BREAST CANCER DETECTION USING HISTOPATHOLOGY IMAGE

A PROJECT REPORT

Submitted by:

RAGHUL RAVIKUMAR (212000173)

SUKESH PERLA (21203957)

Under the supervision of

Dr. Sarp Akcay

School of Mathematics and Statistics

UNIVERSITY COLLEGE DUBLIN

Belfield, Dublin 4, Ireland - D04 V1W8

JULY 2022

ABSTRACT

Keywords - Breast Cancer, Convolution Neural Network, Histopathology, Image Processing, Detection, Accuracy.

Breast Cancer is one of the most common cancers among women. It develops when abnormal cells in the breast divide and multiply. But experts don't know exactly what causes this process to begin in the first place. According to the American Cancer Society, when breast cancer is detected early, and is in the localized stage, the 5-year relative survival rate is 99%. Primarily, breast images emanating from mammograms, X-Rays, or MRIs are analyzed by radiologists to detect abnormalities. But most of the time, a radiologist can't say for sure whether it's cancer or it's not cancer, leading to high false positives and high false negatives. Sometimes multiple screening is required to confirm the screening result.

Recent advancements in image processing and deep learning create some hopes of devising more enhanced applications that can be used for the early detection of breast cancer. This proposed application presents an empirical analysis of the performance of popular convolutional neural networks (CNNs) for identifying whether the specimen is Malignant or Benign to breast cancer. Firstly, A model is trained and validated with 7109 histopathology images, later the trained model is tested with 791 images. The proposed deep learning model has achieved a performance of 91.5% accuracy in predicting whether the specimen is subjected to breast cancer or not. An interactive UI of this application is also created and deployed on the web, which is free to use by any user.

Link to UI - https://nascent-bot-cancer-app-myappwelcome-i66wjm.streamlitapp.com/

Contents

List of Tables

List of Figures

List Of Abbreviations

MRI : Magnetic Resonance Imaging

CNN : Convolutional Neural Networks

UI : User Interface

AI : Artificial Intelligence

ROI : Region Of Interest

ANN : Artificial Neural Networks

SNN : Simulated Neural Networks

FC : Fully Connected

RGB : Red, Green, Blue

ML : Machine Learning

DMR-IR : Database for Mastology Research with Infrared Image

Conv2D : 2D Convolution Layer

iOS : iPhone Operating System

Chapter 1

INTRODUCTION

1.1 Breast Cancer

Breast cancer originates in the breast tissue. It occurs when breast cells mutate (change) and grow out of control, creating a mass of tissue (tumor). Like other cancers, breast cancer can invade and grow into the tissue surrounding the breast [1]. It can also travel to other parts of your body and form new tumours. When this happens, it's called metastasis.

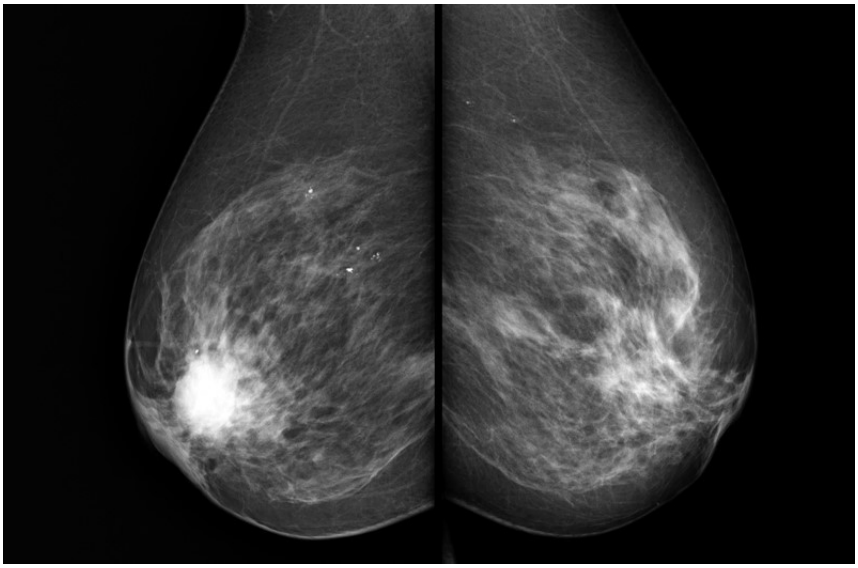

Figure 1.1: Breast Cancer Tumor - Reprinted from "ADCs go head to head in breast cancer" by Karen O'Leary, 2022.

Breast cancer is one of the most common cancers among women, second only to skin cancer. It's most likely to affect women over the age of 50. Though rare, men can also develop breast cancer. Approximately 2,600 men develop male breast cancer every year

in the United States, making up less than 1% of all cases. Transgender women are more likely to develop breast cancer compared to cisgender men. Additionally, transgender men are less likely to develop breast cancer compared to cisgender women.

Research indicates that are several risk factors that may increase the chances of developing breast cancer, it includes.

- Age. Being 55 or older increases the risk of breast cancer
- Sex. Women are much more likely to develop breast cancer than men.
- Smoking. Tobacco use has been linked to many different types of cancer, including breast cancer.
- Obesity. Having obesity can increase the risk of breast cancer.
- Family history and genetics. About 5% to 10% of breast cancers are due to single abnormal genes that are passed down from parents to children.

There are several breast cancer treatment options, including surgery, chemotherapy, radiation therapy, hormone therapy, immunotherapy, and targeted drug therapy. The treatment is based on various factors, including the location and size of the tumor, the results of the lab tests, and whether cancer has spread to other parts of your body. Healthcare providers will tailor the treatment plan according to one's unique needs. It's not uncommon to receive a combination of different treatments, too.

The overall five-year survival rate for breast cancer is 90%. This means that 90% of people diagnosed with the disease are still alive five years later. The five-year survival rate for breast cancer that has spread to nearby areas is 86%, while the five-year survival rate for metastatic breast cancer is 28%. Fortunately, the survival rates for breast cancer are improving as we learn more about the disease and develop new and better approaches to management.

1.2 Neural Network

Neural networks reflect the behavior of the human brain, allowing computer programs to recognize patterns and solve common problems in the fields of AI, machine learning, and deep learning [2].

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

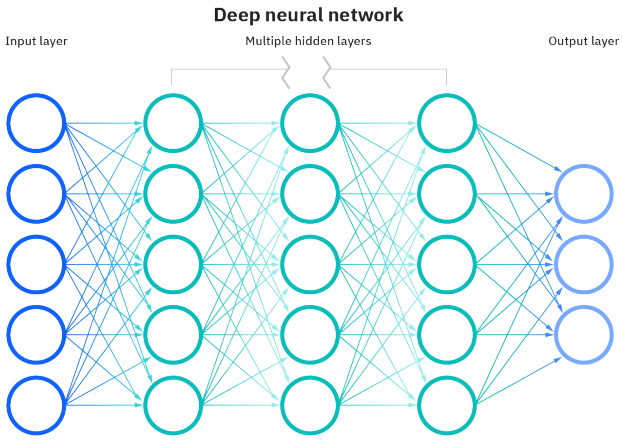

Figure 1.2: Deep Neural Network - Reprinted from “Neural Networks” by IBM Cloud Education, 2020.

1.2.1 Convolution Neural Network

Convolutional neural networks are distinguished from other neural networks by their superior performance with image, speech, or audio signal inputs [9]. They have three main types of layers, which are:

- Convolutional layer : The core building block of a CNN, and it is where the majority of computation occurs
- Pooling layer : Responsible for downsampling, conducts dimensionality reduction, reducing the number of parameters in the input.
- Fully-connected (FC) layer : Task of classification based on the features extracted through the previous layers and their different filters.

1.3 Image Classification

Image classification is where a computer can analyze an image and identify the 'class' the image falls under [3]. (Or a probability of the image being part of a 'class'.) A class is essentially a label, for instance, 'car', 'animal', 'building', and so on. For example, you input an image of a sheep. Image classification is the process of the computer analyzing the image and telling you it's a sheep. (Or the probability that it's a sheep.)

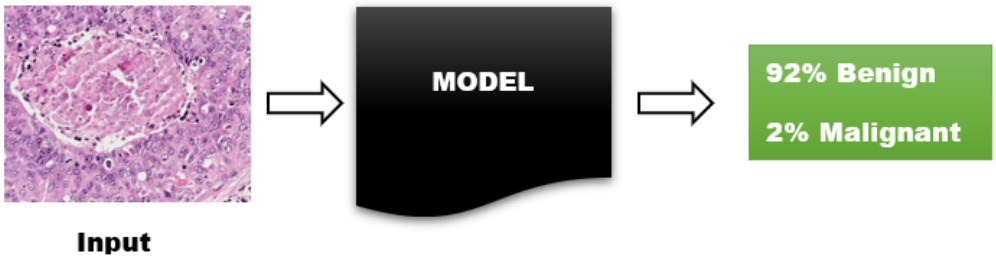

Figure 1.3: Image Classification

It is a supervised learning problem: define a set of target classes (objects to identify in images), and train a model to recognize them using labeled example photos. Early computer vision models relied on raw pixel data as the input to the model. The position of the object, background behind the object, ambient lighting, camera angle, and camera focus all can produce fluctuation in raw pixel data; these differences are significant enough that they cannot be corrected for by taking weighted averages of pixel RGB values.

1.4 Deep Learning In Healthcare

Deep learning in healthcare is a branch of AI that offers transformative potential and introduces an even richer layer to medical technology solutions [4]. Deep learning provides the healthcare industry with the ability to analyze data at exceptional speeds without compromising on accuracy. Deep learning in healthcare has already left its mark. Google has spent a significant amount of time examining how deep learning models can be used to make predictions around hospitalized patients, supporting clinicians in managing patient data and outcomes.

The future of healthcare has never been more exciting. Not only do AI and ML present

an opportunity to develop solutions that cater to very specific needs within the industry, but deep learning in healthcare can become incredibly powerful for supporting clinicians and transforming patient care.

1.5 Objective

Since 1 in 8 women who live to be age 70 will develop breast cancer in her lifetime, Breast cancer is the second most common cancer in women. The objective of this project is to use the image classification technique with the help of deep learning's convolution neural network to detect whether the specimen is malignant or benign to breast cancer. This helps in detecting breast cancer quicker which may help in increasing the survival rate by 93 or more percent.

1.6 Project Flow

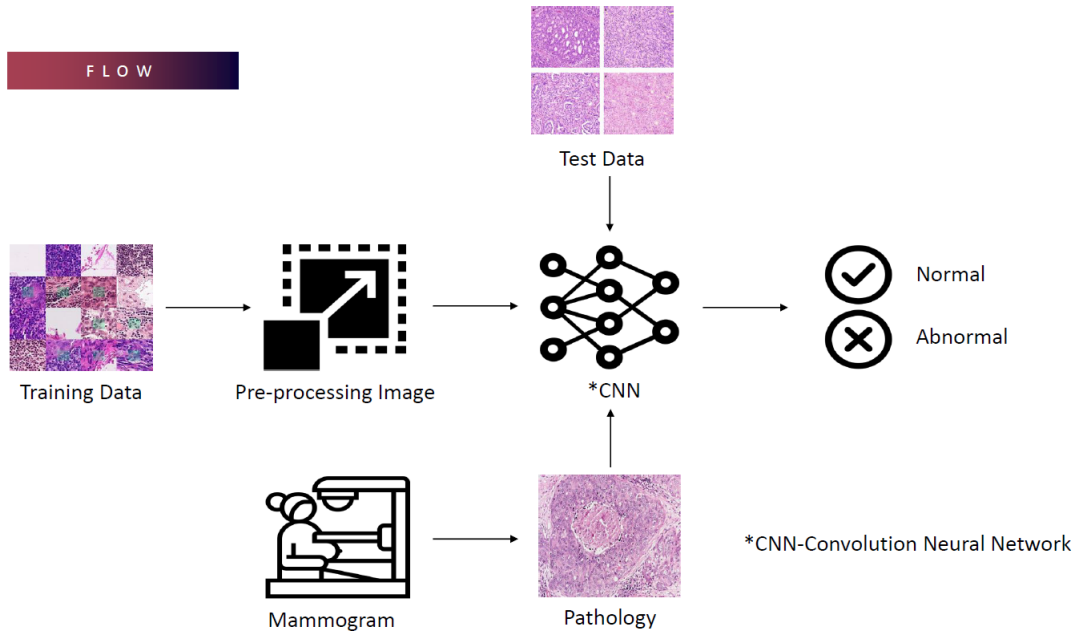

Figure 1.4: Structure Of The Project

Chapter 2

PROPOSED MODEL

2.1 Related Work

The majority of efforts related to the diagnosis of breast cancer from thermograms use web-available DMR-IR data [5]. Several authors have focused on decreasing as much non-relevant information as possible and extracting ROI.

- Mahmoudzadeh et al. [6] used Extended Hidden Markov Models (EHMM), BayesNet and Random Forest for the optimization of breast segmentation techniques.
- S.A. Mojarad; S.S. Dlay; W.L. Woo; G.V. Sherbet [7], used cross-validation approach and multi-layer perceptron neural networks to detect breast cancer

The primary focus of the above work is to extract the tumor region from the image and apply different algorithms to classify the images.

2.2 Proposed Work

The proposed work aims in classifying malignant or benign to breast cancer from histopathology images without segmenting tumor region from the original image. Histopathology is the study of the signs of the disease using the microscopic examination of a biopsy or surgical specimen. We have only used limited number of data to train the model in order to tackle the real life situation where only limited resource is available, this will eventually avoid overfitting and as a result we obtain 91.5% accuracy in predicting the actual result.

Chapter 3

EXPERIMENTAL RESULTS

3.1 Data

The Breast Cancer Histopathological Image Classification (BreakHis) [8] is composed of 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429 malignant samples (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format).

Table 3.1: About Breast Cancer Histopathology Image Dataset

The Breast Cancer Histopathological Image	
Size	4GB
Total Images	9,109 microscopic images
No. Of Patients	82 patients
magnifying factors	40X, 100X, 200X, and 400X
Benign Samples	2,480
Malignant Samples	5,429
Pixel	700 X 460 pixels
Format	PNG

Each image filename stores information about the image itself: method of procedure biopsy, tumor class, tumor type, patient identification, and magnification factor. For example, SOBBTA-14-4659-40-001.png is the image 1, at magnification factor 40X, of a

benign tumor of type tubular adenoma, original from the slide 14-4659, which was collected by procedure SOB.

3.2 Training Model

- The existing folder structure where the images are stored in such a way classifies cancer type. Since the objective here is to only classify whether the specimen is benign or malignant, we first restructure the folder in a simple manner.

```
.
├── ...
├── Train                # Images for training the model
│   ├── Benign           # Class 0
│   ├── Malignant        # Class 1
├── Validation           # Images for validating the model
│   ├── Benign
│   ├── Malignant
├── Testing              # Images for testing the model
│   ├── Benign
│   ├── Malignant
└── ...
```

Figure 3.1: Structure of the folder

- We know that there is a total of 5429 images that are malignant and 2480 images are benign. Here the data is highly imbalanced but this is the case with real-world situations, for example, research says only 1 out of 8 women is diagnosed with breast cancer during their lifetime. Medical data are usually imbalanced because of their nature.

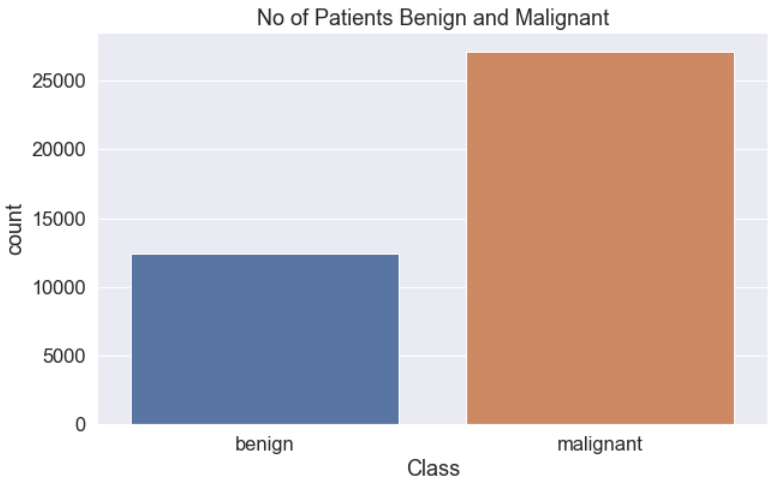

Figure 3.2: Benign vs Malignant

- On viewing the malignant and benign images, it is hard to find out differences as they look similar in nature. This is the whole purpose of using a deep learning algorithm to classify the images that are hard to predict using the naked eye. It takes 5-6 years to become a pathology expert but we could able to obtain much more precision using the latest technologies in a fraction of seconds.

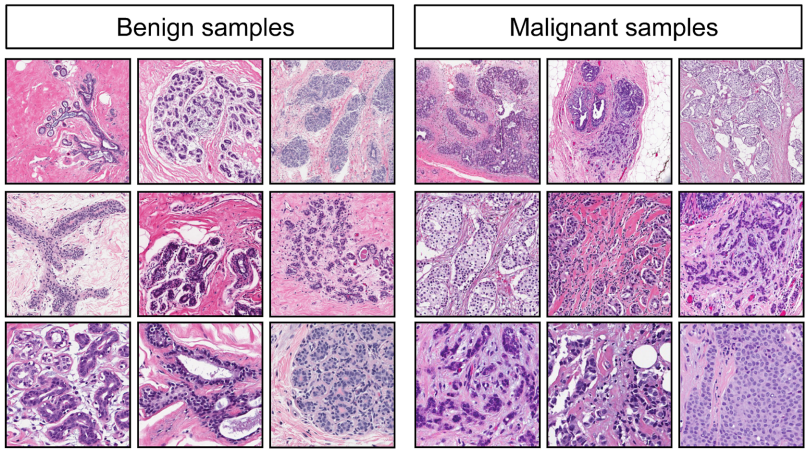

Figure 3.3: Sample Image of Benign and Malignant

- From the original dataset 80% of the data i.e 6407 images are used for training the model. Once the model is trained, 10% from the original data which is 712 images are used to evaluate the model. Finally, 791 images are applied to the model which is not used while training or validation to find the accuracy of the final model.

- In order to increase the accuracy number of images for class 0 (benign) and class 1 (malignant) are balanced for training the model. But, the number of images for benign and malignant are left unbalanced to reflect real-life scenarios.

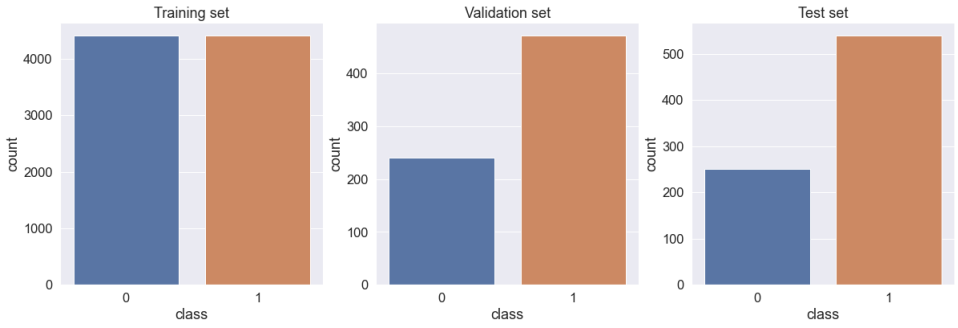

Figure 3.4: Training, Validation and Test data

- Since we have been using only less number of images for training the model, there

are high chances that the model may predict bad when testing on a new image. To avoid this under-fitting of the model, Image Data Generator is introduced which will allow the model to become invariant to the orientation of the object. Image Data Generator class allows users to randomly rotate images through any degree between 0 and 360 by providing an integer value in the rotation_range argument. Below are the list of instruction given to Image Data Generator.

Table 3.2 Shows the list of function that are used to generates batches of tensor image data with real-time data augmentation. CNN use these images with different variations at each epoch to prevent model overfitting.

Table 3.2: Functions Used in Image Data Generator

Image Data Generator
Random rotation by 20 degrees.
Horizontal flip.
Vertical flip.
Rescale image by its pixel value.
Randomly Zoom image by 20%.
Random shear by 20%.

- All the image sizes are set to 128 X 128 to avoid running of model for a long time.
- Now the CNN architecture is introduced and defined using Conv2D layers and Max-pooling layers. Dropout Layers are added to randomly turn off some neurons while training to prevent from over-fitting. Flatten layer is added at the end to form Dense Layers. Relu is used as activation in all the layers and Sigmoid as the activation in the output layer.

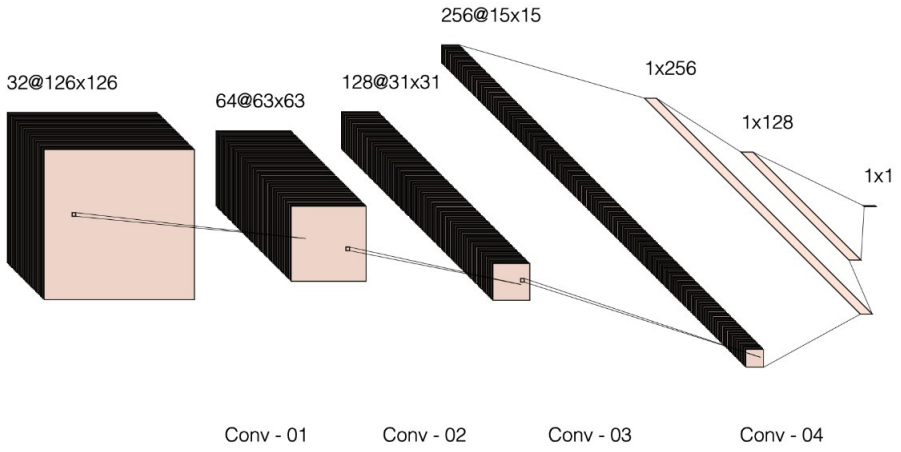

Figure 3.5: Convolution Neural Network with (Relu and Sigmoid)

- Model is compiled by using binary cross-entropy as the loss function and adam optimizer to optimize the weights. Early stopping is involved to monitor validation loss in order to prevent over-fitting with a patience level of 5. Model checkpoint is used to store the best models for every epoch.

- Finally model is fitted using train and validation generators generated using Image DataGenerator. Verbose is set to 1 to monitor accuracy and losses. The model is trained for 200 epochs. Early stopping and checkpoints are used as the callbacks.

- Training accuracy and Validation accuracy is calculated as 93% and 91.5% respectively

Figure 3.6: Accuracy and Loss

- The finalized model is now tested with test data that classify malignant or benign with 91.5% accuracy. And the final model is saved.

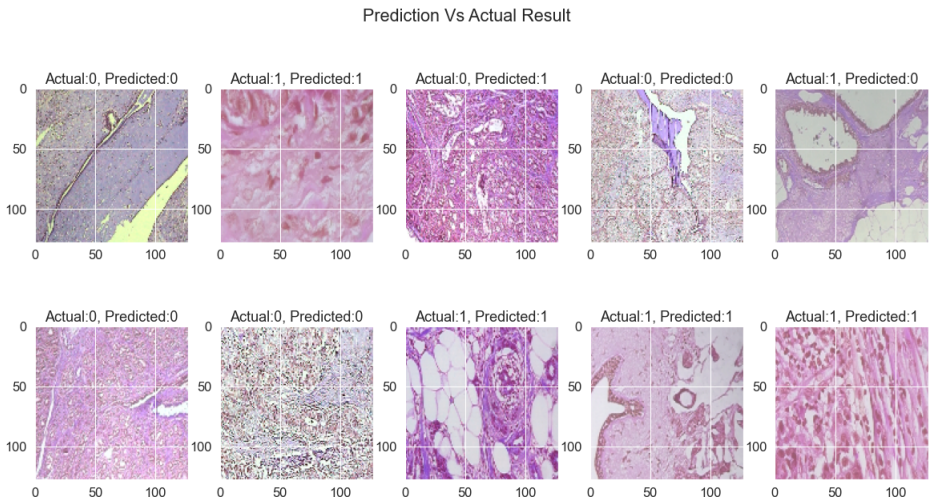

Figure 3.7: Predicted vs Actual Result (0 for Benign & 1 for Malignant)

3.3 Deploying Model as Web Application

We used streamlit [10] to deploy the final model as a web application. It is an open-source python framework for building web apps for Machine Learning and Data Science. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write a python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app.

The designed web page consists of three tabs.

- Welcome: A brief introduction about breast cancer, its causes, treatment, and a list of centres that treat breast cancer.

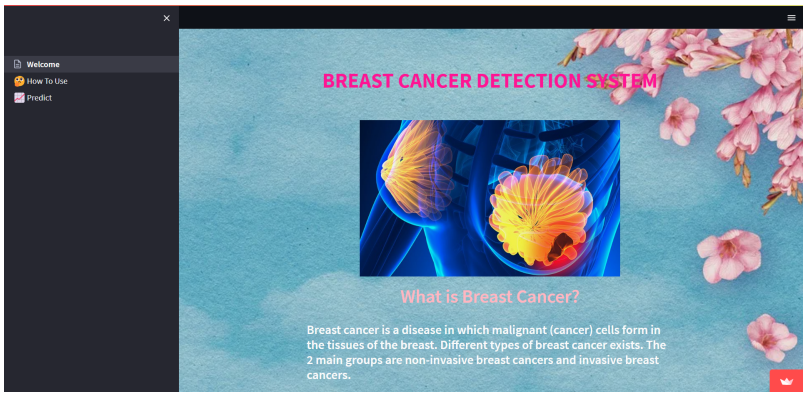

Figure 3.8: Tab 1 - Home Page

- How To Use: Demo and step by step instruction on how to use the application.

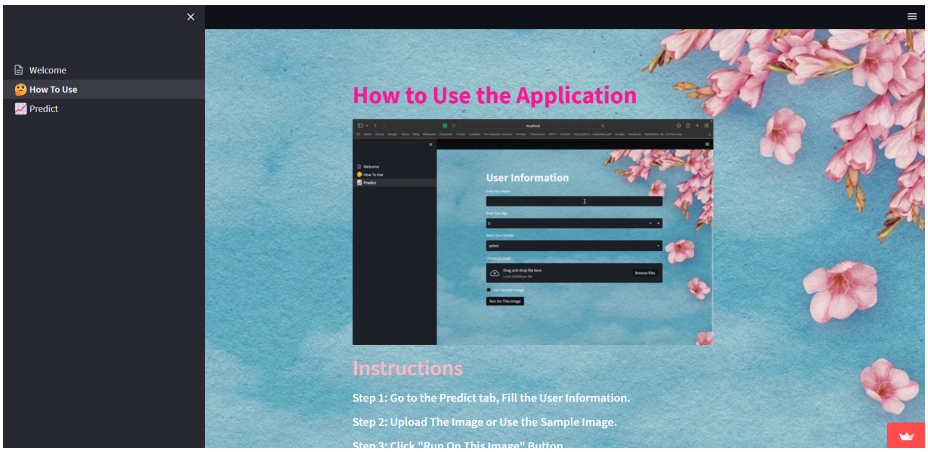

Figure 3.9: Tab 2 - How to Use Page

- Predict: An interactive page that allows users to upload an image and predict whether the specimen is subjected to breast cancer or not.

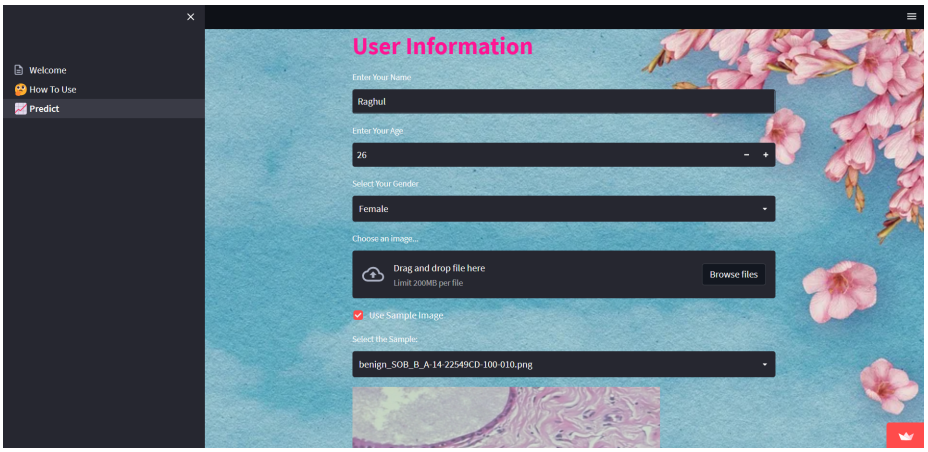

Figure 3.10: Tab 3 - Predict Page

3.3.1 Link To Access Our Application

- Home Page: click here
- How To Use Page: click here
- Predict Page: click here
- To Download Code: click here

Chapter 4

CONCLUSION

Breast cancer is one of the most commonly diagnosed malignancies in women around the world. Several researchers have worked on breast cancer segmentation and classification using a variety of imaging techniques. Histopathology imaging is an effective diagnostic approach that is used for breast cancer detection with the help of a microscope. In this paper, we propose a fully automatic breast cancer detection application. The proposed method is divided into two main stages. First, supervised learning is performed on the number of images, and with the help of image classification using CNN, an effective model is constructed. Second, the trained model is used to verify user uploaded histopathology images to classify whether the specimen is subjected to breast cancer or not.

Based on the experimental results, the proposed model achieved 91.5% accuracy. In addition, the proposed system is deployed as a web application (https://nascent-bot-cancer-app-myappwelcome-i66wjm.streamlitapp.com/) that is free to use by any user across the world. In the future, android and iOS applications of the same will be developed with enhancement.

Bibliography

[1] Breast Cancer, Cleveland Clinic, https://my.clevelandclinic.org/, 2022

[2] Neural Networks, by IBM Cloud Education, 2020

[3] Image Classification in Deep Learning, by Think Automation

[4] Deep Learning In Healthcare, by Marlee Long, 2022

[5] Silva L, Saade D, Sequeiros G, Silva A, Paiva A, Bravo R, et al. A new database for breast research with an infrared image. J. Med. Imaging Health Infor. 2014; 4:92–100.

[6] Mahmoudzadeh E, Montazeri M, Zekri M and Sadri S. Extended hidden Markov model for optimized segmentation of breast thermography images. Infrared Phys. Tech- nol. 2015; 72: 19–28.1

[7] Breast cancer prediction and cross-validation using multilayer perceptron neural networks by S.A. Mojarad; S.S. Dlay; W.L. Woo; G.V. Sherbet.

[8] Breast Cancer Histopathological Database (BreakHis), from Kaggle, 2020

[9] Convolutional Neural Networks, by IBM Cloud Education, 2020

[10] A faster way to build and share data apps, by Streamlit