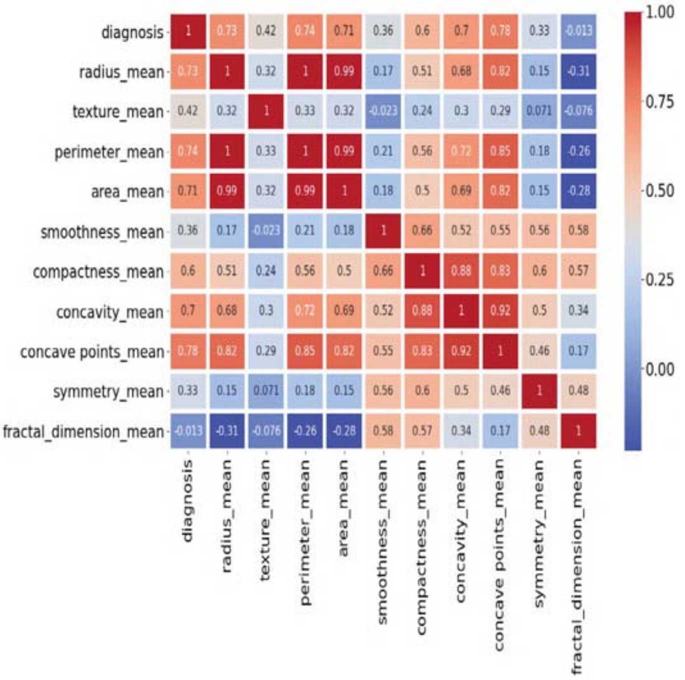
Fig. 2. Mean radius feature of WBCD dataset.

TABLE I
FEATURE DESCRIPTION

Feature Name	Feature Description
Radius	Average of distances from center to circumference points.
Texture	Standard deviation (SD) of gray-scale value.
Perimeter	Gross distance between the snake points.
Area	Total number of pixels on the inside of the snake along with one half of the pixels in the circumference.
Smoothness	Local variance in length of radius, quantified by calculating the length difference.
Compactness	Perimeter$^2/Area$.
Concavity	Intensity of the contour's concave parts.
Concave points	The number of contour concavities.
Symmetry	The difference in length between lines perpendicular to the major axis in both directions to the cell boundary.
Fractal Dimension	Coastline estimation. A higher value leads to a less normal contour representing a higher risk of malignancy.

to the number of measurements that match the dimensional values. The dimensional value (-1 to +1) is calculated from the linearity between the pair of features. If both variables vary and move in the same direction, positive correlation is acquired. In case of negative correlation, increase in one variable is associated with a decrease in the other and vice versa. From the figure, we can see how often one feature affects all the other features in this heat map (e.g radius mean has 32% influence on texture mean).

In Fig. 4, the correlation barplot represents the correlation between the diagnosis outcome and every other feature of the dataset individually. Here, the 'smoothness_se' feature is highly negatively correlated with the diagnosis in the correlation barplot. The 'fractal_dimension_mean', 'texture_se', and 'symmetry_se' features are associated very less negatively and other remaining features are highly positively correlated.



	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
diagnosis	1	0.73	0.42	0.74	0.71	0.36	0.6	0.7	0.78	0.33	-0.013
radius_mean	0.73	1	0.32	1	0.99	0.17	0.51	0.68	0.82	0.15	-0.31
texture_mean	0.42	0.32	1	0.33	0.32	-0.023	0.24	0.3	0.29	0.071	-0.076
perimeter_mean	0.74	1	0.33	1	0.99	0.21	0.56	0.72	0.85	0.18	-0.26
area_mean	0.71	0.99	0.32	0.99	1	0.18	0.5	0.69	0.82	0.15	-0.28
smoothness_mean	0.36	0.17	-0.023	0.21	0.18	1	0.66	0.52	0.55	0.56	0.58
compactness_mean	0.6	0.51	0.24	0.56	0.5	0.66	1	0.88	0.83	0.6	0.57
concavity_mean	0.7	0.68	0.3	0.72	0.69	0.52	0.88	1	0.92	0.5	0.34
concave points_mean	0.78	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1	0.46	0.17
symmetry_mean	0.33	0.15	0.071	0.18	0.15	0.56	0.6	0.5	0.46	1	0.48
fractal_dimension_mean	-0.013	-0.31	-0.076	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	1

Fig. 3. Correlation between the different features.

The less correlated features can be removed as they have too less impact on the target. We have omitted these features for enhancing the accuracy of the implemented classifiers.

C. Training and Testing

Initially, the dataset is read from the CSV file. The data entries from the dataset are analyzed on the basis of their features before they were used for further step. Then, we split the dataset into two portions randomly: training set (75%) and testing set (25%). Not every feature within the dataset is useful and capable of giving same contribution to the result.

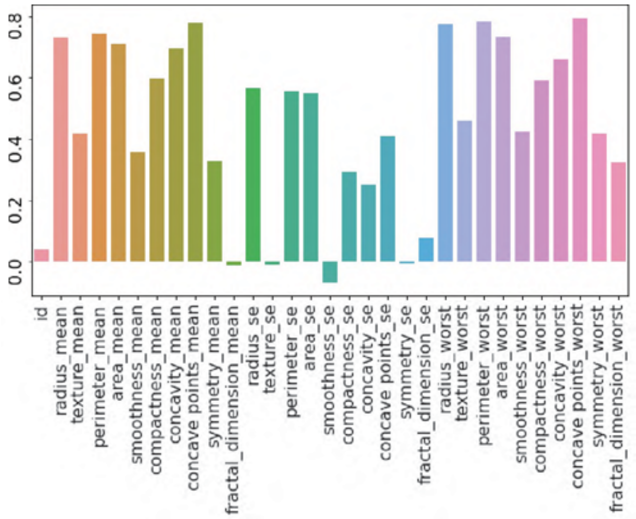


Fig. 4. Correlation of the features with target.

According to the data analysis, we have done feature selection to eliminate less correlated features which will increase the accuracy. Then, the dataset is ready for the application of the ML algorithms to examine their performance. After this step, we have accomplished the performance analysis by the comparative study of the resultant testing and training accuracy. Fig. 5 depicts the overall workflow of the study.

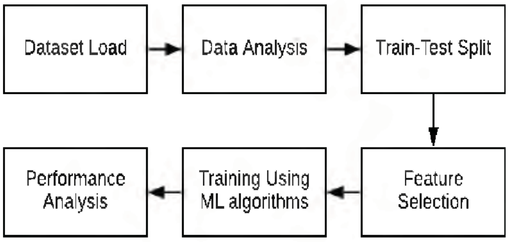


Fig. 5. Overall workflow.

Machine learning is an automated approach to learn where algorithms are programmed to gain experience from past datasets for predicting future. In this project, we have used the following ML algorithms:

- Logistic Regression.
- Support Vector Machine (SVM).
- Random Forest.
- Naive Bayes.
- Decision Tree.
- K-Nearest Neighbors (KNN).

IV. Results

In this section, after implementing ML algorithms, we have analyzed the performance of the algorithms on the dataset. This is performed by executing the algorithms on the formerly set test dataset. 25% of the entire dataset was included in the test dataset. A confusion matrix is generated for the actual and predicted result consisting of TP, FP, TN, and FN for calculating accuracy for each algorithms used. Below, the meaning of the terms is mentioned.

- TP = True Positive (Accurately Identified)
- TN = True Negative (Inaccurately Identified)
- FP = False Positive (Accurately Rejected)
- FN = False Negative (Inaccurately Rejected)

For example, confusion matrix for SVM is shown in Fig 6. The assessment of used ML algorithms is carried out using one of the metrics called accuracy. Accuracy is the prediction fraction. It represents the ratio of the number of accurate predictions over the gross number of predictions done by the model as shown in (1).

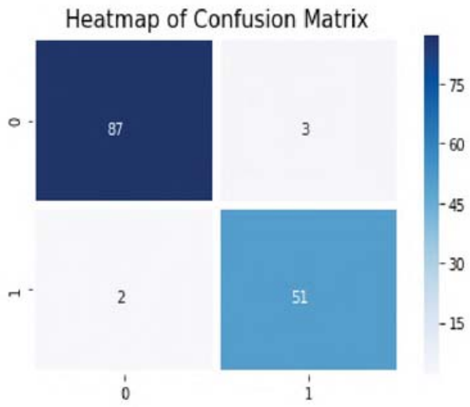


Fig. 6. Confusion matrix.

The formula of accuracy is:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

The result presented in Table II demonstrates the obtained accuracy of training and testing for SVM, Random Forest, Naive Bayes, Decision Tree, KNN and Logistic Regression.

TABLE II
Comparison Among Various Algorithms

Algorithms	Training Accuracy	Testing Accuracy
Logistic Regression	99.1%	94.4%
KNN	97.6%	95.8%
Decision Tree	100%	95.1%
Naive Bayes	95.1%	92.3%
Random Forest	99.5%	96.5%
SVM	98.8%	96.5%

We can see that the maximum accuracy is achieved using Random Forest and SVM on the test set which is 96.5%. Their training accuracy is also higher than other algorithms. Our study shows that SVM is the preferable algorithm for prediction. It is known that SVM performs relatively good if clear separation between classes is present in the dataset and the dataset is high dimensional. Random forest also gave better result alongside SVM. Generally, random forest obtains higher accuracy without any normalization of dataset values.

V. Conclusion

Now-a-days, one of the deadly diseases affecting women is breast cancer. In our work, the Wisconsin Breast Cancer Dataset was utilized and several ML algorithms were applied to assimilate the efficacy and usefulness of these algorithms to find the highest accuracy of classifying malignant and benign breast cancer. The correlation between different features of the dataset has been analyzed for feature selection. The results will assist to pick the best ML algorithm for the construction of an automatic breast cancer diagnostic system. From our study, we can conclude that SVM and Random Forest give the maximum accuracy with an accuracy of 96.5%. We will try to strengthen our work in future by handling a comparatively large dataset and incorporating some more functions such as breast cancer phase detection and so on. We hope that this study will contribute in the clinical application of breast cancer treatment.

References

[1] J. Sivapriya, A. Kumar, S. Siddarth Sai, and S. Sriram, "Breast cancer prediction using machine learning," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, 2019.

[2] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*. IEEE, 2018, pp. 1–5.

[3] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, "Developing a web based system for breast cancer prediction using xgboost classifier," *International Journal of Engineering Research Technology (IJERT)*, vol. 9, 2020.

[4] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair, "A comparative study for breast cancer prediction using machine learning and feature selection," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 1049–1055.

[5] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and k-nearest neighbors," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 2017, pp. 226–229.

[6] M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, "Breast cancer prediction via machine learning," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2019, pp. 121–124.

[7] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018.

[8] N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis," *IEEE Access*, vol. 8, pp. 150 360–150 376, 2020.

[9] A. Toprak, "Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer," *Medical science monitor: international medical journal of experimental and clinical research*, vol. 24, p. 6537, 2018.

[10] D. S. Jacob, R. Viswan, V. Manju, L. PadmaSuresh, and S. Raj, "A survey on breast cancer prediction using data miningtechniques," in *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*. IEEE, 2018, pp. 256–258.

[11] K. L. Kashyap, M. K. Bajpai, and P. Khanna, "Breast cancer detection in digital mammograms," in *2015 IEEE international conference on imaging systems and techniques (IST)*. IEEE, 2015, pp. 1–6.

[12] F.-T. Johra and M. M. H. Shuvo, "Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic," in *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*. IEEE, 2016, pp. 1–5.

[13] O. V. Singh and P. Choudhary, "A study on convolution neural network for breast cancer detection," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE, 2019, pp. 1–7.