Breast Cancer Classification with Deep Learning

Sahil Dinesh Chavan
Mayuresh Bhagwan Nehe
Vaishnavi Kukutlawar
AI Processors and Architecture
(Dr Bahubali Shiragapur)

D. Y. Patil International University,
Akurdi,
Pune, India
https://www.dypiu.ac.in

***Abstract*—Cancer-related fatal illnesses affect both developed and developing countries globally. Particularly, the incidence of female breast cancer cases is increasing every day, in part due to underdiagnosis and late detection. An efficient first-line treatment for breast cancer can only be administered by accurately recognising and diagnosing cancer in its very early stages of development.**

The use of computer-aided diagnosis and medical image analysis technologies may accelerate and automate the identification and categorisation of cancer by instructing and supporting less experienced clinicians. For large datasets of medical images, convolutional neural networks considerably aid in the precise identification and categorisation of cancer.

Keywords—

1. **Electronic diagnostics**
2. **Mammary cancer**
3. **Deep Neural Network**
4. **Transfer learning**
5. **Classification**

• LITERATURE SURVEY

Breast cancer develops when certain breast cells start to grow abnormally, according to a method outlined in the article Ultrasound characterisation of Breast Masses by S. Gokhale.

These cells multiply and spread more quickly than healthy cells do, continue to amass, and eventually form a lump or mass that may begin to hurt.

Rapid cell spread from your breast to your lymph nodes or to other areas of your body is possible. Due to their family history, way of life, obesity, exposure to radiation, and reproductive variables, certain women may be more susceptible to developing breast cancer.

Breast cancer prediction is an active topic of study, according to Pragya Chauhan and Amit Swami's paper, Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach.

• *Problem Definition & Objectives*

1) The second most common cancer in both men and women worldwide is breast cancer. It accounted for roughly 12% of all new cancer cases and 25% of all cancers in women in 2012.

It's recommended to insert figures inside a text box. So your figure would go here (adjusting the text box to the appropriate size to match).

2) Breast cells begin to grow out of control, which is how breast cancer develops. These cells typically develop into tumours that are commonly seen on x-rays or are thought to be lumps. If the tumour's cells have the ability to spread (metastasize) to other parts of the body or enlarge into surrounding tissues, the tumour is considered malignant (cancer).

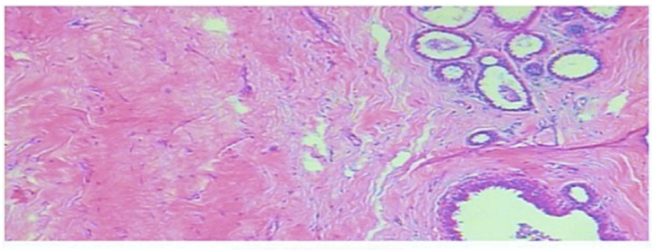
• *Methodology*

There is a tonne of pictures of cancer sufferers accessible. The dataset, often referred to as the Breast Cancer Histopathological imaging Database (BreakHis), was downloaded by completing the online form on the vision laboratory's website.

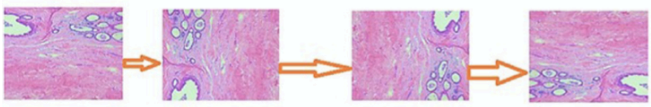
The 7909 total samples that are currently accessible were taken from 82 different patients. Breast cancer photos, both benign and malignant, are included in the dataset utilised in this study. There are four possible magnification factors for each of the accessible photos in the collection.

The dataset contains 2480 benign lesions and 5429 malignant lesions, respectively. Adenosis (AE), Fibroadenoma (FA), Phyllodes Tumour (PT), and Tubular Adenoma are four classifications that are covered by benign pictures (TA).

• ARCHITECTURE



Original Infected Image



Augmented Images

In this project in python, we'll build a classifier to train on 80% of a breast cancer histology image dataset. Of this, we'll keep 10% of the data for validation. Using Keras, we'll define a CNN (Convolutional Neural Network), call it

CancerNet, and train it on our images. We'll then derive a confusion matrix to analyse the performance of the model.

IDC is Invasive Ductal Carcinoma; a cancer that develops in a milk duct and invades the fibrous or fatty breast tissue outside the duct; it is the most common form of breast cancer forming 80% of all breast cancer diagnoses. And histology is the study of the microscopic structure of tissues.

We'll use the IDC_regular dataset (the breast cancer histology image dataset) from Kaggle. This dataset holds 2,77,524 patches of size 50×50 extracted from 162 whole-mount slide images of breast cancer specimens scanned at 40x. Of these, 1,98,738 tested negative and 78,786 tested positive with IDC.

Filenames in this dataset look like this:

8863_idx5_x451_y1451_class0

Here, 8863_idx5 is the patient ID, 451 and 1451 are the x- and y- coordinates of the crop, and 0 is the class label (0 denotes the absence of IDC).

ALGORITHM

1. Convolution Neural Network:

Convolutional neural networks are a form of artificial neural networks used most frequently in deep learning to interpret visual vision.

i) Use of 3X3 Convolution Filters

A kernel, convolution matrix, or mask is a tiny matrix used in image processing. It can be used for edge detection, embossing, sharpening, and more. By performing a convolution between a kernel and an image, this is achieved. In image processing, Kernel, convolution matrix, or mask is a small matrix. It is used for blurring, sharpening, embossing, edge detection, and more. This is accomplished by **doing a convolution between a kernel and an image**.

ii) Stack this filter on top of each other

Can these filters be combined into one 3x3 filter, the order being first T and then W ? (The combination should be possible) And what is the math behind this combination?
I can make a 3x3 filter of all one and after that stride the W filter over that 3x3 and add everything up, is it that simple?

iii) Performing Max Pooling

A pooling operation known as maximum pooling, also known as max pooling, determines the maximum or largest value in each patch of each feature map. The outcomes are down-sampled or pooled feature maps that emphasize the feature that is most prevalent in the patch, as opposed to the feature's average presence in the case of average pooling.

2. Deep Neural Network

At its simplest, a neural network with some level of complexity, usually at least two layers, qualifies as a deep neural network (DNN), or deep net for short. Deep nets process data in complex ways by employing sophisticated math modelling.

Artificial intelligence (AI) and machine learning techniques called deep learning model how people acquire specific types of information. Data science, which also encompasses statistics and predictive modelling, contains deep learning as a key component.

EXPERIMENT RESULT

Here, we are attaching a screenshot of the experiment result of our project which tells us about the accuracy and number of epochs.

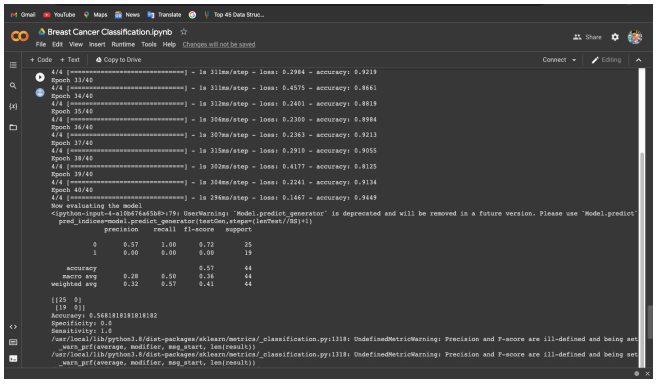
In this script, first, we set initial values for the number of epochs, the learning rate, and the batch size. We'll get the number of paths in the three directories for training, validation, and testing. Then, we'll get the class weight for the training data so we can deal with the imbalance.

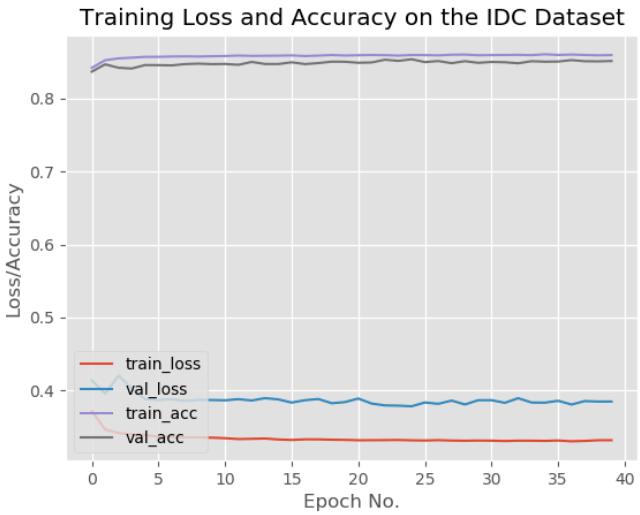
Now, we initialise the training data augmentation object. This is a process of regularisation that helps generalise the model. This is where we slightly modify the training examples to avoid the need for more training data. We'll initialise the validation and testing data augmentation objects.

We'll initialise the training, validation, and testing generators so they can generate batches of images of size batch_size. Then, we'll initialise the model using the Adagrad optimiser and compile it with a binary_crossentropy loss function. Now, to fit the model, we make a call to fit_generator().

We have successfully trained our model. Now, let's evaluate the model on our testing data. We'll reset the generator and make predictions on the data. Then, for images from the testing set, we get the indices of the labels with the corresponding largest predicted probability. And we'll display a classification report.

Now, we'll compute the confusion matrix and get the raw accuracy, specificity, and sensitivity, and display all values. Finally, we'll plot the training loss and accuracy.





CONCLUSION

In this Python project, we developed the network CancerNet and learned how to build a breast cancer classifier using the IDC dataset (containing histology images for Invasive Ductal Carcinoma). To do the same, we made advantage of Keras. I hope this Python project was enjoyable.

RESULT

We have successfully trained our model using the IDC dataset. Then, we evaluate the performance of the model using a confusion matrix and we got 56% of accuracy.

In this project in python, we learned to build a breast cancer classifier on the IDC dataset (with histology images for Invasive Ductal Carcinoma) and created the network CancerNet for the same. We used Keras to implement the same. Hope you enjoyed this Python project.

FUTURE SCOPE

End-to-end Breast Mass Classification

The main objective of an end-to-end computer-aided diagnosis system is to improve classification accuracy while minimising false positives. One can improve our architecture to better simulate the diagnosis procedure followed by radiologists.

GitHub LINK

Link: https://github.com/Sahil2510/Breast_Cancer_Diagnosis_using_Deep_Learning

REFERENCES

1. World Health Organisation, WHO Position Paper on Mammography Screening, World Health Organisation, 2014.
2. S. Ahmad, S. Ur Rehman, A. Iqbal, R. K. Farooq, A. Shahid, M. I. Ullah, Breast cancer research in Pakistan: a bibliometric analysis: 11(3) (2021).
3. M.T. McCann, J.A. Ozolek, C.A. Castro, B. Parvin, J. Kovacevic, Automated histology analysis: opportunities for signal processing, IEEE Signal Process.
4. J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, J. Fan, Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis.
5. Link: https://medium.com/swlh/breast-cancer-classification-using-python-e83719e5f97d
6. Link: https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3
7. Link: https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images