The Journal of Multidisciplinary Research (TJMDR)

Content Available at www.saap.org.in ISSN: 2583-0317

Breast cancer detection using deep learning and cnn-based model

Krishna Banavathu[1], Alikani Vijaya Durga[2], M Ramakrishna[3]

[1] Dept. of ECE, University College of Engineering, Adikavi Nannaya University, Rajamahendravaram, AP.

[2] Dept. of ECE, University College of Engineering, Adikavi Nannaya University, Rajamahendravaram, AP.

[3]Asst. Professor, Department of CSE, Adikavi Nannaya University Rajahmundry

Received: 15 July2022 Revised: 04 Aug 2022 Accepted: 30 Aug 2022

Abstract

The second-most dangerous cancer in the world is breast cancer. Not just in India, but all around the world, breast cancer is the primary cause of death for women. According to the USA in 2011, out of eight one woman had cancer. Inappropriate breast cell division can result in benign or malignant breast cancer. Consequently, this is how breast cancer progresses. Therefore, it is crucial to detect the breast cancer at the early stage. By doing this, many lives can be saved and the sickness can be adequately treated while also being treated as a very serious condition. Breast cancer is most dangerous disease and at present it treated as global disease. Invasive breast cancer will likely affect 246,660 women in the USA in 2016, and 40,450 women will likely pass away from the disease. Mammography continues to be labor-intensive and has acknowledged drawbacks despite its success as a tool for detecting breast cancer, including low sensitivity in women with dense breast tissue. The development of neural networks has been used to breast histopathology images during the past ten years to help radiologists operate more accurately and efficiently. The goal of this study is to use the most recent convolution neural network (CNN) expertise to images of breast histopathology. The first section of the research examines conventional Computer Assisted Detection (CAD) utilising machine learning and a more current CNN-based model for Breast Histopathology Images.

Keywords - Convolution Neural Networks (CNNs), Histopathology Images, Computer Assisted Detection (CAD), machine learning, CNN-based model.

***Corresponding Author**

Krishna Banavathu

Produced and Published by

South Asian Academic Publications

Introduction:

I.INTRODUCTION

Breast cancer (BC) is one of the most common malignancies in women and is responsible for the majority of new cancer cases and cancer-related deaths, according to statistics from throughout the world. The deaths, making it a serious public health issue in today's society. Because it can encourage prompt clinical care for patients, an early diagnosis of BC can considerably enhance the prognosis and likelihood of survival. A more precise classification of benign tumours could spare people from receiving unneeded medical care. As a result, there is a lot of research into the proper diagnosis of patients and the classification of those individuals into benign or malignant groupings. Machine learning (ml) is widely acknowledged as the preferred methodology in BC pattern classification and forecast modelling due to its distinct benefits in essential features discovery from complex BC datasets. Data can be properly categorised using methods like data mining and classification. Particularly in the medical industry, where those techniques are frequently applied to arrive at findings through diagnosis and analysis.

The investigation in this paper tries to identify the features that are most useful in predicting either malignant or benign cancer and to identify broad trends that may help us in model and hyper parameter selection. Finally, to do this, we fitted a function that can predict the discrete class of new input using Deep Learning classification algorithms.

II. LITERATURE WORK

The earlier breast cancer is discovered via a variety of techniques, the better the patient's chances of receiving treatment. In order to combat breast cancer, numerous early detection or prediction approaches are being researched and employed. The purpose of this study was to develop non-invasive, painless techniques for predicting and detecting Breast cancer early. The frequency bandwidth, substrate dielectric constant, electric field, and tumour data from measurements of an antenna were used to test and compare all of the data mining classification methods in Weka for the diagnosis and prediction of breast cancer. The results shows demonstrate simple cart algorithms were the most effective algorithms, which gives accuracy over 90% in identification. This comparative study of various classification methods for the diagnosis of breast cancer provided insight into data mining using data from antenna measurements and a 10-fold cross-validation procedure. By using data mining classification algorithms like bagging, ibk, random committee, random forest, and simple cart, it is possible to identify breast cancer tumours non-invasively, inexpensively, and without subjecting patients to harmful radiation, according to the high accuracy rates of these algorithms.

III.TECHNIQUES FOR MACHINE LEARNING:

The creation of a reliable and computationally effective classifier for medical applications is a significant problem in the data mining and machine learning focal areas of this study. On the Wisconsin breast cancer (original) datasets, we attempted to compare the accuracy, precision, sensitivity, and specificity of support vector classifier, random forest, gradient boosting, naive bayes, cart model, neural network, and linear regression algorithm in order to find the best classification accuracy. The support vector outperforms other techniques with an accuracy of 98.23%. Support vector machine, random forest, and naive bayes classifiers are three proposed supervised machine learning methods to categorise breast cancer.

1) Evaluation of machine learning models:

In contrast, With the advancement of data mining techniques, it is now possible to extract more useful information from complicated databases, and to predict, classify, and cluster data based on the information that has been recovered. In recent years, researchers' focus has shifted more toward breast cancer research and prevention. In this study, two distinct datasets related to breast cancer—the Wisconsin Breast Cancer Database and the Breast Cancer Coimbra Dataset—are classified using five different classification models, including the Decision Tree (DT), Random Forest (rf), Support Vector Machine (SVM), Neural Network (NN), and Logistics Regression (LR). The goal of this study is to investigate the relationship between breast cancer and some attributes in order to reduce the death probability of breast cancer (WBCAD). These five classification models are assessed out of three different indicators, such as prediction accuracy values, f-measure metric, and auc values. The random forest model can perform and better adapt than the other four techniques, according to comparative experiment analysis. The outcome of the prediction will aid in lowering the rate of incorrect diagnoses and aid in developing appropriate therapy initiatives. Give professionals a reference to help them understand the nature of breast cancer. There are still several problems with this study's limitations that should be resolved in recent studies. For instance, even though there are more indices that exist but have not yet been discovered, this study only gathers data for 10 attributes in this experiment. The accuracy of the results is affected by the sparsely populated raw data.

IV.PROBLEM DESCRIPTION:

The main objective of the proposal is to predict Breast Cancer using a Deep Learning Conventional Neural Networks model on Breast Histopathology Images (Breast Cancer (BC) specimens scanned) dataset to identify whether it will be Benign or Malignant. This image dataset is balanced by a collection of 78,786 IDC positive aspects and 198,738 IDC negative aspects. The most frequent form of breast cancer is invasive ductal carcinoma (IDC).

This dataset contains 2 directories each that represents a class, where 0 means NONIDC and 1 indicates IDC.

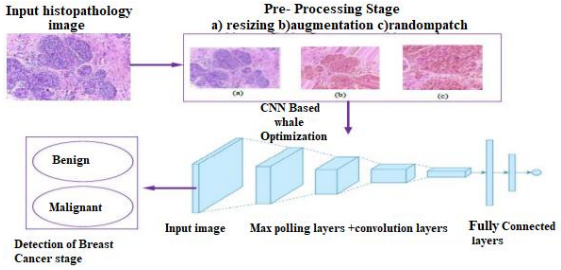


Fig 1: Architecture.

The CNN deep learning method is used to determine the type of malignancy. With the highest level of accuracy, it uses photos as input to forecast cancer and outputs whether it is benign or malignant cancer.

After the model is constructed during the training phase, the classifier uses the test set of data that predict the presence of breast cancer. Eight input attributes are used to make predictions, and predictions are created for any additional attributes.

1) Breast Histopathology Images:

Invasive ductal Carcinoma (IDC), the most prevalent type of breast cancer, is widespread in women. An essential clinical duty is accurately identifying and classifying breast cancer subtypes, and automated techniques can be utilised to speed up the process and minimise mistake. The most prevalent subtype of breast cancer is invasive ductal carcinoma (IDC). Pathologists often concentrate on the areas that contain the IDC when grading the aggressiveness of a whole mount sample. This collection includes 277,524 pictures. 198,738 IDC negative and 78,786 IDC positive patches of size 50 x 50 were extracted, where 0 is NON-IDC and 1 is IDC.

IDC images

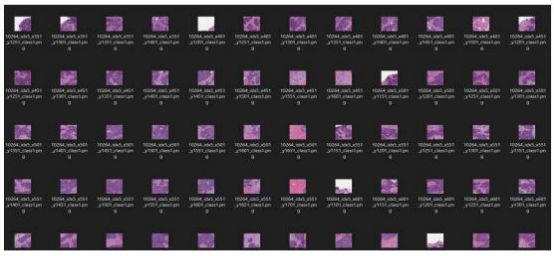


Fig2: IDC Images

NON – IDC images

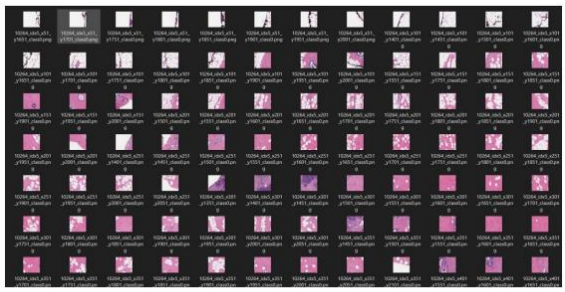


Fig 3: NON-IDC Images

2) Data Pre-Processing:

Data pre-processing and data mining methods are applied to transform the raw data into a format, which is both practical and efficient. Before using machine learning techniques, one step is taken. It transforms the original data to a format that such a specific algorithm can utilise. Data pre-processing activities include things like feature selection.

3) Data Cleaning:

The objective of data cleaning is to create data sets that are streamlined and uniform such that business intelligence and data analytics tools could really easily access and find the right data for each query. Data cleaning is the process of eliminating or changing data that is inaccurate, lacking, unnecessary, duplicated, or formatted incorrectly in order to prepare it for analysis. Data cleaning is a critical step to make sure the results you generate are reliable, regardless of the sort of research or data visualisations you want.

4) Data Transformation:

Transformation is the process of converting data from one format to another, typically the format of such a source system into the necessary form of a destination system.

5) Data Visualization:

It is possible to distinguish rapidly between red and blue, square from circular, as well as other visual aspects in data visualisation, which refers to the graphical representation of information and data using graphs and charts. This piques our interest and keeps us focused on the message. We can immediately see trends and outliers when we examine a chart.

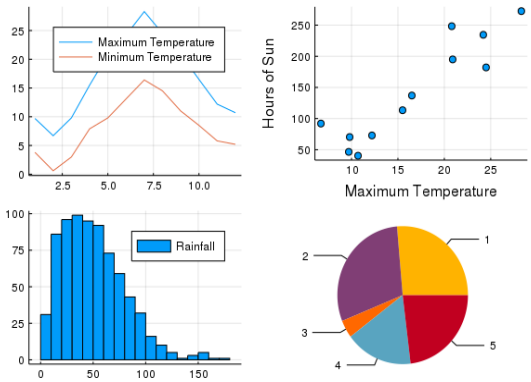


Fig. 4 Data Visualization

SEPARATION OF DATA INTO TRAINING AND TEST SETS:

If you try to analyse your system using training data, you must split the data into training and test sets. This is a crucial step in assessing data mining models. The majority of the data is often used for training, while a smaller portion of the data is utilised for testing when you divide a data set into a training set and testing set.

To help ensure that the training and testing sets are similar, analysis services randomly sample the data. You can reduce the effects of data discrepancies and gain a better understanding of the model's properties by using similar data for training and testing.

Train set: The training set's observations serve as the experience for the algorithm's learning. Each observation in supervised learning problems consists of one or more observed input variables and an observed outcome variable.

Test set: A test set is a collection of observations used to assess the model's effectiveness using certain performance criteria. The test set cannot contain any observations from the training set. It will be challenging to determine if the algorithm has learned to generalise from the training set or has merely memorised it if the test set does contain examples from the training set.

CONVOLUTIONAL NEURAL NETWORKS:

On CNN, we don't employ fully connected (fc) levels until the very last layers of the network. In typical feed-forward neural networks, each neuron in the input layer is connected to every output neuron in the next layer. Thus, a CNN is a neural network that substitutes a specific "convolutional" layer for a "fully-connected" layer for at least one of the network's layers.

The output of these convolutions are then subjected to a nonlinear activation function, such as RELu, and the convolution, activation, and mixture of other layer types process continues in order to help reduce the width and height of the input volume and help reduce over fitting until we finally reach the end of the network and apply one or two fc layers in order to obtain our final output classifications. Every layer in a CNN applies a unique set of filters—typically hundreds or thousands of them—and then mixes the output before passing it on to the following layer.

In the context of image categorization, the CNN may automatically learn the values for these filters during training.

- Convolutional layer
- Pooling layer
- Fully-connected layer
- Dropout layer

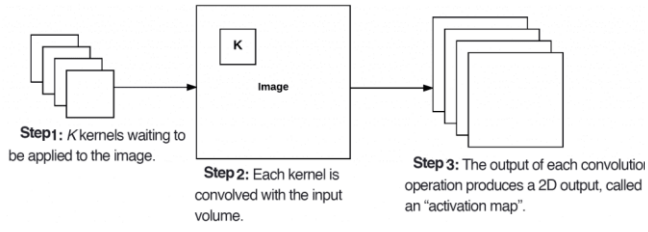


Fig 5. Convolution Operation

Now that all k filters have been applied to the input, we get 2-dimensional activation k maps. The final output volume is subsequently created by stacking our k activation maps along the depth dimension of our array..

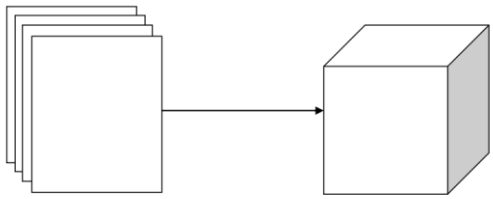


Fig 6 : After obtaining the k activation maps, they are stacked together to form the input volume to the next layer in the network.

An essential element of the layered design is a filter or kernel. The matrix, which has real-valued entries, is lower in size than the image's dimensions. The input volume and kernels are then convolved to produce so-called activation maps. Activation maps show active areas, or areas where the kernel-specific input properties have been identified.

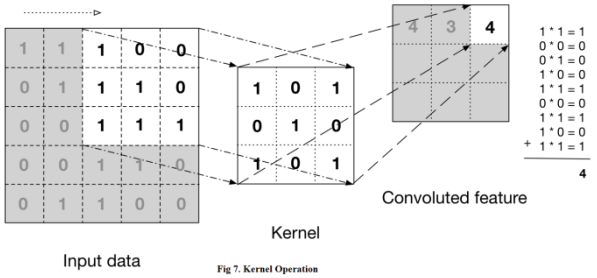


Fig 7. Kernel Operation

Kernel for various functions such as picture identification, edge recognition, and image sharpening.

Stride:

When doing a convolution, we described it as "sliding" a small matrix across a larger matrix, stopping at each coordinate to perform an element-wise multiplication and addition before storing the results. This description reminds me of a sliding window that moves across an image from left to right and top to bottom.

The main purpose of the pool layer is to gradually lower the spatial size of the input volume (i.e., its width and height). By doing so, we may reduce the number of parameters and computations in the network, and pooling also aids in controlling over fitting.

Pool layers separately act on each of the input's depth slices using either the max or average function. While max pooling is often done in the middle of the CNN architecture to reduce the spatial size, average pooling is frequently employed as the network's last layer when we wish to totally avoid employing fc layers. We normally select a 2x2 pool size, while deeper CNNs that use larger input images (>200 pixels) may use a 3x3 pool size early in the network formation. We typically set the stride to either s-1 or s-2 as well.

Applying the pool operation yields an output volume of size woutput x houtput x doutput, where:

Woutput = ((winput-f) =s) + 1

Houtput = ((hinput-f) =s) +1

Doutput = dinput

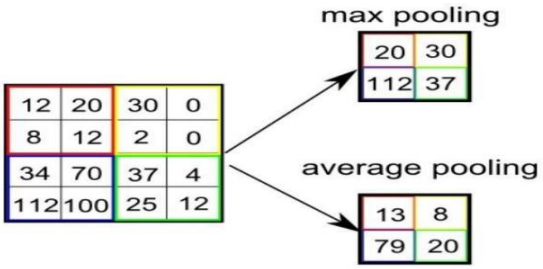


Fig 8: Pool Types

RESULTS:

Importing Data Images:

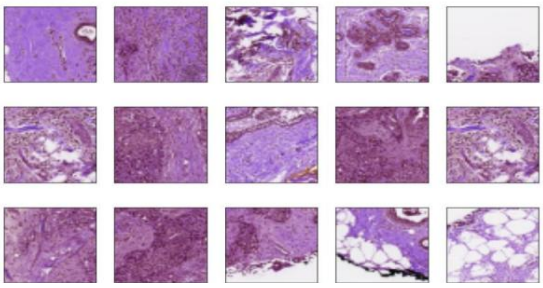
```
data = './10264'
No_breast_cancer = './10264/No'
Yes_breast_cancer = './10264/Yes'
```

```
dirlist=[No_breast_cancer, Yes_breast_cancer]
classes=['No', 'Yes']
filepaths=[]
labels=[]
for i,j in zip(dirlist, classes):
    filelist=os.listdir(i)
    for f in filelist:
        filepath=os.path.join (i,f)
        filepaths.append(filepath)
        labels.append(j)
print ('filepaths: ', len(filepaths), '   labels: ', len(labels))
```

```
filepaths:  1204    labels:  1204
```

Plotting Data Images:

```
#visualize breast tumor images

plt.figure(figsize=(12,8))
for i in range(15):
    random = np.random.randint(1,len(df))
    plt.subplot(3,5,i+1)
    plt.imshow(cv2.imread(df.loc[random,"filepaths"]))
    plt.title(df.loc[random, "labels"], size = 15, color = "white")
    plt.xticks([])
    plt.yticks([])

plt.show()
```



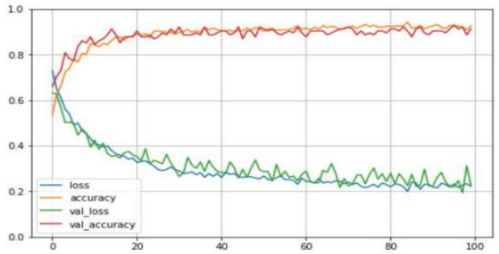
Splitting Training and Testing Data:

```
train_gen = train_datagen.flow_from_dataframe(dataframe = train_new,
                                              x_col = 'filepaths', y_col ='labels',
                                              target_size = (224,224), batch_size = 32,
                                              class_mode = 'binary', shuffle = True)
val_gen = train_datagen.flow_from_dataframe(valid,
                                            target_size=(224,224), x_col = 'filepaths', y_col ='labels',
                                            class_mode='binary',
                                            batch_size= 16, shuffle=True)
test_gen = test_datagen.flow_from_dataframe(test,
                                            target_size = (224,224), x_col = 'filepaths', y_col ='labels',
                                             class_mode = 'binary',
                                            batch_size = 16, shuffle = False)
```

```
Found 1028 validated image filenames belonging to 2 classes.
Found 115 validated image filenames belonging to 2 classes.
Found 61 validated image filenames belonging to 2 classes.
```

Data Accuracy and Data Loss In Graph View:

```
pd.DataFrame(history.history).plot(figsize=(8, 5))
plt.grid(True)
plt.gca().set_ylim(0, 1)
plt.show()
```

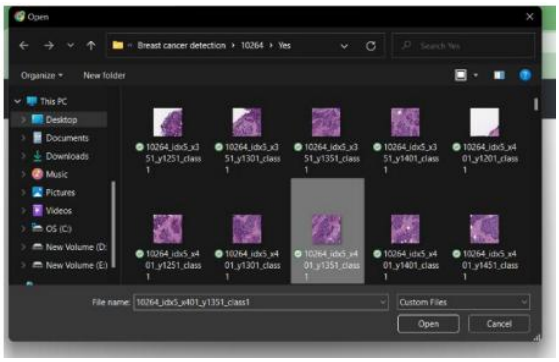


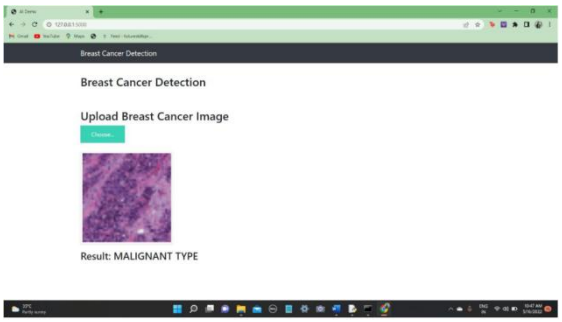
Individual Prediction

```
from PIL import Image
model_path = "model.h5"
loaded_model = tf.keras.models.load_model(model_path)

# import matplotlib.pyplot as plt
import numpy as np

image = cv2.imread("./10264/No/10264_idx5_x1001_y551_class0.png")

image_fromarray = Image.fromarray(image, 'RGB')
resize_image = image_fromarray.resize((224, 224))
expand_input = np.expand_dims(resize_image,axis=0)
input_data = np.array(expand_input)
input_data = input_data/255

pred = loaded_model.predict(input_data)
if pred >= 0.5:
  print("MALIGNANT")
else:
  print("BENIGN")
```

: BENIGN

Choose an image to predict the cancer:





CONCLUSION:

In this Paper, we investigate the use of deep learning classification algorithms to categorise breast cancer. To forecast cancer pictures, we suggest CNN algorithms. Breast Histopathology Images make up the data collection in this case. After loading the dataset, exploratory data analysis is carried out. To forecast breast cancer images and determine accuracy, we create CNN models. The algorithm that predicts the data with the highest degree of accuracy is picked.

REFERENCES

[1] Yi-Sheng Sun, Zhao, Han-Ping-Zhu," Risk factors and Preventions of Breast Cancer" International Journal of Biological Sciences.

[2] AlirezaOsarech, BitaShadgar," A Computer-Aided Diagnosis System for Breast Cancer", International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011

[3] MandeepRana, PoojaChandorkar, AlishibaDsouza, "Breast cancer diagnosis and
recurrence prediction using machine learning techniques", International Journal of
Research in Engineering and Technology Volume 04, Issue 04, April 2015.

[4] VikasChaurasia, BB Tiwari and Saurabh Pal – "Prediction of benign and malignant
breast cancer using data mining techniques", Journal of Algorithms and Computational

Technology
[5] Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction Using Data Mining
Method, IEEE Conference paper
[6] D.Dubey, S.Kharya, S.Soni and –"Predictive Machine Learning techniques for Breast
Cancer Detection", International Journal of Computer Science and Information
Technologies, Vol.4(6),2013,1023-1028.
[7] Nidhi Mishra, NareshKhuriwal.- "Breast cancer diagnosis using adaptive voting
ensemble machine learning algorithm", 2018 IEEMA Engineer Infinite Conference
(eTechNxT), 2018
[8] Chao-Ying, Joanne, PengKukLida Lee, Gary M. Ingersoll –"An Introduction to
Logistic Regression Analysis and Reporting ", September/October 2002 [Vol. 96(No. 1)