Augmenting Cancer Detection

Sreeja Pillai

U1265169

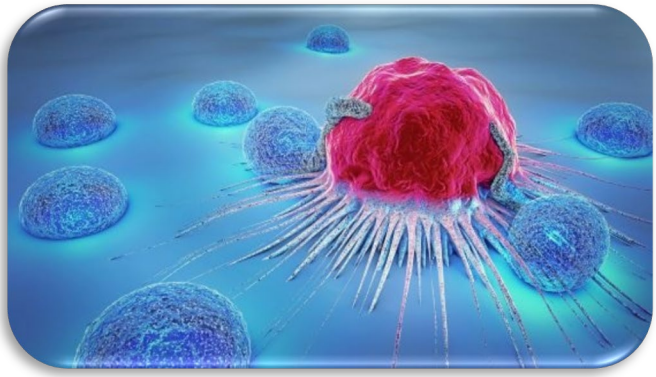
School of computing

Index

Introduction

Every cell in the human body has a DNA, which is responsible for building and maintaining the organism. DNA does this by sending out the code for producing certain kinds of proteins necessary for cell growth and division. DNA cannot directly send out this information to the nucleus; it uses RNA (m-RNA, t-RNA, r-RNA) for this communication. So, at times the DNA mutates and sends signals to divide and grow cell uncontrolled.
This causes the cell to mutate and grow, invading other tissues.

Thus, this disease of controlled cell division is called Cancer.



Advancement in cancer detection has not been rapid. We are still researching on ways to help detect cancer early on. Since the advent, the modus operandi has been "Biopsy" of the tumor to identify if its malignant cell growth or benign. Today we have other methods which do not require a biopsy but still augment doctors in detecting cancerous growth.

Problem Statement:

In this project we are trying to explore and analyze two different datasets, one collected by doing biopsy and another using X-ray images to see if we can get almost the same accuracy in detecting cancer using data with and without biopsy.

Literature Review

In this project we are exploring two datasets.

1. Mammography Dataset from UCI: mammography is the most effective method for breast cancer screening. But it has a high rate of false positive screening, thus causing unnecessary biopsy. Thus, we are developing CAD techniques to assist the doctors in more accurately predicting malignant tumor and thus increasing the chances of going for biopsy when it is truly needed. This dataset was collected using data from mammography done on patients, no biopsy.
   CAD – Computer Aided Diagnosis are methods developed using ML, DL and AI based techniques. For this dataset they have gone with decision trees to predict cancer.

2. RNA-seq Gene Dataset from UCI: DNA does not always reveal the whole picture of the cancer cell growth, but RNA can. Research has showed that, the RNA profiles of cancer patient can help us learn more about the biology of cancer and on how to treat them. This dataset is gene expression of RNA sequence of various cancerous tumors. The main aim of this dataset was to discover molecular aberrations which are common or totally different among different cancer tumors and use them to predict the type of cancer it is.
   This dataset was taken by doing biopsy on many patients with different class of cancers.
   This dataset was built for myriad of uses, we will be using it to see how useful/accurate data is collected using biopsy for predicting cancer.

Descriptive Statistics

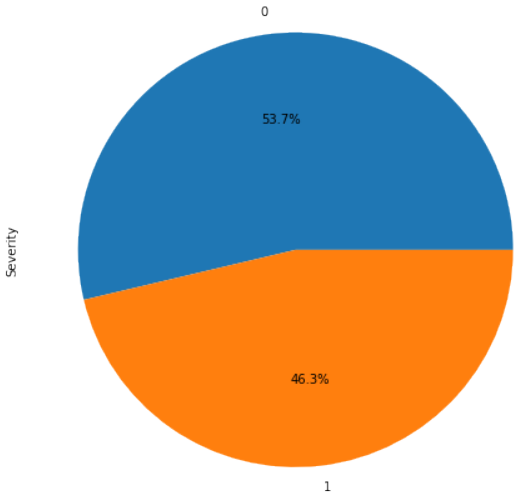
1. Mammography dataset:

- Shape:

The dataset has recorded information for 961 patients.
Captured 6 attributes for each patient.

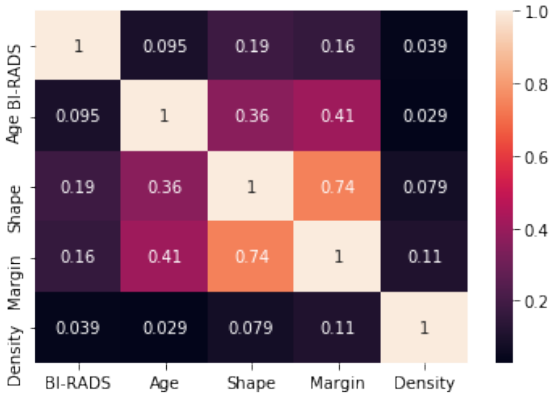
- BI-RADS assessment: 1 to 5 (ordinal, non-predictive!)
- Age: patient's age in years (integer)
- Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
- Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
- Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
- Severity: benign=0 or malignant=1 (binominal, goal field!)

- Imbalance:

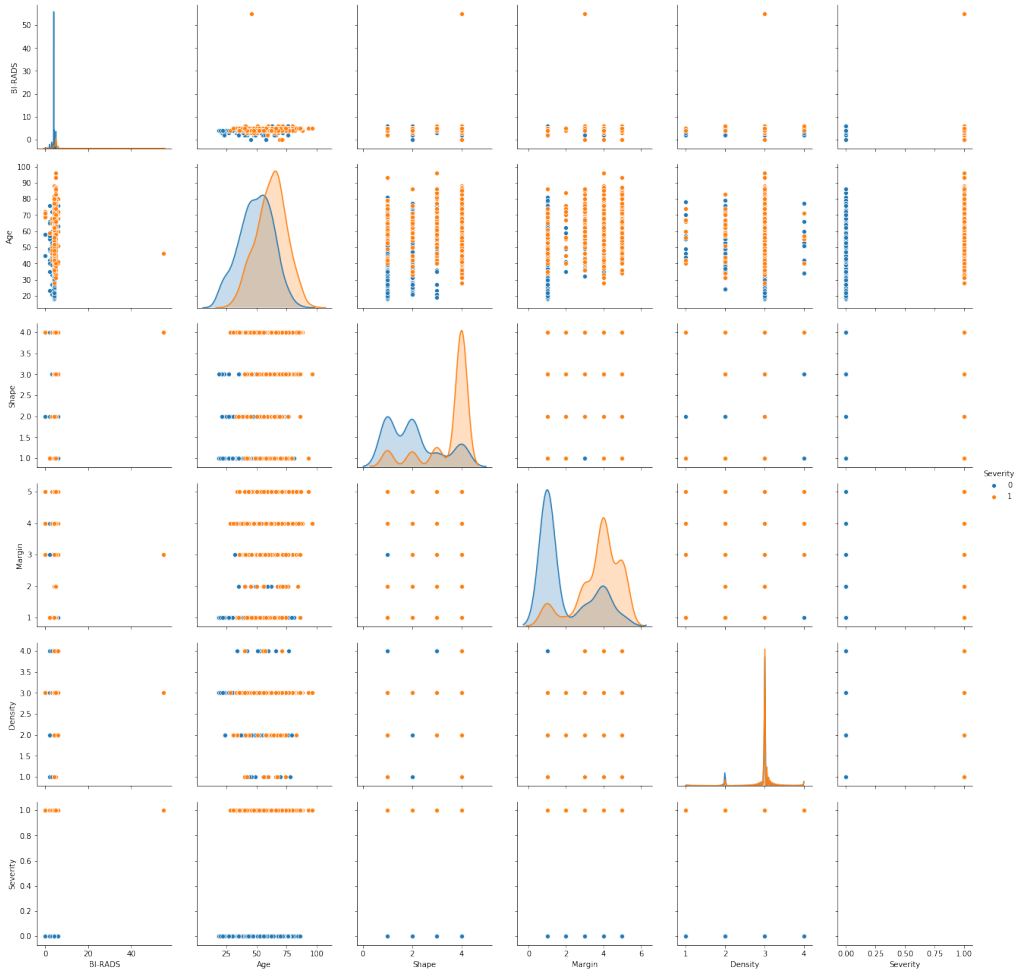


The Target Label we are predicting is Severity. The pie chart shows a Balanced Dataset.

➢ Correlation:



The correlation matrix, indicates.
"margin and "shape" are positively correlated.
"margin" and "age" are positively correlated.
Other attributes do not have high correlation.



The above plot gives us a more detailed knowledge about the correlation between the attributes.
let us take x-axis "shape" and y-axis "age":
An age group of 20 to 40 even if the shape of the tumor increases the cancer is benign.
But as we grow older the shape of the tumor does signify a malignant tumor.

- Missing Values:

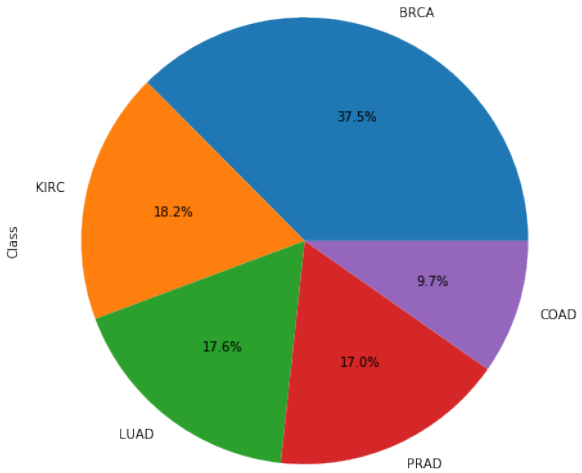
  The dataset had 162 missing values, since we have less samples already, we are not dropping them, we are instead replacing categorical features with mode and numerical ones with mean.

2. RNA-seq gene expression dataset:

- Shape:

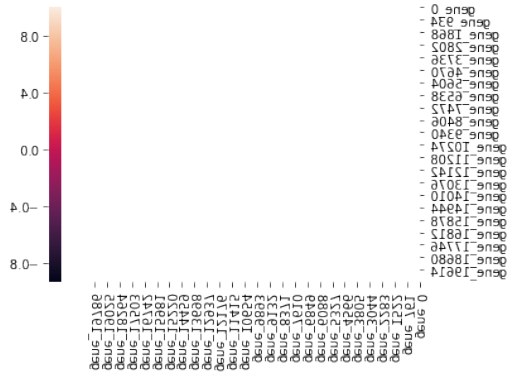
  The dataset has biopsy data for 801 patients. Each recording 20532 gene information.
  The target label we are trying to predict is “Class” , they are:
  Breast invasive carcinoma - BRCA
  Colon adenocarcinoma - COAD
  Kidney renal clear cell carcinoma - KIRC
  Lung adenocarcinoma - LUAD
  Prostate adenocarcinoma - PRAD

- Imbalance:



The dataset is imbalanced we will be using SMOTE to rectify this.

➢ Correlation:



We have 20532 features, hence drawing a correlation matrix is not a feasible solution to understand correlation.

➢ Missing Values:

This dataset had no missing or null values.

Methods & Results

- Data Cleaning:

  Its one of the most crucial steps in analysis. Each dataset needed to be changed before we could apply any algorithm.

  Null vales – can be either dropped or replaced. As the dataset had few samples, I decided to replace them with mode or mean, respectively. We need to take care of them else the model will fail when it encounters null values.

  Encoding categorical features - algorithms like decision trees don't get affected by categorical data, but other ML and DL algorithms need them to be encoded.

  Correlation – correlation between features don't affect classification problems much as it does regression. We have used pandas corr() in the mammography dataset. But RNA -seq has 20532 features, it is not feasible to create a correlation matrix. Hence, we use ANOVA to do the same, which helps us find the gene which contributes to classifying a cancer type. Thus, reducing the feature size by 961 columns.

  Imbalance - This has an impact while training a model. There will be very less data for the model to learn an underrepresented class thus having high generalization error. We use SMOTE to resolve this.

  Train/Test split-Stratify – Stratification is very important while training and testing. To understand how well our model performs on all variations of unseen data.

Method\Dataset	Mammography	RNA-seq
Null values	Replaced categorical with mode. Numerical with mean	No Null values
Imbalance	N/A	SMOTE
Correlation	pandas corr()	ANOVA
Encoding categorical variables	N/A	keras to_categorical
Train/Test split- Stratify	Sklearns's train_test_split	Defined function

- Models:

Non-DL models:

Here we are using 4 non-dl algorithms.

Decision Trees, Logistic Regression, KNN.

- f1_scores without optuna:

Dataset \ Methods	Decision Trees	Logistic Regression	KNN	SVC
Mammography	0.7529	0.799	0.6951	0.7388
RNA-seq Gene	0.9557	1.0	1.0	1.0

- F1_score with optuna:

  Mammography dataset:

  The best model was Logistic Regression with hyper-parameters solver = 'liblinear'

  F1_Score = 0.80874

  RNA-seq Gene dataset:

  The best model was KNN with hyper-parameters n_neighbors=9, weights="uniform", algorithm="brute", metric="euclidean"

  F1_Score = 1.00

Discussion:

i. RNA-seq Gene dataset performs better with KNN as the dataset was constructed to classify similar aberrations among different cancerous tumors. Thus, there was some underlying similarities for KNN to group.
ii. Mammography dataset worked better with Logistic because it was set of metrics and we had to find a relation among them for each sample to predict the output. Hence it worked well.

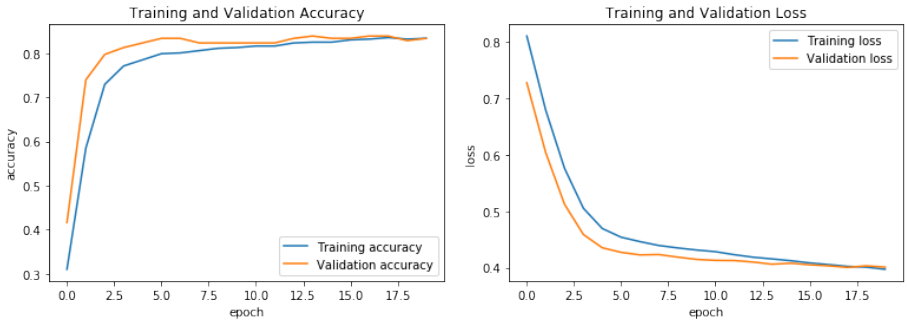
DL models:

We have used Fully Connected NN and Conv 1D

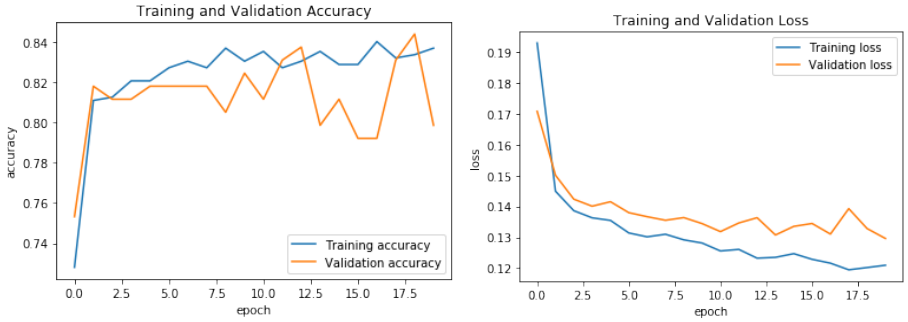
DL models F1_scores:

Dataset \ Methods	Fully Connected NN	Conv 1D
Mammography	0.7937	0.7977
RNA-seq Gene	0.8796	0.8984

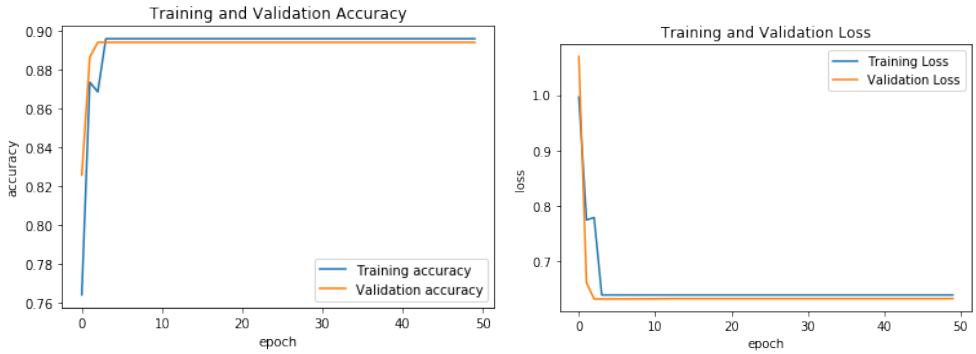
Discussion:



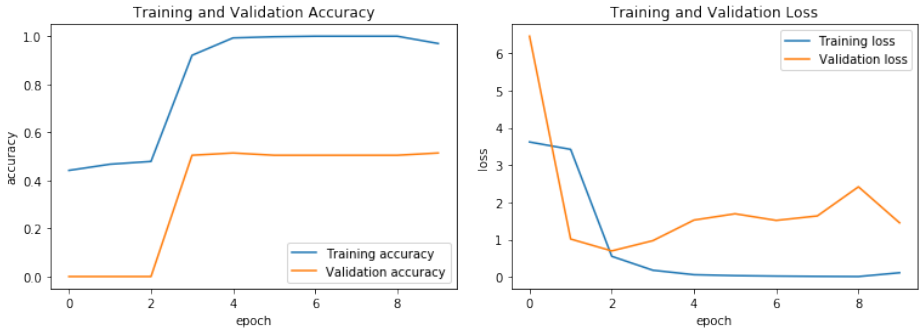
i. The above graphs are from Mammography dataset for Fully connected NN, it is a good fit.



ii. The above graphs are from Mammography dataset for CONV1 D, it is a good fit, but If I run it for more epochs, it might overfit.



iii. These graphs are for RNA dataset for fully connected NN, they are a good fit



iv. These are the graphs from RNA dataset for CONV 1 D, it is clearly an overfit, I have used a 3 hidden layer conv 1d with two dropout layers and since the data is huge in feature space I had to give very small values for output space else the colab gives up on RAM.
I feel if I can use auto encoders and pass its value into the model it might work better.

Challenges Faced:

- The RNA-seq Gene dataset is rich in feature space has 20532 features a normal panda correlation would not work.
- Hence had to perform ANOVA test to find correlation and remove features.
- The multiclass dataset had imbalance and hence the normal train_test_split wouldn't work. Had to write my own stratification function.
- Used SMOTE to overcome imbalance but didn't improve my f1_score much.
- My Conv 1D is overfitting, and since the data has 20532 features the parameters passed from one layer to another becomes too large. If I reduce the output space using filters, the model does not learn much and colab gives up on RAM. I wanted to try auto encoders and feed the conv1 d did not get time.
- I am not able to use OPTUNA for DL based models, I need to work on that.

Conclusions:

- Non-DL models have performed better than DL-models, I feel the reason could be limited rows of data for training.
- Gene Data proves to be more accurate while classifying cancer. As of now within the scope of this project we can state that biopsy will classify/detect cancer almost 98% of the time correctly.
- Mammography dataset along with the CAD we have applied shows good accuracy in detecting cancer, but it is not good as compared to using genome sequence.
- If we can get more data, or augmented images from mammography. It might help in training our models well.