IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BREAST CANCER DETECTION AND PREVENTION USING (Convolutional Neural Network)

[1]Gayatri Suryawanshi, [2]Ritesh Sharma, [3]Nitish Kumar

[1]Student, [2]Student, [3] Assistant professor

[1] Bachelors of computer application (Data Science),

[1]Ajeenkya Dy Patil, Pune, India

Abstract: The effectiveness of the majority of traditional classification systems depends on accurate data representation, and a large portion of the work is focused on feature engineering, a challenging and drawn-out procedure that leverages previous expert domain knowledge of the data to produce valuable features. However, deep learning does not require a domain expert to construct feature extractors; instead, it can extract and arrange the discriminative information from the data. The research community and industry have taken an interest in Convolutional Neural Networks (CNNs), a specific kind of deep, feedforward network that has shown empirical success in tasks like speech recognition, signal processing, object recognition, natural language processing, and transfer learning. To guarantee the efficacy of the model, the project carries out extensive data exploration, visualization, and preprocessing procedures in this work. The suggested model architecture performs well and classifies medical images with a high degree of accuracy. The findings demonstrate the potential of deep learning methods to support medical practitioners in the diagnosis and planning of diseases.

Key words: medicine - breast cancer The World Wide Web. applications in healthcare. algorithm for machine learning. Transfer learning, CNN.

I. INTRODUCTION

Cancer is a multifaceted illness. Given that cancer is the second greatest cause of death worldwide, according to estimates from the World Health Organization (WHO), there were 97,00,000 fatalities and 2,00,00,000 new cases of cancer worldwide in 2016. According to statistics, 670 000 people died worldwide in 2022 and 2.3 million women received a breast cancer diagnosis. There are many distinct forms of cancer, but as breast cancer is more common in women and has a high fatality rate, we shall focus on breast cancer in particular. The strongest risk factor for breast cancer, according to the sex ratio, is female gender. Therefore, early detection is safer and preferable to late detection. incredibly significant, as well as the manual examination procedure might cause a delayed diagnosis, which would mean a later course of therapy and possibly even death.Convolutional neural networks, or CNNs, and the Transfer Learning technique were employed in this thesis to get the most accurate result for the classification of images as malignant or noncancerous. In order to improve accuracy and enable the model to correctly forecast and classify the image, we have also employed a greater number of epochs. Based on the latest data from the World Health Organization's International Agency for Research on Cancer, breast cancer has surpassed lung cancer as the most common cancer, with 2.26 million new cases expected in 2020 Ref[5].It poses a serious risk to women's lives and health. prompt diagnosis is essential in the fight against cancer, and only a dependable detection system can make this happen. Digital pathology and medical image processing are two methods that have been developed to help in the diagnosis of breast cancer. Out of all the cancer types, breast cancer has two particularly concerning characteristics: it is the most common disease in women worldwide and it kills more people than other cancer types due to the fact that the most common method for diagnosing breast cancer is a histopathological examination. Pathologists still frequently diagnose patients by visually assessing histology samples under a microscope.ref[1].

DC stands for "Dermal Ductal Carcinoma." It is a prevalent form of breast cancer worldwide. It begins in the breast's milk ducts and spreads to the surrounding breast tissue. It was invasive because it may have moved outside of the ducts and into other areas of the breast. It disseminates via the bloodstream or lymphatic system. Research in automated histopathological image classification has the potential to expedite and lower the error rate in BC diagnosis. A biopsy is used in histopathology to provide photographs of the damaged tissue. Treatment of the illness and a better prognosis depend on early detection. Clinical breast assessment and tomography tests, such as magnetic resonance imaging, ultrasound, and mammography, are examples of noninvasive BC screening treatments. However, in order to verify the identification of BC, a diagnostician must perform a pathological examination of a part of the suspect area. Glass slides tarnished with hematoxylin and eosin are used to examine the minute intricacies of the tissue under study. There are a number of analytical techniques used to identify BC. A few common techniques include mammography, positron emission tomography (PET), magnetic resonance imaging (MRI), breast ultrasonography, surgery, or fine-needle aspiration to target the nerve of the suspected area (histopathological pictures), among others, as shown in (Fig. i.1) 1.fig imaging modalities for breast tissue: MRI, mammography, and ultrasound (a, b, c). Different techniques are utilized to assess digital pathology images of breast cancer, including rule-based and machine-learning approaches. Recent research has demonstrated that deep learning-based methods, which fully automate the processing, perform better than traditional machine learning techniques on a variety of picture assessment tasks. Convolutional neural networks (CNNs) have shown successful in medical imaging applications, enabling the early identification of diabetic retinopathy and the prediction of age and bone damage, among other issues. Previous research has shown that deep-learning-based functions in histological microscopic misprocessing can be useful in the early identification of breast cancer. Over the past few decades, machine learning has become more and more crucial in the identification of breast cancer. To create a classification model from a dataset using a machine-learning approach, a number of probabilistic, statistical, and optimization techniques need be applied. With 20 primary cancer categories and 18 lectotypes, breast carcinoma is a frequently classified histopathology based on the selection of morphological characteristics of the tumors. The two main histological subtypes of breast cancer are invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC), accounting for about 70–80% of cases. Deep-learning (DL) techniques have the ability to gather information from data, automatically extract features, and acquire complex abstract interpretations of the data. DL methods are effective.They have applications in a variety of fields, such as computer vision and biomedicine, and are capable of resolving common feature-extraction problems. A new BC histopathological Ima category blind inpainting convolutional neural network (BiCNN) model has been developed, which is based on deep convolutional neural networks. It was created to deal with BC's two-class diagnostic classification. The BiCNN model constrains the distance between the properties of distinct BC pathology images by using prior information of the BC class and subclass labels. To accommodate whole-slide image idetification acceptance, a data-augmented approach is offered. The transfer-fine-tuning-training approach is used when it is suitable.Figure 1. Breast tissue medical imaging modalities: ultrasound ,Mammography, MRI (a), (b), and (c)

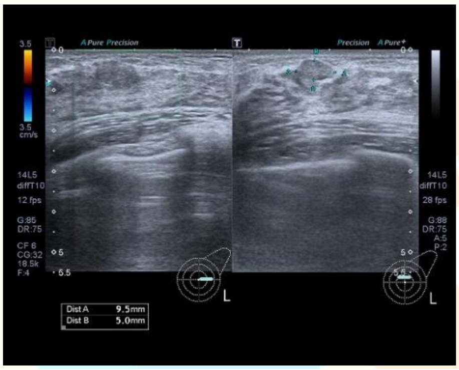


Fig.a

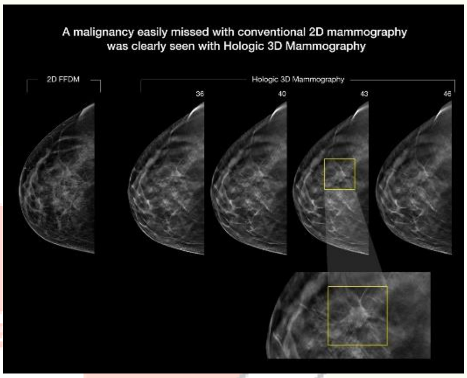

A malignancy easily missed with conventional 2D mammography was clearly seen with Hologic 3D Mammography
2D FFDM
Hologic 3D Mammography
36
40
43
46

Fig.b

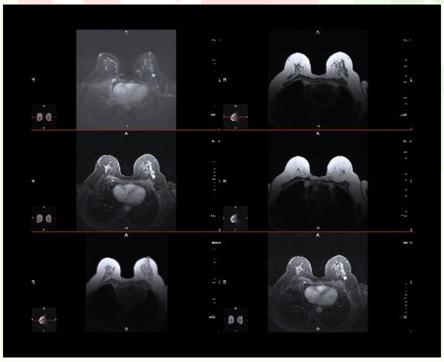


Fig.d

IThe quality of the mammography is increased by applying image enhancement techniques, which also improve the contrast and readability of the image. By enhancing it, it aids the system in identifying mammography lesions with low contrast and visibility. Improving the image quality on low-contrast mammograms is the main objective of mammography enhancement. Small abnormalities in low-contrast regions are frequently hidden by surrounding tissues, which can result in an incorrect diagnosis. The overall quality of the photographs is improved by the image upgrades, which makes it relatively simpler for CAD systems and readers to identify these minute irregularities. The improvements could magnify sounds or introduce distortions to an image's anatomical features. As a result, only techniques that preserve a similar appearance would be permitted.

II. RESEARCH METHODOLOGY

1. Research Objectives: Primary objective: To develop a deep learning model for automated breast cancer detection from Histopathological images. Secondary objectives: To collect and preprocess a dataset of Histopathology images. (Fig. 1.1) To design and train a convolutional neural network (CNN) model for image classification. To evaluate the performance of the trained model on a separate validation dataset. To deploy the trained model as a web application for real-time prediction. Deploy the application: Deploy the Flask application on a web server accessible to users. Monitor performance: Continuously monitor the application's performance and user feedback to identify and address any issues. Update the model: Periodically retrain the breast cancer detection model with new data to improve its accuracy and adaptability to changing conditions. 2. Data Collection & Processing: Invasive Ductal Carcinoma (IDC) - most common subtype of all breast cancers . To assign an aggressiveness grade to a whole mount sample, pathologists typically focus on the region which contains the IDC. (Fig. 1.2). As a result, one of the most preprocessing steps for the automatic aggressiveness grading is to delineate the exact region of IDC inside the whole mount slide. The original data set consisted of 162 whole mounts slide images of breast cancer specimens scanned at 40x magnification for that. 277524 patches of size, 50 by 50 were extracted. The whole data set consists of 198,738 IDC negative and 78,786 IDC, positive. This data set was gathered from Kaggle and near about 1300 images were used for the prediction and for preparing the convolutional neural network.

Fig. 1.1

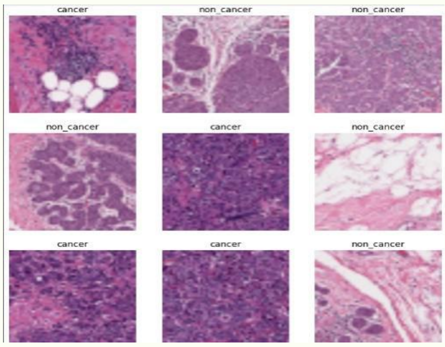


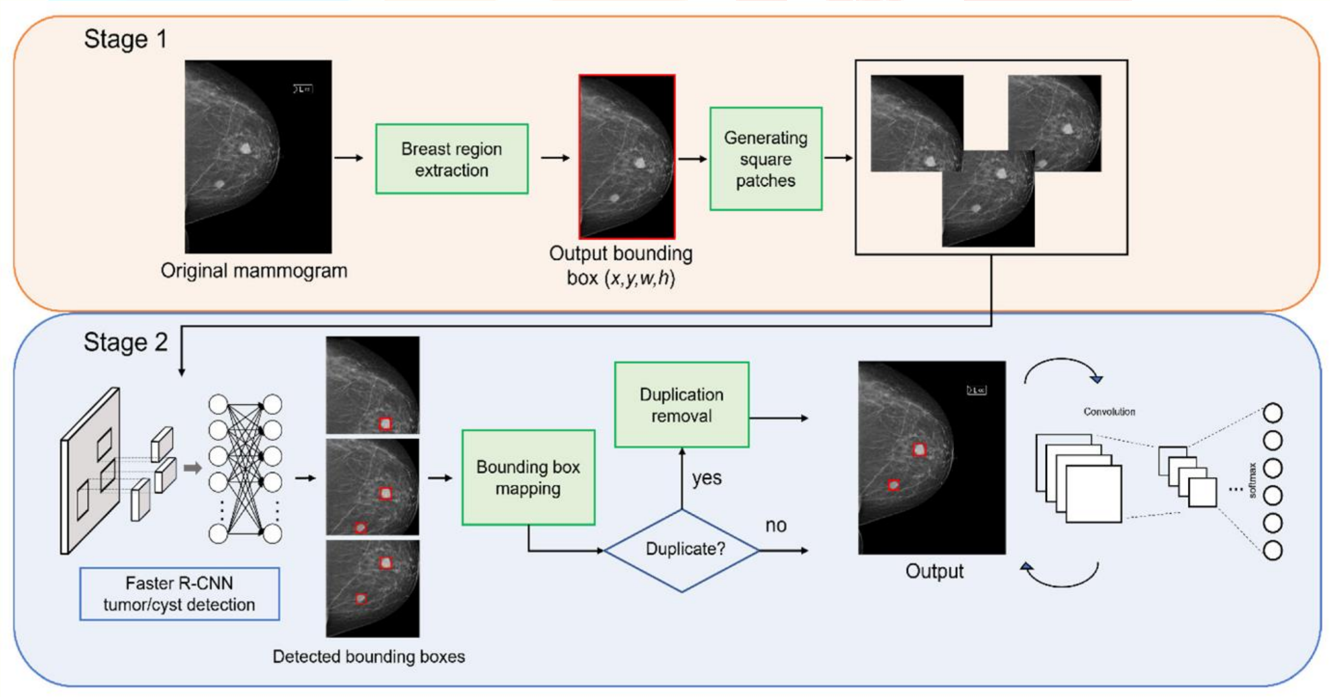
Fig. 1.2

Data Preprocessing – Image Resizing - The process of image resizing was to uniform the dimensions. Here we have done 180x180 pixels to ensure consistency and input data size for the neural network.

Normalization - The scaling pixel value was ranged to zero to one, which aids in the convergence of neural network during training.

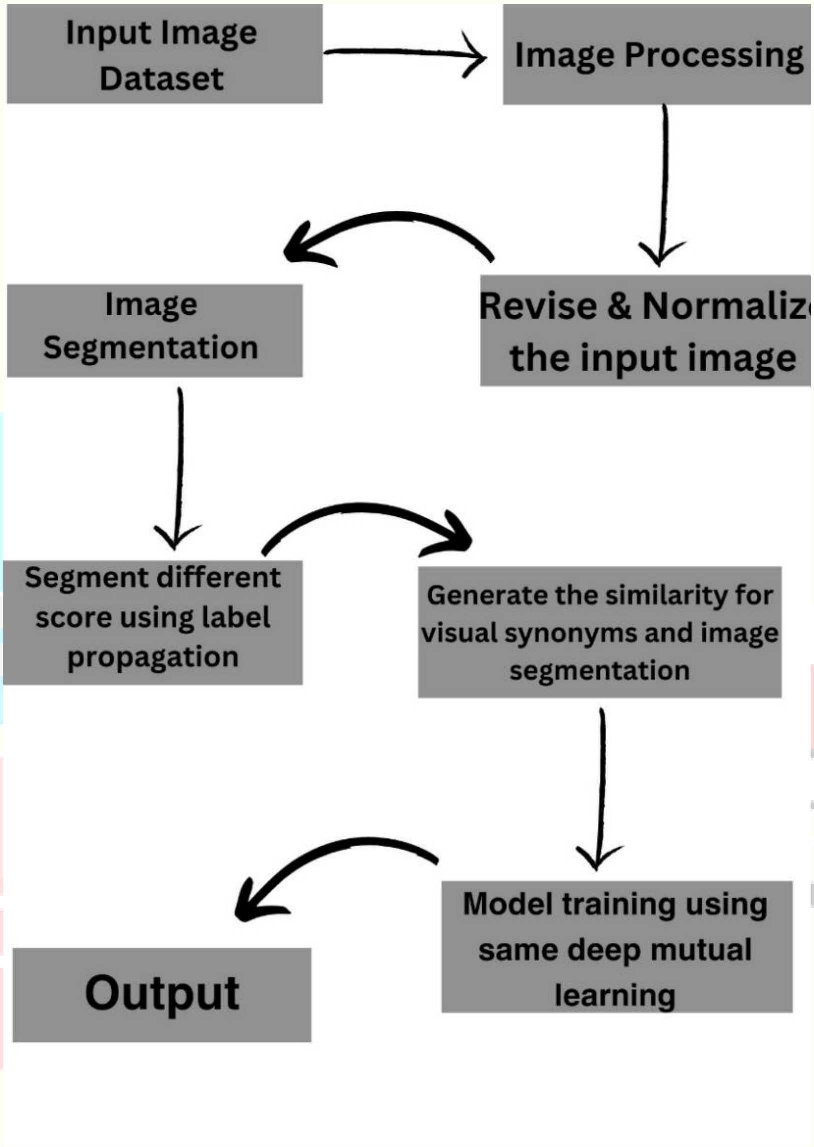
Augmentation – Unique data augmentation techniques were used to expand the training data set artificially and reduce overfitting.

2. Model Architecture:



The architecture that we have outlined employees. A convolutional neural network, which is widely recognized as effective for image recognition task, particularly in the field of medical imaging. Input layer – Rescaling- The model starts with a rescaling layer, which normalizes the pixel value of each image. The normalization process scales pixels, intensities to the range of 0 to one by dividing each pixel by 255. This is crucial because it helps in stabilizing the learning process and generally results in faster convergence. Input Shape – The input layer explicitly expects the image data of shape (180,180,3), where 180 represents the pixel dimensions of the image and 3 stands for the three-color channels, which is RGB.

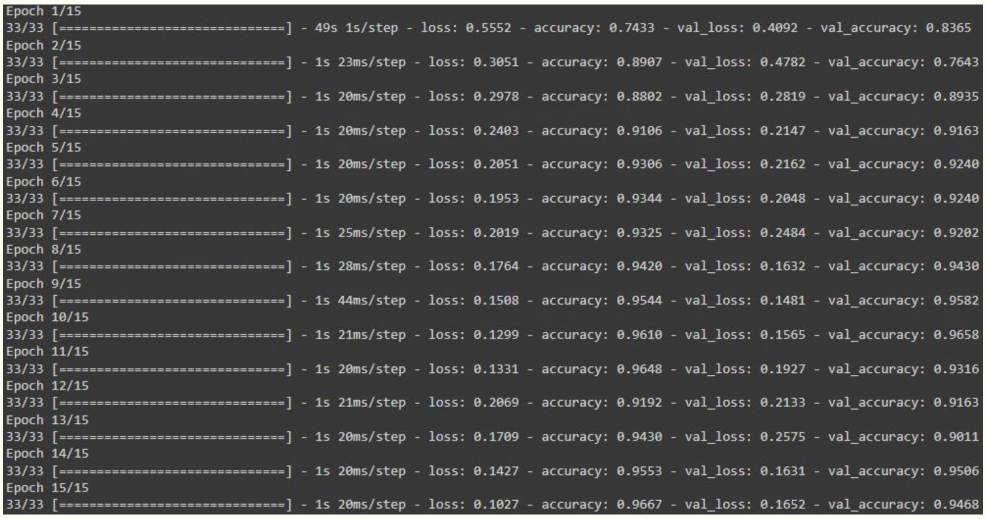
3. Convolutional Layers: First Layer- Filters: 16. Kernel Size: 3x3. Padding: “same”, which means the output of this layer will have the same width and height as the input. This is achieved by padding the input such that the convolutional kernel fits within the input. Activation Function: ReLU(Rectified Linear Unit), which introduces nonlinearity to the model, allowing it to learn more complex pattern in the data. First Max Pulling Layer: Pool size 2X2 which reduces the spatial dimension height and width of the output from the previous layer by half this. Pulling layer helps him reducing the computational load memory usage and the number of parameters, minimizing the risk of overfitting. Second and Third convolutional layer: These layers follow the similar pattern, where the number of filters increases with depth. Second Layer- Filters: 32 Third Layer- Filter: 64 Convolutional layer is followed by a maximum layer with the same specifications as the first Max cooling layer. Increasing the number of filters in deeper layer is a common practice as it allows the reunion network to learn increasingly complex features at each level. Flattening Layer: This layer retains three-dimensional shape, which is then flattened into a 1-dimensional vector. This step is very important to transition from convolutional layer, which handles the 2D data to dense layer which handles 1D data.ref[3]First Dense Layer– Units: 128. Activation Function -ReLU. This layer is densely connected, every neuron in the preceding layer connects to each neuron in this layer design. To further process features extracted from the convolutional base. Output Layer- Units are equal to the number of classes. They are (cancer, non-cancer) Activation Function- Activation function implying linear outputs usually combined with SoftMax activation in the loss function when calculating cross entropy for a classification task.


Input Image Dataset
Image Processing
Revise & Normaliz
the input image
Image Segmentation
Segment different score using label propagation
Generate the similarity for visual synonyms and image segmentation
Model training using same deep mutual learning
Output

4. Compilation: Optimizer: Adam. This is an adaptive learning rate optimizer known for its efficiency in various settings also, with classification. Loss Function: Sparse Categorical Crossentropy, which is suitable for multiclass classification tasks where each class is mutually exclusive. Metrics: Accuracy, which is used to be used to evaluate the performance of the modern during training and testing.

III. RESULTS AND DISCUSSION

The outcome and the results of the model were quite promising. We took him to consideration several performance metrics to see how our classification model has performed and what can be taken as a confident result out of this. We took accuracy precision recall and ROC curve into consideration as our performance metrics and also according to the 15 epochs that it performed and the 15 iteration that the model had, every accuracy and the confidence was mentioned in the given image below.(Fig. 3.1) So the accuracy. went from 0.7433 to 0.9667 while the 15 epoch run time. That was quite the effective technique that we used there and the accuracy was also good. You could figure out that the accuracy was also increasing, and the loss is decreasing simultaneously as the epochs were running.



```
Epoch 1/15
33/33 [==============================] - 49s 1s/step - loss: 0.5552 - accuracy: 0.7433 - val_loss: 0.4092 - val_accuracy: 0.8365
Epoch 2/15
33/33 [==============================] - 1s 23ms/step - loss: 0.3051 - accuracy: 0.8907 - val_loss: 0.4782 - val_accuracy: 0.7643
Epoch 3/15
33/33 [==============================] - 1s 20ms/step - loss: 0.2978 - accuracy: 0.8802 - val_loss: 0.2819 - val_accuracy: 0.8935
Epoch 4/15
33/33 [==============================] - 1s 20ms/step - loss: 0.2403 - accuracy: 0.9106 - val_loss: 0.2147 - val_accuracy: 0.9163
Epoch 5/15
33/33 [==============================] - 1s 20ms/step - loss: 0.2051 - accuracy: 0.9306 - val_loss: 0.2162 - val_accuracy: 0.9240
Epoch 6/15
33/33 [==============================] - 1s 20ms/step - loss: 0.1953 - accuracy: 0.9344 - val_loss: 0.2048 - val_accuracy: 0.9240
Epoch 7/15
33/33 [==============================] - 1s 25ms/step - loss: 0.2019 - accuracy: 0.9325 - val_loss: 0.2484 - val_accuracy: 0.9202
Epoch 8/15
33/33 [==============================] - 1s 28ms/step - loss: 0.1764 - accuracy: 0.9420 - val_loss: 0.1632 - val_accuracy: 0.9430
Epoch 9/15
33/33 [==============================] - 1s 44ms/step - loss: 0.1508 - accuracy: 0.9544 - val_loss: 0.1481 - val_accuracy: 0.9582
Epoch 10/15
33/33 [==============================] - 1s 21ms/step - loss: 0.1299 - accuracy: 0.9610 - val_loss: 0.1565 - val_accuracy: 0.9658
Epoch 11/15
33/33 [==============================] - 1s 20ms/step - loss: 0.1331 - accuracy: 0.9648 - val_loss: 0.1927 - val_accuracy: 0.9316
Epoch 12/15
33/33 [==============================] - 1s 21ms/step - loss: 0.2069 - accuracy: 0.9192 - val_loss: 0.2133 - val_accuracy: 0.9163
Epoch 13/15
33/33 [==============================] - 1s 20ms/step - loss: 0.1709 - accuracy: 0.9430 - val_loss: 0.2575 - val_accuracy: 0.9011
Epoch 14/15
33/33 [==============================] - 1s 20ms/step - loss: 0.1427 - accuracy: 0.9553 - val_loss: 0.1631 - val_accuracy: 0.9506
Epoch 15/15
33/33 [==============================] - 1s 20ms/step - loss: 0.1027 - accuracy: 0.9667 - val_loss: 0.1652 - val_accuracy: 0.9468
```

Fig.3.1

Take reference of ROC Curve in the below image. (Fig 3.2) We also took manual evaluation metrics into consideration so that we could see that in real life situation when the images are provided for classification, how well does our model perform. Not only by looking at the graph, but we were able to see how accurate our model is. This was done to check the dependency of the model and for the manual consideration, so that we as humans can take machines into consideration when taking medical decisions. For this, we performed an experiment where we took 50 specimens of each where 50 images were provided to the model for both cancerous and non-cancerous specimens. Here we were able to predict how accurate our model is through calculating the recall, the precision and the F1 value. Using these values, we can calculate the accuracy, precision, recall, and F1 score of the model as follows:

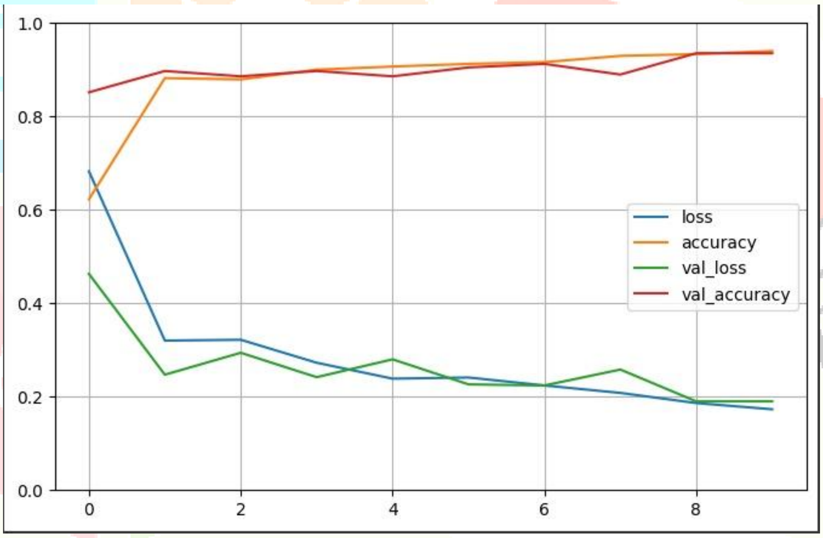


Fig 3.2

	Predicted Positive	**Predicted Negative**
Actual Positive	48	2
Actual Negative	5	45

Accuracy: (TP + TN) / (TP + FP + TN + FN) = (48 + 45) / (40 + 2 + 45 + 5) = 0.93 or 93%

Precision: TP / (TP + FP) = 48 / (48 + 2) = 0.96 or 96%

Recall: TP / (TP + FN) = 48 / (48 + 5) = 0.905 or 90.5%

F1 score: 2 * (precision * recall) / (precision + recall) = 2 * (0.96 * 0.905) / (0.96 + 0.905) = 0.93 or 93%

Conclusion:

Deep Learning in Medical Imaging: Several studies have shown how well CNNs perform when analyzing medical images, especially when it comes to tasks like tumor identification and categorization. CNNs have outperformed more conventional machine learning methods in terms of performance because they can automatically extract pertinent characteristics from unprocessed picture data. Data Quality and Availability: Having access to sizable datasets is essential for successfully training CNN models. To build and assess breast cancer detection algorithms, numerous researchers have made use of publicly accessible medical imaging datasets, including the Digital Database for Screening Mammography (DDSM) and the Curated Breast Imaging Subset of DDSM (CBIS-DDSM). Robust model training and generalization require diverse and high-quality data.Techniques for Preprocessing Data: Preprocessing data is essential to optimizing input data for neural network models. Preprocessing methods that are frequently used are shrinking photos to a uniform resolution, normalizing pixels, and augmenting datasets to improve diversity and reduce overfitting.

The model that we created from the hysto pathological data. was certainly well structured and well defined. We could see over the time while the epochs were running, our accuracy and the confidence number increases and our loss and validation loss also decreases. through this we can conclude that our model was successfully able to classify between the cancerous and noncancerous specimens that were provided to it through the images. We also took the manual prediction into consideration and. made a manual matrix for evaluation. We took samples for both cancerous and non cancerous and provided it to the model and calculated the recall and the precision, so that we can know how precise our model is. Through the above experiment, we were able to see that cancerous class precision was 0.9796 and the recall was also 0.9796 . But for the non-cancerous class, the precision was 0.96 & the recall was 1.0.

Acknowledgment

I wish to extend my deepest gratitude to our Professor Mr.Nitish Kumar contributed to the completion of our project.

References

[1]. Siegel, R.L.; Miller, K.D.; Fedewa, S.A.; Ahnen, D.J.; Meester, R.G.; Barzi, A.; Jemal, A. Colorectal cancer statistics, 2017. CA Cancer, J. Clin. 2017, 67, 177–193. [CrossRef] [PubMed]

[2]. Spanhol, F.A.; Oliveira, L.S.; Cavalin, P.R.; Petitjean, C.; Heutte, L. Deep features for breast cancer histopathological image classification. In Proceedings of the 2017 IEEE International Conference on Systems, Man and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 1868–1873.

[3]. Desai, M.; Shah, M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network, (MLP) and Convolutional neural network (CNN). Clin. eHealth 2021, 4, 1–11. [CrossRef]

[4]. Alanazi, S.A.; Kamruzzaman, M.M.; Sarker, N.I.; Alruwaili, M.; Alhwaiti, Y.; Alshammari, N.; Siddiqi, M.H. Boosting Breast Cancer Detection Using Convolutional Neural Network. J. Health Eng. 2021, 2021, 5528622. [CrossRef] [PubMed]

[5]. Lakhani, S.R.; Ellis, I.O.; Schnitt, S.; Tan, P.; van de Vijver, M. WHO Classification of Tumors of the Breast, 4th ed.; WHO Press: Geneva, Switzerland, 2012; Volume 4.