International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 10 | Oct 2023 www.irjet.net p-ISSN: 2395-0072

A Comprehensive Evaluation of Machine Learning Approaches for Breast Cancer Categorization

Venkata Siddarth Gullipalli[1], Katakam Hemanvitha[2], P. Naga Jyothi[3]

[1]*B. Tech, CSE, GITAM University, Visakhapatnam, India*
[2]*B. Tech, CSE, GITAM University, Visakhapatnam, India*
[3]*Assistant Professor, Dept. of CSE, GITAM University, GST, Visakhapatnam, India*

Abstract - *Breast cancer is the most common disease among women. There were approximately 2,00,000 deaths from breast cancer worldwide in 2022. As we progress to 2023, an even more alarming projection looms, with an anticipated 627,000 precious lives at risk of being lost to this cancer. Early detection and accurate classification of breast cancer are imperative for effective treatment and patient outcomes. This research paper aims to compare the performance of three models—Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM)—in the context of breast cancer classification. The study seeks to determine which model yields the highest accuracy and reliability in diagnosing breast cancer from Breast Histopathology Images. The CNN model demonstrated the highest classification accuracy at 87%, followed by SVM at 76% and ANN at 71%. CNN also exhibited superior sensitivity in detecting malignant cases, while ANN had the fastest training time. Our findings suggest that CNN is the most promising model for breast cancer classification due to its high accuracy and sensitivity.*

Key Words: **Breast cancer classification, CNN, ANN, SVM, Histopathology images**

1.INTRODUCTION

According to a Medical News Today survey, the expected number of new breast cancer cases by 2023 is 2.87 lakhs. These have been increasing since 2011; by 2030, they will have increased by more than 50%. According to the UICC (Union for International Cancer Control), one out of every eight persons is diagnosed with breast cancer. According to a WHO report, breast cancer affects people of both sexes in 158 countries. Breast cancer will impact one in every three (30%) new females in the United States, according to the American Cancer Society's predictions. Since resources are unjustly distributed, the mortality rate continues to climb. Breast cancer cannot be realized early if it has advanced. Breast cancer cells frequently develop a tumor, which can be visible on an x-ray or felt as a bump. These often appear in lining (epithelium) cells in the glandular tissue of the breast. It is localized initially to a lobule or duct, called stage 0, which can be treated with various systematic treatments. It efficiently saves lives and prevents cancerous cells from developing and expanding. If these cells advance to blood and lymph arteries, the cancer develops to a metastatic stage and may damage the lungs, liver, bones, and brain. In such a way, one form of cancer might spread to different body organs. This proposed work aims to help people become aware of illness symptoms as they progress from early to late stages. Expert knowledge is needed to decide what type of treatment a patient or clinical practitioner might get.

Every day marks another step forward in developing prediction and treatment methods that might assist the public in every individual health result. The planned effort for therapy helps in the context of breast cancer. Early diagnosis and treatment will be critical in improving the patient's prognosis. Breast cancer is anticipated to be the leading cause of death among women in the following years. A clinical breast examination, breast imaging modalities, and biopsy are the existing and regularly utilized imaging tools in the diagnostic process. Mammography and ultrasound are the most often used diagnostic methods, while other modern procedures such as PET, PET-CT, SLNB, BSGI, and others need effective data collecting.

2. REVIEW OF RESEARCH AND DEVELOPMENTS IN THE DOMAIN

Most of the researcher's work focuses on whether the cancer is malignant or benign. The reviews discussed the type of tool, input clinical data, and use of computer-aided algorithms. The paperwork defines not only provides a decision on the cancer types and its follow-up treatment.

Duan et al. (2021) used Wisconsin Breast cancer tumor prediction using Random Forest and AdaBoost algorithms. The data collection includes 569 random samples, of which 357 are benign and 212 are malignant. They considered factors for results evaluation like breast mass concavity, density, tumor, cell nucleus radius, texture, circumference, area, etc. The data standardization is done without missing their correlation by using various techniques of Python packages. Ensemble learning algorithm they had used to aggregate the decision with appropriate prediction accuracy. RF with AdaBoost used search parameter optimization is selected to improve the performance by hyperparameters optimization still. The limitation is there is knowledge with

expertise, and experience. It is a computer-aided decision and cannot be justified as accurate for the diagnosis[1].

Huang et al. (2021) proposed the Hierarchical Clustering Random Forest Algorithm(HCRF) and explained decision trees were weak in classification tasks, so they added a component clustering. The outcome of the representative trees will be clustered with low similarity and high accuracy and use variable Importance Measure(VIM) to optimize the feature selection process. The model HCRF includes the standard Wisconsin Diagnosis Breast Cancer dataset (WDBC) and WBC of 569 and 669 samples with 39 attributes. Selecting the most discriminating feature provides biomarker information using the VIM method of eliminating the least important attributes. Applying HCRF will create diversified DTs, which are grouped according to high similarity measure, then calculating the similarity measure and comparing and grouping it multiple times to build a decision with the highest AUC value and ignoring the other DTs' decision. The similarity measure is calculated by the Disagreement Measure (DIS). The comparative results were obtained using different techniques: DT, AdaBoost, RF, and HCRF. The future direction of the researcher's work is using visualization DTs and heuristic algorithms[2].

Xiaomin Zhou et al. (2020) discussed a review on breast cancer based on Breast Histopathological Image Analysis (BIA), Artificial Intelligence(AI) and deep learning approaches, input datasets, and transfer learning techniques of various researchers' works. The limitations per the review mentioned are a collection of data, input image analysis, overfit of data, time of execution, comprehensive knowledge of input data and feature extraction, and suitable algorithm application. The future direction of his work is to focus on the analysis of microscopic images and the precision of the drug for proper treatment[3].

Mazo et al. (2020) proposed a systematic review of the Clinical Decision Support Systems(CDSSs), which assists medical staff in decision-making. He discussed the works collected from various databases and illustrated their limitations, benefits, and opportunities for individual results. DSS majority supports the medical diagnosis and improves decision-making and medical practitioner education. According to the researcher summary, CDSSs are not many based on breast cancer treatment as reported by global standards. His literature screened 1000 articles to explore the availability of tools and how they assist medical organizations[4].

Sara et al. (2019) discussed using machine learning algorithms to diagnose breast cancer with the Mammographic mass dataset. To predict the severity, models used six attributes of BI-RADS(Breast Imaging-Reporting Data Stream), which has 516 benign and 455 malignant images, which they collected from the Radiology of University Erlangen-Nuremberg. ANN, KNN DT, and Binary SVM classifiers are used, and their results are compared with different metrics [5].

Babak et al. (2017) proposed work to classify hematoxylin and eosin-stained breast specimens using Convolutional Neutral Networks. They used nearly 646 breast tissue biopsies. The discriminative evaluations use AUC(Area under ROC) on normal and abnormal tissue stromal regions. The briefing of the classification work includes pre-processing, feature selection like epithelium, stroma, and fat, and extraction for stromal areas. The model is based on two primary constructs for breast tissue component classification: the classification of stromal regions and the framework for the model parameters. A VGGnet (CNN) with 11 layers and a ReLU activation function, uses 12 filters composed of CNN1 followed by CNN2 network to optimize the classification. The results were 95% on classifying tissue into epithelium, stroma, and fat. Pixel level accuracy is 92% for discriminating. The future work uses biomarkers to predict using advanced computer-aided techniques[6].

Benny et al. (2021) proposed their work on protein segments in the membrane region for HER2 expression of breast cancer using segmentation methods to identify ROI. The data of HER2 are tissue images and categorized into three groups based on the scores. The methodology includes various segmentation techniques like FCN, SegNet, and U-Net. Segment the protein structure area and quantify for parameter evaluation between target and predicted masks. The higher the loss of images, the lower the prediction values[7].

Bhangu et al. (2020) discussed the classification of cancerous and benign tumors by applying various machine-learning techniques. The digitized images of FNA (Fine needle aspiration) consist of 569 samples with 32 variables (dependent and independent) and are categorized into 3, each containing 10 features based on the breast mass and radius. Preprocessing was carried out with a correlation matrix and models applied for classifying into M and B classes: SVM, LR, K-NN, Naive Bayes, DTs, RF, Multilayer Perceptron, Linear Discriminant Analysis, AdaBoost, Gradient Boosting, XGBoost classifiers, and comparative analysis is done. The limitation of their work is that it does not support clinician expertise for decisions, and the future scope is to use ensemble models[8].

Karthiga et al. (2018) discussed the K-Means clustering for image segmentation of cell nuclei, and the DWT transformation technique is applied for the segmented image. The BreakHis dataset includes 40X to 200X images of 300 histopathological images of ductal carcinoma, lobular carcinoma, and mucinous carcinoma images. The image is partitioned into clusters using K-Means to k clusters and applying the coifelt wavelet and wavelet transform to generate sub-band images such as LH, LL, HL, and HH. Using Shannon entropy and Log energy Entropy for feature

extraction to quantify the image by analyzing and classifying the benign and malignant subband of images [9].

Naga Jyothi P et.al (2021) discussed a model to extract the fraudulent behavior from the claims data, which the claimants submit by using different machine learning algorithms and various data preprocessing techniques to optimize the results. The application was developed to assist the healthcare organization and worked on CMS data, which is an authenticated dataset from a U.S. organization. [10] The work is preceded (Naga Jyothi P et al. (2020)) by the results obtained from the supervised approach for identifying the outliers of the claims based on the hospital location and category of the surgery. The results are obtained from the fraud and legitimate records for using different classification rules [11].

Shanti Latha P et al. (2021) The present review focuses on the importance of exosomal tetraspanins in regulating tumor cell progression and its mechanistic role in contributing to metastasis. Exosomes are small vesicular proteins that are important for cell communication and subsequent signaling. Literature from numerous studies demonstrates that exosomal tetraspanin proteins such as CD9, CD63, and CD81 serve to sort and selectively recruit biomolecules, select targets, cell-specific entry, capture angiogenesis, and vasculogenesis [12].

Shanti Latha P et al. (2018) The present research investigates the expression of angiogenesis marker VEGF, chemokines, and its specific receptor in prostate cancer stem cells (CSCs). CCL5 is a classical chemotactic cytokine that plays a crucial role in tumor growth, proliferation, and angiogenesis. However, their role in the development of stemness was not studied. Briefly, 5- fluorouracil-resistant prostate CSCs were sorted using FACS, and the expression of VEGF, CCL5, and its associated receptor CXCR4 was evaluated by western blotting [13].

3. PROPOSED METHODOLOGY

The decision-making process in healthcare is critical and plays a superficial role in everyday life. A clinical decision support system constantly improves the quality of treatment with lower costs for better health. The general approach to treating a patient is an umbrella approach where collecting the symptoms and, based on the reports, diagnosis is carried out.[10] The precision-based system considers many factors, and the level of granularity is at optimum to know the patient's status. Diagnosis is intertwined in a more précised way to every individual. The framework of many healthcare organizations is to optimize and personalize their services for patients to progress continuously. Perhaps personalized healthcare service is possible with advanced computational models nowadays [11].

The paper's objective is to compare three different models and find out which model provides more accurate results for our problem statement.

3.1. DATASET DESCRIPTION

In this study, pictures of breast histopathology are used to carry out the classification.

The dataset comprises 280 main folders, each of which contains two subfolders labeled '0' and '1'. Within the '0' subfolder, there are 780 non-cancerous images, while the '1' subfolder contains 780 cancerous images. Here, '0' typically denotes the absence of cancer (non-cancerous), and '1' indicates the presence of cancer (cancerous) in the medical images. Fig 1. Shows the number of non-cancer and cancer cases in our dataset.

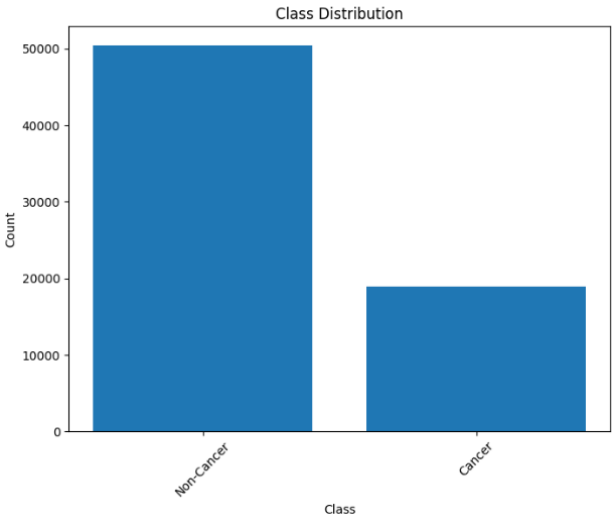

Fig -1: Class distribution of the dataset

3.2. ALGORITHM USED

The procedure followed is as follows:

- Loading the data
- Performing preprocessing.
- Building the model architecture
- Train the model
- Evaluating the model using various evaluation metrics.

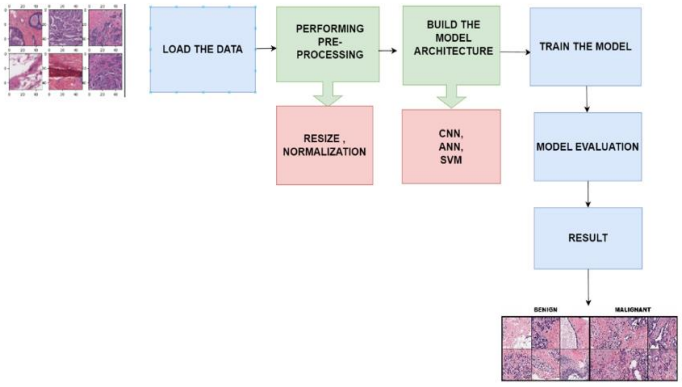

Fig -2: Architecture of Proposed Methodology

3.2.1. Convolutional Neural Networks

Three layers of neural network make up a CNN model: convolutional layers, pooling layers, and fully-connected layers. We structured the CNN model as a sequential stack of layers with four convolution layers`, four max-pooling layers, and two fully connected layers.

Convolutional layers help use to capture patterns and features of the input image. Max-pooling layers reduce the spatial dimensions of the features generated by the convolutional layers. They are critical in reducing the computational load and helping the model focus on the most relevant features. Also, there are two fully connected layers. Fig.3 shows the architecture of CNN.

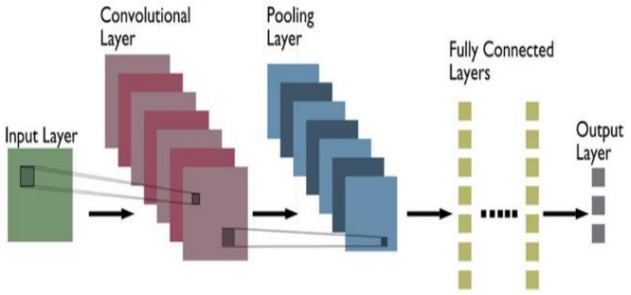

Fig -3: CNN Architecture

3.2.2. Artificial Neural Networks

An ANN comprises three layers of neurons: an input layer, one, two, or three hidden layers, and an output layer. Figure 2 displays a common layout with neurons connected by lines. The weight of each connection is a numerical value. Artificial neural networks, a useful model for classification, pattern recognition, grouping, and prediction, are becoming increasingly popular among today's young. Fig.4 shows the architecture of ANN.

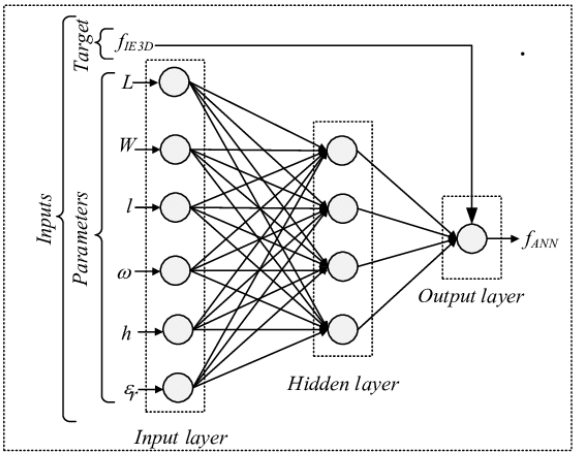

Fig -4: ANN Architecture

3.2.3. Support Vector Machine

SVM is a machine learning approach for image classification, also known as Support Vector Machine. It works by selecting the best hyperplane for dividing the classes in the feature space. SVM can handle multidimensional data and is resistant to noise and outliers.

3.3. PERFORMANCE ANALYSIS WITH METRICS

Evaluation metrics used to examine the three models in this study are: accuracy, recall, f1 score and precision.

Table.1 shows the comparison between the three models with respect to Accuracy, recall, f1 score and precision.

MODEL	*Accuracy (in percentage)*	*Precision*	*Recall*	*F1-score*
CNN	88%	0.80	0.73	0.77
ANN	71%	0.69	0.64	0.67
SVM	59%	0.50	0.45	0.47

Table -1: Comparison of CNN, ANN and SVM

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Here, TP - true positives, TN - true negatives, FP-false positives, FN – false negatives.

Here, TP - true positives, TN - true negatives, FP-false positives, FN – false negatives.

True Positives are the cases where the model correctly predicted the positive class. True Negatives are the cases where the model correctly predicted the negative class. False Positives are the cases where the model incorrectly predicted the positive class when the actual class is negative. False Negatives are the cases where the model incorrectly predicted the negative class when the actual class is positive.

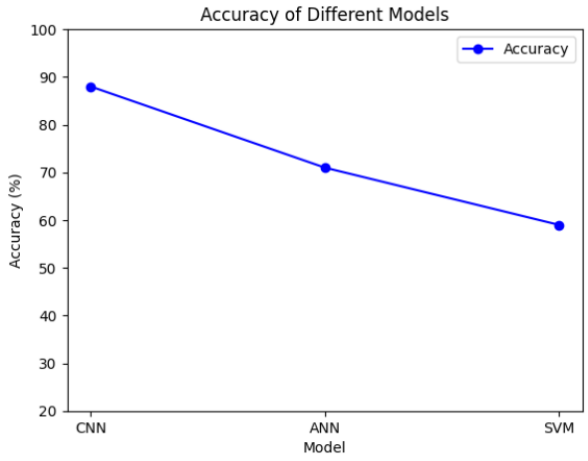

Fig -5: Performance Comparison

Figure.5 provides a clear and insightful representation of the performance of three distinct machine learning models – Convolutional Neural Network, Artificial Neural Network, and Support Vector Machine in the context of breast cancer classification. These models have been tested on a dataset consisting of histopathological data acquired from breast tissue samples. The primary objective of these models is to determine whether the tissue samples are benign or malignant based on the provided data.

The graph highlights the accuracy achieved by each of these models, which is a fundamental metric for evaluating their classification capabilities. Accuracy is calculated by dividing the number of properly categorized cases by the total number of occurrences in the dataset.

4. CONCLUSION AND FUTURE WORK

When performing breast cancer classification using the Breast Histopathology image dataset, this study focuses on a complete and in-depth evaluation of Convolutional Neural Networks, Artificial Neural Networks, and Support Vector Machines. The primary goal of this study was to compare and contrast the accuracy of these three unique models. These models were evaluated based on accuracy, F1-Score, Recall, and precision. Our analysis of the three models showed that the CNN model achieved the highest accuracy of 88%, followed by ANN with 71%, and SVM with 59%.

In this challenge, the CNN model fared better than the other models because it was built to process and analyze picture data. For the classification of breast cancer, its capacity to automatically extract relevant information from the picture data proved helpful. Although ANN also demonstrated a respectable degree of accuracy, its performance fell short of CNN's. The least accurate machine learning algorithm was support vector machines (SVM), indicating the program's shortcomings.

In conclusion, CNN has transformed into a powerful instrument for categorizing photos of breast cancer, and model selection plays a vital role in medical image analysis. Finally, the study's findings will help patients and doctors by supporting continuing efforts to improve breast cancer detection and treatment.

Additional research into medical image analysis can be launched from this study. Future research initiatives might look at more complicated neural network topologies, massive datasets, and other aspects to improve the overall performance of the breast cancer classification mode.

REFERENCES

[1] Huang, Zexian, and Daqi Chen. "A breast cancer diagnosis method based on VIM feature selection and hierarchical clustering random forest algorithm." IEEE Access 10 (2021):3284-3293.

[2] Yifan, Duan, Lu Jialin, and Feng Boxi. "Forecast Model of Breast Cancer Diagnosis Based on RF-AdaBoost." 2021 International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE, 2021.

[3] Zhou, Xiaomin, et al. "A comprehensive review for breast histopathology image analysis using classical and deep neural networks." IEEE Access 8 (2020): 90931-90956.

[4] Mazo, Claudia, et al. "Clinical decision support systems in breast cancer: a systematic review." Cancers 12.2 (2020): 369.

[5] Laghmati, Sara, Amal Tmiri, and Bouchaib Cherradi. "Machine learning based system for prediction of breast cancer severity." 2019 International Conference on Wireless Networks and Mobile Communications (WINCOM). IEEE, 2019.

[6] Bejnordi, Babak Ehteshami, et al. "Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images." 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017). IEEE, 2017.

[7] Benny, Stephy, and Satishkumar L. Varma. "Semantic Segmentation in Immunohistochemistry Breast Cancer Image using Deep Learning." 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3). IEEE, 2021.

[8] Bhangu, Kamalpreet S., Jasminder K. Sandhu, and Luxmi Sapra. "Improving diagnostic accuracy for breast cancer using prediction-based approaches." 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2020.

[9] Karthiga, R., and K. Narasimhan. "Automated diagnosis of breast cancer using wavelet based entropy features." 2018 Second international conference on electronics, communication and aerospace technology (ICECA). IEEE, 2018.

[10] P.Naga Jyothi, Rajya Lakhmi D, Rama Rao KVSN, Identifying Fraudulent behaviors in healthcare claims using random forest classifier with SMOTEechnique, International Journal of e-Colloboration (IJEC), Vol.16,Issue.4,pp:30-47.ESCI, WOS,SCOPUS,Jan-Mar 2021.

[11] P.Naga Jyothi, Rajya Lakhmi D, Rama Rao KVSN, A Supervised Approach for Detection of Outliers in Healthcare Claims Data, JESTR,SCOPUS Vol.13, pg:204-213, Mar 2020.

[12] Dr.P.Santhi Latha et.al A Study on the Expression of CCL5,CXCR4 and Angiogenic Factors by Prostate Cancer Stem Cells.Annals of R.S.C.B., Scopus, APR-2021, 25, 1020-1028.

[13] Dr.P.Santhi Latha et.al Exosomal tetraspanins as regulators of cancer progression and metastasis and novel diagnostic markers.Asia Pac J Clin Oncol., Scopus, 2018, 2018 Dec, 383-391.