

BREAST CANCER DETECTION FROM HISTOPATHOLOGY IMAGES USING DEEP LEARNING

Research Project Submitted

to

SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH

(Category – I Deemed to be University)

[Accredited by NAAC with 'A++' Grade]

In partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

in

MEDICAL BIOINFORMATICS

By

Abirami Shree.P

(Reg. No. E8120014)



Sri Ramachandra Faculty of Engineering & Technology

August 2022



SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai

SRI RAMACHANDRA FACULTY OF ENGINEERING & TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that, the dissertation/research project titled "**Breast Cancer Detection From Histopathology Images Using Deep Learning**" is based on the bonafide work carried out under my supervision by **Abirami Shree P** (Registration No: **E8120014**) during the academic year 2021-2022 as a part of his/her M.Sc. Medical Bioinformatics course. This dissertation/research project has not previously formed the basis for the award of any degree, diploma, associateship, fellowship or similar title.

Research Supervisor

Date: 24.08.2022

Prof. V. RAJU

Provost

Sri Ramachandra Engineering & Technology

SRI RAMACHANDRA

Institute of Higher Education and Research

(Deemed to be University)

Provost Chennai - 600 116.

Sri Ramachandra Faculty of
Engineering & Technology

Submitted for Examination held
on.....

Internal Examiner

External Examiner



SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH

(Category - I Deemed to be University) Porur, Chennai

SRI RAMACHANDRA FACULTY OF ENGINEERING & TECHNOLOGY

DECLARATION BY THE CANDIDATE

1. I **Abirami Shree.P** post graduate student hereby declare that the dissertation/research project work titled “**Breast Cancer Detection From Histopathology Images Using Deep Learning**” is a bonafide and original work done by me under the guidance of **Prof. Chiranjeevi N** in Sri Ramachandra Faculty of Engineering & Technology, Sri Ramachandra Institute of Higher Education and Research (Category – I Deemed to be University), Porur, Chennai 600 116 during the period from April 2022 to August 2022.
2. I further declare that only after obtaining permission of the Publication and Oversight Committee of Sri Ramachandra Institute of Higher Education and Research (Deemed to be University) that publication of the said dissertation/research project work, either in part or in full, will be made by me and that such publication will include only the names of individuals who has/ have actually contributed to the said dissertation work.
3. I agree to publish the said dissertation/ research project work within a period of one year from the date of qualifying for the M.Sc. Medical Bioinformatics degree, failing which I hereby give my consent to transfer my rights to the Guide of the Faculty of Engineering & Technology, Sri Ramachandra Institute of Higher Education and Research (Category – I Deemed to be University).

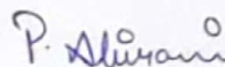
4. I hereby declare that Sri Ramachandra Institute of Higher Education and Research (Category – I Deemed to be University) shall have the rights to preserve, use and disseminate this dissertation/ research project work in print or electronic format for academic / research purposes.

5. I also declare and undertake that this dissertation/research project work either in part or in full, will not be utilized by me subsequently for any purpose, without the written prior permission and consent of the University.

6. I also understand and agree that this declaration made by me is final and irrevocable.

Place: Chennai

Date: 24.08.2022



Signature of the Candidate

Name: Abirami Shree P

M.Sc. Medical Bioinformatics

ACKNOWLEDGEMENT

I would like to show my immense gratitude to my guide **Prof. Chiranjeevi N**, Lecturer Sri Ramachandra Faculty of Engineering and Technology. His enthusiasm has been infectious and I have derived confidence from it every time I interacted with him. He has been a pillar of support during the course of my academic project. I am grateful to him for the encouragement that he continued to provide along every step of the way. He had constantly given his technical advice and expertise and thereby enabled the successful completion of the project.

I place on record my sincere thanks to the Chancellor, Pro-chancellor, Vice-Chancellor, Dean of Faculties, Trustees and the Management for providing an environment that was conducive and the necessary facilities to carry out my research.

My heartfelt gratitude is extended to the **Prof. V. Raju**, Provst - Sri Ramachandra Faculty of Engineering and Technology, for providing the freedom to work and also whole-heartedly promoting research interest.

My heartfelt thanks to my Teachers Dr. G. Dicky John Davis, Mrs. C. R Hemalatha, Dr. S. Venkatesean, Mrs. M. Arundathi, Mrs. M. Premavathi and Dr. G. Jayanthi for all the help.

Hearty thanks to my parents for the unceasing encouragement, support and attention. I am also grateful to my brother Mr. S. Pradeep Kumar who supported me through this venture. I also place on record, my sense of gratitude to one and all, who directly or indirectly, have let their hand in this venture

Abirami Shree P

CONTENTS

<u>S.NO</u>	<u>TOPIC</u>	<u>PG-NO</u>
1	ABBREVIATION	5
2	LIST OF FIGURES	6
3	ABSTRACT	9
4	INTRODUCTION	10
5	REVIEW OF LITERATURE	29
6	AIM & OBJECTIVE	33
7	MATERIALS AND METHODS	35
8	RESULTS	50
9	DISCUSSION	63
10	REFERENCE	67

ABBREVIATIONS

1. **AI** – Artificial Intelligence
2. **NLP** – Natural Language Processing
3. **CNN** – Convolutional Neural Network
4. **IDC** – Invasive Ductal Carcinoma
5. **ANN** – Artificial Neural Network
6. **ML** – Machine Learning
7. **WSI** – Whole Slide Image
8. **DL** – Deep Learning
9. **HPE** – Histopathology Images
10. **ResNet** – Residual Networks
11. **Pytorch** – Python Torch

LIST OF FIGURES

S.NO	TOPIC	PG-NO
01	Exploratory Data Analysis	51
02	Displaying of the Healthy patches	52
03	Displaying of the Cancer patches	53
04	Binary data visualization per tissue slice	54
05	Grid Highlighting the Cancerous part being spotted from the patches	55
06	Target distributions difference of the datasets	56
07	Data augmentation using the Pytorch data loaders	57
08	Sklearn module to highlight the grid that is spotted with the cancer cells	58
09	Search for an optimal cyclic learning rate	59
10	Loss convergence of training, developing & test dataset	60
11	Loss convergence – The running losses	61
12	Sklearn plots – statistical representation of the normal vs the cancerous dataset	62
13	ROC curve for the ResNet model	63

ABSTRACT

Breast cancer being the most type of cancer identified especially women, thereby it is much advised to be on the earliest diagnosis as possible as it is safer to know a disease at the initial level to terminate it at the initial stage. Thereby to invade into the treatment to prohibit the further final circumstances will become much easier to terminate it in the bud. At present there are many testing methods available but the level of accuracy, intensity and especially the time taken for a correct examination is all a tedious process. Further on the other hand the world being turning around on the other side is essential for the evolution of the medical domain to boost-up its technologies as well, as it is not a choice but the only way to survive its liveliness. Here in my work, I have used the artificial intelligence-based techniques via., its deep learning architectures to stand out the trend. The major idea is to compare two CNN model to check the level of accuracy and the time consumed. The dataset of 249 patients with both cancer and non-cancerous tissues in the form of HPE images is being taken as the input data with 2,77,512 patches is taken where the patches are being segregated as binary data initially by pre-trained data, so as to check the model for both training and testing. Here the CNN model being used with both the architecture of ResNet-18 & ResNet-50, whereby finally the comparison as to which ResNet architecture of the CNN model is better than the other, and also thereby proving the same with the help of heat maps, Roc curves & Model summary to put forth the exact differences being existing between the two. Also as known to all commercializing the diagnosis with a fully automated AI based diagnosing tool makes it an essential advancement when compared to the existing procedures as it will also provide us with diagnostic results which might also not be visible or missed by the naked eye by the physicians i.e., pathologists or even by a well-trained medical-staffs.

1. INTRODUCTION

1.INTRODUCTION

Cancer starts when cells in the body begin to grow out of control. Cells in nearly any part of the body can become cancer cells, and can then spread to other areas of the body. In most cases, activation of oncogenes and/or deactivation of tumor suppressor genes lead to uncontrolled cell cycle progression and inactivation of apoptotic mechanisms. As opposed to benign tumors, malignant cancers acquire metastasis, which occurs in part due to the down-regulation of cell adhesion receptors necessary for tissue-specific cell–cell attachment, and up-regulation of receptors that enhance cell motility.(1)

Cancer is an umbrella term for a large group of diseases that can affect any part of the body. Other terms used are malignant tumors and neoplasms. A characteristic feature of cancer is the rapid formation of abnormal cells that can grow beyond their usual limits and then invade neighboring parts of the body and spread to other organs; this latter process is known as metastasis. Widespread metastases are the leading cause of cancer death. (2)

Cancer is one of the leading causes of death worldwide, responsible for almost 10 million deaths in 2020. The most common in 2020 (in terms of new cancer cases) were:

- Breast (2.26 million cases);
- Lungs (2.21 million cases);
- Colon and rectum (1.93 million cases);
- Prostate (1.41 million cases);
- Skin (no melanoma) (1st20 million cases);

The leading causes of cancer deaths in 2020 were:

- lung (1.80 million deaths);
- colon and rectum (916,000 deaths);

- Liver (830,000 deaths);
- Stomach (769,000 deaths); and
- Breast (685,000 deaths).

400,000 children are being tested with cancer every year. The most common types of cancer vary from country to country with cervical cancer being the more common in 23 countries (3).

1.2 Cancer Causes:

Cancer arises from the conversion of normal cells into tumor cells in a multi-step process that generally progresses from a precancerous lesion to a malignant tumor(4). These changes are the result of the interaction between a person's genetic factors and three categories of external agents, including:

- physical carcinogens such as ultraviolet and ionizing radiation;
- chemical carcinogens such as asbestos, components of tobacco smoke, alcohol, aflatoxin (a food contaminant) and arsenic (a drinking water contaminant);
- Biological carcinogens, such as infections caused by certain viruses, bacteria, or parasites.

1.3 BREAST CANCER

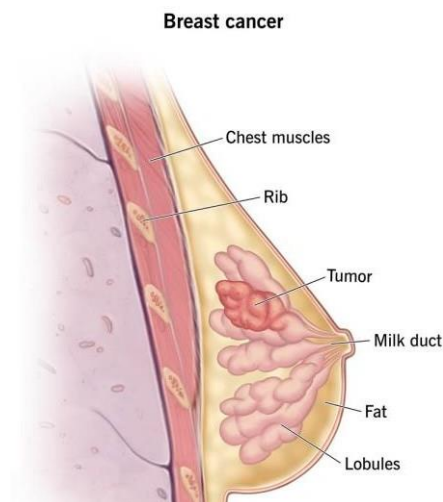


Fig: 1.1 Breast parts – basic outline

Breast cancer forms due to random dividing of cells in the breast. Breast skin cancer, breast cancer is the most diagnosed cancer in women in the United States. Extensive support for breast cancer awareness and research funding have helped drive advances in the diagnosis and treatment of breast cancer.(5) Breast cancer survival rates have increased and the number of deaths related to the disease has steadily decreased, mainly due to factors such as earlier detection, a new personalized treatment approach and a better understanding of the disease.(6)

1.4 Symptoms:

- A breast lump or thickening that feels different from the surrounding tissue
- Change in the size, shape or appearance of a breast
- Changes to the skin over the breast, such as dimpling
- A newly inverted nipple
- Peeling, scaling, crusting or flaking of the pigmented area of skin surrounding the nipple (areola) or breast skin
- Redness or pitting of the skin over your breast, like the skin of an orange.

1.5 TESTING METHODS

Micrograph displaying a lymph node invaded with the aid of using ductal breast carcinoma, with an extension of the tumour past the lymph node. Breast ducts are the passageways wherein milk from the milk glands (lobules) flows to the nipple.

Invasive ductal carcinoma is most cancers (carcinoma) that takes place whilst ordinary cells developing with inside the lining of the milk ducts alternate and invade breast tissue past the partitions of the duct. Once that takes place, the most cancers cells can unfold. They can spoil into the lymph nodes or bloodstream, wherein they could tour to different organs and

regions with inside the body, ensuing in metastatic breast most cancers. Several exams can assist your physician pick out and diagnose IDC.

Digital mammography is an advanced approach for breast imaging this is carried out just like a everyday mammogram. However, it's far higher than traditional mammography in detecting most cancers in more youthful sufferers and in people with dense breast tissue. Electronic photographs may be more suitable with pc-aided detection structures to identify masses, calcifications and abnormalities related to most cancers.

Breast ultrasound makes use of sound waves to take a look at the breast tissue and gauge blood flow. It is secure for inspecting pregnant sufferers, and does now no longer use radiation.

Breast magnetic resonance imaging (MRI) makes use of a huge magnet, radio waves and a pc that could stumble on small breast lesions, and can be especially beneficial in inspecting sufferers with an excessive danger of breast most cancers, along with people with BRCA1, BRCA2 or different gene mutations related to most cancers.

A breast biopsy includes taking a pattern of breast tissue from a suspicious region and sending it to a laboratory for microscopic exam with the aid of using a pathologist, a physician who makes a speciality of figuring out symptoms and symptoms of disease. A biopsy can affirm or rule out the presence of most cancers and, if most cancers is present, monitor its characteristics.

Staging workup - When most cancers are detected, the subsequent step is staging. "Staging determines how a way the most cancers cells have unfolded," Wright says. "Staging is primarily based totally on the dimensions of the tumour and whether or not or now no longer it has unfolded to the lymph nodes or entered the bloodstream and unfold everywhere else in body."(7)

1.6 HISTOPATHOLOGY:

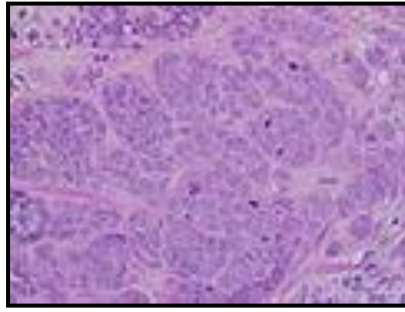


Fig: 1.2 HISTOPATHOLOGY IMAGE OF IDC

Histopathology is the study of signs of disease through microscopic examination of a biopsy or surgical specimen that is processed and fixed on glass slides. In order to make different tissue components visible under a microscope, sections are stained with one or more dyes. The goal of staining is to reveal cellular components; Counterstains are used to create contrast.(8) Hematoxylin-eosin (HandE) staining has been used by pathologists for over a hundred years which stains the nuclei of the cell blue, while eosin stains the cytoplasm, connective tissue in pink. Because of HandE's long history, well-established methods, and a large body of data and publications, many pathologists firmly believe that HandE will remain common practice for the next 50 years.(9)

1.7 Epidemiology

By 2020, 2.3 million women worldwide were diagnosed with breast cancer and 685,000 died. At the end of 2020, there were 7.8 million women who had been diagnosed with breast cancer in the past 5 years, making it the most common cancer in the world. Globally, women lose more disability-adjusted life years (DALYs) to breast cancer than any other cancer.(10)Breast cancer occurs in women of all ages after puberty in every country in the world, but with increasing frequency later in life. Breast cancer mortality changed little from the 1930s to the 1970s Survival improvements began in the 1980s in countries with early

detection programs combined with different treatment modalities to eradicate invasive diseases. Who is at risk? Breast cancer is not a communicable or contagious disease. Unlike some cancers that have infection-related causes, such as For example, human papillomavirus (HPV) infection and cervical cancer, no viral or bacterial infections are known to be associated with the development of breast cancer.(11)

About half of breast cancers occur in women who have no identifiable risk factors for breast cancer other than gender (female) and age (over 40 years). Certain factors increase breast cancer risk, including increasing age, obesity, harmful alcohol use, family history of breast cancer, history of radiation exposure, reproductive history (such as age at onset of menses and age at first pregnancy), tobacco use, and postmenopausal hormone therapy.(12)

Behavioral choices and related interventions that reduce breast cancer risk include:

- Prolonged breastfeeding;
- Regular physical activity;
- Weight control;
- Avoidance of Harmful Alcohol Use;
- Tobacco Smoke Exposure Avoidance;
- Avoidance of long-term use of hormones; and
- Avoid excessive exposure to radiation.

Unfortunately, even if all potentially modifiable risk factors could be controlled, this would only reduce the risk of developing breast cancer by no more than 30%. Female gender is the strongest risk factor for breast cancer.(12) About 0.5-1% of breast cancers occur in men. The treatment of breast cancer in men follows the same treatment principles as in women. A family history of breast cancer increases the risk of breast cancer, but most women diagnosed with breast cancer have no family history of the disease. A lack of a known family history does not necessarily mean a woman is at lower risk.

Certain inherited 'high penetrance' gene mutations significantly increase breast cancer risk, with mutations in genes BRCA1, BRCA2 and PALB-2 being the most potent. Women with mutations in these important genes might consider risk reduction strategies such as: B. the surgical removal of both breasts. Consideration of such a highly invasive approach affects only a very limited number of women, should be carefully weighed against all alternatives, and not rushed.(13)

1.8 Breast Metastasis

Metastatic breast cancer is the most advanced stage of breast cancer. Breast cancer occurs when abnormal cells in the breast start dividing uncontrollably. A tumour is huge collection of these abnormal cells. Metastases refer to cancer cells that have spread to a new area of the body.(14) In metastatic breast cancer, cells can spread to:

- Bones
- Brain
- Liver
- Lungs

Healthcare providers name cancer after its primary cause. This means breast cancer that spreads to other parts of the body is still considered breast cancer. The cancer cells are yet continued to be breast cancer cells despite their spread.(15)Your treatment team will use breast cancer therapies even if cancer cells are found in other areas. These two terms – breast metastasis as well as stage 4 breast cancer means essentially the same thing. Breast cancer classified as stage 4 has spread outside the breast, or metastasized, to other parts of the body.(16)

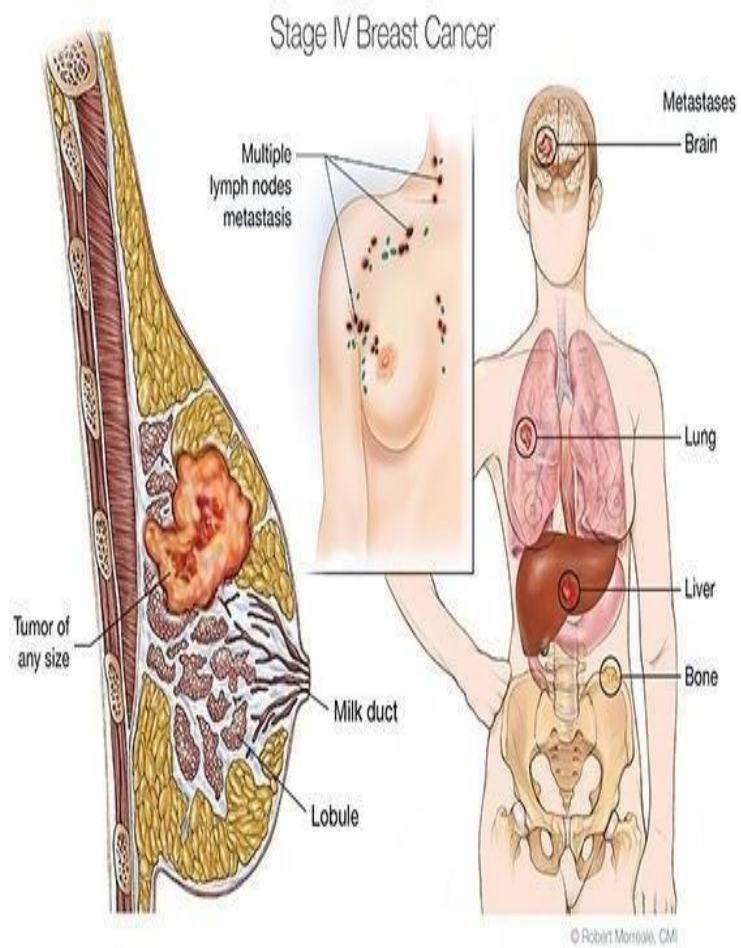


Fig:1.3 Stage 4 Breast Cancer (or) Breast Metastasis

1.9 Treatment

There are different types of treatment for patients with breast cancer.

Seven types of standard treatment used are:

- Watchful waiting or active surveillance
- Surgery
- Radiation therapy and radiopharmaceutical therapy
- Hormone therapy
- Chemotherapy
- Immunotherapy
- Bisphosphonate therapy

There are treatments for bone pain caused by bone metastases or hormonetherapy.

New types of treatment are being tested in clinical trials.

- Cryosurgery
- High-intensity–focused ultrasound therapy
- Proton beam radiation therapy
- Photodynamic therapy
- Treatment for breast cancer may cause side effects.
- Patients may want to think about taking part in a clinical trial.
- Patients can enter clinical trials before, during, or after starting their cancer treatment.(17)

1.10 Bioinformatics

Molecular medicine requires the integration and analysis of genomic, molecular, cellular, as well as clinical data and it thus offers a remarkable set of challenges to bioinformatics.

Bioinformatics nowadays has an essential role both, in deciphering genomic, transcriptomic,

Information gathered from traditional biology and medicine.(18)

The evolution of bioinformatics, which started with sequence analysis and has led to high-throughput whole genome or transcriptome annotation today, is now going to be directed towards recently emerging areas of integrative and translational genomics, and ultimately personalized medicine.

Therefore considerable efforts are required to provide the necessary infrastructure for high-performance computing, sophisticated algorithms, advanced data management capabilities, and-most importantly-well trained and educated personnel to design, maintain and use this environment. This review outlines the most promising trends in bioinformatics, which may play a major role in the pursuit of prostate cancer and metastasis diagnosis.(19)

1.11 Neural network

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons.(20)

Neural network simulations appear to be a recent development. However, this field was established before the advent of computers, and has survived at least one major setback and several eras.

Many important advances have been boosted by the use of inexpensive computer emulations following an initial period of enthusiasm; the field survived a period of frustration and

disrepute. During this period when funding and professional support was minimal, important advances were made by relatively few researchers. These pioneers were able to develop convincing technology which surpassed the limitations identified by Minsky and Papert. Minsky and Papert neural networks among researchers, and was thus accepted by most without further analysis. Currently, the neural network field enjoys a resurgence of interest and a corresponding increase in funding.(21)

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze.(22) This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

Other advantages include:

- Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.
- Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time.
- Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.
- Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.(24)

1.12 Neural Network Structure

Although neural networks impose minimal demands on model structure and assumptions, it is useful to understand the general network architecture. The multilayer perceptron (MLP) or radial basis function network is a function of predictors (also called inputs or independent variables) that minimize the prediction error of target variables (also called outputs). (25)

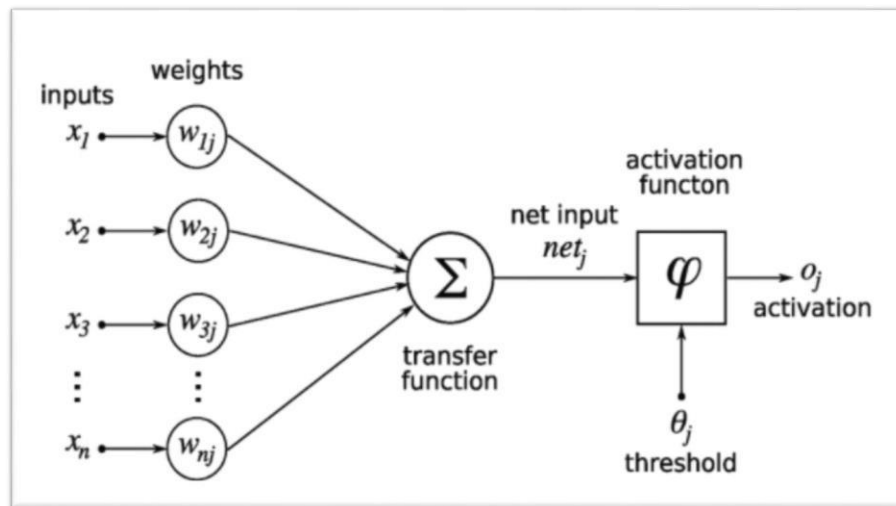


Fig:1.4 Architecture of a single neuron in a neural network

1.13 WHAT IS DEEP LEARNING?

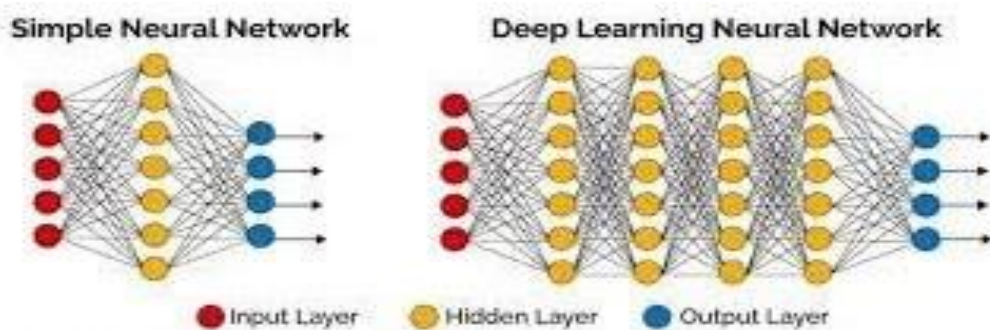


Fig:1.5 Deep Learning Network

Deep learning is a type of machine learning and artificial intelligence (AI) that mimics the way humans acquire certain types of knowledge. Deep learning is an important element of data science, which includes statistics and predictive models. It is extremely beneficial for data scientists tasked with collecting, analyzing large amounts of data; which is made more feasible and easier with deep learning. In its simplest form, deep learning can be viewed as a way to automate predictive analytics.(26) On the other hand, ML algorithms are linear, whereas DL algorithms are stacked in a hierarchy of increasing complexity. To understand deep learning, imagine a child whose first word is dog. The young child learns what a dog is and isn't by pointing at objects and saying the word dog. The parent says, "Yes, that's a dog" or "No, that's not a dog. As the toddler continues to point, he becomes aware of the qualities that all dogs possess. What the little boy is unknowingly doing is clarifying a complex abstraction, the concept of a dog, by constructing a hierarchy in which each level of abstraction is created using the knowledge gained from the previous level of the hierarchy(27).

1.14 CONVOLUTIONAL NEURAL NETWORK:

A convolutional neural network (CNN) is a type of artificial neural network used in image recognition and processing, specifically designed to process pixel data. CNN are powerful imagery and artificial intelligence (AI) processors that use deep learning to perform both generative and descriptive tasks, often using computer vision, including image and video recognition, as well as natural language processing (NLP) and recommender systems). A neural network is a hardware and/or software system designed to adapt to the functioning of neurons in the human brain. Traditional neural networks are not ideal for image processing and must receive images in reduced-resolution blocks. CNNs have their "neurons" arranged more like those in the frontal lobe, the area responsible for processing visual stimuli in

humans and other animals. The layers of neurons are arranged to cover the entire field of view, avoiding the image processing problem that parts of traditional neural networks exhibit. A CNN uses a multilayer perceptron-like system designed for low processing requirements. CNN layers - an input layer, an output layer, and a hidden layer, which includes multiple convolution layers, pooling layers, fully connected layers, and normalization layers. Removing constraints and increasing the efficiency for image processing results in a system that is much more effective and easier to train, and is limited to processing images and natural language.(28)

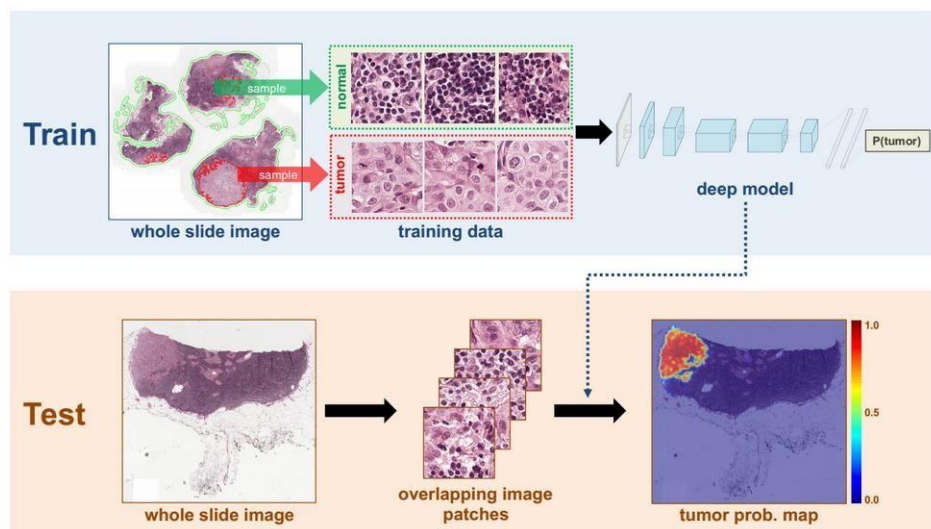


Fig:1.6 Outline working of a CNN model

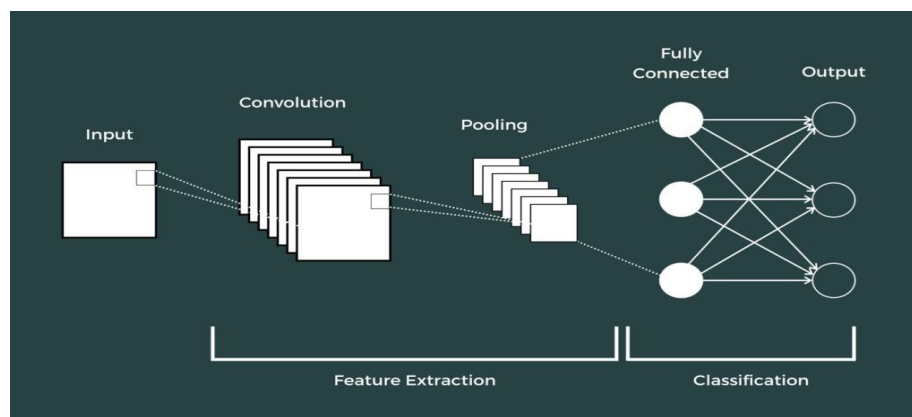


Fig:1.7 Basic ResNet architecture

1.15 ResNet:

ResNet, short for Residual Networks is a classic neural network used as a backbone for many computer vision tasks. This model was the winner of ImageNet challenge in 2015. The fundamental breakthrough with ResNet was it allowed us to train extremely deep neural networks with 150+layers successfully. Prior to ResNet training very deep neural networks was difficult due to the problem of vanishing gradients.(29)

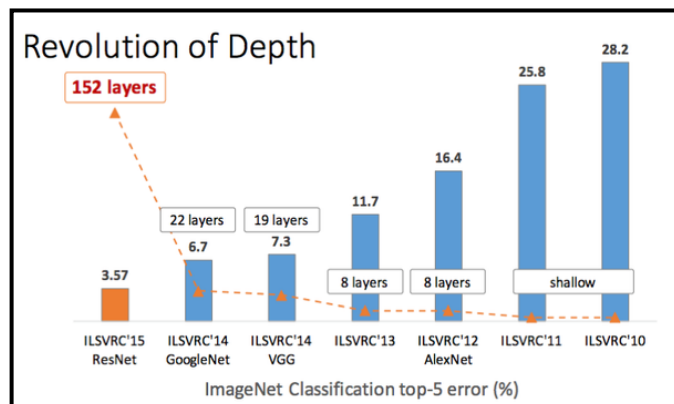


Fig:1.8 Revolution of Depth

1.16 ResNet-18:

ResNet models were actually introduced in “Deep Residual Learning for Image Recognition”. ResNet-18 is a convolutional neural network that is trained on more than a million images from the ImageNet database. There are 18 layers present in its architecture. It is very useful and efficient in image classification and can classify images into 1000 object categories. The network has an image input size of 224x224.(30)

Layer Name	Output Size	ResNet-18
conv1	$112 \times 112 \times 64$	$7 \times 7, 64, \text{stride } 2$
conv2_x	$56 \times 56 \times 64$	$3 \times 3 \text{ max pool, stride } 2$ $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
conv3_x	$28 \times 28 \times 128$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
conv4_x	$14 \times 14 \times 256$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
conv5_x	$7 \times 7 \times 512$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$
average pool	$1 \times 1 \times 512$	$7 \times 7 \text{ average pool}$
fully connected	1000	$512 \times 1000 \text{ fully connections}$
softmax	1000	

Fig;1.9 Workflow of a RsNet-18 architecture

From this diagram we can see how layers are configured in the ResNet-18 architecture. First there is a convolution layer with 7×7 kernel size and stride 2. After this there is the beginning of the skip connection. The input from here is added to the output that is achieved by 3×3 max pool layer and two convolution layers with kernel size 3×3 , 64 kernels each. This was the first residual block.

Then from here, the output of this residual block is added to the output of two convolution layers with kernel size 3×3 and 128 such filters. This constituted the second residual block. Then the third residual block involves the output of the second block through skip connection and the output of two convolution layers with filter size 3×3 and 256 such filters. The fourth and final residual block involves output of third block through skip connections and output of two convolution layers with same filter size of 3×3 and 512 such filters.(31)

ResNet-18 – ADVANTAGES:

Networks with large number (even thousands) of layers can be trained easily without increasing the training error percentage.

ResNet help in tackling the vanishing gradient problem using identity mapping.(32)

1.17 ResNet-50:

The ResNet-50 model consists of 5 stages each with a convolution and Identity block. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers. The ResNet-50 has over 23 million trainable parameters.(33)

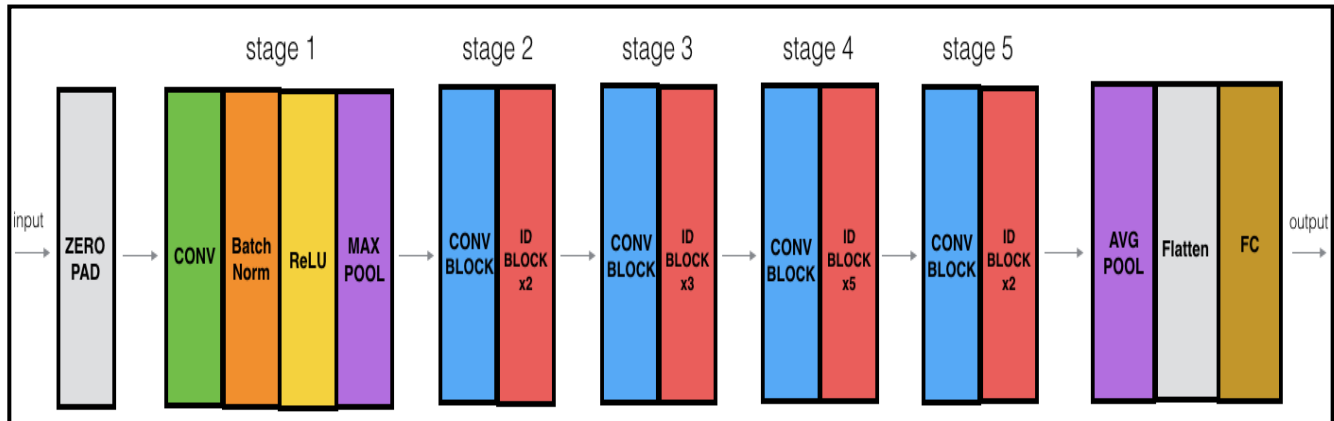


Fig:1.10 Resnet-50 depicting its layers and workflow

2.REVIEW OF LITERATURE

2.1.1 International evaluation of an AI system for breast cancer screening

Mammography screening aims to detect breast cancer early in the disease stage, when treatment may be most successful. Despite the existence of screening programs worldwide, the interpretation of mammograms is hampered by high rates of false positives and false negatives. Here we present an artificial intelligence (AI) system capable of outperforming human experts in predicting breast cancer. To assess its performance in the clinical setting, we selected a representative large data set from the UK and a large enriched data set from the US. We showed an absolute reduction of 5.7% and 1.2% (US and UK) for false positives and 9.4% and 2.7% for false negatives.(34) We provide evidence of the system's ability to generalize from the UK to the US. In an independent study of six radiologists, the AI model outstood every human reader: the area under the Receiver Operating Characteristic Curve (AUC-ROC) for the AI system was greater than the AUC-ROC for the average radiologist of for an absolute Margin of 11.5%. We conducted a simulation where the AI system participated in the double reading method used in the UK and found that the AI system maintained no worse performance and reduced the workload of the second reader by 88%. This robust assessment of the AI system paves the way for clinical trials to improve the accuracy and efficiency of breast cancer detection.(35)

2.1.2 Deep Learning to Improve Breast Cancer Detection on Screening Mammography

The rapid development of deep learning, a family of machine learning techniques, has sparked great interest in its application to medical imaging problems. Here we develop a deep learning algorithm that can accurately detect breast cancer in screening mammograms using an “end-to-end” training approach that efficiently leverages training datasets with full clinical annotations or just the cancer status (label) of the entire image. With this approach, injury annotations are only required at the initial training stage, and

later stages only require image-level annotations, thereby eliminating the reliance on infrequently available injury annotations. Our fully convolutional network method for classifying screening mammograms has performed exceptionally well compared to previous methods.(36) In an independent test set of digitized film mammograms from the Digital Database for Screening Mammograms (CBIS-DDSM), the best single model achieved an AUC per frame of 0.88 and four models with a mean of improved the AUC to 0.91 (sensitivity: 86.1%, specificity: 80.1%). In an independent test series of full-field digital mammography (FFDM) images from the INbreast database, the best single model achieved an AUC per image of 0.95 and the mean of four models improved the AUC to 0.98 (sensitivity: 86.7%, specificity: 96.1%). We also show that a full image classifier trained with our end-to-end approach on digitized CBIS-DDSM film mammograms can be transferred to INbreast-FFDM images using only a subset of the INbreast data for fine-tuning will and without having to rely further on the availability of infringement notes. These results show that deep machine learning methods can be easily trained to achieve high accuracy on heterogeneous mammography platforms and show promise for improving clinical tools to detect false-positive and false-negative screening mammography results to reduce.(37)

2.1.3 Boosting Breast Cancer Detection Using Convolutional Neural Network

Breast cancer forms in breast cells and is considered a very common type of cancer in women. Breast cancer is also a very life-threatening disease for women after lung cancer. In this study, a convolutional neural network (CNN) method is proposed to advance automatic breast cancer detection by analysing hostile ductal carcinoma tissue zones in whole-slide images (WSI).(38) The paper examines the proposed system, which uses different Convolutional Neural Network (CNN) architectures, to automatically detect breast cancers and compares the results to those of machine learning (ML) algorithms. All architectures

were guided by a large data set of approximately 275,000 50 × 50pixel RGB tiles. Validation testing of the quantitative results was performed using the performance indicators for each method. The proposed system was found to be successful, yielding results with an accuracy of 87%, which could reduce human error in the diagnostic process. Furthermore, our proposed system achieves an accuracy of, which is higher than the accuracy of 78 of machine learning (ML) algorithms. Therefore, the proposed system improves the accuracy by 9 over the results of machine learning (ML) algorithms.(39)

2.1.4 IDC Breast Cancer Detection Using Deep Learning Schemes

In recent years, Deep Learning (DL) architectures have been deployed in many potential areas such as: B. Object recognition, face recognition, natural language processing, medical image analysis and other related applications. In these applications, DL has achieved remarkable results that match the performance of human experts. This article presents a new convolutional neural network (CNN)-based approach to diagnose breast cancer in IDC tissue regions using WSI. It has been established that breast cancer is one of the leading causes of death in women. It also remains a difficult task for the pathologist to find the malignant regions of WSI. In this research, we implement different CNN models including VGG16, VGG19, Xception, Inception V3, MobileNetV2, ResNet50 and DenseNet. The experiments were performed on a standard WSI slide dataset comprising 163 IDC patients. For performance evaluation, the same dataset was split into 113 and 49 images for training and test, respectively. The test was performed for each model separately and the obtained results showed that our proposed CNN model achieved an accuracy of 83, which is better than the other models.(40)

3. AIM AND OBJECTIVE

3.1 AIM:

To detect the presence of cancer cells if any, from the provided histopathology images of human breast using deep learning model.

3.2 OBJECTIVE:

Classify - a histopathology image to detect the presence of cancer.

Formulate – a deep neural network model for the detection of normal & abnormal breast cells.

Evaluate - the accuracy of the model & report the metrics.

4.MATERIALS AND METHODS

4.1 TABLE : TOOLS/SOFTWARE/DATABASE

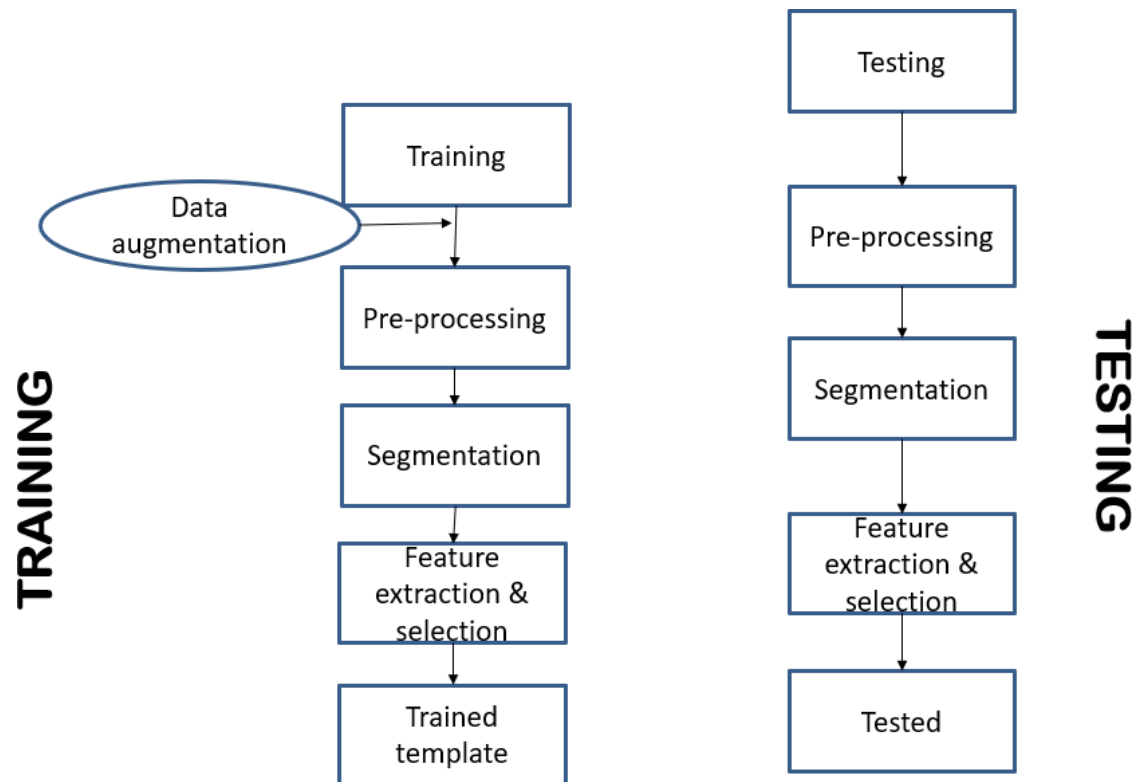
S. No	TOOLS/ SOFTWARES/DATABASES	APPLICATIONS	APPLICATION IN MY PROJECT
1	Kaggle Dataset	To find and publish datasets, uses GPU integrated notebooks, build and publish models in a web – based data science environment.	To find and build models for implementing machine learning models using breast cancer dataset.
2	Python	A Programming Language. To create web – server applications.	An open – source library used for image preprocessing functions.
3	PyTorch	An open-source machine learning framework that accelerates the path from research prototyping to production deployment.	An open – source library used for image preprocessing functions.
4	Pandas	An open- source python package, used for data analysis/ data science and machine learning tasks.	It is mainly used for working with relational and labelled data. Used to provide data structures and operations for manipulating numerical data and time series.

5	Seaborn	Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.	It is used for viewing the statistical distribution of dataset being used for the classification of cancer and non-cancerous HPE images
6	Matplotlib	A python library used to create 2D graphs and plots by using python scripts.	To perform Data Visualization and Graph plotting Library.
7	Tensor Flow	It is an open – source library primarily for deep learning applications and also supports machine learning models.	Used to develop and train machine learning models.

8	Sklearn	Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistency interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.	To classify the dataset as of based on the presence of cancer tissues s binary data and to graphically represent them to visualize the range of dataset being used for the model development
9	PIL	Python Imaging Library is a free and open-source additional library for the Python programming language that adds support for opening, manipulating, and saving many different image file formats.	Here the conversion of file formats and segmentation as well as augmentation takes place in a wide range of array, thereby PIL helps in controlling the issues thus being faced

10	GLOB	<p>In Python, the glob module is used to retrieve files/pathnames matching a specified pattern. The pattern rules of glob follow standard Unix path expansion rules. It is also predicted that according to benchmarks it is faster than other methods to match pathnames in directories. With glob, we can also use wildcards ("*,?,) [ranges]) apart from exact string search to make path retrieval more simple and convenient.</p>	<p>To easily retrieve the files according to the requirement in any cause of time and at any instant the GLOB module is being utilized</p>
----	-------------	--	--

4.2 Methodology



4.3 KAGGLE :

Kaggle is an online based community platforms for data scientists and enthusiasts. It allows users to collaborate with other users to find and publish datasets and used to solve data science problems. It is used to explore and build models in web-based data science environment. Kaggle works with other data scientists, machine learning engineers and enter competitions to challenge data science problems. Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public platform and now also offers a public data platform, a cloud-based workbench for data science. (41)

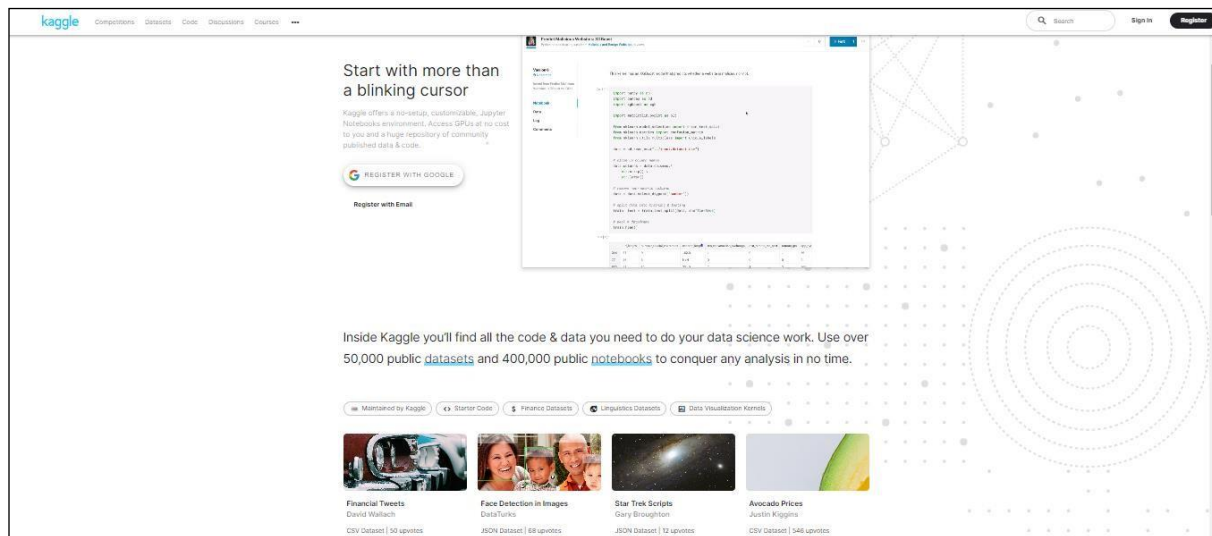


Fig:4.1 Kaggle Homepage

4.4 JUPYTER NOTEBOOK:

Jupyter notebook is the free – software, open source and web services for interactive computing across all programming languages. It also contains Jupyter lab which is the latest web- based interactive development environment for notebooks, code and data. Its flexible interface allows users to configure and arrange workflows in datascience, scientific computing, computational journalism and machine learning . It is the original web – application for creating and sharing computational documents.(42)

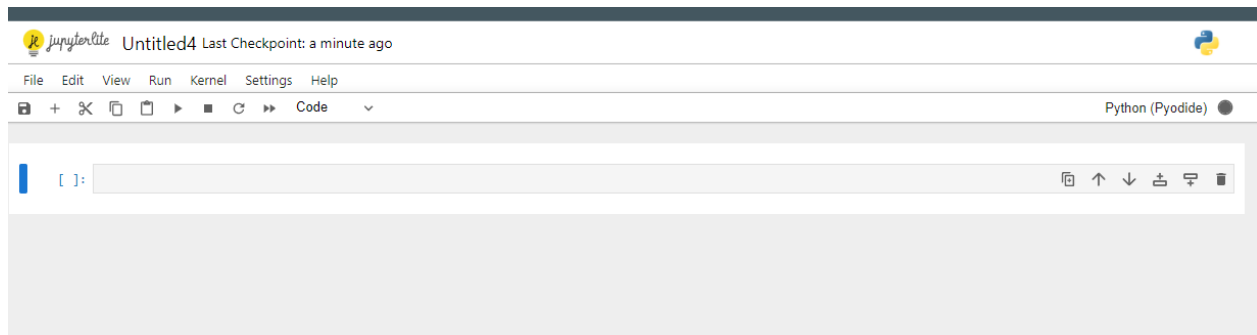


Fig:4.2 Jupyter Notebook Homepage

For Implementing Machine Learning Models:

Step 1 : Importing the Libraries:

The required packages has and libraries has to be installed and imported .

```
In [1]: # importing libraries
import numpy as np # Linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
```

After installing numpy and pandas package, we are ready to fetch data using pandas package.

Step 2 : Data Collection:

The data has to gathered and collected and used to read into the dataframe.

```
Data Collection

In [2]: # reading data into the dataframe
df = pd.read_csv('data.csv')
```

```
In [3]: df
```

```
Out[3]:
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15960	0.19740	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	...
...
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	...
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	...
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	...
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	...
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	...

569 rows x 33 columns

Step 3 : Describing the dataframe:

Shape () :

It tells us the number of rows and columns of a given dataframe.

```
In [4]: # Cell 2
df.shape
```

```
Out[4]: (569, 33)
```

Head():

It is used to access the first n rows of the dataframe.

```
In [5]: # displaying first five rows
df.head()
```

Out[5]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	te
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27790	0.3001	0.14710
1	842517	M	20.57	17.77	132.90	1326.0	0.09474	0.07864	0.0869	0.07017
2	84300903	M	19.69	21.25	130.00	1203.0	0.10990	0.16990	0.1974	0.12790
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

5 rows x 33 columns

Tail():

It is used to display the last n rows of the dataframe.

```
In [6]: df.tail()
```

Out[6]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	te
964	926424	M	21.56	22.39	142.00	1479.0	0.11500	0.11590	0.24390	0.13890
965	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791
966	926954	M	16.80	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302
967	927241	M	20.80	29.33	140.10	1265.0	0.11790	0.27700	0.35140	0.15200
968	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000

5 rows x 33 columns

Step 4 : Data Exploration

Summary of the dataframe:

Information of the dataframe.

```
In [7]: # concise summary of dataframe
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 969 entries, 0 to 968
Data columns (total 33 columns):
 #   Column              Non-Null Count  Dtype
---  -
 0   id                  969 non-null    int64
 1   diagnosis           969 non-null    object
 2   radius_mean         969 non-null    float64
 3   texture_mean        969 non-null    float64
 4   perimeter_mean      969 non-null    float64
 5   area_mean           969 non-null    float64
 6   smoothness_mean     969 non-null    float64
 7   compactness_mean    969 non-null    float64
 8   concavity_mean      969 non-null    float64
 9   concave points_mean 969 non-null    float64
10   symmetry_mean       969 non-null    float64
11   fractal_dimension_mean 969 non-null    float64
12   radius_se           969 non-null    float64
13   texture_se          969 non-null    float64
14   perimeter_se        969 non-null    float64
15   area_se             969 non-null    float64
16   smoothness_se       969 non-null    float64
17   compactness_se      969 non-null    float64
18   concavity_se        969 non-null    float64
19   concave points_se   969 non-null    float64
20   symmetry_se         969 non-null    float64
21   fractal_dimension_se 969 non-null    float64
22   radius_worst        969 non-null    float64
23   texture_worst       969 non-null    float64
24   perimeter_worst     969 non-null    float64
25   area_worst          969 non-null    float64
26   smoothness_worst    969 non-null    float64
27   compactness_worst   969 non-null    float64
28   concavity_worst     969 non-null    float64
29   concave points_worst 969 non-null    float64
30   symmetry_worst      969 non-null    float64
31   fractal_dimension_worst 969 non-null    float64
32  Unnamed: 32          0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

Displaying the columns:

The total number of columns are displayed.

```
In [8]: # column names
df.columns

Out[8]: Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
              'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
              'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
              'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
              'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
              'fractal_dimension_se', 'radius_worst', 'texture_worst',
              'perimeter_worst', 'area_worst', 'smoothness_worst',
              'compactness_worst', 'concavity_worst', 'concave points_worst',
              'symmetry_worst', 'fractal_dimension_worst', 'unnamed: 32'],
              dtype='object')
```

Dropout the columns:

The unwanted columns are removed using dropout command .

```
In [9]: # dropping 'unnamed: 32' column.
df.drop("unnamed: 32", axis=1, inplace=True)

In [10]: # dropping id column
df.drop('id', axis=1, inplace=True)
```

Describing the dataframe:

It counts / describes the number of columns but not the empty values.

```
In [11]: # descriptive statistics of data
df.describe()

Out[11]:
```

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fract
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	
mean	14.127292	19.269649	91.969033	654.859104	0.096360	0.104341	0.085799	0.048919	0.181162	
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038603	0.027414	
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	
50%	13.370000	18.840000	86.240000	551.100000	0.096870	0.092630	0.061540	0.033600	0.179200	
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.428800	0.201200	0.304000	

8 rows x 30 columns

Checking the Missing values:

This function takes scalar or array like an object and indicates whether the values are missing.

```
In [12]: df.isna()
```

Out[12]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False
...
564	False	False	False	False	False	False	False	False	False	False
565	False	False	False	False	False	False	False	False	False	False
566	False	False	False	False	False	False	False	False	False	False
567	False	False	False	False	False	False	False	False	False	False
568	False	False	False	False	False	False	False	False	False	False

569 rows x 11 columns

```
In [13]: df.isna().any()
```

Step 5: Data Visualization:

It is the graphical representation of data and information in a pictorial (or) graphical format.

The visualization of data can be done using the matplotlib library.

```
In [19]: import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go

%matplotlib inline
sns.set_style('darkgrid')
```

```
In [20]: x = df.iloc[:,1:11] #independent columns
y = df.iloc[:,0] #target column i.e price range

correlation
```

```
In [21]: # with the following function we can select highly correlated features
# it will remove the first feature that is correlated with anything other feature

def correlation(dataset, threshold):
    col_corr = set() # Set of all the names of correlated columns
    corr_matrix = dataset.corr()
    for i in range(len(corr_matrix.columns)):
        for j in range(i):
            if (corr_matrix.iloc[i, j]) > threshold:
                colname = corr_matrix.columns[i] # getting the name of column
                col_corr.add(colname)
    return col_corr
```

```
In [22]: corr_features = correlation(x, 0.85) #85% is used generally
len(set(corr_features))
```

Out[22]: 11

```
In [23]: corr_features
```

Out[23]: {'area_mean',
 'area_se',
 'area_worst',
 'compactness_worst',
 'concave points_mean',
 'concave points_worst',
 'concavity_mean',
 'concavity_worst',
 'perimeter_mean',
 'perimeter_se',
 'perimeter_worst',
 'radius_worst',
 'texture_worst'}

Step 6 : Implementing Deep Learning Models:

Implementing Deep learning Model:

Importing the libraries:

```
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib as mpl
import matplotlib.pyplot as plt

import tensorflow as tf
from tensorflow import keras
```

Loading the Dataset:

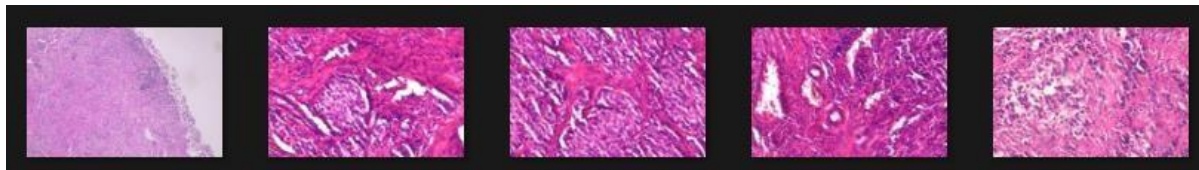
The train, test and validation data has been successfully uploaded from their appropriate path.

```
train_dir='C:/Users/Vamsi/Documents/breast/data/train'
validation_dir='C:/Users/Vamsi/Documents/breast/data/validation'
test_dir='C:/Users/Vamsi/Documents/breast/data/test'

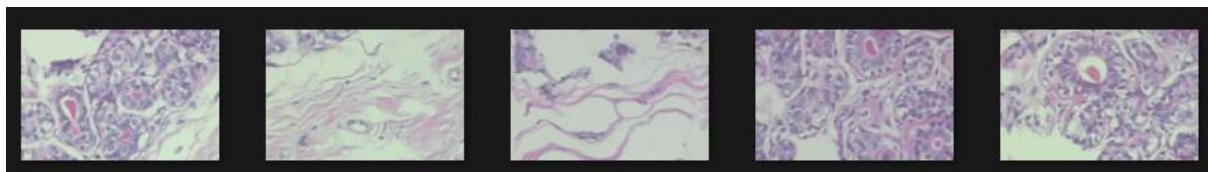
from tensorflow.keras.preprocessing.image import ImageDataGenerator
```

Visualization of Dataset:

Malignant Images



Benign Images



Data Augmentation:

It is a set of techniques which is artificially used to increase the amount of data by generating new data points from existing data.

Predicting the Model:

RELU Activation Factor:

The rectified linear activation unit (RELU) for short is a piecewise linear function that will output the input directly if it is positive , otherwise the output will be zero. This activation function in neural network is easier to train the model and achieves the better performance.

F1- Score:

The F1 score combines the precision and recall score of a classifier into a single metric by taking their harmonic mean. It is mainly used to compare the performance of two classifiers.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where,

- TP – true positive
- TN – true negative
- FP – false positive
- FN – false negative

Heat Map:

It represents the coefficients to visualize the strength of correlation among variables. It helps find features that are best for building machine learning model. It transforms the correlation matrix into color coding the dark color being the highest area of interaction and the cool colors point to the section with the lowest interaction.

ROC curve:

Displays an ROC (Receiver Operating Characteristic) curve for each categorical dependent variable. It also displays a table giving the area under each curve. For a given dependent Variable, the ROC chart displays one curve for each category.

If the dependent variable has two categories, then each curve treats the category at issue as the positive state versus the other category. If the dependent variable has more than two categories, then each curve treats the category at issue as the positive state versus the aggregate of all other categories.

Synaptic weights

Displays the coefficient estimates that show the relationship between the units in a given layer to the units in the following layer. The synaptic weights are based the training sample even if the active dataset is partitioned into training, testing, and holdout data. Note that the number of synaptic weights can become rather large and that these weights are generally not used for interpreting network results

Model summary - Displays a summary of the neural network results by partition and overall, including the error, the relative error or percentage of incorrect predictions, the stopping rule used to stop training, and the training time. The error is the sum-of-squares error when the identity, sigmoid, or hyperbolic tangent activation function is applied to the output layer. It is the cross-entropy error when the softmax activation function is applied to the output layer.

5. RESULTS



Fig: 5.1 - Exploratory Data Analysis

The distribution of the dataset as per the requirement is thus split and shown in the statistical form, as that of cancerous and non-cancerous data

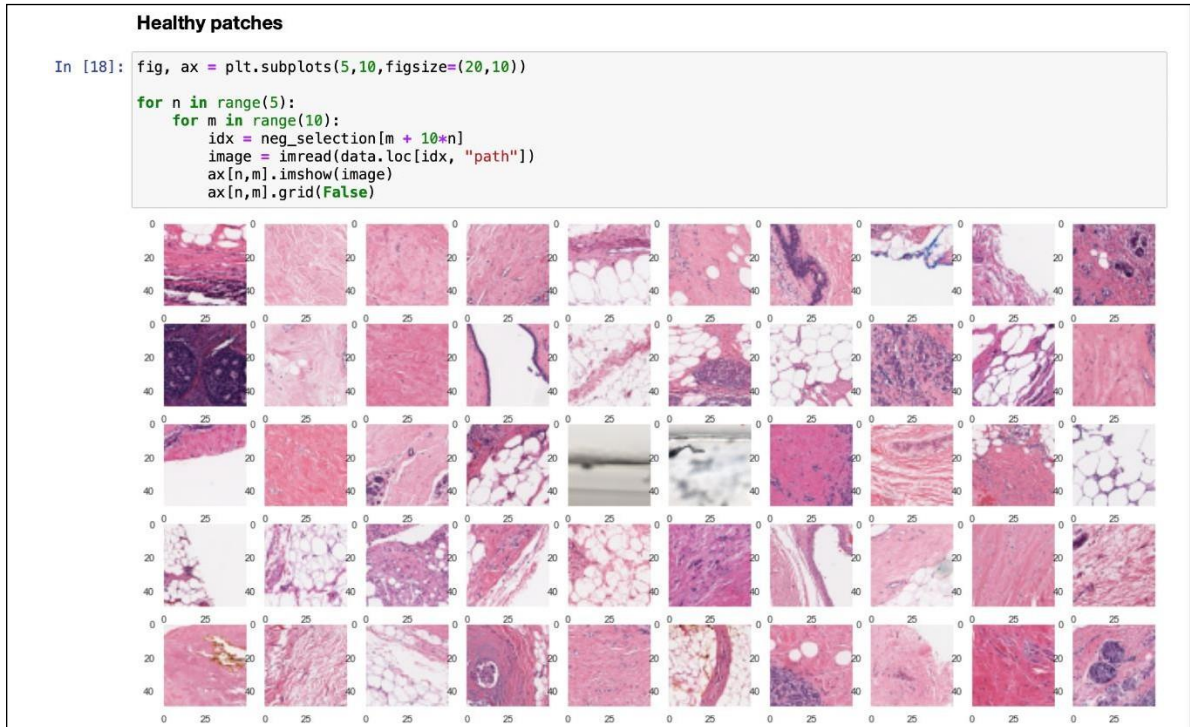


Fig: 5.2 - Display of Healthy Patches

The dataset thus depicting as of how the healthy patches of the segmented data looks like is thus viewed by displaying a maximum of 50 healthy patches of the tissues

```
In [21]: fig, ax = plt.subplots(5,3,figsize=(20, 27))

patient_ids = data.patient_id.unique()

for n in range(5):
    for m in range(3):
        patient_id = patient_ids[m + 3*n]
        example_df = get_patient_dataframe(patient_id)

        ax[n,m].scatter(example_df.x.values, example_df.y.values, c=example_df.target.values, cmap="coolwarm", s=20)
        ax[n,m].set_title("patient " + patient_id)
        ax[n,m].set_xlabel("y coord")
        ax[n,m].set_ylabel("x coord")
```

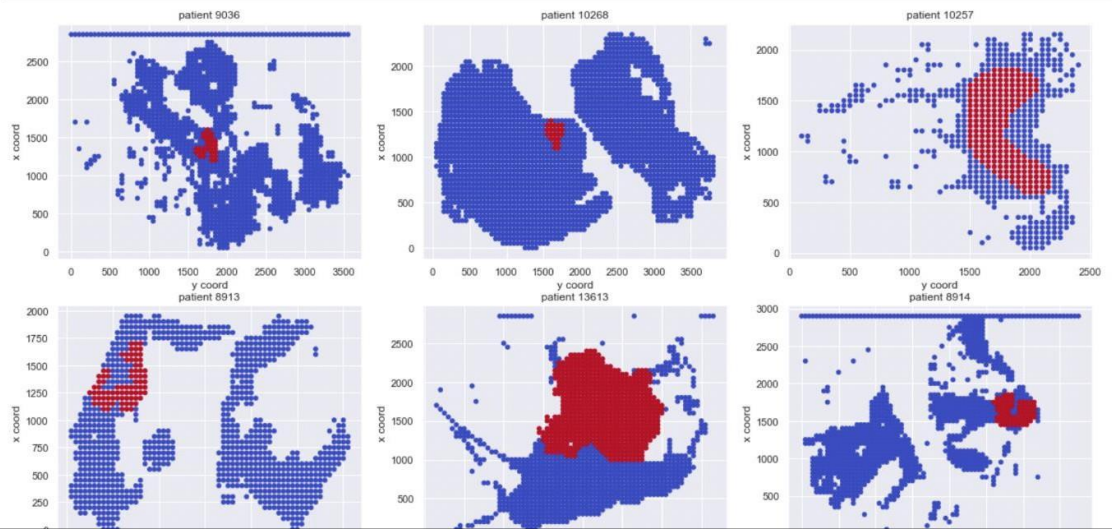


Fig: 5.3 - Binary Data Visualization Per Tissue Slice

Hence the connected patches of both healthy and cancerous patches which when segmented and trained after joining together shows exactly the spot of the cancer in a WSI

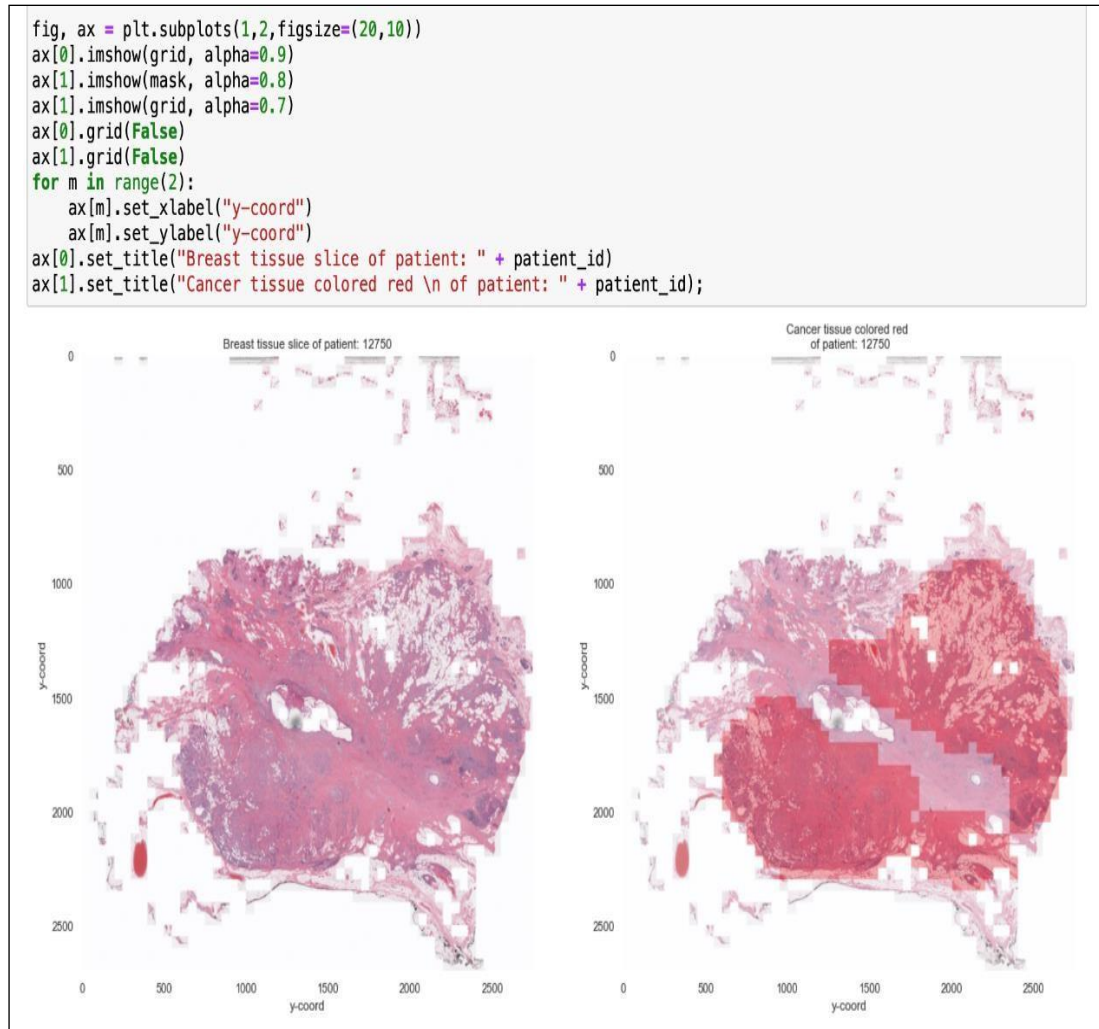


Fig: 5.4 - Grid Highlighting the Cancerous part being spotted from the patches

Hence the connected patches of both healthy and cancerous patches which when segmented and trained after joining together shows exactly the spot of the cancer in a WSI is thus predominantly viewed by zooming in a WSI to view the HPE of a patient

```
In [32]: fig, ax = plt.subplots(1,3,figsize=(20,5))
sns.countplot(train_df.target, ax=ax[0], palette="Reds")
ax[0].set_title("Train data")
sns.countplot(dev_df.target, ax=ax[1], palette="Blues")
ax[1].set_title("Dev data")
sns.countplot(test_df.target, ax=ax[2], palette="Greens");
ax[2].set_title("Test data");
```

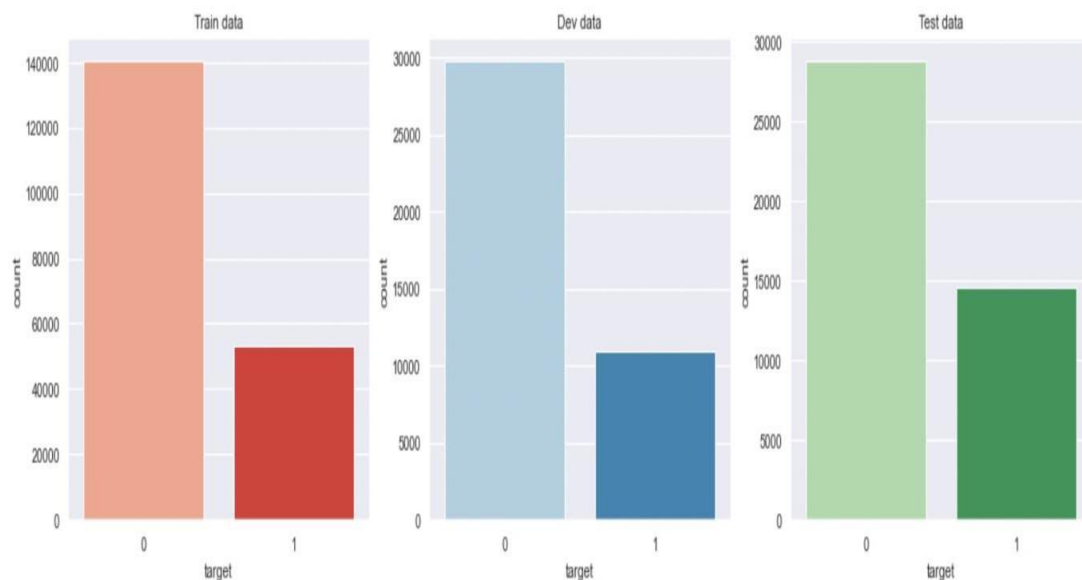


Fig: 5.5 - Target distributions difference of the datasets

Image displays the involvement of the test, develop and the training dataset and their segregation in the model training and testing to evaluate the efficiency

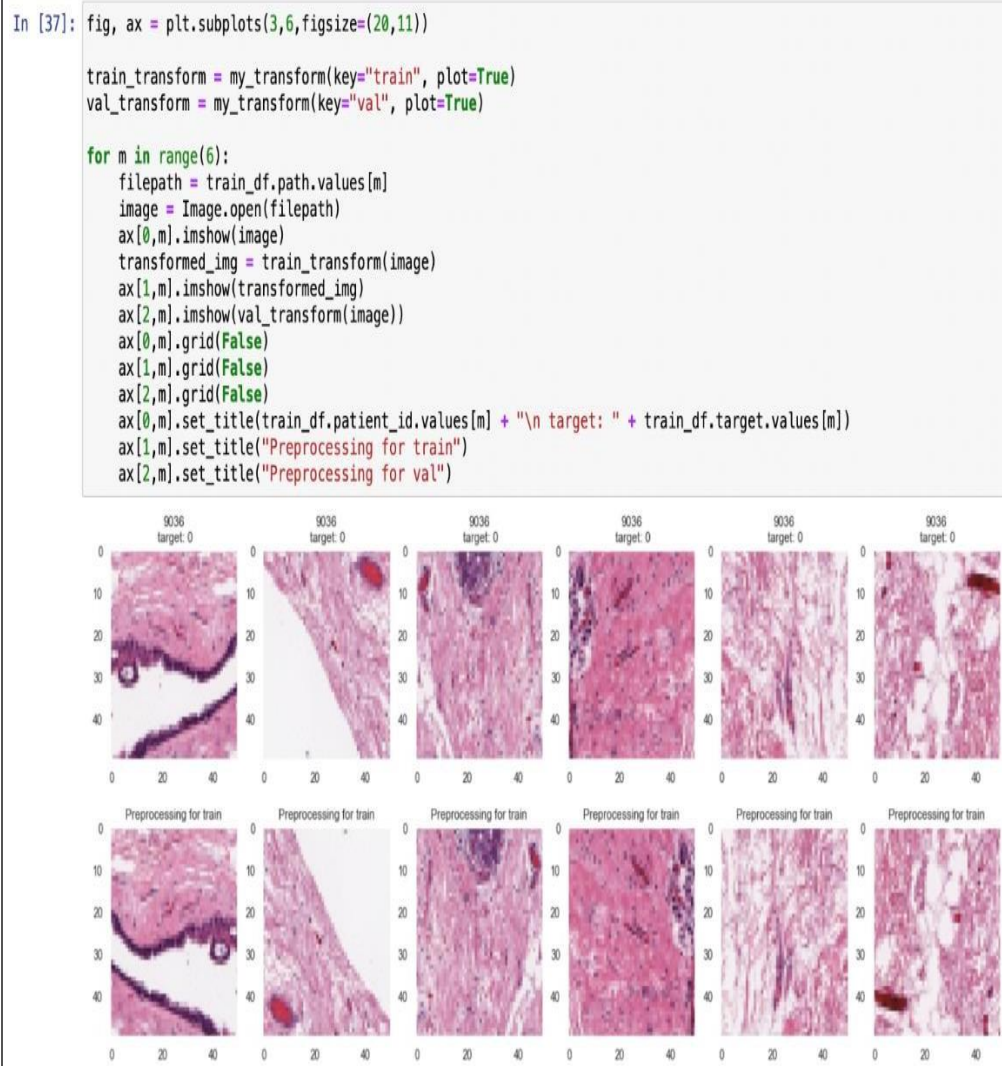


Fig: 5.6 - Data augmentation using the Pytorch dataloaders

Deep learning involves the main task to train a huge amount of dataset to test a model's efficiency, hence flipping of data for the process of data augmentation is thus displayed

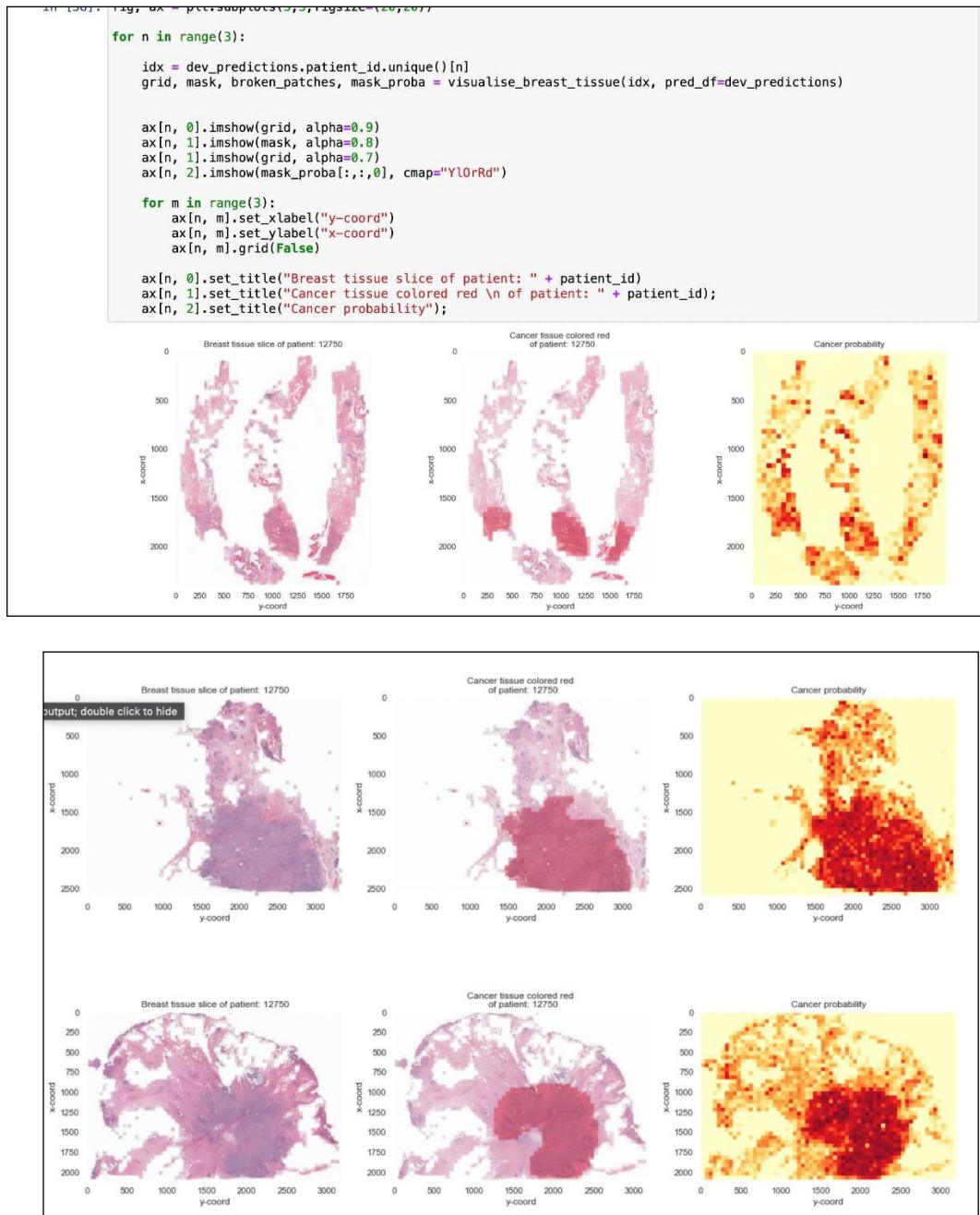


Fig: 5.7 - Sklearn module to highlight the grid that is spotted with the cancer cells

Data clustering, segmentation and regression is thus the terminal process involved wherein the segmented patches is thus collected together as a whole to know the amount of cancer spread to review its metastatic condition that is taking place

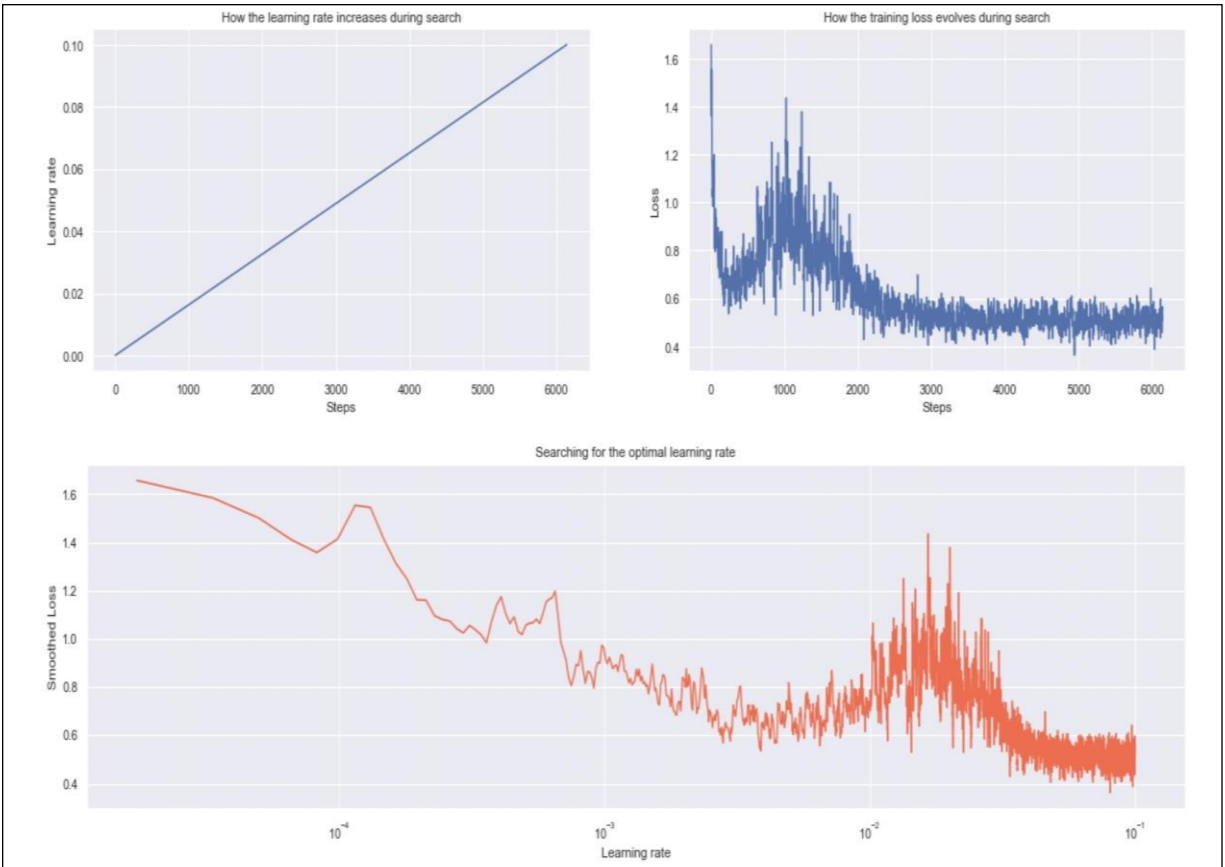


Fig: 5.8 - Search For An Optimal Cyclic Learning Rate

Displaying the learning rate of the training model and thus checking its efficiency after it has been pre-trained by the developing data and thus by improving the epochs, an observing is been seen that how far the model is able to learn and depict its results

```
In [54]: plt.figure(figsize=(20,5))
```

```
plt.plot(losses_df["train"], '-o', label="train")  
plt.plot(losses_df["dev"], '-o', label="dev")  
plt.plot(losses_df["test"], '-o', label="dev")  
plt.xlabel("Epoch")  
plt.ylabel("Weighted x-entropy")  
plt.title("Loss change over epoch")  
plt.legend();
```

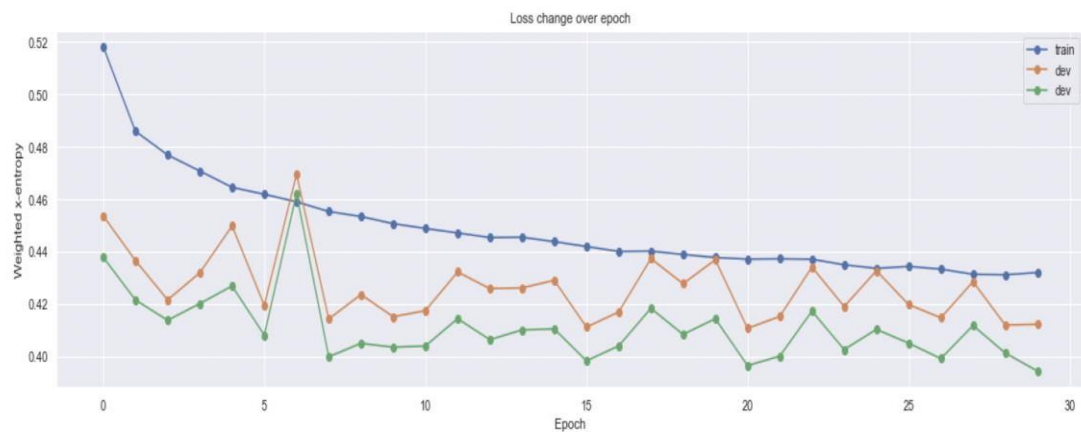


Fig: 5. 9 - Loss convergence of training, developing & test dataset

The loss or the noise that is still being captured inspite of the training that is being done is thus noted in a single statistical view to compare the train, developing and test model

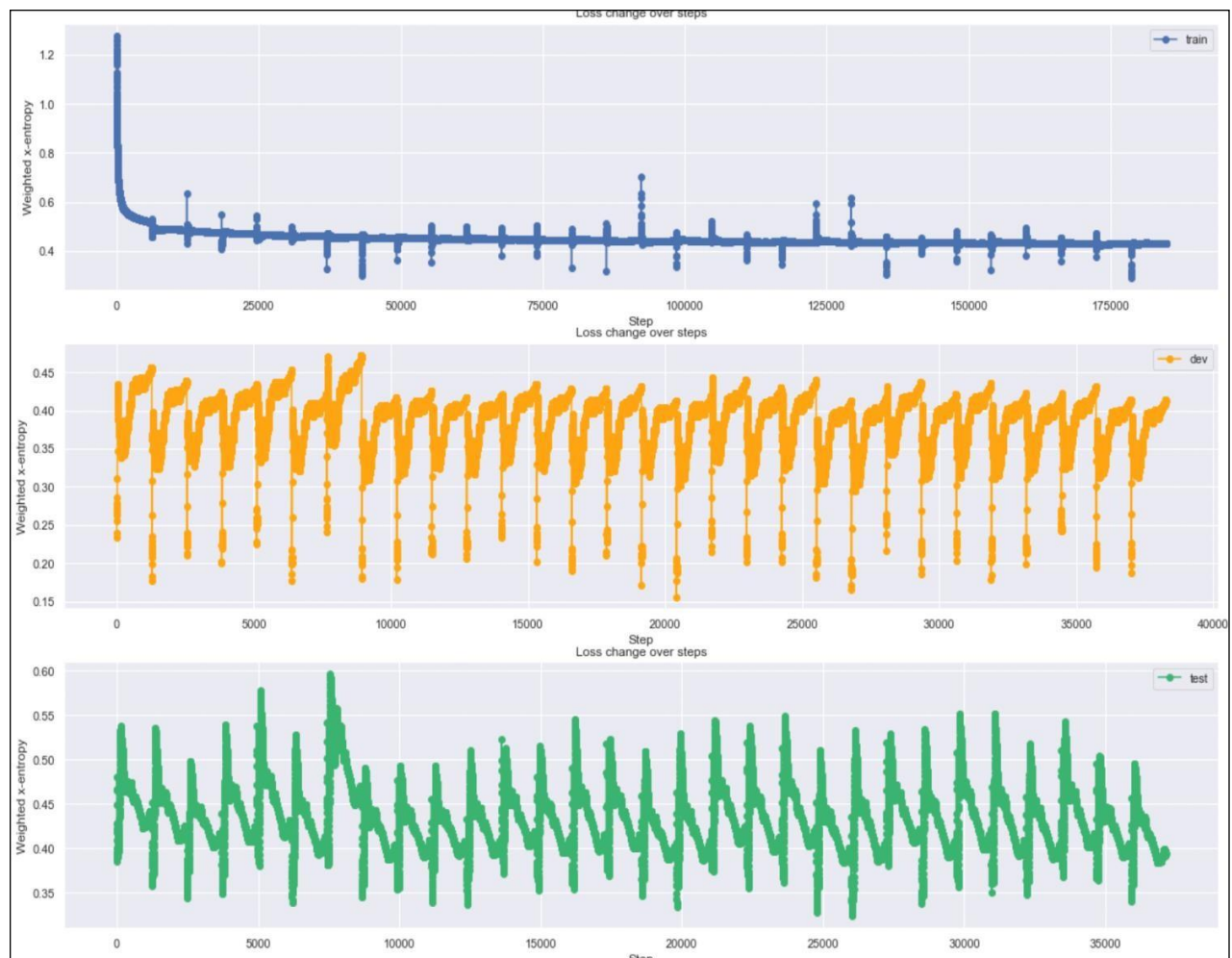


Fig: 5.10 - Loss Convergence – The Running Losses

The loss or the noise that is still being captured inspite of the training that is being done is thus noted in a single statistical view to compare the train, developing and test model

```
In [60]: fig, ax = plt.subplots(1,3,figsize=(20,5))
sns.countplot(dev_predictions.true.astype(np.float), ax=ax[0], palette="Reds_r")
ax[0].set_title("Target counts of dev data");
sns.distplot(dev_predictions.proba.astype(np.float), ax=ax[1], kde=False, color="tomato")
ax[1].set_title("Predicted probability of cancer in test");
sns.countplot(test_predictions.true.astype(np.float), ax=ax[0], palette="Reds_r")
ax[0].set_title("Target counts of dev data");
sns.distplot(test_predictions.proba.astype(np.float), ax=ax[2], kde=False, color="mediumseagreen");
ax[2].set_title("Predicted probability of cancer in test");
```

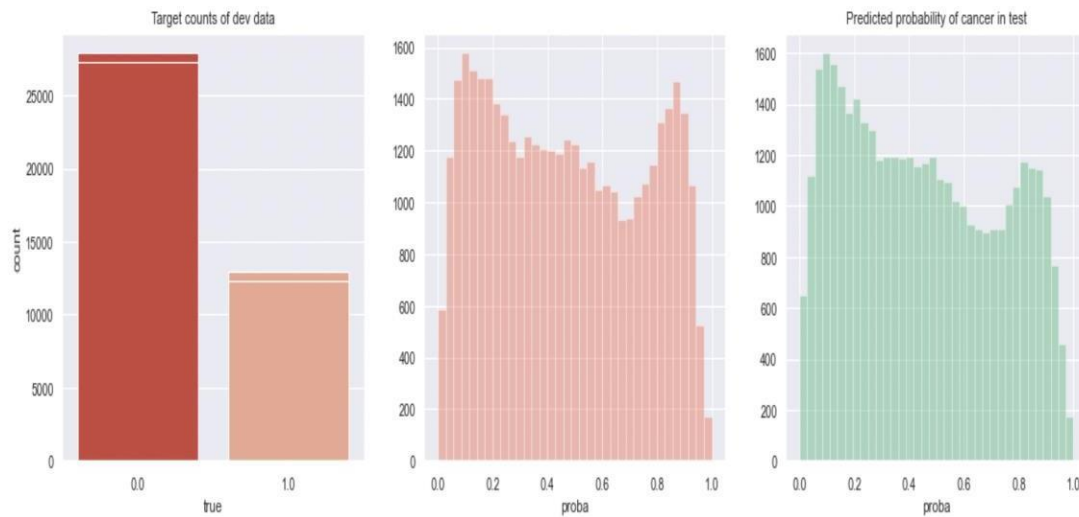


Fig: 5.11 - Sklearn Plots

Statistical Representation Of The Normal Vs The Cancerous Dataset

The probability distribution of the dataset as per the requirement is thus split and shown in the statistical form, as that of cancerous and non-cancerous data

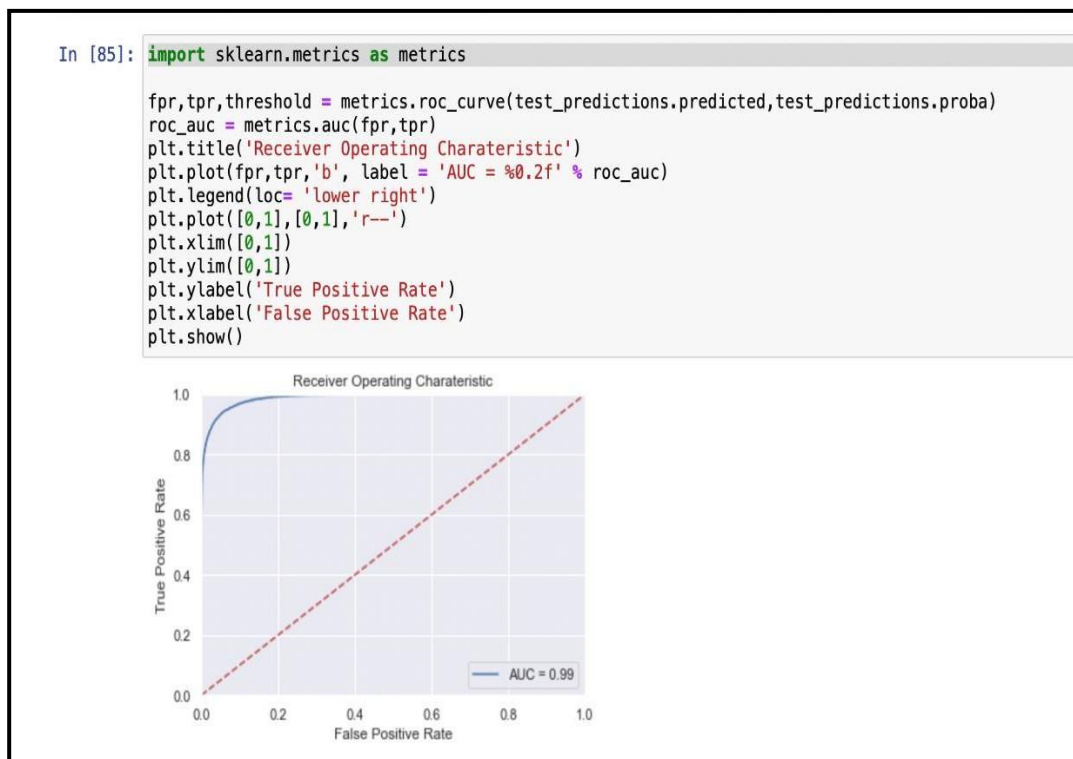


Fig: 5.12 - ROC curve for the ResNet model

ROC curve depicting a deep learning model's architecture is thus noted here which is being displayed here in this graphical view

6. DISCUSSION

6.1 Breast Metastasis

Metastatic breast cancer is the most advanced stage of breast cancer. Breast cancer occurs when abnormal cells in the breast start dividing uncontrollably. A tumour is huge collection of these abnormal cells. Metastases refer to cancer cells that have spread to a new area of the body. In metastatic breast cancer, cells can spread to:

- Bones
- Brain
- Liver
- Lungs

Healthcare providers name cancer after its primary cause. This means breast cancer that spreads to other parts of the body is still considered breast cancer. The cancer cells are yet continued to be breast cancer cells despite their spread. Your treatment team will use breast cancer therapies even if cancer cells are found in other areas. These two terms – breast metastasis as well as stage 4 breast cancer means essentially the same thing. Breast cancer classified as stage 4 has spread outside the breast, or metastasized, to other parts of the body.

6.2 Need of the study

- Four major challenges in breast cancer treatment and management— diagnostic accuracy; risk stratification, particularly for application of active surveillance and focal therapy
- At present standard imaging techniques, such as ultrasound, MRI, CT, and nuclear medicine, cannot detect early disease, besides which they also provide limited information in terms of disease staging.

7. CONCLUSION

7.1 Conclusion

The Convolutional Neural Network Model for differential diagnosis of Normal, Breast Cancer Invasive Ductal Carcinoma with a classifying accuracy of 95% was found to perform better than the former model for the same purpose with a classifying accuracy of 86%.

Thereby implementing DL models for the diagnosis purpose in the clinical practice using image based data will be an upper hand in the domain of both diagnosis as well as prevention of diseases to further proliferate.

8. REFERENCES

Journal References:

1. Singh, D., Singh, A.K.: Role of image thermography in early breast cancer detection-past, Present and future. *Computed Methods Programs Biomed.* (2020)
2. E. Halim, P.P. Halim, M. Herbart, Artificial intelligent models for breast cancer early detection, in 2018.
3. Rawla P. Epidemiology of Breast Cancer. *World J Oncol.* 2019; 10(2):63–89.
4. Aladdin, E.: *Introduction to Machine Learning*, 2nd edh. MIT Press, Cambridge (2010) (<https://www.ncbi.nlm.nih.gov/books/NBK284958/>).
5. Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, Peng X (2021) Predicting breast cancer 5-year survival using machine learning: A systematic review.

Web References:

6. <https://www.pcf.org/about-breast-cancer/diagnosis-staging-breast-cancer/breast-cancer-/>.
7. <https://oncohemakey.com/anatomy-and-pathology-of-breast-cancer/>.
8. Isaacs W, Kainu T. Oncogenes and Tumor Suppressor Genes in BreastCancer. *Epidemiologic Reviews.* 2001 Jan 1;23(1):36–41.
9. <https://breastcancernewstoday.com/breast-cancer-overview/>.
10. <https://www.cancer.net/cancer-types/breast-cancer/risk-factors-and- prevention>.
11. <https://www.cancer.net/cancer-types/breast-cancer/symptoms-and-signs>.
12. <https://speciality.medicaldialogues.in/breast-cancer-cases-to-double-by- 2020- study>.
13. <https://www.slideshare.net/FarazaJaved/breast-cancer-79088572>.
14. <https://blog.crownbio.com/breast-cancer-preclinical-models>. In.
15. <http://genesdev.cshlp.org/content/32/17-18/1105/F2.expansion.html>.
16. <https://www.intechopen.com/books/pathophysiology-altered-physiological-states/an-overview-on-breast-pathophysiology-new-insights-into-breast- cancer->

clinical-diagnosis.

17. <https://www.ncbi.nlm.nih.gov/books/NBK65915/>.

18. <https://pubmed.ncbi.nlm.nih.gov/14580855/>.

19. https://www.sas.com/en_in/insights/analytics/machine-learning.html.

20. Breast cancer: diagnosis and management. Breast cancer. :54.

21. <https://ai.plainenglish.io/different-types-of-machine-learning-algorithms-28974016e108>.

22. <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02231/full>.

23. <https://www.javatpoint.com/machine-learning>.

24. <https://www.geeksforgeeks.org/neural-network-advances/#:~:text=More%20precision%20in%20cancer%20treatment,many%20different%20kinds%20of%20tumour>.

25. <https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207>.

26. https://www.researchgate.net/figure/Architecture-of-a-single-neuron-in-a-neural-network-5_fig2_326826129.

27. <https://www.sciencedirect.com/science/article/abs/pii/S0893608003001163>.

28. Gurney K. Introduction to Neural Networks. Oxford: Taylor & Francis;

29. <https://www.kaggle.com/datasets/sebastianrieichert/lbreastcancer-diagnosis>

30. <https://www.verywellmind.com/what-is-a-dependent-variable-2795099>.

31. <https://www.advernesia.com/wp-content/uploads/2018/02/nominal-ordinal-danscale-pada-SPSS.png>.

32. <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>

33. <https://bookdown.org/rdpeng/advstatcomp/random-number-generation.html>.

34. <https://machinelearningmastery.com/how-to-develop-convolutional-neural-models-for-image>

classification/.

