Breast Cancer Classification by Implementation of Deep-Learning with Dataset Analysis

Saheel Patil
Department of Electronics Engineering
KJSIEIT
Mumbai, India
saheel.patil@somaiya.edu

Akshay Pashte
Department of Electronics Engineering
KJSIEIT
Mumbai, India
akshay.pashte@somaiya.edu

Satyam Rai
Department of Electronics Engineering
KJSIEIT
Mumbai, India
satyam.r@somaiya.edu

Prof. Sejal Shah
Department of Electronics Engineering
KJSIEIT
Mumbai, India
sshah@somaiya.edu

2022 5th International Conference on Advances in Science and Technology (ICAST) | 978-1-6654-9263-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAST55766.2022.10039480

***Abstract*—Cancer is a fatal disease recognized and researched about, around the globe. Researchers and scientists have been investing their time and imparting their expertise, and knowledge for the advancements of traditional methods and treatments to tackle it. Recent surveys reveal that the mortality rate among the female populous, over the world, is also one of the results of breast cancer. The definition of breast cancer can be described as an uncontrolled aggressive growth of old cells which thereby aid the formation of a pernicious mass in the tissue of a breast. Gradually, this may result in the formation of a tumor of malignant nature. Deep learning, considered a sub-field of Machine Learning, enables experts to analyze, model, and study complicated or rather complex scientific data over a comprehensive list of medical applications. This study aims to create a user-friendly, adept system to perform the classification of breast tumors of malignant or benign nature. The proposed system is divided into two halves or stages. The initial stage is the pre-processing and analysis of the acquired dataset which also involves training of the neural network. The next and final stage is the classification of breast tumors by utilizing the created model and loading it onto an API through which users can upload tissue images and check what type of breast cancer the tissue contains. This would eliminate the time spent on studying every particular data using traditional clinical methods. This project would help support the radiologists in training, research, and diagnostic aspects and overall support the entire process of cancer diagnosis and treatment.**

Keywords—Machine Learning, Convolution Neural Networks (CNN), convolution

I. INTRODUCTION

The in-depth study of microscopic slides of the different human organs and their tissues, in biomedical research and studies play an important role in the knowledge of the human physiology. The annotation of various image slides (such as molecules, ducts, etc.) is of notable relevance amongst the microscopic image examination rules. Various applications have been created for the purpose of categorization of microscopic images. Breast cancer is one of the most pervasive major reason of mortality around the globe in women of the young adults to mid-aged demographic i.e., 20 to 59 years. The survival rate from this disease can subject to be increased to 80%, if discovered in initial stages. There are different diagnosis methods used for the finding of breast cancer in a patient. Two major procedures of diagnosis are mammography and biopsy. In mammography, a breast is subjected to controlled radiation under the supervision of a radiologist. The breast is flattened as much as it could, considering the discomfort of the patient, under two plates and the image is taken. Through this, early signs of cancer can be diagnosed. The use of mammogram for examination has noted to be crucial in reducing the fatality ratio. Biopsy is another immensely effective and solid groundwork and approach for a diagnosis of breast cancer. This procedure involves extracting a sample of cellular tissue from the mass affected of the breast which is then inspected with the help of a scientific microscope by an experienced, knowledgeable pathologist for the evaluation of cellular tissue extracted and categorization of tumor present. Biopsy is being known to be a crucial method of diagnosis for breast cancer and other forms of cancer as well. Through these detection methodologies, a medical practitioner can determine and provide clarification on two categories of lesion i.e., benign and malignant. Lesion of benign nature is less-likely to be considered carcinogenic. There happen to be anomalies in the epithelial cells which are unable to evolve to be a root cause of breast cancer. Though the malignant or carcinogenic cells have an aberrant division and irregular expansion among the breast tissue, it is an exceptionally tedious and complicated procedure to examine the histopathology images manually, considering the peculiar and unusual impression of normal and malignant cells. In past years, various field-practitioners and researchers advised alternative methodologies for automated classification to identify cancerous cells in cytology images. Several medical researchers have aimed to achieve the analysis of nuclei by extraction of attributes from the nuclei to obtain substantial information for the categorization of cells into benign or malignant type. Likewise, methods based on clustering amalgamated with circular Hough Transform and several statistics are also dabbled with to achieve segmentation and categorization of the nucleus. Algorithms about the analysis of histopathological images are rapidly growing albeit the need for an automated system, achieving highly accurate and efficient results.

Deep learning is used to counterbalance the drawbacks of classic machine learning methods by extracting important information from the crude dataset and utilize it proficiently for the process of classification. Features are not adapted manually but rather accomplished by using the datasets to provide learning of the model through the implementation of general-purpose learning framework. Past few years, Convolution Neural Networks (CNN) have set an impressive foothold in the scope of analysis of biomedical imagery which includes mitotic cell recognition from microscopic images, identification of tumor, neural membranes segmentation, categorization of skin disease, and mass quantization in mammograms.

In the designed framework, we deploy a deep-learning based neural network model which is structured to balance out the ineffectiveness of current systems of detection and classification of tumors of benign and malignant nature along with a full analysis on the dataset being utilized. The fundamental contribution of this study can be summarized with the following:

- To develop a deep-learning based system to detect and classify cancerous breasts.
- To analyze the dataset being utilized to provide a thorough study into the breast cancer classification and deep learning model efficiency.
- To provide a web-based API functionality for the usage of the system for any health organization personnel or educational authorities.

Further narrative is sectioned such as: Section 2 presenting a descriptive analysis of the suggested methodology which specifies subsections like basic study of dataset acquired, visualization tissue slices of each patient in Binary Objective Visualization to determine dataset integrity and data pre- processing v/s normal data to see its effect on proposed model metrics. Followed by Section 3 which explains the experimental results realized post-application of the proposed framework as well as the web API created for the purpose of utilizing the created model. Lastly, section 4 discusses the conclusion of the study and provides future scope and scalability of the application.

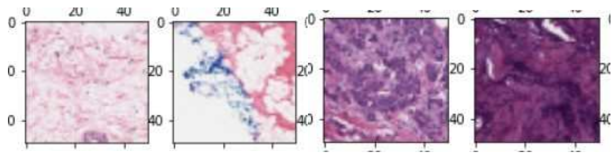
II. LITERATURE SURVEY

Feng Gao et al., 2019 construe a cancer classifying framework, known as deep cancer subtype classification (DeepCC), which uses functional spectra of neural networks and specifying functionalities of biological pathways into studies about colorectal and classification of malignancy. SanaUllah Khan et al., 2019 introduce a system in which image components of detailed nature are obtained using CNN architectures that are trained beforehand and then processed in a fully connected layer for classification of cells to bifurcate malignancy using average pooling classification.

Sebastien Jean Mambou et al., 2016 delve into digital imaging using infrared technology, which is based upon an assumption that an underlying comparison of thermal heat map amongst a normal breast and a cancer-affected breast, exhibits an increment in thermic activity of the pre-malignant tissues along with the surrounding-areas spreading breast cancer. This is a comparative study of considerable detection techniques using cutting-edge techniques of computer-vision and models of deep-learning. Nawaz et al., 2016 present a breast cancer histology images classification with the use of AlexNet thereby basing a transfer learning-based groundwork for the classification of cancerous images of breast.

III. DATASET ANALYSIS

In this section, the dataset utilized is analyzed to understand its integrity and completeness while also gauging its effective- ness for the deep-learning model proposed. The initial dataset comprised of 162 whole mount slide images of specimens of Breast Cancer which were scanned at 40x. 277,524 patches were extracted from the slide images of size 50 x 50 with 198,738 being IDC negative and 78,786 being IDC positive. IDC here stands for Invasive Ductal Carcinoma which is considered to be one of the most frequent subtypes of all breast cancers. The format of each patch's filename consists of the ID of the subject of examination, x-coordinate of the patch specifying the x-position of where it was cropped from, y coordinate of the patch specifying the y-position of where it was cropped from, and the class indicating if the patch is non-IDC or IDC.



(a) Benign Patches (b) Malignant Patches

Fig. 1. Types of Patches

1. *Basic Dataset Analysis*

With the acquired dataset being segmented already, it is apt for the training and testing of any deep-learning model. Though it is highly imperative for a dataset to be as complete as possible since a proper complete dataset would yield a considerably better accuracy and other important metrics which determine the efficiency of any deep-learning model. We try to resolve a few questions regarding the dataset which is to be used for the training and testing of the proposed mode which are as follows:

- How many patches do we have per patient?

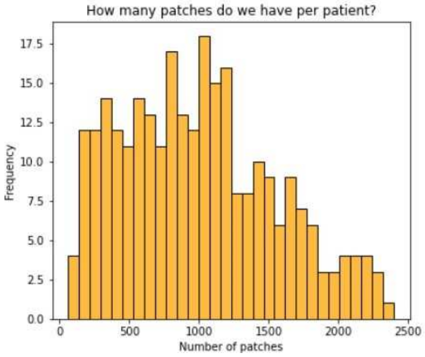


Fig. 2. Number of patches per patient

How much percentage of an image (classified as 1) is covered with IDC?

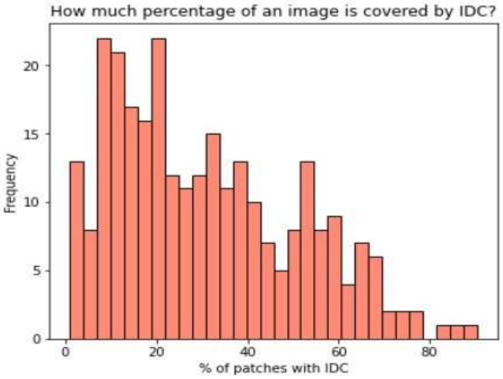


Fig. 3. Percentage of patches with IDC

- How many patches are IDC in contrast to non-IDC?

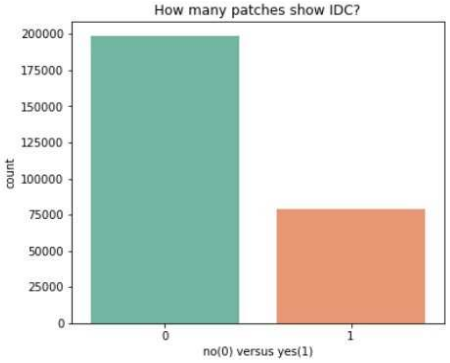


Fig. 4. Non- IDC versus IDC

As we can see, the quantity of patches we have per patient, varies quite a lot. Meaning the integrity of the data is not quite high yet we can use this dataset for training purposes of the proposed model. We also check for the affection of malignant cancer in the patches diagnosed and classified as IDC and it is clear that some patients have highly malignant patches (reaching up to 80%) while most of them are affected less. And lastly, we see the number of classified patches in contrast to each other which determines that the number of malignant patches are quite small than the benign patches. This can affect the ability to determine the classification between the patches since one of two classes has lesser data. Nonetheless, the data still holds quite a good integrity for the purpose of this application and we use it for our neural network.

2. *Shape of whole tissue slices using Binary Objective Visualization*

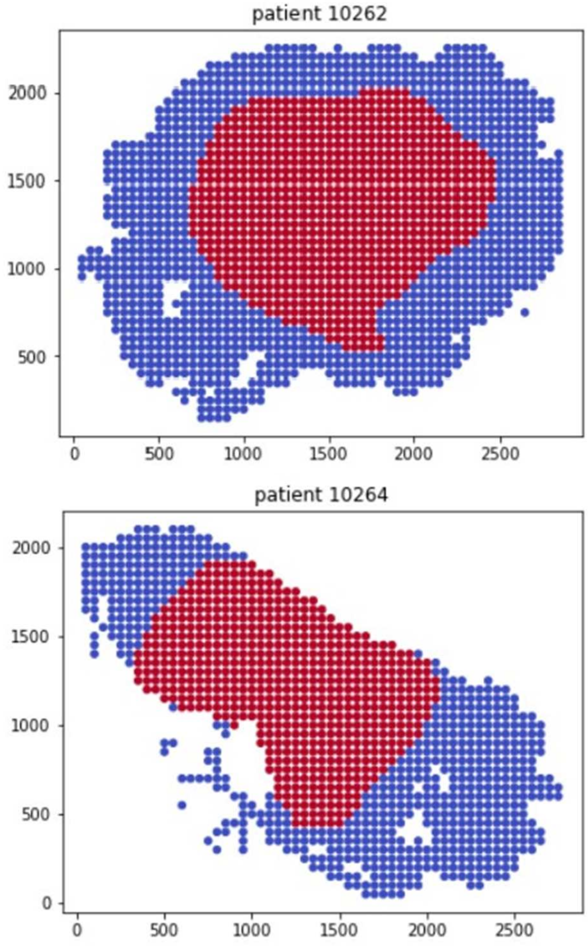


Fig. 5. Segmented Patches of 2 patients

The distilled patches are of breast tissues which are from whole slide images for the purpose of segmentation and creating a larger dataset. We use Binary Objective

Visualization to check how the whole slide images from in accordance to the patches extracted. We can see that some patients have whole of their slides formed quite effectively. This determines that there has been no loss of data in that patient's patches extraction. Though we can also notice some visualized images with blank spots in between or missing out most parts of the formed visualization. This might be because of loss of tissue slice information during the segmentation i.e. some patches may have been lost or damaged during the process of segmentation.

3. *Pre-processed v/s Normal Images*

The dataset images are pre-processed to check whether pre- processed images are beneficial for the performance of the model. We apply various filters to the images and train two different models of the same configuration as the proposed model to see which fares better. Here, for example, we take a random image and then apply Gaussian Filter to it. Gaussian Filter reduces noise i.e., high frequency components by using a low pass filter thereby blurring regions of the image to which the filter is applied.

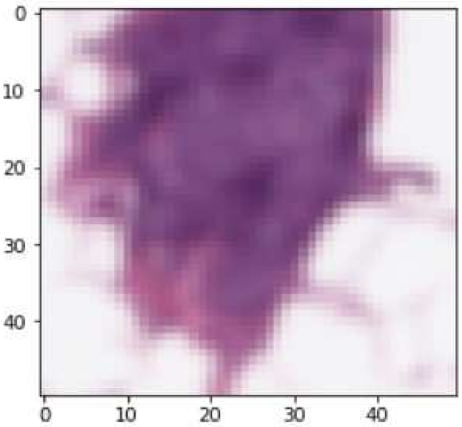


Fig. 6. Gaussian Image

It is then followed by a random noise function. Random noise can be defined as a function to add random noise of different types to a floating-point image.

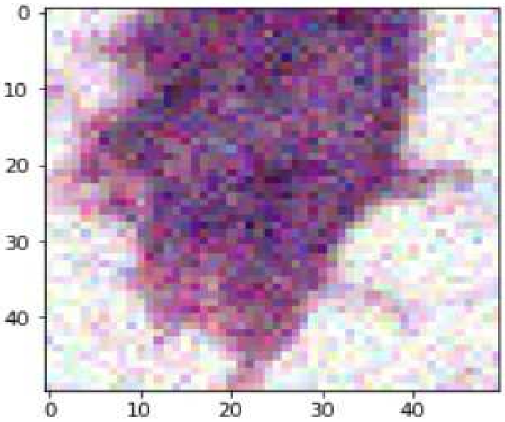


Fig. 7. Random Noise Image

This offers a relatively sharper image than the normal images. We then sort out 20,000 images from both the pre-processed and normal image dataset and use it to train two different models each for pre-processed and normal images and see which performs better. The train-test splitting is done such as the 18000 images (i.e. 90%) are used for training while rest 2000 images (i.e. 10%) are used for testing which thereby provides us the accuracy metric. The accuracy obtained by the model which is trained on pre-processed images is 0.7670 (76.70%) while the accuracy on the model trained on normal images is 0.7860 (78.60%). With this we can reach to a conclusion that normal image dataset would be better for the modelling of our proposed model.

IV. OBSERVATION AND EXPERIMENTAL RESULTS

A. *Creating the model*

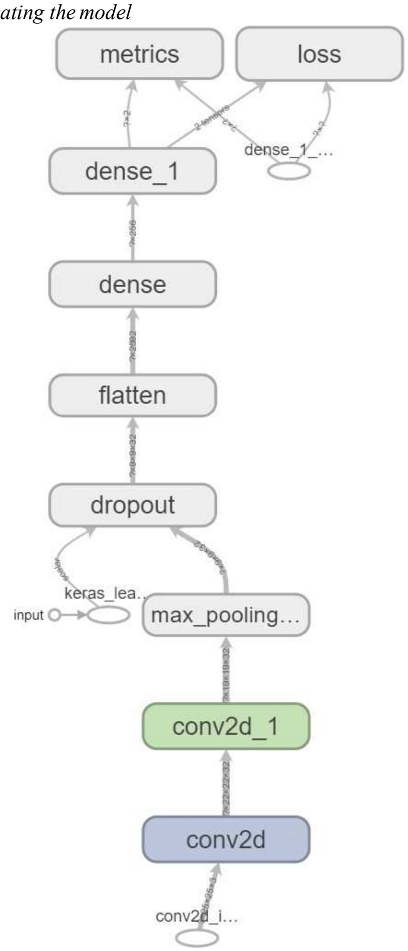
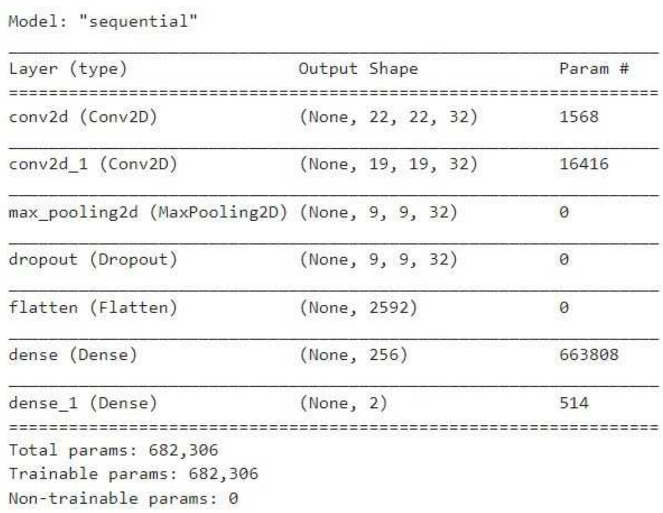


Fig. 8. Model Diagram



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 22, 22, 32)        1568
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 19, 19, 32)        16416
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 9, 9, 32)          0
_________________________________________________________________
dropout (Dropout)            (None, 9, 9, 32)          0
_________________________________________________________________
flatten (Flatten)            (None, 2592)              0
_________________________________________________________________
dense (Dense)                (None, 256)               663808
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 514
=================================================================
Total params: 682,306
Trainable params: 682,306
Non-trainable params: 0
```

Fig. 9. Model Summary

The model being used is a Sequential model. This model is called in by using the Keras library. In a Sequential model, each layer has precisely one tensor for both input and output. This is quite applicable for a simple stack of layers. This sequential model can be tweaked according to the purpose of the application and other down sampling layers can be added such as:

1. Conv2D

2D convolution is applied over an input signal which comprises of several planes of input where a kernel “slides” through the 2D input data thereby performing an element-wise multiplication. The results get added into an output pixel which is lone and distinguished. The kernel will execute the same functioning for every segment it slides over which modifies the original 2D matrix of features into a unique 2D features matrix.

2. MaxPooling2D

This layer performs down sampling of the input with respect to the spatial dimensions (i.e., height and width). A maximum value is taken over an input window for each input channel. Along each dimension, the window shifts by strides.

3. Dropout

Dropout layer avoids over-fitting of the model. This is done by aimlessly bringing the outer edges of hidden components (neurons establishing hidden layers) to 0. This is done at the phase of training, at each update. Dropout tends to make the process of training seem noisy by making the nodes within a layer to probabilistically take responsibility, more or less, for the inputs.

4. Flatten

In the flattening layer, the data is altered and processed as an 1-dimensional array then forwarded to the next layer. The Output of the convolutional layers is smoothed to form a long single vector of feature which is then connected to final model of classification which also happens to be a layer which is fully-connected. Whole of the pixel data is now a single line and ready to build links with the final layer.

5. Dense

A dense layer extensively connects with its preceding layer, for any neural network. This implies that the neurons of the dense layer are connected to every neuron of the layer preceding it. Every neuron of the preceding layer sends output to the neuron of dense layer. The dense layer neurons perform matrix-vector multiplication.

6. Model Performance

The created neural network is then trained on about 2,44,771 image patches and then being tested using 27,753 image patches. The model is implemented with the ’EarlyStopping’ callback function which stops the training after a chosen metric (here ’val loss’) reaches a point of saturation. This prevents the model from overfitting. Model performance can degrade on latest data by learning the noise and information consisting in the training data. This is known as overfitting.

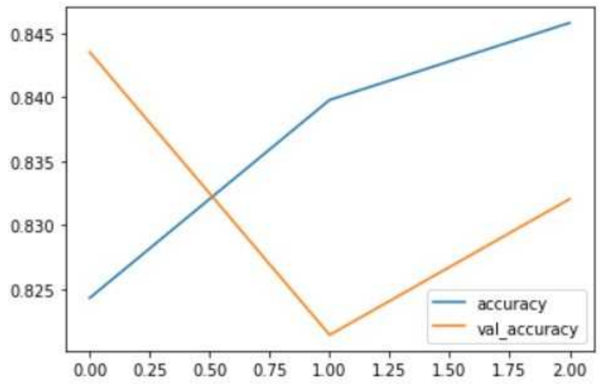


Fig. 10. Accuracy Plot

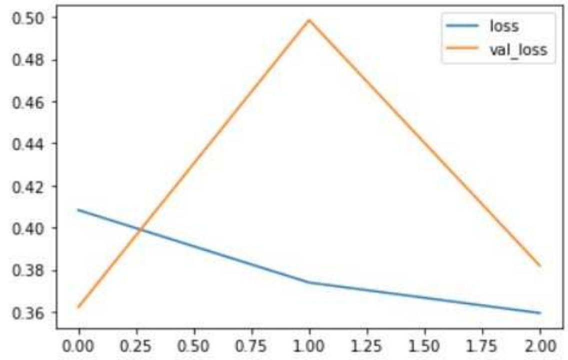


Fig. 11. Loss Plot

The proposed model gives us a accuracy of 0.8354 (83.54%) which is very good for mid-end systems. Since the malignant patches look more violet and crowded than benign ones, it can be thought that the model is able to detect hidden patterns and the color grading in these images that enable us to determine the state of each image. Overfitting is immensely less, which shows the trained model is ready to be saved and loaded onto a web API, for its usage.

V. API AND WEB LOCALIZATION

The model is then saved and converted into a JSON file for the purpose of API and web localization. This is done so as to provide a smooth graphical user interface (GUI) to a user who wants to make use of the application. The web application is created using Flask library in Python which lets developer create lightweight web applications by using different tools and integrating HTML, CSS files. The GUI of the web application is created using HTML and CSS and provides an outlook of a hospital or a medical institute to show that such applications can be implemented on a larger scale and provide help in diagnosis and classification.

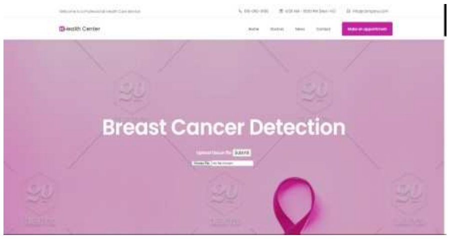


Fig. 12: Webpage

The user is required to “Choose file” by clicking the respective button which prompts the user with a browsing window. Then, user has to choose a tissue patch or tissue slice (since the model training is based on histopathology images) and press “OK” and then hit on the “Submit” button which would take the user to another page displaying the results.

VI. CONCLUSION

Through the examination and supported narrative of our study, we deployed a deep-learning based model which efficiently classifies and detects breast cancer while analyzing the dataset used for the training-testing of the neural network. In this model, the features are determined by the network which provides a good accuracy with less overfitting which is apt for API and web localization. Similarly, we also provided an analysis on the dataset, showing that the integrity of a dataset is an imperative pre-requisite for deep learning applications. Finally, the model is saved and loaded onto an API for the user to access and make use of the model to determine the type of cancer through a web-based interface. It can be observed that the designed framework gives considerably good results for mid-end systems and affordable systems and can be implemented in hospitals or medical educational institutes of all likes. In future, further classifications like in-situ carcinoma, invasive carcinoma, etc. remain imminent.

REFERENCES

[1] Sana Ullah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, and Joel JP C. Rodrigues.” A novel deep learning based framework for the detection and classification of breast cancer using transfer learning.” Pattern Recognition Letters 125 (2019): 1-6.

[2] Gao, Feng, Wei Wang, Miaomiao Tan, Lina Zhu, Yuchen Zhang, Evelyn Fessler, Louis Vermeulen, and Xin Wang. ”DeepCC: a novel deep learning-based framework for cancer molecular subtype classification.” Oncogenesis 8, no. 9 (2019): 1-12.

[3] Mambou, Sebastien Jean, Petra Maresova, Ondrej Krejcar, Ali Selamat, and Kamil Kuca.” Breast cancer detection using infrared thermal imag- ing and a deep learning model.” Sensors 18, no. 9 (2018): 2799.

[4] Nawaz, Wajahat, Sagheer Ahmed, Ali Tahir, and Hassan Aqeel Khan.” Classification of breast cancer histology images using alexnet.” In International conference image analysis and recognition, pp. 869-876. Springer, Cham, 2018.

[5] Viale, Giuseppe.” The current state of breast cancer classification.” Annals of oncology 23 (2012): x207-x210.

[6] Chukwu, Jennifer K., Faisal B. Sani, and Aliyu S. Nuhu.” Breast Cancer Classification Using Deep Convolutional Neural Networks.” FUOYE Journal of Engineering and Technology 6, no. 2 (2021): 35-38.

[7] Hatt, Mathieu, Chintan Parmar, Jinyi Qi, and Issam El Naqa.” Machine (deep) learning methods for image processing and radiomics.” IEEE Transactions on Radiation and Plasma Medical Sciences 3, no. 2 (2019): 104-108.

[8] Nazeri, Kamyar, Azad Aminpour, and Mehran Ebrahimi.” Two-stage convolutional neural network for breast cancer histology image classi- fication.” In International Conference Image Analysis and Recognition, pp. 717-726. Springer, Cham, 2018.

[9] Ahmad, Hafiz Mughees, Sajid Ghuffar, and Khurram Khurshid. ”Clas- sification of breast cancer histology images using transfer learning.” In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 328-332. IEEE, 2019.

[10] Mehra, Rajesh.” Breast cancer histology images classification: Training from scratch or transfer learning?” ICT Express 4, no. 4 (2018): 247-254.

[11] Pouyanfar, Samira, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S. Iyengar.” A survey on deep learning: Algorithms, techniques, and applications.” ACM Computing Surveys (CSUR) 51, no. 5 (2018): 1- 36.

[12] M. Veta, J.P. Pluim, P.J. Van Diest, M.A. Viergever, Breast cancer histopathology image analysis: a review, IEEE Trans. Biomed. Eng. 61 (5) (2014) 1400–1411.

[13] M.T. McCann, J.A. Ozolek, C.A. Castro, B. Parvin, J. Kovacevic, Automated histol ogy analysis: opportunities for signal processing, IEEE Signal Process. Mag. 32 (1) (2015) 78–87.

[14] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.