International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Breast Cancer Diagnosis using Ensemble Learning Algorithms Deployed in a Website for Easy Access to Rural Areas

Krupa Mistry[1], Disha Pasari[2], Supriya Tiwari[3], and Sudha Gupta[4]

[1,2,3,4] KJ Somaiya College of Engineering, Mumbai, India
{krupa. mistry, d.pasari,supriya.rt,sudhagupta}@somaiya.edu

ABSTRACT.

Purpose: Breast cancer has become a leading cause of can- cer deaths among women, and the death rate is relatively high in rural areas due to the lack of awareness and knowledge about the disease. India has a poorer survival rate for Breast cancer, compared to other West- ern nations primarily due to a lack of early detection and assessment facilities.

Methods: This research outlines a three-stage method for detecting breast cancer that includes risk assessment for early-stage detection of breast cancer, classification of breast cancer tumors, and website cre- ation for easy access to females in rural areas. At the initial stage, the risk assessment tool helps females understand their risk of having breast cancer through questionnaires.

Results: A convolution neural network algorithm is used at the classi- fication stage of breast cancer into malignant and benign tumors using ultrasound images. At the final stage, the website brings both risk assess- ment and classification together for easy use by females in rural areas. The website also suggests the nearest cancer care facility for additional diagnosis and treatment in the event of a positive diagnosis.

Conclusions: This strategy is the need of the hour to provide better breast cancer screening and care especially in rural India.

Keywords: Breast Cancer, Convolution Neural Network, DenseNet, Risk assessment, Decision Tree, Naive Bayes, Website development, Rural Area

1. Introduction

Breast cancer is one of India's most common cancers. It accounted for 13.5% of all cancer cases and 10.6% of all deaths, with a cumulative risk of 2.81% in India [1]. Patients with breast cancer in India exhibit a shorter chance of survival than in Western nations because of early onset, a delay at the beginning of definitive management, and insufficient treatment facilities [2]. Early detection and aggressive treatment are the most efficient methods to combat breast cancer, as the WHO World Cancer Report 2020 has correctly noted. Making people, especially females, aware of breast cancer, its symptoms, and risk factors such as family history, early periods, delayed menopause, and obesity is equally crucial. Lack of awareness leads to late detection which in turn increases mortality.

Early detection can lead to saving many lives of females in our country, especially with the help of advanced technologies where people even from remote areas have access to the internet and mobile phones. To combat the same, a website trained with the best of machine learning algorithms which take different features of a patient and predict whether they are prone to breast cancer or not will help in early detection and guide them for further diagnosis can prove to be very helpful. A classification tool to distinguish between malignant and benign tumors using ultrasound images is also deployed on the website using a deep learning method. This website will provide an all around guidance and assistance in the early detection of breast cancer for females residing especially in rural India.

The entire paper is organized as Existing work in section 2, Flow diagram of Proposed work in section 3, Proposed methodology in section 4, Result analysis in section 5 and conclusion and future scope in section 6.

2. Existing Research

There are various machine learning algorithms for breast cancer prediction and diagnosis, like logistic regression, support vector machines, naive Bayes, etc. For image-based classification of tumors into malignant and benign, there are many deep learning methods like convolutional neural networks, recurrent neural networks, and some pre-trained models like AlexNet, Google Net, VGG16, VGG19, and ResNet. During our research, we found that the convolutional neural net- work gives accurate results for image classification of breast cancer in most cases. As shown in Table 1. In the comparison of different existing research, some results are shown.

According to Neslihan Bayramoglu et al., who proposed a framework using CNN for learning histopathology images on the BreaKHis database, an average recognition rate of 83.25% was achieved in a single training session per fold. A comparative study of single-task CNN and multitask CNN. Multi-task CNN predicts both the image magnification level and its benign/malignancy or malignant properties simultaneously. It concluded that, when estimating the magnification factor of an input image or when there is a lack of training data from a single magnification level, multi-task models may be more beneficial [3].

Ref. No.	Year	Summary of the Paper	Algorithm and Technique used	Technical Gap/Future Scope	Remarks
1	2016	Deep Learning for Magnification Inde- pendent Breast Cancer Histopathology Image Classification.	Deep Learning, CNN	Investigate stain nor malization, deeper architectures, and separating the network before the last fully-connected layer.	The study magnifies breast cancer data using single-task and multi- task CNN.
2	2018	Breast Cancer Detection From Histopathological Images Using Deep Learning.	CNN, Deep Learning, SVM, Random Forest	To improve cancer diagnosis, the paper suggests testing the approach with new features and an actual photo dataset. The method works for other cancers.	The study compares deep learning, CNN, and other machine learning algorithms for categorization.
3	2015	Breast Cancer Risk Prediction Using Data Mining Classification Techniques	Naïve bayes, J48 Decision tree	By expanding the training set sample size, the classifier can cover more attributes and create a more accurate model.	Based on question naires, this study assessed risk using naive bayes and decision tree algorithms.
4	2023	A model for predicting both breast cancer risk and non breast cancer death among women.	Gail Model for risk asessment, NHS analysis	Risk assessment parameters like breast density from mammography re- quire prior diagnosis. Asian and Hispanic ladies have not tested the model.	This study predicts cancer and non- cancer deaths in 55-year-old women of diverse races and origins. Risk competitive model for same.

Table 1. Comparison of different existing researches

According to Naresh Khuriwal et al., who used the median filter and histogram equalization method for pre-processing breast cancer data sets of histopathology images and implemented a convolution neural network algorithm on the 12 features, like mean, standard deviation, kurtosis, skewness, entropy, energy, etc., to achieve 98% accuracy [4].

According to Mohamad Mahmoud Al Rahhal, using CNNs and deep learning to classify breast cancer in histopathology images. The data set employed in the study is made up of a number of convolution layers, followed by fully connected and max-pooling layers. A cross-entropy loss function and the Adam optimizer were used to train the CNN, and data augmentation methods were applied to increase the model's generalized. Five-fold cross-validation was used to evaluate the suggested approach, which resulted in an accuracy of 83.25% in CNN. An accuracy of 86.60 is attained with VGG [5].

According to Kehinde Williams et al., who has used data mining techniques for the assessment of risk factors involved in breast cancer in Nigerian patients with a primary focus on the two data mining methods, Naive Bayes and J48 decision tree, the J48 decision trees showed a higher accuracy of 94.2% with lower error rates compared to those of the Naive Bayes, which showed an accuracy of 82.6%. Some of the risk factors that were taken into consideration are family history of breast cancer, age at first birth, age at menopause, and smoking frequency [6].

Mara A. Schonberg et al., have used competing risk regression and data from 83,330 women greater than 55 years of age. The women here have completed answering NHS 2004 questionnaires. This is a Black Women's Health Study(BWHS). Age, race, family history of breast cancer, body mass index, smoking status, and comorbidities were all taken into account by the algorithm when calculating risk. Since most models only estimate breast cancer risk or take non-breast cancer death into account as a competing risk, the work fills a vacuum in the literature. The suggested model significantly improves the ability to forecast breast cancer risk in women 55 years or older [7].

3. Flow Diagram of Proposed work

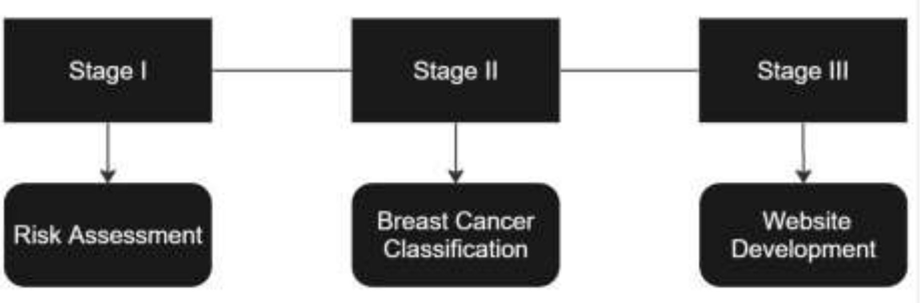


Fig. 1. Stages of Proposed work

3.1 Risk assessment

The following steps shown in the flowchart Fig. 2 are used by our Machine Learning model based on the Decision Tree algorithm to perform the risk analysis of a patient taking up the test on our website.

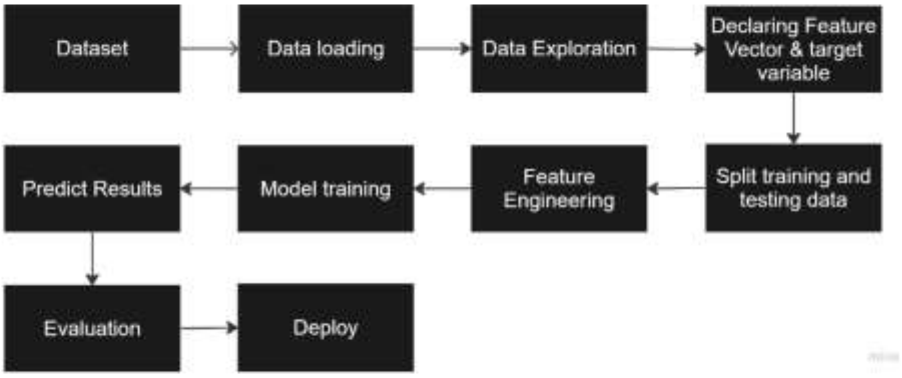


Fig. 2. Steps for Risk Assessment using K-Nearest Neighbor

3.2 Classification

The flowchart in Fig. 3 shows the methodology followed for the classification of ultrasound breast images into malignant and benign, the data set gathered from Kaggle was first loaded into the model and then pre-processed to increase the fea- ture extraction accuracy, and then the Convolution Neural Network layers were defined and finally the results very cross-validated using k-folds cross-validation method.

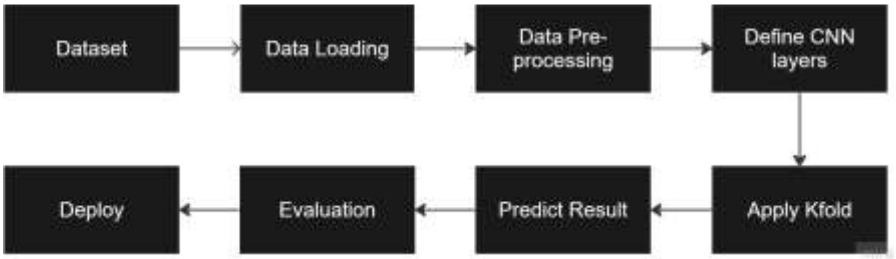


Fig. 3. Steps for Classification using CNN

4. Proposed Methodology

4.1 Risk Assessment

Data Collection and Preparation - The data set from Breast Cancer Surveil- lance Consortium(BCSC) was used for the risk assessment purpose [8]. There are 395297 rows and 11 columns in the data set with attributes, age group, race/ethnicity, history of breast cancer in a first-degree relative, age at first men- struation, age at first birth, use of hormone replacement therapy, menopausal status, body mass index, and previous breast biopsy. In Table 2 the example questions based on various factors asked in the risk assessment stage are shown. The data set is partitioned into two classes 1 and 0 where 1 indicates Cancer cases and 0 indicates non-cancerous cases. 72,468 (18.3%) of the data set has can- cer cases and 81.6% of the data set is non-cancerous. For balancing the data set Synthetic Minority Over-sampling Technique(SMOTE) technique is used which works by creating synthetic samples of the minority class by interpolating be- tween the minority class samples. The train test split is 70:30 in the data set, 70% training data, and 30% testing.

An individual's chance of acquiring breast cancer is determined by a num- ber of factors, including age, family history, personal health history, and lifestyle, these factors are considered while making the risk assessment model using differ- ent Machine Learning algorithms like Decision Tree, Naive Bayes, and K nearest neighbor.

Sr. No.	Questions	Options
1	Age Group in Range of 5 years	18-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84,≥ 85
2	Age(years) at firs period/menstruation	≤ 12, 12 − 13, ≥ 14
3	Age(years) at first birth	≤ 20, 21 − 24, 25 − 29, ≥ 30, *Nulliparous*
4	Menopausal Status	Pre-menopausal stage, Post- menopausal stage, Surgical Menopause, Unknown
5	Mody Mass Index(BMI)	10-24.99,25-29.99, 30-34.99, ≥ 35

Table 2. Example questions asked in the risk assessment stage.

Decision Tree - A decision tree is a machine-learning algorithm that is com- monly used for classification and regression tasks. In order to build a decision tree, the data first needs to be recursively divided into subsets according to the values of one of the input features. In order to achieve maximum homogeneity with regard to the target variable, splits must be designed. Decision trees can be constructed using a variety of methods, including ID3, C4.5, and CART. Each algorithm has a unique way of choosing the best feature to divide the data into different groups and build a tree. A decision Tree can be utilized for binary and multi-class classification issues and can handle both category and numerical in- put features [9]. The decision tree uses entropy which means the quantity of data required to accurately characterize a sample. Another parameter of the decision tree is the Gini index which measures the inequality in the sample [10].

$Gini = 1 - \Sigma(p)^2(ci)$ (1)

$Entropy = \Sigma(-p(ci) * (\log 2(p(ci))))$ (2)

Naive Bayes - Naive Bayes is based on Bayes' theorem, which explains the likelihood of an event based on knowledge of potential circumstances that may be relevant to the event and is commonly used for classification tasks. The term naive is derived as it assumes that the input features are conditionally indepen- dent given the class, which means that the presence or absence of one feature does not affect the probability of another feature. In order to use the Naive Bayes technique, training data must first be used to estimate the prior proba- bility of each class. Then, based on the frequency of the feature in each class, it determines the conditional probability of every attribute given in each class, also known as the likelihood. The class with the highest probability is selected as the predicted class after applying Bayes' theorem to determine the subsequent likelihoods of each class given the input features [11]. The Gaussian Naive Bayes theorem is predicated on the premise that continuous values are drawn from a Gaussian distribution [12]. It also assumes the following:

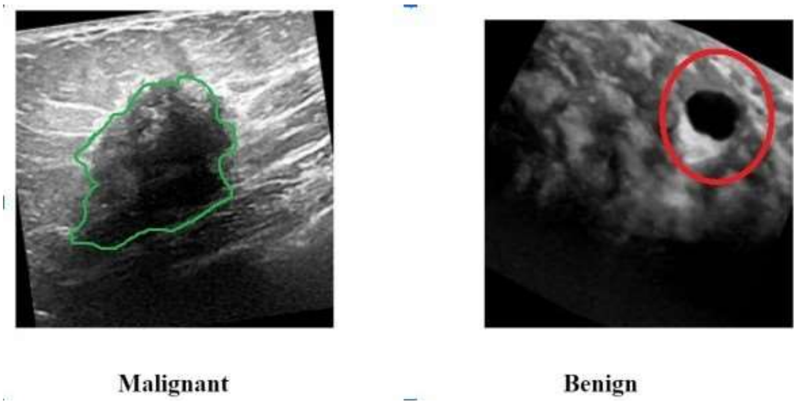
$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \qquad (3)$$

K-nearest neighbor - The K-Nearest Neighbors (K-NN) algorithm is a type of supervised machine learning used in classification and regression tasks. Its ability to recognize patterns and do predictive analysis without depending on any assumptions about the distribution of the underlying data is one of its strong points. The K-NN algorithm recognizes the data points that are closest to a new data point when it is presented, accounting for any features that might have scale variations. The nearest data points are sorted in order of increasing distance once the distance between these data points and the new point is computed, often using Euclidean distance. The algorithm then chooses a predetermined number of nearby data points, typically an odd number if there are two classes, and classifies the newly added point according to the category with the greatest number of data points [13]. The formula for computing the distance in KNN is given as:

$$P(y=j \mid X=x) = \frac{1}{K} \sum I(y^i = j) \tag{4}$$

4.2 Breast Cancer Classification

Data Collection and Preparation - We have used a data set of Ultrasound Breast Images of Kaggle for the classification purpose which consists of 8116 images for training of which 4074 (50.19%) are Benign instances and (49.80%) are Malignant cases. There are 500 malignant and 400 benign images used for testing [14]. The area in Fig. 4 encircled with red in the benign image shows a non-cancerous tumor and is concentrated whereas the green marked area in the Fig. 4 malignant image shows how breast cancer cells are spread which is one of the significant features extracted by the machine learning for image classification into benign and malignant.



Malignant **Benign**

Fig. 4. Image of Malignant and Benign tumor cells

Breast Cancer Classification into malignant and benign can be performed us- ing deep learning by training a neural network model. Commonly used methods for image classification in deep learning are Convolution Neural Networks, Re- current Neural Networks, DenseNet121, and DenseNet169.

Convolution Neural Network - The CNN architecture consists of several layers, including convolution layers, pooling layers, and fully connected layers. In order to identify image characteristics like edges, corners, and textures, the convolution layers conduct a sequence of convolutions on the input image using a collection of filters or kernels. The pooling layers downs ample the output of the convolution layers, reducing the dimensional of the feature maps and making the model more computationally efficient. The flattened output of the pooling layers is used by the fully connected layers in predicting the class of the input image. CNNs are trained using labeled training data, where each image in the training set is associated with a label indicating its class. The neural network's weights are tuned during training to reduce the discrepancy between expected and actual labels. Typically, a loss function like categorical cross-entropy is used for this [15].

The work done here uses ReLU deep learning function as an activation func- tion which is simple and does not require heavy processing as shown in Fig 5. The formula for ReLU is: Relu(x) = max(0, x)

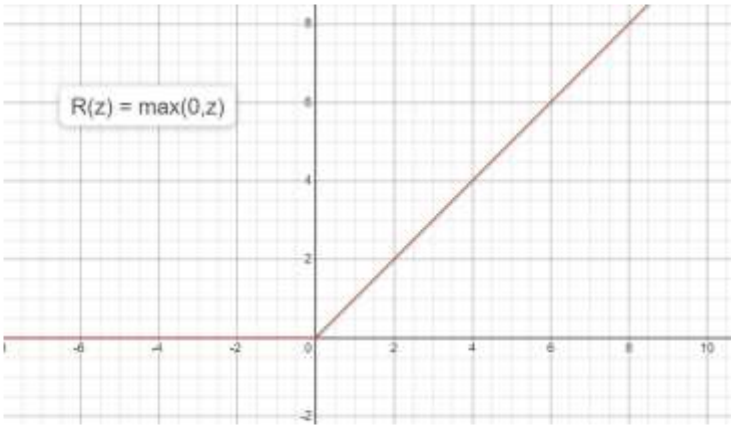


Fig. 5. Graph Representing the ReLU function

Another function that is used is the Softmax activation function. It is applied on the outer layers having two neurons (Fig 7).

$$Softmax(x_i) = \sum \frac{\exp(x_i)}{_j \exp(x_j)} \qquad (5)$$

DenseNet121 - DenseNet121 is a pre-trained convolution neural network ar- chitecture for image classification tasks. It consists of a total of 121 layers, in- cluding 4 dnse blocks and 3 transition layers. A batch normalization layer, a ReLU activation function, and a number of convolution layers are all included in each dense block. The output of each dense block is then fed into a transition layer, which includes a 1x1 convolution layer and a 2x2 average pooling layer to reduce the spatial dimension of the feature maps. DenseNet121 was pre-trained on the Image-Net data set, which consists of over 1 million images across 1000 classes, and achieved state-of-the-art performance on the classification task [16]

As known, densenet does not sum the output feature maps of the layer with the incoming features but concatenates them.

$$(x_l) = H_l([x_0, x_1, ..., x_l - 1]) \qquad (6)$$

4.3 Website Development

The web application has been developed using HTML, CSS, and Javascript, for the frontend development and PHP for the backend development. The website is divided into four broad pages, which are the home, risk assessment tool, clas- sification tool, and cancer care center suggestion page. The website's home page contains different facts about breast cancer for the overall general awareness of the user. Images and texts are used to display the facts with higher ease espe- cially for people living in rural areas. The risk assessment tool allows users to assess their risk of developing breast cancer. PHP is used in the backend to de- tect the risk based on user input age, family history, medical history, menopause age, etc factors. This tool then gives the result whether the patient is prone to breast cancer or not. The breast cancer classification tool allows users to upload their ultrasound images and classify the tumor as Benign or Malignant based on the input image. The classification .h5 deep learning model and risk assessment pickle machine learning model are deployed on the website using Python and PHP for the interface of models with the website. We have used Google Map API for showing the nearest breast cancer hospitals.

5. Result Analysis

5.1 Risk Assessment

As shown in Table 3, Risk Assessment was performed using three machine learn- ing algorithms, decision trees, naive bayes, and k-nearest neighbors. The Naive Bayes algorithm showed an accuracy of 65% and that of the Decision tree showed 75%.

Table 3. Evaluation of different risk assessment algorithms.

Sr. No.	Algorithm	Accuracy
1	**K-Nearest Neighbor**	85.99%
2	**Decision Tree**	75.91%
3	**Naive Bayes**	63.66%

We deployed our website with the k-nearest model as it showed the best result with an accuracy of 85%. The system gives the output from two classes, whether the user is prone to breast cancer or not.

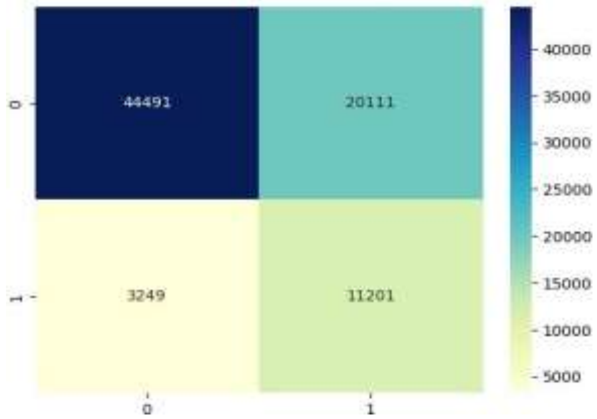


Fig. 6. Confusion Matrix of Risk Assessment of Breast Cancer

As shown in Fig 6, the confusion matrix has been plotted on the k-nearest neighbor model, giving the highest accuracy of 85%. There are 3249 cases that have been classified incorrectly. as not prone to breast cancer of the total data we had used to train and test the model and 20111 instances that are classified incorrectly. as prone to breast cancer.

5.2 Classification

According to the Confusion Matrix(Fig 7) plotted using the matplotlib library available in Python after the model has been trained and tested using Convolutional Neural Network and has been cross-validated using k-folds cross- validation. The results are quite promising as there are only 20 malignant cases misclassified as benign and 35 benign instances misclassified as malignant. The accuracy achieved on classification is 97%.

In the Fig. 8, graphs represent the accuracy and loss obtained on the train- ing and testing data after every epoch in the convolution neural network, k-folds

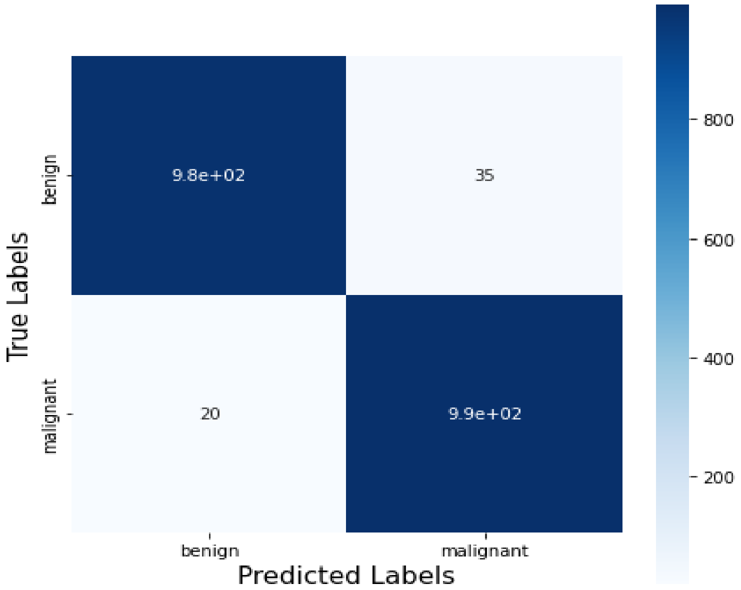


Fig. 7. Confusion Matrix of Classification of Breast Cancer

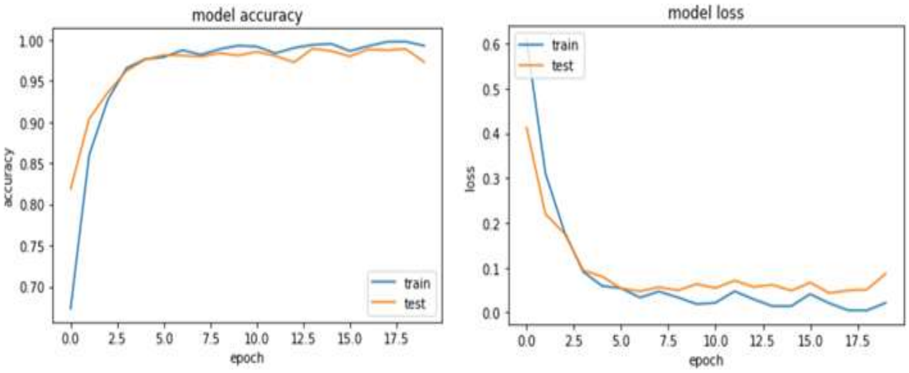


Fig. 8. Training model accuracy and loss

cross-validation with 10 folds was performed to get to the accuracy of 97%. We can also notice in the graph that the model is well generalized as the gap be- tween test and train accuracy is less. The graph which has been plotted using the matplotlib library from Python represents the accuracy.

From Table 4 and Table 5, we can see the accuracy and loss of the graphs. The accuracy after the first 5 folds and from Table 5 the average accuracy after the 10 folds is also calculated and listed in the table below for a comparative understanding of the model performance.

Table 4. Accuracy after first 5 fold and average accuracy calculated. Average accuracy after 10 fold is 97.422 % and loss is 0.111%.

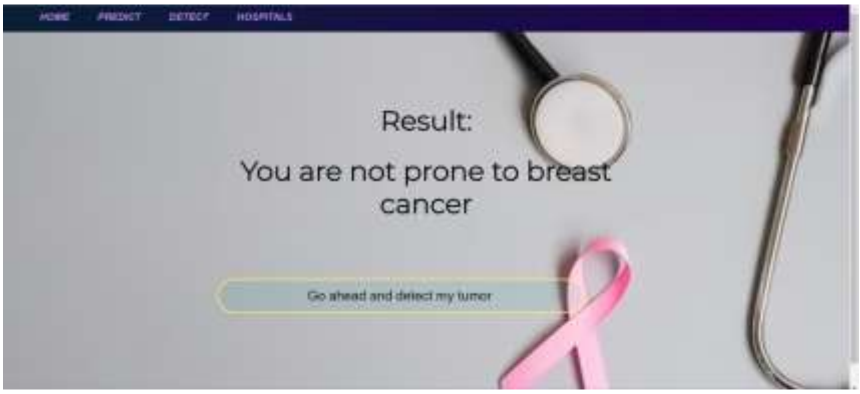
Folds	Accuracy	Loss
1	**98.331%**	0.109
2	**98.808%**	0.049
3	**98.687%**	0.071
4	**92.601%**	0.243
5	**97.494%**	0.097

Table 5. Accuracy after first 5 fold and average accuracy calculated. Average accuracy after 10 fold is 97.422 % and loss is 0.111%.

Sr. No.	Algorithm	Accuracy
1	**CNN using k-fold cross validation**	97.4227%
2	**DenseNet121**	98.489%
3	**DenseNet169**	96.979%
4	**CNN**	97.289%

5.3 Website Results

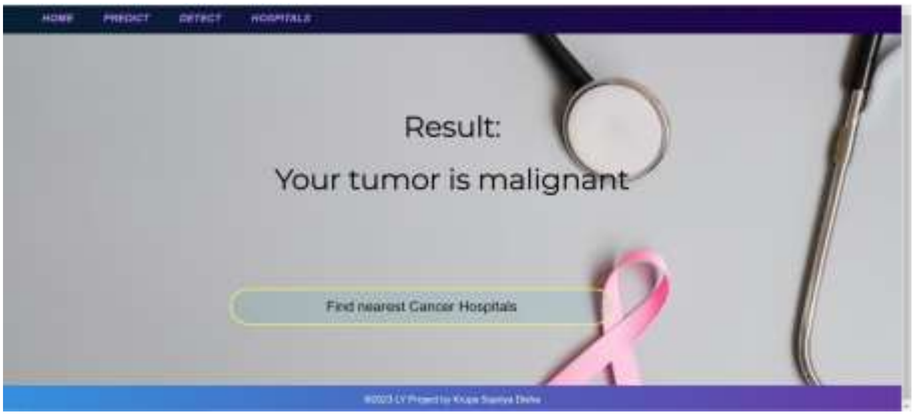
This study used a total of 10 features to identify breast cancer risk. However, the model might gain from including fresh risk factors as they are identified, such as genetic markers or way of life elements like diet and exercise. Fig. 9 and Fig. 10 show us the results of the output on the website.



Result:

You are not prone to breast cancer

Fig. 9. Results of Risk Assessment on Website


HOME
PREDICT
DETECT
HOSPITALS
Find nearest Cancer Hospitals
©2023 LY Project by Krupa Supriya Disha

Result:

Your tumor is malignant

Fig. 10. Results of Classification on website

6. Conclusion and Future Work

This study used a total of 10 features to identify breast cancer risk. However, the model might gain from including fresh risk factors as they are identified, such as genetic markers or way of life elements like diet and exercise. The KNN model utilized in this study has an opportunity for future optimization, which might improve its accuracy, which is now 85%. In conclusion, the method for assessing the risk of breast cancer reported in this study holds promise for further study because it has the ability to incorporate more risk factors and improve accuracy. Although CNN has produced impressive results in the classification of breast can- cer, there is still much room for further research and development. Investigating new architectures, hyperparameters, or regularization methods may improve ac- curacy. Another strategy to increase breast cancer categorization accuracy is to train the CNN using multi-modal data, such as mammograms, MRI scans, and CT scans. Future studies might concentrate on creating models that can success- fully combine various forms of data to increase accuracy. The CNN model used in this work also has the potential to be applied in actual clinical settings. As a result, the suggested techniques may eventually be used to treat other cancers as well as breast cancer.

References

1. Priya Ranganathan, Manju Sengar, Girish Chinnaswamy, Gaurav Agrawal, Ra- jkumar Arumugham, Rajiv Bhatt, Ramesh Bilimagga, Jayanta Chakrabarti, Arun Chandrasekharan, Harit Kumar Chaturvedi, et al. Impact of covid-19 on cancer care in india: a cohort study. The Lancet Oncology, 22(7):970–976, 2021.

2. Ajeet P Maurya and Swagata Brahmachari. Association of hormonal and repro- ductive risk factors with breast cancer in indian women: A systematic review of case–control studies. Indian Journal of Cancer, 60(1):4–11, 2023.

3. Neslihan Bayramoglu, Juho Kannala, and Janne Heikkil¨a. Deep learning for mag- nification independent breast cancer histopathology image classification. In 2016 23rd International conference on pattern recognition (ICPR), pages 2440–2445. IEEE, 2016.

4. Naresh Khuriwal and Nidhi Mishra. Breast cancer detection from histopathological images using deep learning. In 2018 3rd international conference and workshops on recent advances and innovations in engineering (ICRAIE), pages 1–4. IEEE, 2018.

5. Mohamad Mahmoud Al Rahhal. Breast cancer classification in histopathological images using convolutional neural network. International Journal of Advanced Computer Science and Applications, 9(3), 2018.

6. Kehinde Williams, Peter Adebayo Idowu, Jeremiah Ademola Balogun, and Adeni- ran Ishola Oluwaranti. Breast cancer risk prediction using data mining classifica- tion techniques. Transactions on Networks and Communications, 3(2):01, 2015.

7. Mara A Schonberg, Emily A Wolfson, A Heather Eliassen, Kimberly A Bertrand, Yurii B Shvetsov, Bernard A Rosner, Julie R Palmer, and Long H Ngo. A model for predicting both breast cancer risk and non-breast cancer death among women¿ 55 years old. Breast Cancer Research, 25(1):8, 2023.

8. Bcsc.

9. Decision tree, Jul 2022.

10. Audrea says:, Admin Says:, Balle says:, 3M 3200 says:, Evangelina Goodrow says:, Jason Roy says:, Mia Park says:, and Harrell Cannon says:. Gini index vs entropy information gain: Decision tree: No 1 guide, Apr 2022.

11. Turing. Naive bayes algorithm in ml: Simplifying classification problems, Mar 2022.

12. Prateek Majumder. Gaussian naive bayes, Feb 2020.

13. K-nearest neighbor(knn) algorithm for machine learning - javatpoint.

14. Vuppala Adithya Sairam. Ultrasound breast images for breast cancer, Nov 2022.

15. Manav Mandal. Introduction to convolutional neural networks (cnn), Apr 2023.

16. Arjun Sarkar. Creating densenet 121 with tensorflow, Jul 2020.