## STAT 542 / CS 598: ****Group Project: Skin Cancer Diagnostics****

## ****Ashwin Amrutphale (ashwina4) (Team Lead)****

## ****Pratik Patwari (patwari3)****

## ****Rohit Bansal (bansal3)****

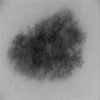
1. **Project Description and Summary**

## 1.1 Goal:

## Construct classification models for identifying malignant moles based on the pixels (RGB colors) of these images and propose new data processing/feature engineering approaches that improve the interpretability and model accuracy. This project compares pixel-based approach with feature-based approach for the image and compare results for the same.

## 1.2 Approach:

**In the first pixel-based approach, we have used EBIMAGE library in r to read the image and then resized the image into 100 x 100 pixel for resampling and for reducing channels. For the image classification, 3 classifier models were developed which are 1. Random Forest 2. KNN 3. SVM**

**Original image After processing**

**In 2nd approach, feature based extraction was used. ABCD** (Asymmetry, Border Irregularity, Color, Diameter) **Features** are features that dermatologists commonly use. The significance of these features is that, based on them, a dermatologist decides whether an observed skin lesion is a melanoma or not. Based on these features we have classified using 1. 2.

**1.3 Results:**

As shown below, feature based classification has more accuracy as compared to 1st approach

**Approach 1 (Pixel Based):**

|  |  |
| --- | --- |
| **Classification Model** | **Accuracy** |
| **Random Forest** | **72.22** |
| **KNN** | **65.55** |
| **SVM** | **71.11** |

**Approach 2 (Features Based):**

|  |  |
| --- | --- |
| **Classification Model** | **Accuracy** |
|  |  |
|  |  |

1. **Data Processing for Question 1**

EBIMAGE library was used to perform Data Processing and Analysis on the benign and malignant images. Images were resized to 100x100 and turn them into greyscale so that we can load them into R easily and reduce training time. Each image will be turned into a vector of length 784, with each element representing the value in a pixel.

The values of the pixels from the resized and greyscale images where then stored into an array. The benign class was labeled as 1 whereas malignant class was labelled 0.

lable <- c(rep(0,150),rep(1,150))

imageData <- rbind(benign\_data,malignant\_data)

This data was then randomly split into a training and test set in the ratio 70:30 respectively with PCA applied to the feature train set in order to determine the features that accounted for 85% of the variance. It was found that there were total 29989 features present in the data set.

extract\_feature <- **function** (dir\_path, width, height, is\_benign = TRUE, add\_label = FALSE) {

img\_size <- width\*height

*## List images in path*

images\_names <- list.files(dir\_path)

*## This function will resize an image, turn it into greyscale*

feature\_list <- pblapply(images\_names, **function**(imgname) {

img <- readImage(file.path(dir\_path, imgname)) *## Read image*

img\_resized <- resize(img, w = width, h = height) *## Resize image*

dim(img\_resized)

img\_vector <- imageData(img\_resized)

**return**(img\_vector)

})

feature\_matrix <- do.call(rbind, feature\_list) *## bind the list of vector into matrix*

feature\_matrix <- as.data.frame(feature\_matrix)

names(feature\_matrix) <- paste0("pixel", c(1:img\_size)) *## Set names*

**return**(feature\_matrix)

}

1. **Classification Models Based on Pixels**

After data processing was done, next step is to process the classification model on pixel and find the accuracy with each model. Three classifiers were then chosen: K- Nearest Neighbor (KNN), Random Forest and Support Vector Machines (SVM).

**KNN:** KNN classification was used as the 1st classification model and it was fitted with transformed train data and test data. The performance of the model was measured on accuracy, recall, f-1 score and precision. For kNN algorithm, the tuning parameters are ‘k’ value and number of ’features/attributes selection. We have used k 1:20 to improve performance of the mode. Its performance was 72% accuracy, 0.8 recall, 0.742 f-1 score and 0.692 precision.

**for** (i **in** 1:20)

{

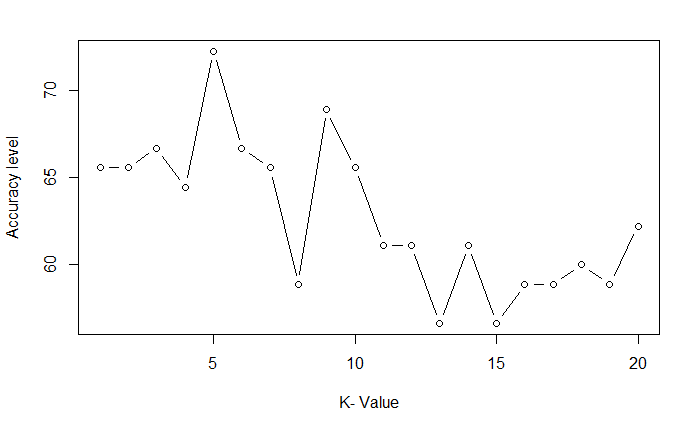
knn.model = knn(train=train\_images, test=test\_images, cl=as.factor(train\_labels), k=i)

optm[i] = (100 \* sum(as.factor(test\_labels) == knn.model)/NROW(test\_labels))

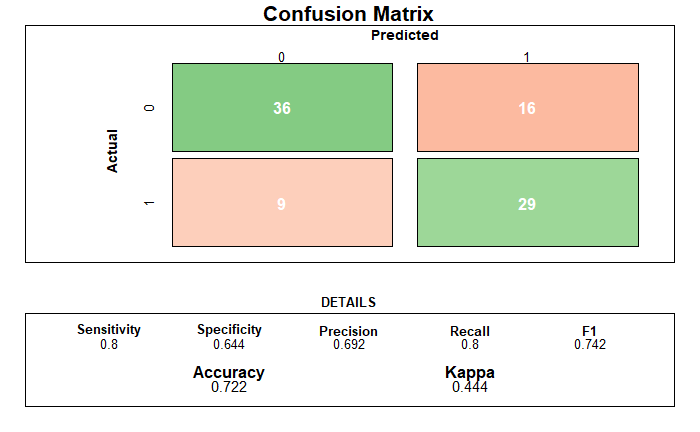
cat(i,'=',optm[i],'\n') *# to print % accuracy*

}

##### As shown in below chart,a**t k=5, maximum accuracy achieved which is 72%, after that, it seems increasing K increases the classification but reduces success rate.**



**Confusion matrix was created to describe the performance of a classification model.**  it predicted malignant correctly 36 times and benign correctly 29 times. It misclassified benign as malignant 16 times and vice versa 9 times.



**Random Forest:**

Radom forest has been used as seconds classification model; it was also fitted with the train set to predict the test set. The performance of the model was also measured on accuracy, recall, f-1 score and precision. For Random forest, the tuning parameters ntree and random state was used to tune the mode. Its performance was 72% accuracy, 0.689 recall, 0.713 f-1 score and 0.738 precision.

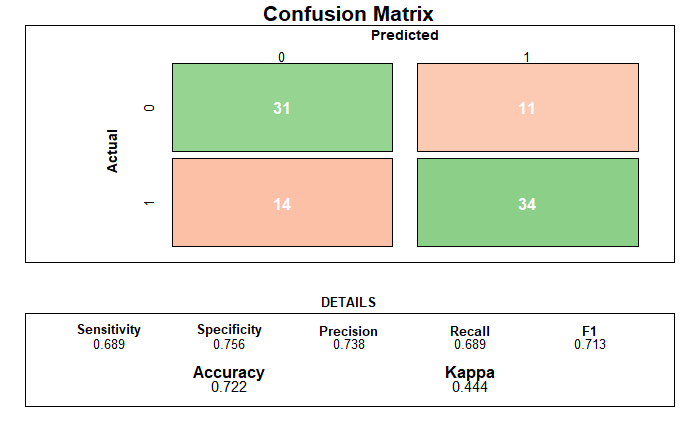
**Confusion matrix was created to describe the performance of a classification model** on a set of test data**.**  it predicted malignant correctly 31 times and benign correctly 34 times. It misclassified benign as malignant 11 times and vice versa 14 times.

**library**(ggplot2)

rf <- randomForest(x = train\_images, y = as.factor(train\_labels)

, xtest=test\_images, ytest=as.factor(test\_labels),keep.forest=TRUE,ntree = 500, random\_state = 0)

printpaste("Random Forest Accuracy:", (sum(ifelse(rf$test$predicted == as.factor(test\_labels),1,0)) / length(rf$test$predicted)) \* 100))



**SVM:**

SVM has been used as third classification model; it was also fitted with the train set to predict the test set. The performance of the model was also measured on accuracy, recall, f-1 score and precision. For Random forest, the tuning parameters ntree and random state was used to tune the mode. Its performance was 73% accuracy, 0.689 recall, 0.721 f-1 score and 0.756 precision

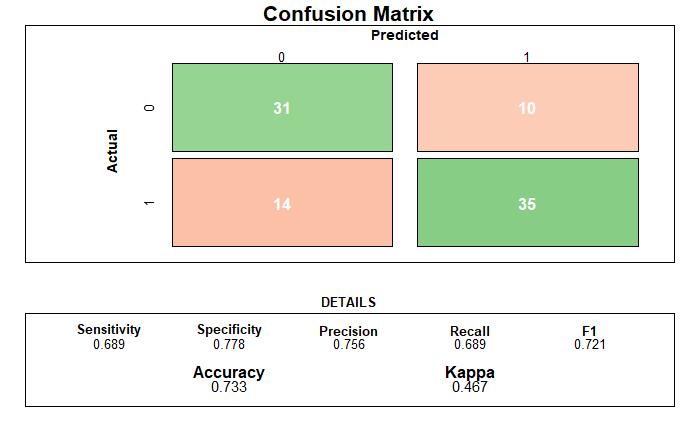
**Confusion matrix was created to describe the performance of a classification model.**  it predicted malignant correctly 31 times and benign correctly 35 times. It misclassified benign as malignant 11 times and vice versa 14 times.

**library**(e1071)

svm.model <- svm(x=train\_images,y=as.factor(train\_labels),kernel = 'radial', type = 'C-classification', cost=10, scale=FALSE)

y\_pred = predict(svm.model, newdata = test\_images)

print(100 \* sum(as.factor(test\_labels) == y\_pred)/NROW(test\_labels))



**Conclusion:**

Below table outlines the summary of classifier on Accuracy, Precision, Recall and F-1 Score. SVM performed best overall while KNN and Random Forest were mostly equal in performance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy %** | **Precision** | **Recall** | **F-1 Score** |
| **KNN** | 72% | 0.692 | 0.8 | 0.742 |
| **Random Forest** | 72% | 0.738 | 0.689 | 0.713 |
| **SVM** | 73% | 0.756 | 0.689 | 0.721 |

Based on the confusion matrix as well, we can say that SVM is best predictor for benign as it predicted correctly 35 times and KNN is best predictor for malignant as it predicated correctly 36 times.

1. **New Feature Engineering Approaches**

**4.1 Literature review**

**ABCD Features**

These are the features that dermatologists commonly use. The significance of these features is that, based on them, a dermatologist decides whether an observed skin lesion is a melanoma or not. The ABCD rule followed by the doctors has been mathematically modeled and implemented in this project. The details of these features are as follows:

**Asymmetry:** is an important feature for detection of skin cancer. It is represented in terms of Asymmetry Index. In order to find this feature, it is necessary to find the true axis of symmetry around which the image of lesion area is folded. This was done by first finding the center of rotation of the image. The center of image is taken as the point placed in the middle row and middle column of a rectangular region that encompasses the image. The image is then folded at different axes, each passing through the center. A total of 18 axes at increments of10◦, for a total of180◦, were tried. The true axis of symmetry is the one at which the two halves, when folded, result in maximum overlap.

**Border Irregularity:** Another important feature for the identification of melanoma or non-melanoma is the shape of the lesion. A regular border is usually benign while an irregular border indicates melanoma.

**Color:** It is the most important feature in melanoma and can be identified by using the statistical features i.e., mean, median and standard deviation of each plane (red, green and blue).

**Diameter:** The main feature that has been often ignored is the diameter. Diameter changes with distance between the capturing device and the lesion. For fair comparison, the distance between camera and the lesion should remain the same for all images in a data-set.

**4.2 Feature Engineering**

**4.3 Classification Models Based on New Features**