Breast Cancer in Data Analytics

Preethi Balasubramaniam
balasubramaniampree@cityuniversity.edu

CS612- DATA EXPLORATION/VISUALIZATION
Sept 10th, 2020
Dr. Brian Maeng
City University of Seattle

Abstract

Breast Cancer is generally seen in women, and maybe the main reason for increasing the rate of death among women. The diagnosis of breast cancer is time-consuming and due to the lesser availability of frameworks, it is fundamental to create a framework that can accordingly investigate breast cancer in its early stages (Goel, V. 2018). Different Deep Learning Algorithms and Machine Learning have been applied for the group of malignant and benign tumors. The most common subtypes of all breast cancer are Invasive Ductal Carcinoma (IDC) (Narayan, B. 2020). Pathologists normally attention to regions that involve IDC to decide whether a quiet suffers from breast cancer or not. Breast Histopathology images dataset is being used in this project from Kaggle for society.

The key purpose of this paper is to develop a robust data analytics model which may support in (i) a higher understanding of breast cancer survivability in presence of lost data, (ii) finding cohorts of patients that share similar properties , and (iii) providing a better, visions into factors connected with the patient survivability (Shukla, N. 2017).

Keywords:

Breast Cancer, Deep Learning, Histopathology Images, IDC Data set.

I. Introduction

Breast cancer is one of the foremost dangerous diseases because most of the women's died every year. Some tumors present within the breast will be cancerous (malignant) and noncancerous (benign). Benign tumors cannot be extended to the remaining components of the body and these tumors are not harmful to the body. After removing these tumors, they do not grow again. Malignant tumors are serious, and these are spread to the remaining parts of the body, after removing this tumor it will grow again (K.Shailaja, 2018). Among women, breast cancer is that the most often diagnosed cancer and the number one reason for death, within the developed world, between 1 in 8 and 1 in 12 women will have breast cancer during her lifespan. Predicting breast cancer risk is important in combating this disease (Stark.G, 2019).

There are two main forms of breast cancer risk. The first form represents the probability that specific will contract breast cancer during a specified period. The second form reflects the probability that a mutation will occur during a high-risk gene. This hopes to develop a model that accurately predicts the previous 4 of risk. Current breast cancer screening guidelines are limited (Stark.G, 2019). The Preventative Services Task Force recommends screening for ladies ages 50-74 and does not provide conclusive guidelines for ladies outside this age range. Exactly predicting breast cancer risk can both inspire screening for high-risk women who would not be screened and promote observation to screening guidelines among those that will not follow them (Stark.G, 2019). A statistical model evaluates that breast cancer risk may inform chemoprevention and other actions to decline one's risk. The diagnosis of breast cancer is comprehensive by analyzing histopathological images which play an important role in patients and their diagnosis. By presenting it to analyze histopathological images of breast cancer via supervised and unsupervised deep convolutional neural networks (CNN). Deep learning techniques can automatically abstract features, regain information from data automatically, and learn advanced abstract representations of data (X.Juanying, 2019).

II. Literature Review

In literature [1], the paper was published in the year 1999. The main objective of this book is access to early diagnosis of BC by primary healthcare professionals. This book includes 232 studies of which 33 are focused on evaluating diagnostic tools to identify the BC. Mammography, clinical examination, a fine- needle biopsy may be useful for discovering BC, especially when used in combination. Mammography is used for diagnosing the BC in women of all ages and it is more sensitive than clinical examination. According to the evidence, screening improves survival in women aged 50 to 65 years and benefits for women aged 65 to 75 years. There are fewer benefits of mammography for younger women (<50 years). Fine- needle biopsy sensitivity is normally high, but the specificity varies. Both the specificity and sensitivity change depend on the position of the needle. Ultrasound is advised as the first radiological investigation for younger women (< 35 years). There is no sufficient evidence that breast lesions self-examination improves survival.

In literature [2], the paper was published in the year 2007. This paper mainly focuses on routine clinical check-ups after the treatment for breast cancer has benefits for the patients. Apart from survival, there is a lack of inquiry relating to long term clinical review of patients in terms of quality of life, patient satisfaction, psychological outcomes, access to specialty advice regarding management of symptoms, and reassurance. Regardless of supporting evidence, UK hospitals continue to have a routine check-up every six months for patients up to a minimum of five years.

In literature [3], the paper was published in the year 2010. The purpose of this paper is to have a complete understanding of breast cancer screening studies using ultrasound or breast magnetic resonance imaging (MRI), mammography and also helps nurse practitioners to choose the most appropriate screening tool for their individual patients. In all the studies, compared to mammography and ultrasound, breast MRI has a higher sensitivity and lower specificity. Also, the ultrasound has higher sensitivity compared to mammography.

In literature [4], the paper was published in the year 2014. The main objective of this paper is to solve an imbalanced breast cancer dataset. This paper suggests a two-steps approach to increase the quality of class prediction imbalance breast cancer dataset. The two- steps approach follows two main techniques: 1) using selection techniques to filter out unimportant features from the dataset.To filter out three evaluated functions ChiSquareAttribute-Evaluation, ConsistencySubsetEvaluation, and InfoGainAttribute. The ConsistencySubset Evaluation is the best attribute set, which consists of 9 attributes. Those nine attributes are fed into an oversampling phase to adjust the size of the minority class. 2) Using the oversampling technique to modify the size of the minority class to the size of the majority class. The dataset used in this step was the best final attribute set from feature selection in the initial step with the over-sampling technique, SMOTE (Synthetic Minority Oversampling technique). After receiving a new dataset from the two-step approach, it is learned the data with three different classification algorithms: decision tree(C4.5), artificial neural network (MLP), Naïve Bayes which gave the accuracy of 83.80%, 78.54%, and 75.62% respectively. The results showed that the decision tree was more suitable to classify this dataset.

In literature [5], the paper was published in the year 2015. The focus of this paper is to assess whether Urbanization is a major Culprit of Rampant Breast Cancer. The research concluded that westernized life raised economic standards and modified food habits led to obesity which is the main culprit for breast cancer. Urbanization also includes the use of late childbirth, late marriages, oral contraceptives, shorter breastfeeding, night shift working, alcohol consumption, smoking, and a sedentary lifestyle which all lead to increased risk of breast cancer.

In literature [6], the paper was published in the year 2018. The main goal of this paper is to impact mammography screening programs on the incidence of advanced BC on ABCR (advanced breast cancer rate) and to summarize their limitations and contribute to evidence on screening effectiveness. This paper concluded that the mammography service screening programs are due to observational data that was collected or analyzed with methodological approaches. Improving the knowledge of limitations earlier studies will help to establish consensus on the correct methodology.

In literature [7], the paper was published in the year 2019. This paper mainly discusses the prevention, treatment, and recurrence of breast cancer. This literature suggests a “healthy” dietary pattern such as fruits, vegetables, and whole grains, low moderate intake of dairy products, no more than three portions of red meat per week, and very little consumption of sugar, sweet, processed meat and alcohol. Breast Cancer (BC) patients should be encouraged to improve their dietary habits before, after, and during treatment to have a better quality of life and long-term survival.

In literature [8], the paper was published in the year 2020. The main objective of this paper is to Predict BC Using Deep Learning and Machine Learning Techniques. The BC diagnosis is very time consuming and due to the lesser availability of the system; it is required to develop a system that can automatically diagnose breast cancer in its starting stages. This paper mainly focuses on different models that are implemented such as Support Vector Machine (SVM), Logistic Regression, and K Nearest Neighbor (KNN), Multi-Layer perceptron classifier, Artificial Neural Network (ANN), etc. on the dataset taken from the Kaggle repository. The experiments have shown that Random Forest Classifier and SVM are the best for predictive analysis with an accuracy of 96.5%. Deep learning algorithms such as ANN and CNN have been implemented, to increase the precision of prediction. The maximum accuracy shown in the case of CNN and ANN is 97.3% and 99.3% respectively.

III. Dataset

As breast cancer is the most common type of cancer in woman and just in 2012 there are 1.70 million cases. As there is abundant data for the early stage of prediction with cancer positive or negative.

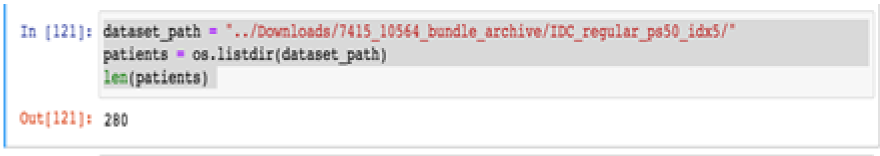
Download and Build Dataset:

Python Packages:

1. Numpy, Pandas
2. Matplotlib - visualizing various graphs or images
3. opencv-python - computer vision task for reading and changing imaging properties like color
4. scikit-learn python machine learning library
5. keras – python deep learning classifier model

Downloaded the Dataset from the Kaggle which consists of image files of the scans and need to do the data cleaning. This constitutes to pick only png or jpeg file extensions.

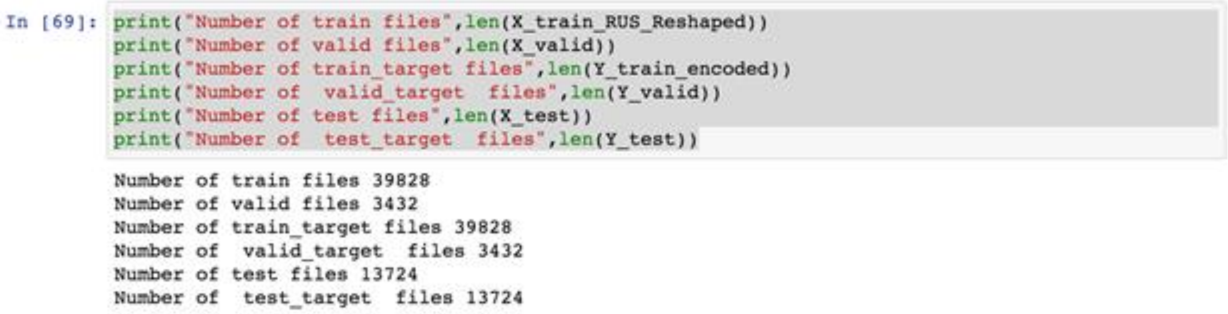
Total number of patients using in the model are:



```
In [121]: dataset_path = "../Downloads/7415_10564_bundle_archive/IDC_regular_ps50_idx5/"
          patients = os.listdir(dataset_path)
          len(patients)

Out[121]: 280
```

Data Cleaning:

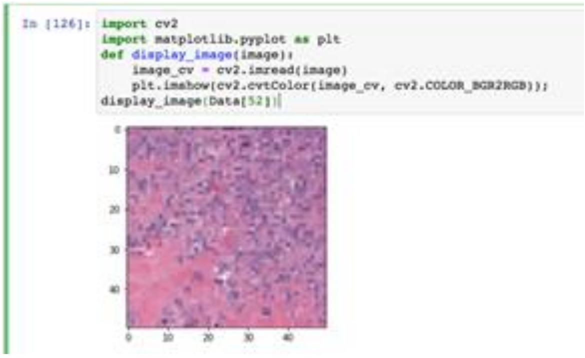
- Data cleaning with the image set would be discarding the images or enhancing the pixels.
- Most of the images are not bright enough with a smaller number of pixels so if needed discard all the images with less pixel values using the OpenCV-python (cv2) library.
- All the images might not show the same resolution and it varies between different patients because for some patients the full image shows IDC positive and for some a corner piece of the image. Processing the hue image to figure this out is a big task.
- Another important aspect is some images are not clear or incomplete patches that have a smaller size than the defined pixels (50 x 50).
- After the data cleaning and dividing the data sets to training sets and testing sets resulted in about 40K Images for training each IDC(+) and IDC(-).



```
In [69]: print("Number of train files",len(X_train_RUS_Reshaped))
         print("Number of valid files",len(X_valid))
         print("Number of train_target files",len(Y_train_encoded))
         print("Number of  valid_target  files",len(Y_valid))
         print("Number of test files",len(X_test))
         print("Number of  test_target  files",len(Y_test))

         Number of train files 39828
         Number of valid files 3432
         Number of train_target files 39828
         Number of  valid_target  files 3432
         Number of test files 13724
         Number of  test_target  files 13724
```

Data Visualization:

- Display of Input Image:

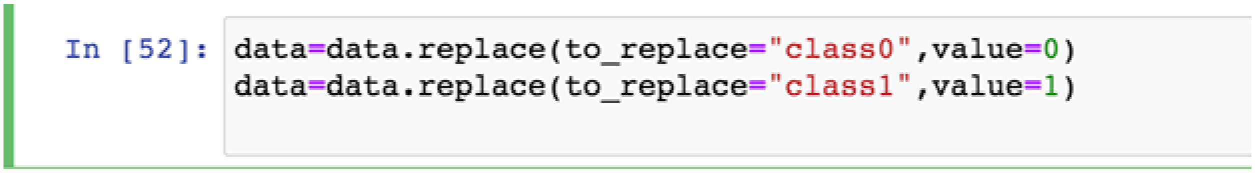


```
In [126]: import cv2
          import matplotlib.pyplot as plt
          def display_image(image):
              image_cv = cv2.imread(image)
              plt.imshow(cv2.cvtColor(image_cv, cv2.COLOR_BGR2RGB));
          display_image(Data[52])
```

- Classifying the input images as 0 and 1:

As per Kaggle the given dataset Class1 folder represents IDC (+) and Class0 represents IDC (-).

- o 0 - IDC+ Patient Image
- o 1 - IDC- Patient Image



```
In [52]: data=data.replace(to_replace="class0",value=0)
         data=data.replace(to_replace="class1",value=1)
```

As for the Breast Cancer Classification needs to run various models with numerous times till will get good robust cross-validation scores as it is a matter of life and death and also reduce the **False Negatives**.

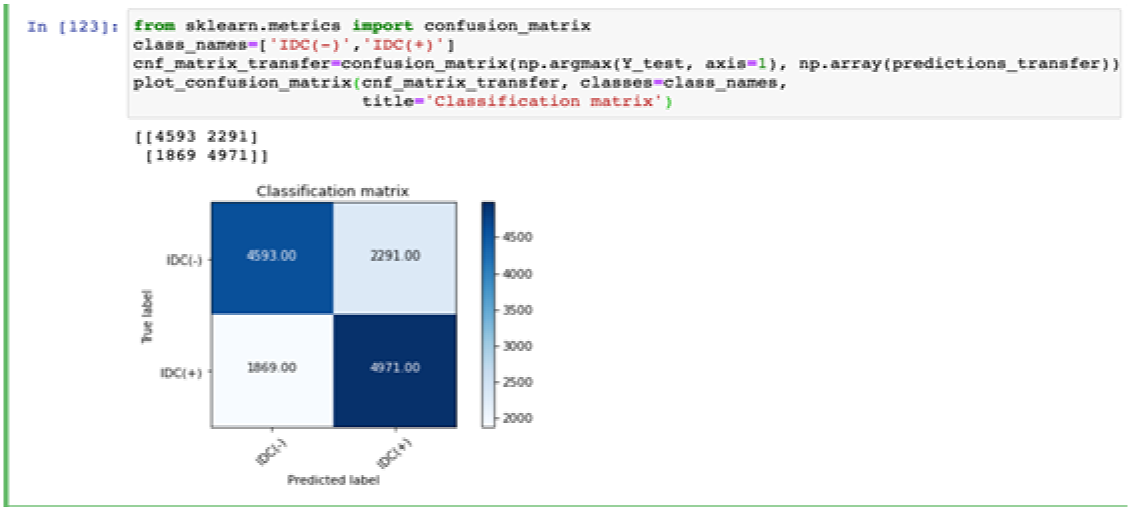
Keras Deep Learning Model as our first prediction training model.

Keras Deep Learning Model:

Keras learning model is most popularly used, with less data for the positive cases compared to the negative and the Keras argument model will give out more specificity accurate and confidence values using the Sklearn package metrics output. Also trained this model with refined data set multiple times even though could not process large data as it needs high computing machines, the model predicted with good results.

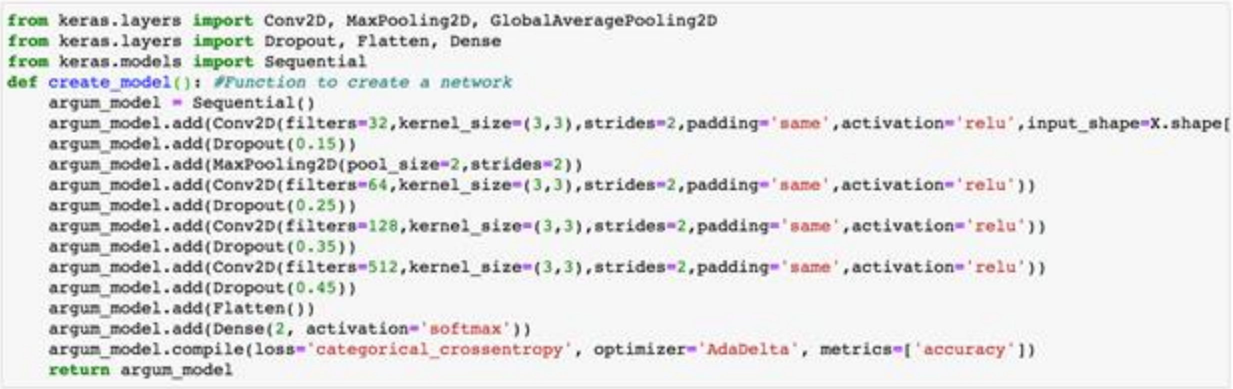
Although there are different models that have different outputs of accuracy with the batch sizes and loss of data to train the model. The Robust model would be the difference in validation scores running the model multiple times is less. Needed to run with different models with the Dense configuration of Keras Classifier by different dropout ratios.

Our Project includes a comparison of models to build scores for False Positives, False Negatives, True Positive, True Negative.



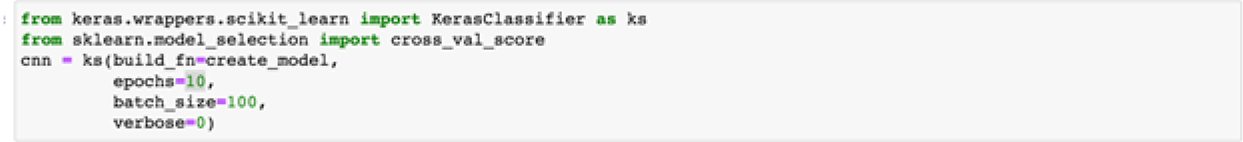
```
In [123]: from sklearn.metrics import confusion_matrix
          class_names=['IDC(-)','IDC(+)']
          cnf_matrix_transfer=confusion_matrix(np.argmax(Y_test, axis=1), np.array(predictions_transfer))
          plot_confusion_matrix(cnf_matrix_transfer, classes=class_names,
                                title='Classification matrix')

          [[4593 2291]
           [1869 4971]]
```

Using the Sequential model and creating a total of 10 models as batch size and pass build the classifier. Applied scikit learn function cross_val_score to evaluate our model and after running multiple times the average is:



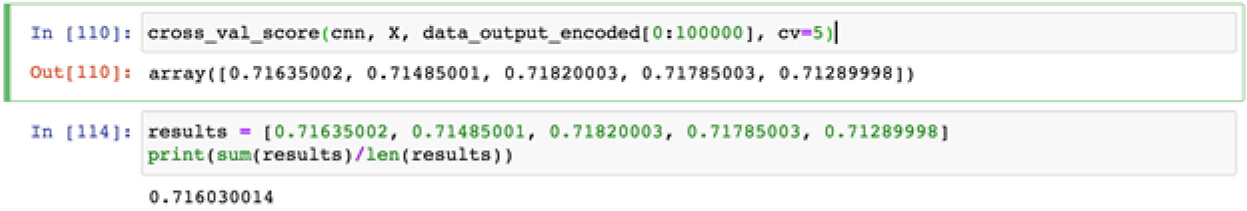
```
from keras.layers import Conv2D, MaxPooling2D, GlobalAveragePooling2D
from keras.layers import Dropout, Flatten, Dense
from keras.models import Sequential
def create_model(): #Function to create a network
    argum_model = Sequential()
    argum_model.add(Conv2D(filters=32,kernel_size=(3,3),strides=2,padding='same',activation='relu',input_shape=X.shape[
    argum_model.add(Dropout(0.15))
    argum_model.add(MaxPooling2D(pool_size=2,strides=2))
    argum_model.add(Conv2D(filters=64,kernel_size=(3,3),strides=2,padding='same',activation='relu'))
    argum_model.add(Dropout(0.25))
    argum_model.add(Conv2D(filters=128,kernel_size=(3,3),strides=2,padding='same',activation='relu'))
    argum_model.add(Dropout(0.35))
    argum_model.add(Conv2D(filters=512,kernel_size=(3,3),strides=2,padding='same',activation='relu'))
    argum_model.add(Dropout(0.45))
    argum_model.add(Flatten())
    argum_model.add(Dense(2, activation='softmax'))
    argum_model.compile(loss='categorical_crossentropy', optimizer='AdaDelta', metrics=['accuracy'])
    return argum_model
```

KerasClassifier is one of the deep learning models and evaluate the models using the resampling k-fold cross validation score.



```
from keras.wrappers.scikit_learn import KerasClassifier as ks
from sklearn.model_selection import cross_val_score
cnn = ks(build_fn=create_model,
         epochs=10,
         batch_size=100,
         verbose=0)
```

After running with a single model and the mean of the average accuracy is 0.716030014 [0.71635002, 0.71485001, 0.71820003, 0.71785003, 0.71289998] which says this model was a robust model with a difference in cross-validation score of 0.01 only.



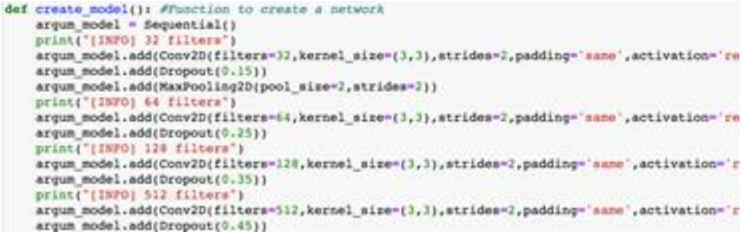
```
In [110]: cross_val_score(cnn, X, data_output_encoded[0:100000], cv=5)
Out[110]: array([0.71635002, 0.71485001, 0.71820003, 0.71785003, 0.71289998])

In [114]: results = [0.71635002, 0.71485001, 0.71820003, 0.71785003, 0.71289998]
          print(sum(results)/len(results))

          0.716030014
```

Furthermore, will be doing more classification and train the data on different training models for more accuracy and better prediction results for IDC analysis.

Data Model:

Building the breast cancer classifier with the IDC dataset and create a Keras Conv2D deep learning model with different parameters.

As per the complexity of the given dataset images, applied (32, 64, 128) filters and used Max pooling to reduce the spatial dimensions of output with a kernel size of (3x3) for filtering throughout the network. For matching the output volume sizes with the input volume had used the padding as *'same'.*

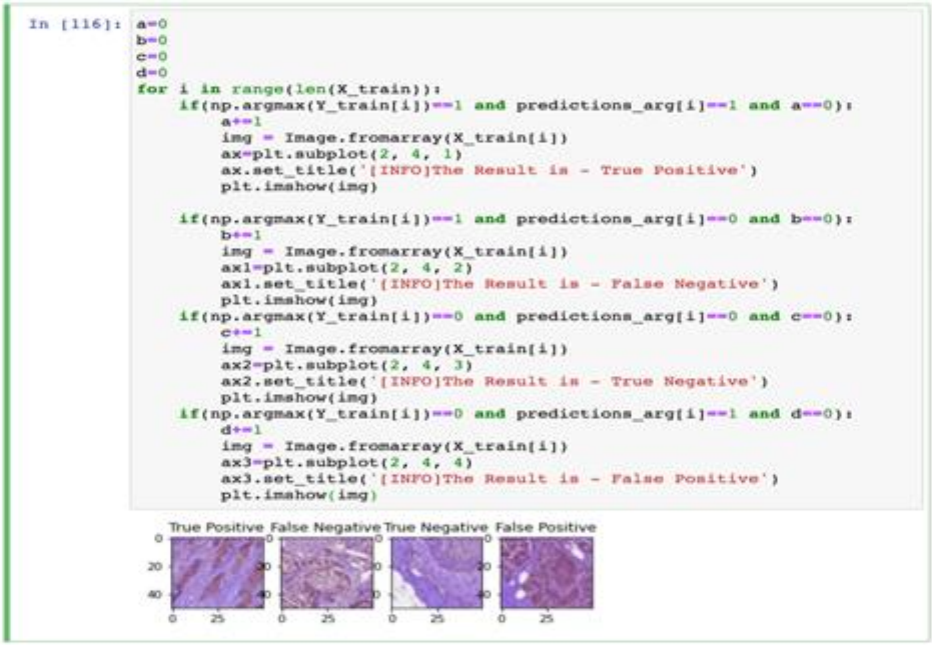


```
def create_model(): #Function to create a network
    argum_model = Sequential()
    print("[INFO] 32 filters")
    argum_model.add(Conv2D(filters=32,kernel_size=(3,3),strides=2,padding='same',activation='re
    argum_model.add(Dropout(0.15))
    argum_model.add(MaxPooling2D(pool_size=2,strides=2))
    print("[INFO] 64 filters")
    argum_model.add(Conv2D(filters=64,kernel_size=(3,3),strides=2,padding='same',activation='re
    argum_model.add(Dropout(0.25))
    print("[INFO] 128 filters")
    argum_model.add(Conv2D(filters=128,kernel_size=(3,3),strides=2,padding='same',activation='r
    argum_model.add(Dropout(0.35))
    print("[INFO] 512 filters")
    argum_model.add(Conv2D(filters=512,kernel_size=(3,3),strides=2,padding='same',activation='r
    argum_model.add(Dropout(0.45))
```

Loading the data, defined a neural network in Keras, compiled and trained the model. After building the Keras deep learning model and training it multiple times by shuffling the data achieved great results with an average mean of weights - 0.716. The sensitivity and specificity mean of 71% score using a number of positive cases and negative cases.

After building the classification matrix for False Positives, False Negatives, True Positive, True Negative.

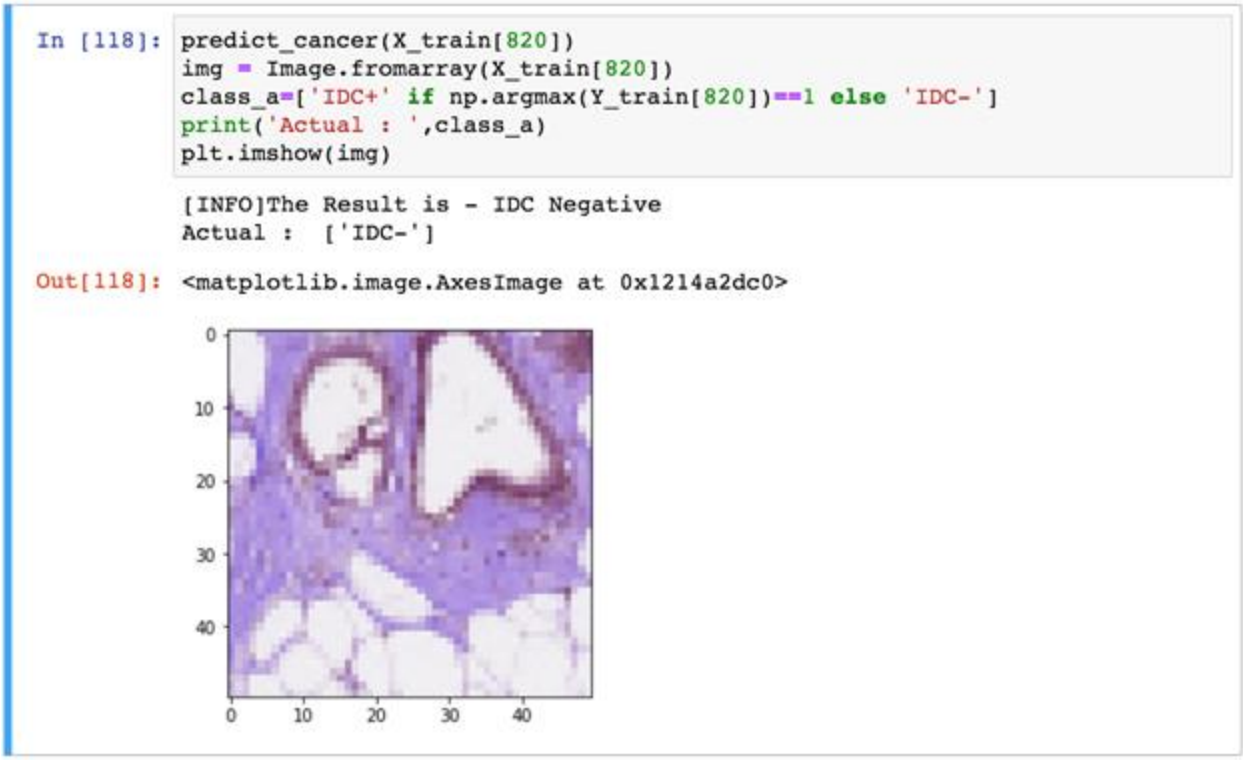
- True Positive: Observation is a Positive class and model classified as Positive
- False Positive: Observation is a Negative class and model classified as Positive
- True Negative: Observation is a Positive class and model classified as Negative
- False Negative: Observation is a Negative class and model classified as Negative



```
In [116]: a=0
          b=0
          c=0
          d=0
          for i in range(len(X_train)):
              if(np.argmax(Y_train[i])==1 and predictions_arg[i]==1 and a==0):
                  a+=1
                  img = Image.fromarray(X_train[i])
                  ax=plt.subplot(2, 4, 1)
                  ax.set_title('[INFO]The Result is - True Positive')
                  plt.imshow(img)

              if(np.argmax(Y_train[i])==1 and predictions_arg[i]==0 and b==0):
                  b+=1
                  img = Image.fromarray(X_train[i])
                  ax1=plt.subplot(2, 4, 2)
                  ax1.set_title('[INFO]The Result is - False Negative')
                  plt.imshow(img)
              if(np.argmax(Y_train[i])==0 and predictions_arg[i]==0 and c==0):
                  c+=1
                  img = Image.fromarray(X_train[i])
                  ax2=plt.subplot(2, 4, 3)
                  ax2.set_title('[INFO]The Result is - True Negative')
                  plt.imshow(img)
              if(np.argmax(Y_train[i])==0 and predictions_arg[i]==1 and d==0):
                  d+=1
                  img = Image.fromarray(X_train[i])
                  ax3=plt.subplot(2, 4, 4)
                  ax3.set_title('[INFO]The Result is - False Positive')
                  plt.imshow(img)
```

Defined the *predict_cancer()* function for making the prediction of IDC+ or IDC- on the Keras trained model. Applied the sigmoid the function and have the predictions of the probability 0 and

Tested with different examples of the Keras dataset and the predicted result is 70% absolute correct. As an example, calling our function with a result of IDC Negative.



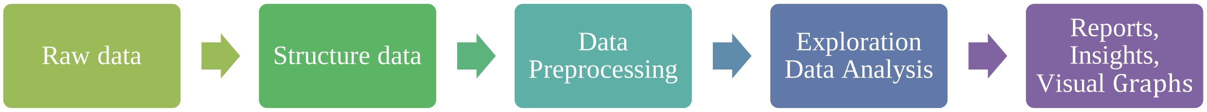
```
In [118]: predict_cancer(X_train[820])
          img = Image.fromarray(X_train[820])
          class_a=['IDC+' if np.argmax(Y_train[820])==1 else 'IDC-']
          print('Actual : ',class_a)
          plt.imshow(img)

          [INFO]The Result is - IDC Negative
          Actual :  ['IDC-']

Out[118]: <matplotlib.image.AxesImage at 0x1214a2dc0>
```

IV. Methodology

KNN could be a supervised learning technique which means the label of the info is identified before making predictions. KNN algorithm does not have a training phase. Predictions are made supported the Euclidean distance to k-nearest neighbors. This method is applied to the prediction of breast cancer dataset since it already has labels like malignant and benign. In our IDC dataset, the result variable or variable specifically having only two sets of values, either M (Malign) or B(Benign). So, it applies the Classification algorithm of supervised learning.

A deep learning algorithm is often executed for the smoothing of knowledge during Data Preprocessing. Data preprocessing is performed to enhance the standard of an IDC dataset to induce clean data which may be useful for modeling. There are several processes involved in data preprocessing. These processes involve data cleaning, feature selection, feature extraction, etc. Data cleaning involves the inconsistencies which are present within the data, thereby improving the standard of the information. During the information preprocessing stage, the information is partitioned into the IDC dataset and validation of the dataset. The IDC dataset is used in preparing the deep learning model, while the validation of the dataset is used during the prediction stage. Other preprocessing operations performed on the dataset are centering and scaling.

Data preprocessing is a method that is used to change raw data into a clean dataset. In simple words, whenever the data is collected from different sources it is usually collected in raw format which is not feasible for the analysis. So, there are few steps that should be executed to convert the data into a small clean data set. This process is performed before the execution of the iterative analysis. Data preprocessing includes data integration, data cleaning, data transformation, and data reduction. The figure below gives a better understanding of Data preparation.



Data Preparation

The Crucial part of data science is data preprocessing. It includes concepts such as feature engineering and Data Cleaning. These two are mandatory for achieving better performance and accuracy in the Deep Learning and Machine learning. Data preprocessing is important because of the existence of unformatted data in real-world data. Mostly real-world data comprises of Jagreet, K. G. (2018, December 23).

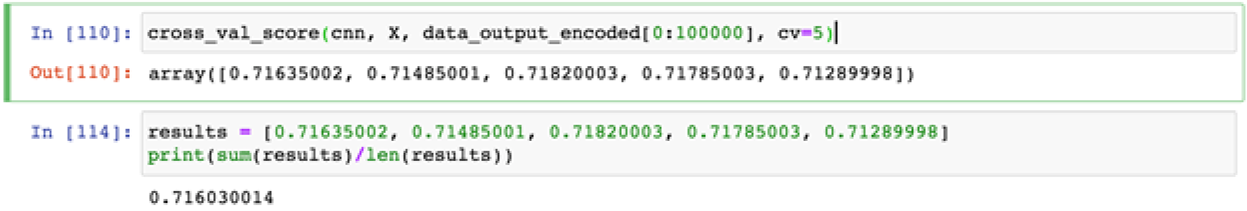
- Missing data (Inaccurate data)
- Erroneous data and outliners
- Inconsistent data.

Data Wrangling can handle the issue of data leaking while it is executed in machine learning and data Manipulation is another way to refer this process. There are different ways to perform it. Jagreet, K. G. (2018, December 23).

- Cleaning the data
- Transformation the data
- Enrichment the data

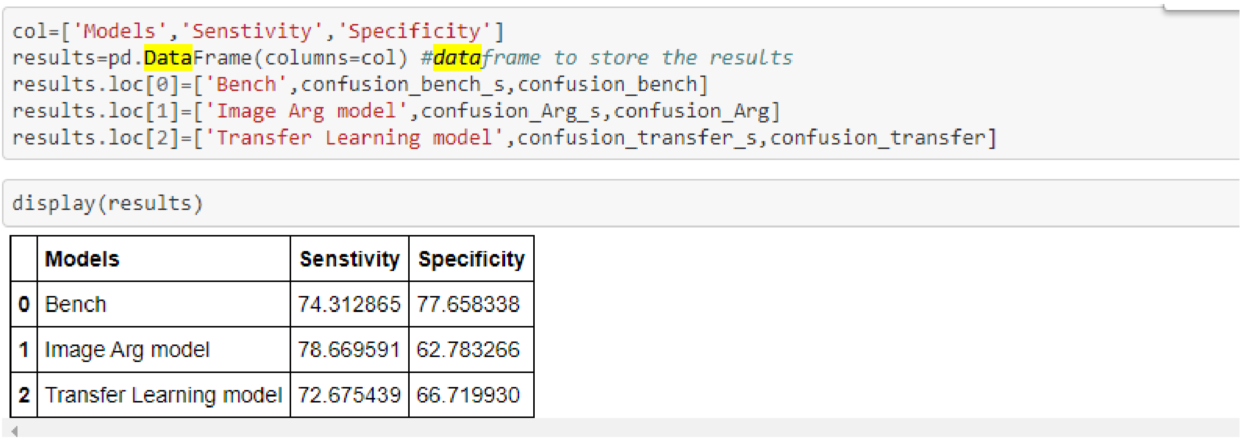
V. Predicting breast cancer results

The goal of the project is to identify IDC when it is present in unlabeled histopathology images.



```
In [110]: cross_val_score(cnn, X, data_output_encoded[0:100000], cv=5)
Out[110]: array([0.71635002, 0.71485001, 0.71820003, 0.71785003, 0.71289998])

In [114]: results = [0.71635002, 0.71485001, 0.71820003, 0.71785003, 0.71289998]
          print(sum(results)/len(results))

          0.716030014
```

By looking into the screenshot above it is understood that the output of our model has achieved ~ 71% accuracy, however, the raw accuracy of benign/no cancer is correct most of the time. To understand performance at a deeper level, the execution of sensitivity and specificity is necessary. By the screenshot below its understood that the sensitivity and specificity changes based on the models used.



```
col=['Models','Senstivity','Specificity']
results=pd.DataFrame(columns=col) #dataframe to store the results
results.loc[0]=['Bench',confusion_bench_s,confusion_bench]
results.loc[1]=['Image Arg model',confusion_Arg_s,confusion_Arg]
results.loc[2]=['Transfer Learning model',confusion_transfer_s,confusion_transfer]
```

```
display(results)
```

	Models	Senstivity	Specificity
0	Bench	74.312865	77.658338
1	Image Arg model	78.669591	62.783266
2	Transfer Learning model	72.675439	66.719930

The Bench model has the sensitivity of ~74% and specificity of ~77%, Image Arg model has ~78% of sensitivity and ~62 % of specificity, and whereas the transfer learning model sensitivity is ~72% and Specificity is ~66%.

It is necessary to be careful when it is false-negative – these patients cannot be put under “No cancer” when are in fact “Cancer positive. It also important to know about false positive – these patients should not mistakenly classify as “Cancer Positive” and then the patients might go through expensive and painful treatment when patients do not need them.

A machine learning/ deep learning engineer and practitioner must manage the balance between specificity and sensitivity because the balance becomes extremely important when it comes to deep learning and healthcare treatment.

VI. Conclusion

Breast cancer is that the second important reason for cancer deaths all over the world and screening mammography has been observed to decrease mortality. BC using deep learning is extremely useful for detecting the BC. Deep learning is encouraged by the workings of the human brain and its biological neural networks. In this research paper, people learned how to use the Keras deep learning library to coach a Convolutional Neural Network for breast cancer classification. The image size on the disk and spatial dimensions of the histology images are very huge when it's loaded into memory. A complete of 277,524 images belonging to two classes are included within the dataset:

1. Positive (+): 78,786
2. Negative (-): 198,738

The class imbalance, together with the challenging nature of the dataset result ins us obtaining ~70% classification accuracy, ~71% sensitivity, and ~71% specificity.

The usual machine learning methods give confined styles limited to either particular density type or datasets. Even though deep learning methods demonstration auspicious improvements in breast cancer diagnosis but still there is some limitation of data scarcity and computational cost which have been overcome to a vital extent by applying data augmentation and improved computational power of deep learning algorithms.

GitHub Address: https://github.com/karnatisravya/Cancer-Prediction

References

Cicco. P., Catani. M., Gasperi. V., Sibilano. M., Quaglietta. M., & Savini. I. (2019, Jul 11). *Nutrition and Breast Cancer: A Literature Review on Prevention, Treatment and Recurrence*. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6682953/

Goel, V. (2018). Building a Simple machine learning Model on breast cancer data. Retrieved From: https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

Jamdade. V. (2015, April). Urbanization: A Major Culprit of Rampant Breast Cancer. Retrieved From: https://www.researchgate.net/publication/278038581_Urbanization_A_Major_Culprit_of_Rampant_Breast_Cancer

Jjano. J., Srivihok. A. (2014, Nov) Duo Bundling Algorithms for Data Preprocessing: CaseStudy of Breast Cancer Data Prediction. Retrieved From:https://www.academia.edu/27460944/Research_paper_for_data_processing_breast_cancer

K.Shailaja., & M, Jabber. (2018). Prediction of Breast Cancer Using Big Data Analytics. Retrieved from: https://www.researchgate.net/publication/334456222_Prediction_of_Breast_Cancer_Using_Big_Data_Analytics

M. J. M. Broeders, P. Allgood, S. W. Duffy, S. Hofvind, I. D. Nagtegaal, E. Paci, S. M. Moss & L. Bucchi. (2018, Sept 3). Retrieved From: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-018-4666-1#Sec15

Narayan, B. (2020). Convolutional Neural Network for Classification of Histopathology Images for Breast Cancer Detection. Retrieved From: https://ieeexplore.ieee.org/abstract/document/9058279

P Hider and B Nicholas. (1999). Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews. (Original worked published 1999). https://www.ncbi.nlm.nih.gov/books/NBK67849/

Ravert. P., Huffaker.C. (2010, Dec 22,). Breast cancer screening in women: An integrative literature review. Retrieved From: https://pubmed.ncbi.nlm.nih.gov/21129075/

Sheppard. C. (2007, Sept). Breast cancer follow-up: Literature review and discussion. Retrieved From: https://www.scienceadirect.com/science/article/pii/S1462388906001761

Shukla, N. (2017). Breast cancer data analysis for survivability studies and prediction. Retrieved from: https://pubmed.ncbi.nlm.nih.gov/29512500/

Stark.G., Hart.G., Nartowt.B., & Deng.J. (2019). Predicting breast cancer risk using personal health data and machine learning models. Retrieved from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226765

Tiwari.M., Bharuka.R., Shah. P., Lokare. R. (2020, April 10). Breast Cancer Prediction Using Deep Learning and Machine Learning Techniques. Retrieved From: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3558786

X.Juanying., L.Ran., L.Joseph., & Z.Chaoyang. (2019). Deep Learning Based Analysis of Histopathological Images of Breast Cancer. Retrieved from https://doi.org/10.3389/fgene.2019.00080

Jagreet, K. G (2018, December 23). Xenonstack [Blog post]. Retrieved from https://www.xenonstack.com/blog/data-preparation/#:~:text=Need%20of%20Data%20Preparation%20Process%20and%20Preprocessin g&text=Some%20specified%20Machine%20Learning%20and,the%20original%20raw%20data%20set.