# Melanoma Skin Binary Imagine Classification



**Group Members**

* Achal Suresh -asuresh6@jhu.edu
* Mark Rao -jrao7@jh.edu
* Maitreiyi Maheshwari -mmahesh5@jh.edu
* Maria Vargas -mvarga15@jhu.edu
* Lu Cao -lcao12@jh.edu
* Ruchika Raikar -rraikar1@jh.edu

Introduction

Melanoma is one of three types of skin cancers that starts in skin cells called melanocytes. Melanoma is a type of cancer that could be visible on top of the skin like a mold or spot in the surface of the skin.  Cancer is always something that completely affects the person but also family and friends of the affected.  Treatments are also long, painful, and tedious. However, as a human being we have to see the positive side even in the dark moments. The good thing about this type of cancer is that if it is detected in an early stage it could be treatable and it will give a high probability of rate of survival. However, if it’s not detected at an early stage it could spread very easily to other parts of the body and be fatal.

The objective of our product is based in a Machine Learning application which will try to reduce mortality by having an early detection. The tool we have created is an application for medical professionals to detect Melanoma in an early stage which will be extremely helpful for dermatologists to detect the cancer on time and treat it as soon as possible.

EpiDermis Application

Epidermis is the name of our application which the objective of the dermatologist is to have the availability to see and compare multiple skin pictures to analyze the suspicions skin lesions and determine if it is melanoma or not. The application will give the opportunity to the dermatologist to analyze the suspected skin lesion and determine if the person needs a biopsy by taking a sample of skin for testing and then taking it to the lab for examination.

EpiDermis Application Benefits

The EpiDermis application will add an incredible amount of value to the dermatologist and the American Cancer society to help detect cancer on time so it could be treatable.  The application could train a set of pictures in a database that could have 100,000 or more  images and span the breadth of potential melanoma cases for certified dermatologists to distinguish skin melanomas from benign moles. In our research and solution, we developed the random forest, xgboost and CNN ,and leaning CNN models. This application will give outstanding results that a computer system trained over a short period of time could outperform human experts who had spent multiple years in training plus the experience of each dermatologist.

The advantage of the application is that the computer system and its diagnostic performance will continue to learn.

XGBoost Model

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. The name XGBoost, though, actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms.

First, the dataset was split into a 70% training set and a 30% testing set. Then the XGBoost model was defined and fit to the training set. Then the model was tested on the testing set and it had a training F1 Micro Average of 0.847, test F1 Micro Average of 0.82, and test Accuracy of 0.82. The confusion matrix was as the following.And the roc curve shows a great learning ability of the model.

The parameter was further set up by adding restrictions for the model in the number of classes, evaluation metric, and objective. Then the model was tested on the testing set and it had a training F1 Micro Average of 0.938, test F1 Micro Average of 0.839, and test Accuracy of 0.84. The feature importance was analyzed by the Cohen Kappa quadratic score.Point 13757 was the most important point on the picture in the XGBoost model. It might be the most likely spot for the Malignant. 5-fold cross-validation was also applied to improve the credibility of the model. The training F1 Micro Average was 0.858. Test F1 Micro Average was 0.827 and Test Accuracy was 0.826. The most important feature was 13757 and 21613.

Basic Classification Model

Classification is the cognizant way to categorize the cases as benign or malignant.We have explored different classifiers to identify the accuracy of the acquired dataset. Initially the entire dataset was split into 80:20 ratio for training and testing data. Naive Bayes classification algorithm is more suitable for binary and multiclass classification. In this case the accuracy is lower and the misclassification rate is comparatively higher. Predictive supervised decision trees that consist of nodes that are driven by decision rules that help generate outcomes. Here, while the trained model was oversimplified and the trees are effective on both continuous and categorical variables, the test results did not generate higher accuracy. Support vector machine tries to maximize the hyperplane to generate a clear distinction between the classes. Here the accuracy is higher on the training data however, it drops for the test data set. Multi-layer perceptron classifier is feedforward and works best on categorical variables but here its training loss did not improve beyond tol=0.010000 for 10 consecutive epochs.

In this ensemble tree based algorithm it generates a set of decision trees and performs averaging to improve prediction and control overfitting. Here, the accuracy is little above 80% and misclassification rate is lower than the above two classifiers. We further used 5-fold cross validation to evaluate the random forest model and used grid search to identify the optimal hyperparameters for identifying the best model with best scores, best params and highest accuracy. A receiver operating characteristic (ROC) curver is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. We generated two ROC, one is for five different classifiers and the other is for cross validation.

Convolutional Neural Network (CNN):

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.

To train our image classification model we started with a basic CNN model built on Keras Framework and decided to iterate the CNN models continuously to improve the accuracy of the model. The iterations performed on the model are:

oAdding Dropout layer to increase the speed of the epoch and also possibly prevent overfitting of data.

oAdding Class weights to prevent Oversampling as the dataset has more benign images than malignant tumor images

oAdjusting the learning Rate.

oAdding Additional Neural Layers

oAdding Extra data to the training set.

From these we observed that Adding more layers had an initial impact on the performance of the model through improved accuracy however the accuracy was coming up to stagnant and slowly regressing.

A method to counter this we adopted the ResNet50 with additional data to create a more balanced dataset to improve the accuracy of the model. ResNet50 is a pretrained convolutional neural network that is 50 layers deep.

ResNet utilizes identity shortcut connections to skip layers to prevent the vanishing gradient problem by connecting output layer to the intermediate layer and skipping some other layers.

By Comparing the accuracy of the CNN models we identified that the best models are:

· ResNet50 with 20 epochs (with extra data)

· Deep CNN with extra data performed better than with class weights

Results:

Classification is the cognizant way to categorize cases as benign or malignant

* *Gaussian Naive Bayes -* Naive Bayes classification algorithm is more suitable for binary and multiclass classification
  + Accuracy: 0.639
  + Training set score: 0.700000
  + Test set score: 0.638889
* *Decision Tree -* consist of nodes that are driven by decision rules that help generate outcomes
  + Accuracy: 0.700
  + Training set score: 1.000000
  + Test set score: 0.700000
* *Random Forest -*  set of decision trees and performs averaging to improve prediction and control overfitting
  + Accuracy: 0.811
  + Training set score: 1.000000
  + Test set score: 0.811111
* *Support Vector Machine -* maximize the hyperplane to generate clear distinction between the classes
  + Accuracy: 0.794
  + Training set score: 0.836111
  + Test set score: 0.794444
* *Multi-Layer Perceptron:*

Feedforward and works best on categorical variables, it generates linear combination using its input weights

* + Accuracy: 0.806
  + Training set score: 0.808333
  + Test set score: 0.805556
* *Pipeline of Decision Tree* - helps you create and optimize a baseline Decision Tree

model for Binary Classification

* + Accuracy - 0.794 (79%)
* *Receiver Operating Characteristics:* AUC is 0.66, Illustrates the diagnostic ability of a binary

classifier system as its discrimination threshold is varied

* *Cross Validation Score* - effectiveness of your model,

particularly in cases where you need to mitigate over-fitting

* + Mean Accuracy - 0.776388 (77%)
* *XG Boost Performance*
  + Training F1 Micro Average was 0.858.
  + Test F1 Micro Average was 0.827.
  + Test Accuracy was 0.826.
  + The Cohen Kappa quadratic score
  + The most important feature was 13757 and 21613
* *CNN Model:*
  + *The ResNet50 which is pretrained transfer learning model showed the best accuracy among the CNN Models*

Appendix

**Basic Classification Model**

*logistics*

Training Data : (720, 16382)

Test Data : (180, 16382)

Training Target  Data : (720,)

Test Target Data : (180,)

*Explore - Gaussian Naive Bayes*

Predictive supervised binary classifier

Misclassified : 65

Accuracy: 0.639

Training set score: 0.700000

Test set score: 0.638889

*Explore - Decision Tree*

Misclassified : 54

Accuracy: 0.700

Training set score: 1.000000

Test set score: 0.700000

*Explore - GridSearch*

{'bootstrap': False, 'criterion': 'gini', 'n\_estimators': 250}

0.8152777777777779

*Explore - Random Forest*

Misclassified : 34

Accuracy: 0.811

Training set score: 1.000000

Test set score: 0.811111

*Explore - Support Vector Machine*

Misclassified : 37

Accuracy: 0.794

Training set score: 0.836111

Test set score: 0.794444

*Explore - Multi-Layer Perceptron*

Misclassified : 35

Accuracy: 0.806

Training set score: 0.808333

Test set score: 0.805556

*Generate - Confusion Matrix*

Confusion Matrix (Random Forest)

[[143   2]

 [ 32   3]]

Confusion Matrix (SVC)

[[145   0]

 [ 35   0]]

Confusion Matrix (Decision Tree)

[[119  26]

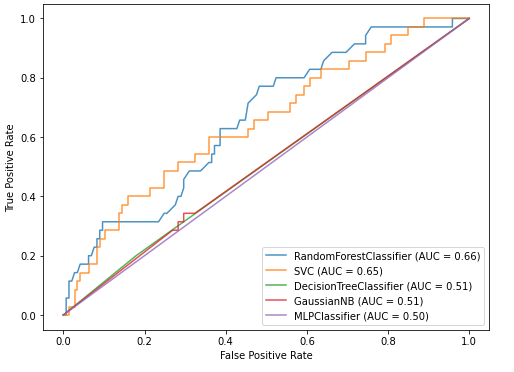
 [ 28   7]]

Confusion Matrix (MLP)

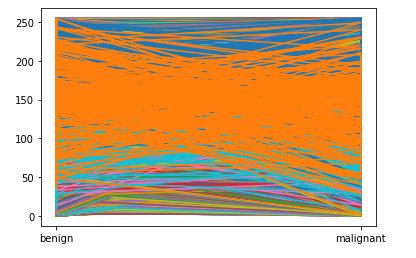
[[145   0]

 [ 35   0]]

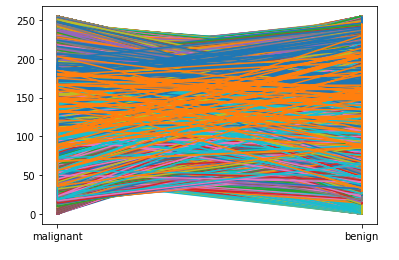
*Explore - ROC*



Basic plot of train predictor and response data



Basic plot of test predictor and response data



*Explore - Cross Validation Score*

Estimator - Random Forest

All Accuracies, Mean, Std -

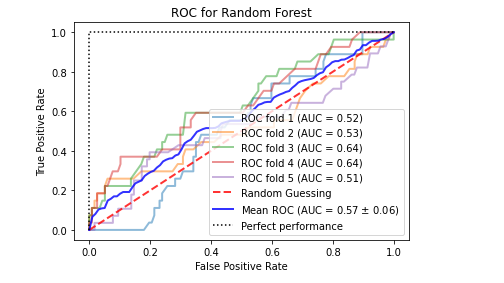
[0.74305556 0.80555556 0.76388889 0.78472222 0.78472222]

0.7763888888888889

0.02124591463996994

Explore - ROC of Random Forest

Split = 5



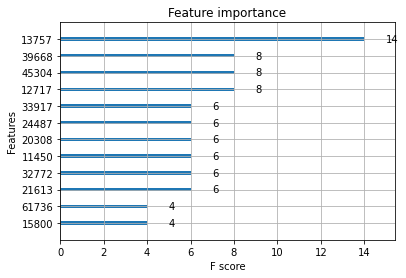
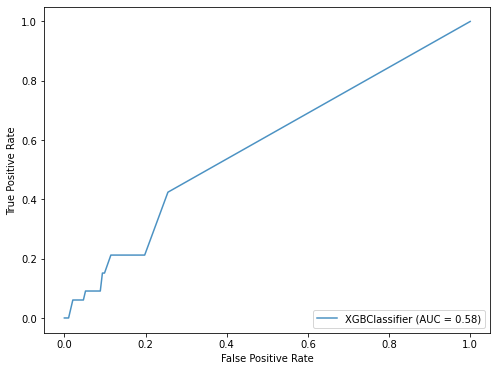
*Explore - Pipeline of Decision tree*

StandardScaler, PCA (components = 10), RandomForest(max\_depth = 10)

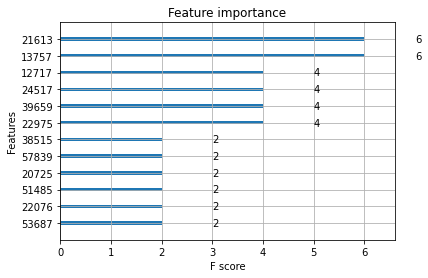
Accuracy = 0.794

Confusion Matrix for XGBoost

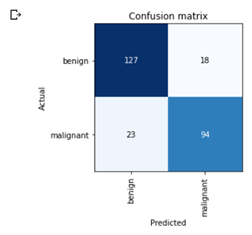
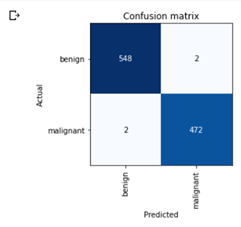
|  |  |
| --- | --- |
| 182 | 10 |
| 30 | 3 |



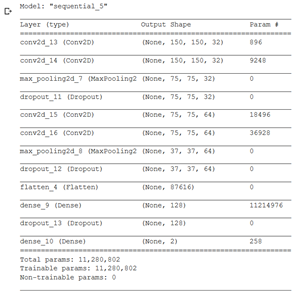
(Figure 1: ROC Curve for XGBoost)  (Figure 2: Feature Importance for XGBoost)



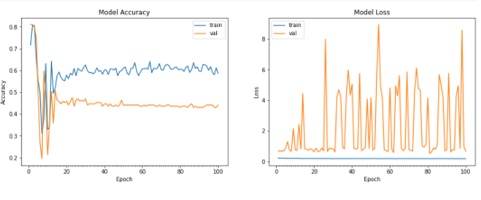
(Figure 4: Feature Importance for XGBoost after Improvement)



(Figure 5: Confusion Matrix of ResNet50)



(Figure 6: CNN Model summary)



(Figure 7: Accuracy and Loss for the CNN Model)

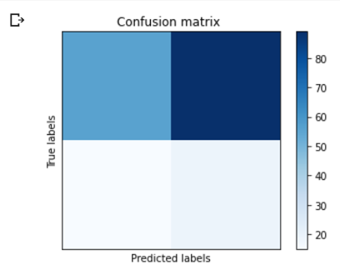


Figure 8: Confusion Matrix for the CNN model

Resources:

<https://www.mayoclinic.org/diseases-conditions/melanoma/diagnosis-treatment/drc-20374888>

<https://www.cancer.org/cancer/melanoma-skin-cancer/about/what-is-melanoma.html>

[https://medium.com/@tylereyarnell/melanoma-skin-binary-image-classification-using-xgboost-algorithm-f747c133851](https://medium.com/@tylereyarnell/melanoma-skin-binary-image-classification-using-xgboost-algorithm-f747c1338511)