iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50413

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Breast Cancer Prediction Using Machine Learning

Riddhi R. Gujar[1], Apurwa A. Rajurkar[2], Yash G.Naik[3] , Shreya D. Bhise[4] , Dr. Alam N. Shaikh[5]
[1, 2, 3, 4, 5]*Electronics and Telecommunication Department Vasantdada Patil Pratishthans College of Engineering, Mumbai, Maharashtra, India*

Abstract: Worldwide, breast cancer is the most common cause of death for women, and increasing survival rates require early identification. Medical scans like mammograms and ultrasounds can be used to predict breast cancer using convolutional neural networks (CNNs). The architecture of CNNs, training data and methods, and performance evaluation criteria are all reviewed in this work as well as the current state of research on CNNs for breast cancer prediction. Also, the benefits and drawbacks of CNNs against conventional techniques for the identification of breast cancer are explored. Although though CNNs have the potential to lower false-positive outcomes and have higher accuracy rates, further testing and study are required to assure their dependability. Also, it's critical to consider moral concerns like data privacy and bias in machine learning algorithms. Thus, using CNNs to predict breast cancer has enormous potential to increase early detection and, eventually, save lives.So, our objective is to develop a model that can predict breast cancer from mammography images, which will help patients choose between various tests as well as assist medical students in validating their study.

Keywords: CNN (Convolutioal Neural Networks), RGB (Red Green Blue), FCL (Fully connected Layer), Benign, Malignant

I. INTRODUCTION

Breast cells are the origin of the malignancy known as breast cancer. It can afflict men as well as women and is the most common cancer in women globally. Breast cancer can begin in a variety of locations, such as the glands that generate milk or the ducts that bring milk to the nipple (lobules). Although the actual causes of breast cancer are unknown, there are a number of risk factors that can raise one's risk of getting the illness, including age, gender, family history, genetics, and lifestyle choices. All women should have regular breast cancer screenings because early detection and treatment are crucial for managing breast cancer successfully. Mammography and physical exams are two common approaches for finding breast cancer, however, they are not always reliable. Mammography has long been the go-to method for detecting breast cancer since it employs X-rays to produce images of breast tissue. It does have limitations, though, especially for women with dense breast tissue. Mammography is less effective at finding early-stage tumors in women with thick breasts because it can miss up to 20 percent of breast cancers. Physical examination, in which a doctor feels the breasts for lumps or other anomalies, is also not always accurate. Finding tiny tumors or ones that are hidden might be challenging. Additionally, false-positive results from mammography and physical exams might result in unnecessary biopsies and patient worry. To overcome these limitations, more advanced technologies are being developed, such as 3D mammography and breast MRI. They cost more than conventional procedures and are not as generally accessible. Thus traditional method has a high level of demonstrated inaccuracy. As a result, we apply artificial intelligence to detect things that will give us the most accurate results possible. Develop an algorithm that can analyze mammography images and predict if a patient has breast cancer or not. With human lives at stake, the algorithm has to be highly accurate. While performing ANN the first layer neurons can be up to millions and hidden layer neurons are also in a range of millions in case of a large-size RGB image. So the total weights that have to calculate come up to more than 20 million which is a lot to compute for a computer. Thus ANN is not feasible for such types of images also it is sensitive to the location of an object in the image. The neurons in our brain work on different features of an image then aggregate the results and identify the object. Similarly, the computer recognizes these features with the help of filters.

Many studies have demonstrated that CNNs can perform better than conventional techniques for finding breast cancer, like mammography and ultrasound. In contrast to conventional approaches, which had an accuracy of 77.3% when predicting breast cancer using mammography pictures, Wang et al(2016)'s study indicated that a CNN model had a 90.2% accuracy rate. Convolutional neural networks have been used in a number of research to predict breast cancer. In one study, Phan et al. (2021) used a complex convolutional neural network without region-of-interest labelling to predict the recurrence of breast cancer. Deep convolutional neural networks and support vector machines were applied for theidentification of breast cancer in a different study by Ragab et al. (2019). Moreover, studies have examined the application of trained convolutional neural networks for the diagnosis of breast cancer (Chen et al., 2020).

Also, an integrated machine learning framework for classifying SEER breast cancer using enhanced convolution neural networks has been created (Liet al., 2023). These papers show how convolutional neural networks can be used to predict and detect breast cancer.To help patients decide whether to proceed with the costly biopsy test or not, we have developed a model in this research that analyses breast cancer using mammography images rather than biopsy images

II. CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural networks are a subgroup of artificial neural networks used in computer vision, image and video processing, and other related fields. In order to locate patterns and features in the data, convolutional filters are used to dynamically and interactively learn spatial hierarchies of features from input data. High accuracy and data handling capacity are demonstrated by CNNs. Convolutional neural networks are made up of crucial layers like convolution layers, max-pooling, padding, and fully connected layers.

A. Convolution Layer

Convolutional neural networks' central processing unit is the convolution operation. The primary factor in the convolutional neural network's successful performance is convolution operation, which is in charge of recognising the image's edges and features. Let's imagine we have a little image that is 6x6 pixels in size. These numbers indicate the image's pixel values, and if it is a grayscale or black-and-white image, each pixel value spans from 0 to 255. Between 0 and 255, the spectrum's extremes are black and white, respectively. Moreover, we have a 3x3 pixel filter. By superimposing these filters on their image and utilising it as a sliding window, we can acquire the values of the 4x4 resultant matrix produced by the convolution operation between their image and the filter. It is a process of using a filter to extract the output matrix from the input image. Every image with a size of n*n and a filter of f*f will produce

$(n\times n)*(f\times f)=(n-f+1)\times(n-f+1)$

Thus convolutional operator acts as a feature detector or edge. We will use a large number of these filters in a single layer of a convolutional neural network, and these various filters will be identifying different properties of these images. If we employ c such filters, then c photos will be included in the output as follows

$(n-f+1)\times(n-f+1)\times c$

Red, Green, and Blue are the three channels that make up a coloured image. Hence, one coloured image will be nxnx3 in size. In order to conduct convolution operation on a coloured image, a filter with three channels (Fxfx3) is also required. This filer will now be superimposed over the coloured image, and the values will be multiplied by each individual cell. We will use numerous such filters in a single layer of the convolution neural network, and these filters may be able to detect the edges all across our image. These edges will now be transmitted to a subsequent layer that will identify features related to the image line texture and other features. And in a subsequent layer, it would be possible to see the full image's structure when merging textures, features.

B. Padding Layer

The multiple procedures carried out by numerous such convolution filters result in a size reduction of the resultant image that is so significant that we risk losing crucial information. Also, while the sliding window advances across the image, the pixels in the middle undergo multiple rounds of filtering while those in the corners only receive one. As a result, the image's borders on all four sides have zero padding. A 6x6 image becomes an 8x8 image when we pad it.

C. Stride

During the convolution process, we move our filter by one pixel to the right or down. This is referred to as having a stride length of 1. With a stride, we can move our filter by 2 pixels, 3 pixels, or any other number we like.

D. Maxpooling

The purpose of the pooling layer is to lower the layer's size and dimensions while maintaining its features. A filter with a fixed size must be taken into consideration along with strides when performing a pooling operation. Place the filter in the window's upper left corner to execute the max pooling operation. Next, we will extract the maximum value from the window. Next, move the window back by 2. As a result, the image size is reduced, which lowers the computational expense. This helps the model learn more quickly and improves and sharpens the image. Convolutional layer is always applied before the max pooling layer.

Once the features have been extracted, the Max pooling layer is utilised to improve the features while reducing the output image size from the Convolutional layer. It performs a downsampling procedure along the spatial dimensions, producing an image that is smaller and has fewer dimensions. Fig.1 demonstrates the convolution and maxpooling operation.

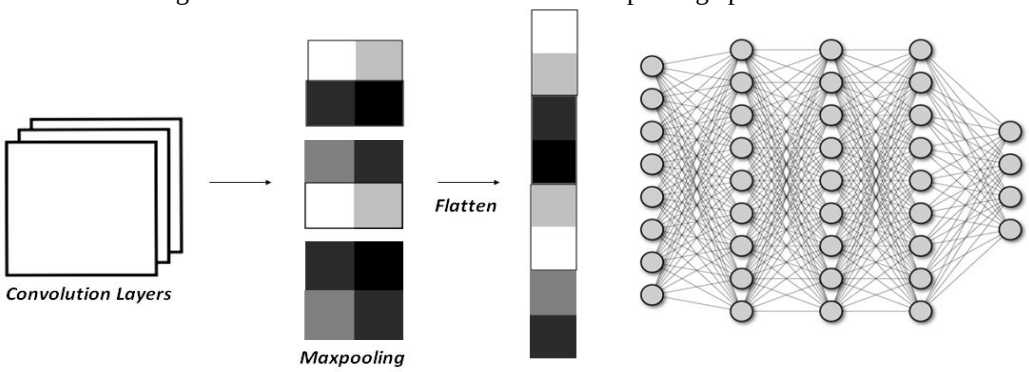


Fig. 1. Convolution and Maxpooling Operation

E. FCL

Fully connected layer is nothing but a dense network of neutrals. It is used to assign a picture to a certain category after its features have been extracted. The image with enhanced features from the max pooling layer is flattened into 1D vectors or arrays and used as an input to the FCL. Every single neural in a layer is linked to every neural in the layer before it and to the layer after it. The number of neurons in the final, fully linked layer will be equal to the number of categories that we currently have. Moreover, it links characteristics to the specific label.

F. CNN Architecture

Typical neural network is made up of the components mentioned above. Convolutional layer output is equal to the number of filters used after scaling the values with ReLU functions. Fig 2 above will help understand the CNN operation. In the Figure 2 represented above, the input image has the dimensions 32x32x3, where 3 denotes the use of an RGB or colour image. The image is convolved using a 5x5x3x4 filter, where 4 represents the total number of filters employed. We obtain 28x28x4 from the formula to remove the convolutional layer. We pass it to the maximum pooling layer after convolving. Let's say a filter with a stride of 2 is utilised for maximum pooling. The size of the output image is 14x14x4. The output is then once more sent to the conv2 layer with a filter size of 3x3x4x3 and 8 filters. The outcome of the maximum pooling operation of the 22 filters is 668. Depending on the application, this layer may be applied again. Convolutional layer and maximum pooling layer operations are completed, and then the fully connected layer is contacted.

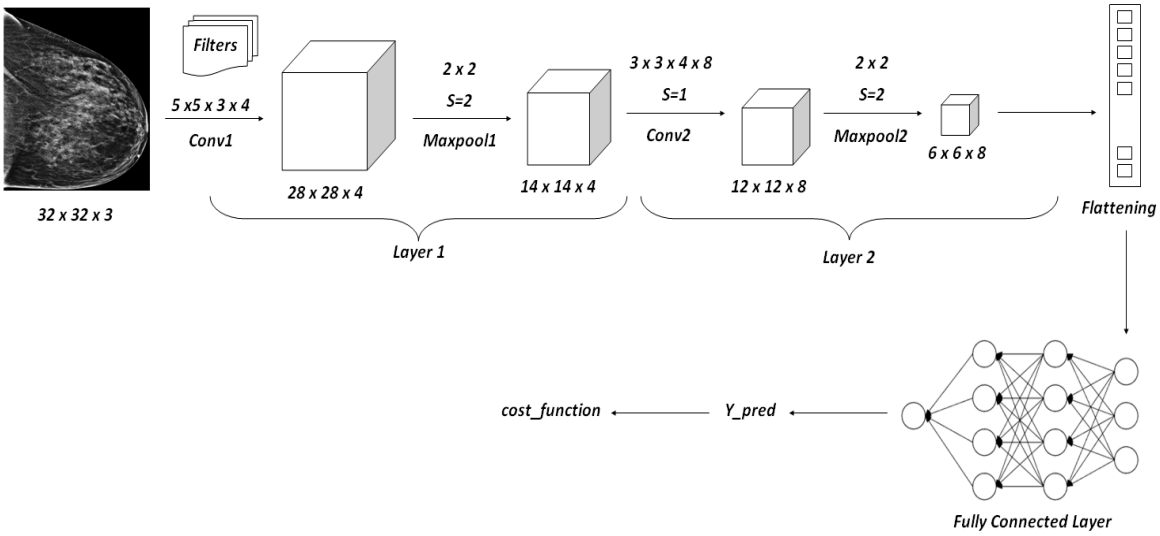


Fig. 2. CNN Architecture

First, a flattened 1D output of 288 is created from the final 6x6x8 image output. Next, the Fully connected layer is connected to the 288 weights. The sigmoid activation function is used in the top layer. The output of y-pred is presented in this last layer. The category of the image is predicted by the Y-pred. The cost function is later calculated using the y-pred. The cost function tells the errors, the model is receiving while making predictions.

III. MATERIALS AND METHODS

Fig.3 represents the methodology used by our proposed model to achieve maximum accuracy through CNN. The process is divided into 9 groups explained in detail below.

A. Dataset

The Dataset used in this project is Dataset BUSI with GT which consists of the breast Ultrasound Images Dataset by Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A, obtained from Kaggle. The baseline data includes breast ultrasound pictures obtained of women among the ages of 25 and 75. This data was compiled in 2018. There are 600 female patients in all. The collection consists of 780 photos, each measuring 500 by 500 pixels on average. The pictures are PNG files. Together with the original photos, real-world images are displayed. The pictures are divided into benign and malignant categories. The dataset, which consists of nearly 780 whole slide images from the five different medical centers, was organized to achieve the following two goals: first, to determine the probability that a given lymph node tissue is cancerous and to predict tumor regions in the whole slide image of the lymph node tissue.

B. Convolutional Neural Network to Predict Breast Cancer

1) Step 1- Upload Dataset

The information examines ultrasound-generated medical images of breast cancer. Images from the breast ultrasound dataset are divided into two groups: benign images and malignant images. Using Kaggle to Learn, you may access the dataset for Breast Ultrasound Pictures. Use fetch mldata to upload it.

2) Step 2: Data Pre-processing

Resizing photos, normalising data, and augmenting data are all examples of data pre - processing stage, which involves cleaning and preparing the data for analysis. The design of the model's architecture entails choosing the ideal CNN architecture, such as VGG or ResNet, and modifying the number of layers and neurons for speed optimisation.Data is fed into the CNN model during training, and weights and biases are changed to reduce the loss function. The model's efficacy in predicting breast cancer is then assessed using criteria including accuracy, sensitivity, and specificity.

3) Step 3 - Input layer

The data gets reshaped in this step. The square root of the pixel count determines the form. Here, for example, the shape is 125x125x3. You must indicate whether or whether the image is coloured. If so, you have three to the shape—three for RGB—instead of one.

4) Step 4 - Convolution layer

Using the same padding, the first convolutional layer has 32 filters with a 3x3 kernel size in the first two convolution layers. When there is equal padding, the input and output tensors' height and width should match. To verify that the rows and columns are of the same sizes, Tensorflow will add zeros to both. The Relu activation function is utilized.

5) Step 5 - Pooling layer

The computation for pooling comes after the convolution. The data's dimensionality will be decreased by the pooling computation. Use the max pooling2d module with a size and stride of 2x2 respectively. The previous layer serves as your input.

6) Step 6 - Dense layer

The completely connected layer must then be defined. Prior to connecting the feature map and thick layer, the feature map must be flattened. A Relu activation function is added. Moreover, a dropout regularisation term with a rate of 0.3 is added, resulting in 30 percent of the weights being set to 0.

The argument mode in the cnn model fn function is used to decide whether the model has to be trained or evaluated. Then these 2D maps are flattened into 1D array which is further connected to fully connected dense neural network. ReLU operation is used later to replace the negative values in output to zero and thus makes the model non linear.

7) Step 7 - Logistic layer

The final layer in the TensorFlow image classification example can be defined using the model's prediction. The batch size and 10, the total number of photos, determine the output shape. The classifications and the probabilities associated with each class can be put into a dictionary. The highest value from the logit layers is returned by the module tf.argmax(). The probability for each class is returned by the softmax function. The next step is to calculate the model's loss. The model must be optimised, or the best weight values, as the last step in the TensorFlow CNN example. A gradient descent optimizer with a learning rate of 0.001 is used for that. The goal is to reduce the loss.

8) Step 8 - Image Classifier

The input consists of 1164 photos from our training dataset, each of which has been assigned to one of two classes.The classifier is then trained using this training set to discover the characteristics of each class.Finally, we assess the classifier's performance by asking it to forecast labels for a fresh batch of photos that it has never seen before. Afterward, we will contrast the actual labels on these pictures with those that the classifier predicted whether Benign or Malignant.

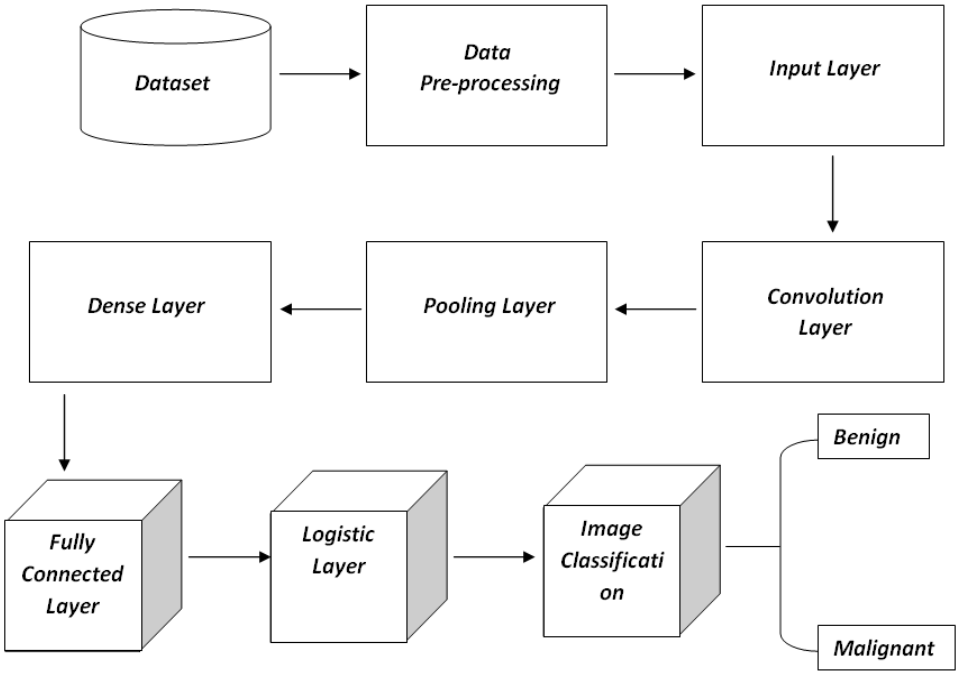


Fig. 3. Flow of Methodology

IV. RESULTS AND DISCUSSIONS

Comparison of CNN with other Machine Learning Algorithms

A. Decision Tree Algorithm

The decision tree method divides the dataset recursively using decision nodes until there are only pure leaf nodes left. By maximizing the entropy gain, it determines the ideal split. A data sample advances to the left side if the decision node's condition is met; otherwise, it moves to the right until it reaches a leaf node, when a classable is assigned to it. The strong sensitivity of decision tress to the training data, however, may lead to High Varience. Due to the inability of the Decision Tree Algorithm Model to generalize, it only provides an accuracy of 73.33 %.

B. Random Forest Algorithm

Random Forest Algorithm is employed to solve the drawback of high variance in Decision tree algorithm. It is made up of a number of different random decision trees and is considerably less sensitive to the training set of data. Building new datasets from the original data is the first stage in applying the Random Forest algorithm. Then, random rows are chosen at random from the original dataset to create additional datasets, each of which has the same amount of rows as the original dataset. This process is known as bootstrapping. Additionally, decision trees are trained independently on each bootstrapped dataset, but the subset of features used for each tree is chosen at random. The accuracy of the model has boosted to 77.44% by using the Random Forest algorithm.

C. K-Nearest Neighbor Algorithm

K-Nearest Neighbor is simply a supervised classification algorithm that uses some data points or data vectors that have been divided into several categories in order to try and predict how quickly new samples will be classified. KNN does not learn on its own; instead, it memorises the process. KNN uses either the Euclidean distance or the Minkowski distance to determine the neighbourhood of a given point by measuring its distance from the example points. The model has attained accuracy of by using the KNN method is 67.69 %.

D. Naïve Bayes Classifier

Naïve bayes classifier works on the principle of Conditional Probability as given by the bayes theorem. The conditional probability of an event A given B equals to the probability of B given A times probability of A over the probability of B. This type of algorithm handles both continuous and discrete data and is highly scalable with number of predictors and data points.However the accuracy reached by model through Naïve Bayes Classifier is 56.92%.

E. Support Vector Machine Algorithm

A supervised learning approach called Support Vector Machine is utilised for both classification and regression models. In order to easily place new data points in the appropriate category in the future, SVM is used to generate the optimal line or Decision Boundary that can divide n-dimensional space into classes. SVM selects extreme points to aid in the creation of the hyperplane. These extreme points are called Support vectors. Through SVM 78.97% accuracy is obtained.

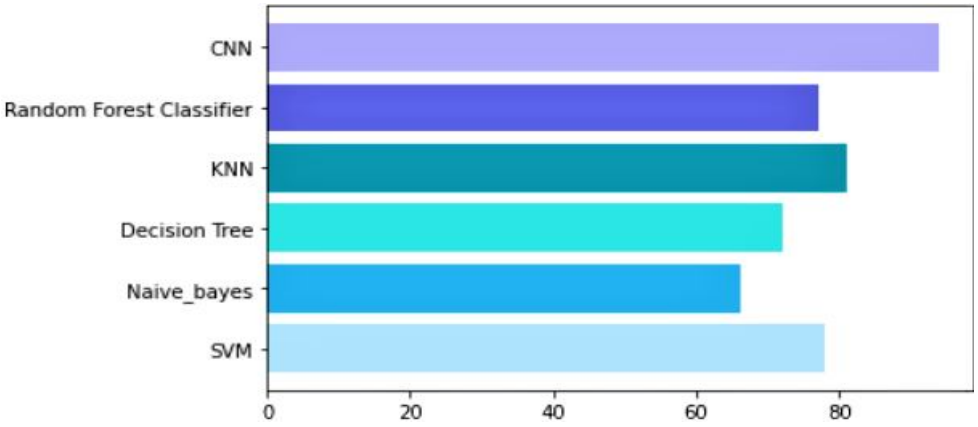


Fig. 4. Comparison Graph

Table. 1. Comparison Table

Algorithm	Training Accuracy	Testing Accuracy
1. Decision Tree	99.77%	72.31%
2. Random Forest	98.23%	77.95%
3. K- Nearest Neighbor	82.78%	81.03%
4. Naïve Bayes	62.25%	66.15%
5. Support Vector Machine	99.77%	78.97%
6. Convolutional Neural Networks	96%	93.85%

As obtained in Fig.4. it is observed that the best accuracy for model is achieved by the proposed methodology by using Convolutional Neural Networks. Table.1. gives detailed comparison between Testing and traing Accuracy of the Machine learning algorithms used for the model. It is seen that SVM and CNN have received high accuracy than others wherein CNN has achieved the the highest accuracy.

F. Convolutional Neural Networks

Since we are dealing with images, the model we are employing here is sequential to add layers this means that our images must be pre-processed. A statistic used to assess how well a deep learning model. Further layers, such as conv2D, are gradually added, along with a number of filters, kernel sizes, and activation functions that are specifically employed for this model, and then input image shape and size.By including the dropout rate, the data will be decreased at that rate. We used an image data generator to provide the input data, and we used a model fit generator to train our model around 10 iterations using validation data. The most common metric for evaluating model performance is accuracy. The accuracy attained by the model through Convolutional Neural Networks is 93.85 percent. Through model graphical interpretation we have seen how well our model has performed in Fig.5. represents the plot of accuracy & val-accuracy. It is seen that val-accuracy performed well throughout the model and accuracy is neither overfitting nor underfitting.The test accuracy is almost close to training accuracy.

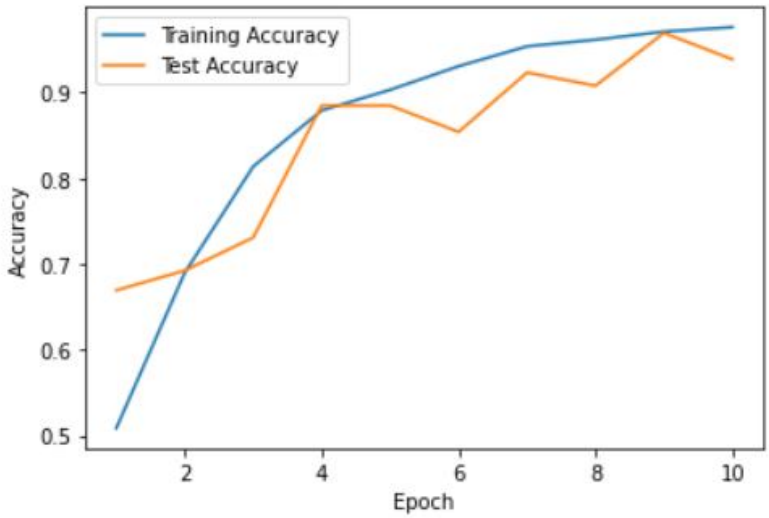


Fig. 5. Accuracy vs Validation Accuracy

The plot of loss and val-loss is shown in Fig.6. A statistic used to assess how well a deep learning model matches the training data is called the training loss, which is calculated by adding the total number of errors for every instance in the training set. On the other side, validation loss is a statistic that is used to assess how well a deep learning model accomplished on the validation set.In the image above, the training loss and validation loss both reduce and level at a specific point. This denotes a ideal fit that is model is neithe overfitting or underfitting.

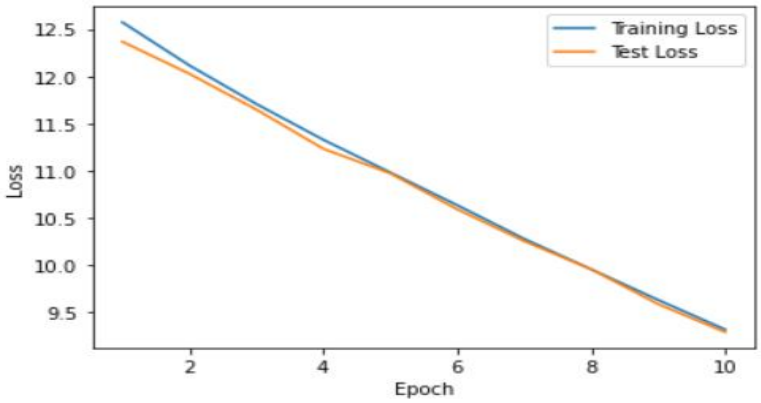


Fig. 6. Loss vs Validation Loss

The confusion matrix obtained in Fig.7. provides a graphical representation of the expected versus actual classifications by listing the number of true positives (TP), false negatives (FN), false positives (FP), true negatives (TN), in a table format. demonstrates that while the model correctly identified 100% of malignant cases as malignant (TP) and 90% of benign cases as benign (TN), it incorrectly labelled 9.88% of benign cases as malignant and 0% of malignant cases as benign (FN). (FP). As a result, the overall accuracy would be 93.85%, the sensitivity would be 100 percent, the specificity would be 90.12%, and the precision would be 85.96%. There are many limitations and difficulties when utilising CNNs to forecast breast cancer. Lack of reliable datasets for CNN model training and testing is a serious barrier that can lead to overfitting and incorrect conclusions.

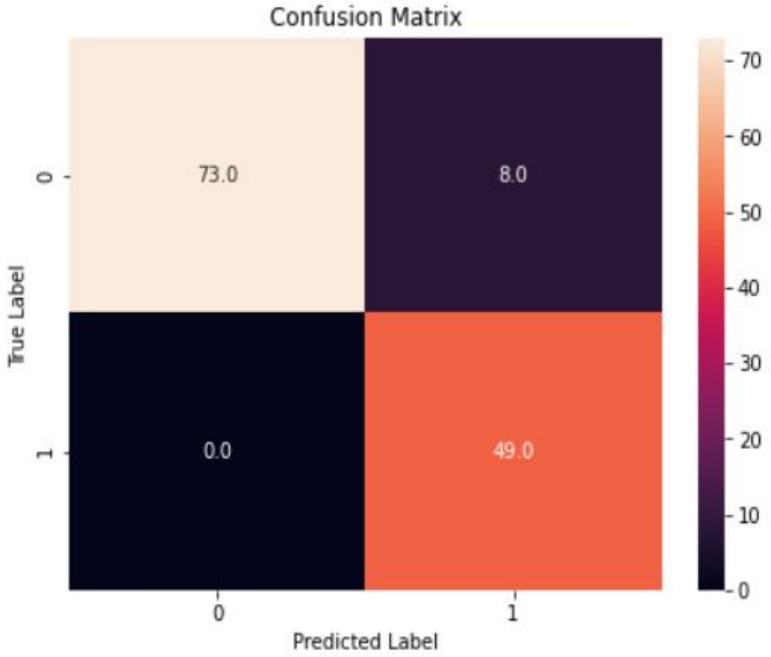


Fig. 7. Confusion Matrix

When using CNNs to predict breast cancer, there are a variety of constraints and challenges. Lack of consistent datasets for training and testing CNN models is a significant obstacle that can produce overfitting and inaccurate findings. Concerns about data privacy and bias in machine learning algorithms are also present. Further testing and research are also required to guarantee the accuracy and generalizability of CNN models in clinical contexts

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

V. CONCLUSIONS

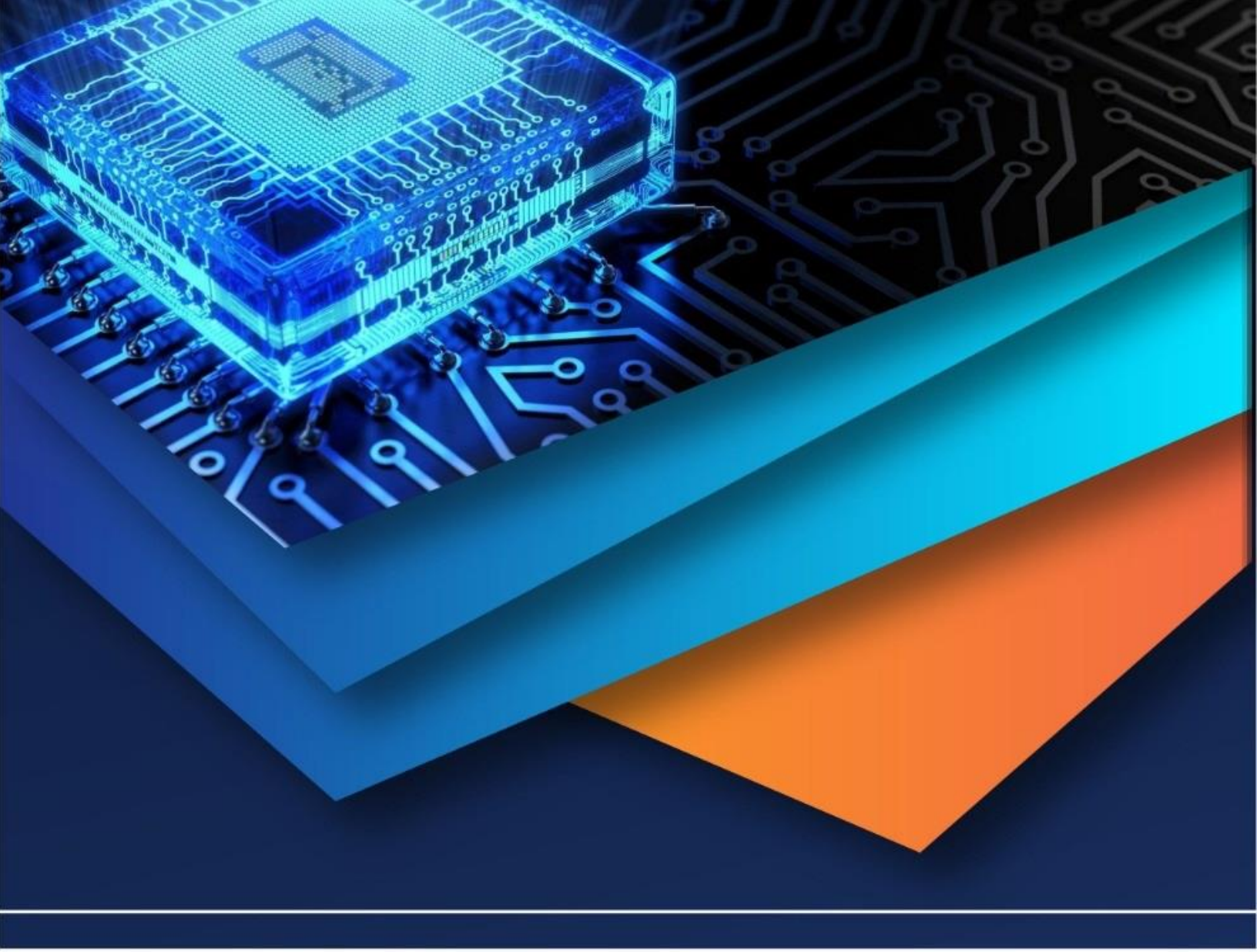
The Convolutional neural networks (CNNs) have demonstrated promising results in the prediction of breast cancer. CNNs are able to precisely identify patterns and features suggestive of breast cancer by examining medical pictures such as mammograms and ultrasounds. CNNs have demonstrated higher accuracy rates when compared to conventional techniques, and they may also lessen the amount of false-positive findings, enabling more precise and effective screening and diagnosis. To guarantee the accuracy and applicability of CNNs in predicting breast cancer, additional study and validation are required. It's also important to take into account ethical issues like bias in machine learning algorithms and data privacy. Overall, the use of CNNs to the prediction of breast cancer has enormous potential to enhance early diagnosis and eventually save lives. Using the idea of transfer learning is another way to continue to enhance the current model. The performance of CNN models for predicting breast cancer can be enhanced using the potent method known as transfer learning. Transfer learning can decrease the quantity of training data needed while increasing the model's accuracy and effectiveness by drawing on knowledge from previously trained models

VI. ACKNOWLEDGMENT

The Prof. Alam N. Shaikh provided encouragement for this research, thus we would like to show our appreciation and thank him for being able to work on it. Throughout the course of working on this project, we learned a lot of new terms and concepts, for which we are grateful

REFERENCES

[1] Mr. Madhan S, Priyadharshuini P, Brindha C, Bairavi B, Department of Computer Science and Engineering, University College of Engineering Thirukkuvalai,Predicting Breast Cancer using Convolutional Neural Network ”Special Issue ICMR Mar 2019
[2] Maleika Heenaye-Mamode Khan ,Nazmeen Boodoo-Jahangeer ,Wasimah Dullull ,Shaista Nathire ,Xiaohong Gao ,G. R. Sinha ,Kapil Kumar Nagwanshi, ”Multi- class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN)”, August 26, 2021
[3] Apoorva V1, Yogish H K, Chayadevi M L,”Breast Cancer Prediction Using Machine Learning, Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication Security (ICIIC 2021)Techniques”
[4] Apparna Allada, Ganaga Rama Koteswara Rao, Prasad Chitturi, M.S.N. Prasad, ”Breast Cancer Prediction using Deep Learning Techniques”,12 April 2021,Doi-.2021.9395793
[5] David Solti, Oregon Episcopal School, Haijun Zhai, Cincinnati Children’s Hospital Medical,”Predicting Breast Cancer Patient Survival Using Machine Learning”,Doi-2506583.2512376
[6] Ramik Rawal,School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology,”Breast cancer prediction using machine learning”,2020 JETIR May 2020, Volume 7, Issue 5
[7] Rohith Reddy1, Shrushti Gupta1, Shashank Ala1, Rasamsetti Sampath1,Chinta Sumanth1. Students, Department of Computer Science Engineering, GITAM Deemed to be University, Visakhapatnam, India-5300451,”Breast Cancer Detection using Convolutional Neural Network”,2022 JETIR April 2022, Volume 9, Issue 4
[8] Marcos Pinto and Ouri Alkada, Computer Systems Technology Department, NYC College of Technology, CUNY, Hsinrong Wei, Business and Information Systems Department, US Marine Merchant Academy (USMMA), ”Health Care AI : Predicting Breast Cancer with Machine Learning”,JCSC 34, 2 (December 2018)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY