# Data Mining

# Project Report

**Name:** Ali Kamal

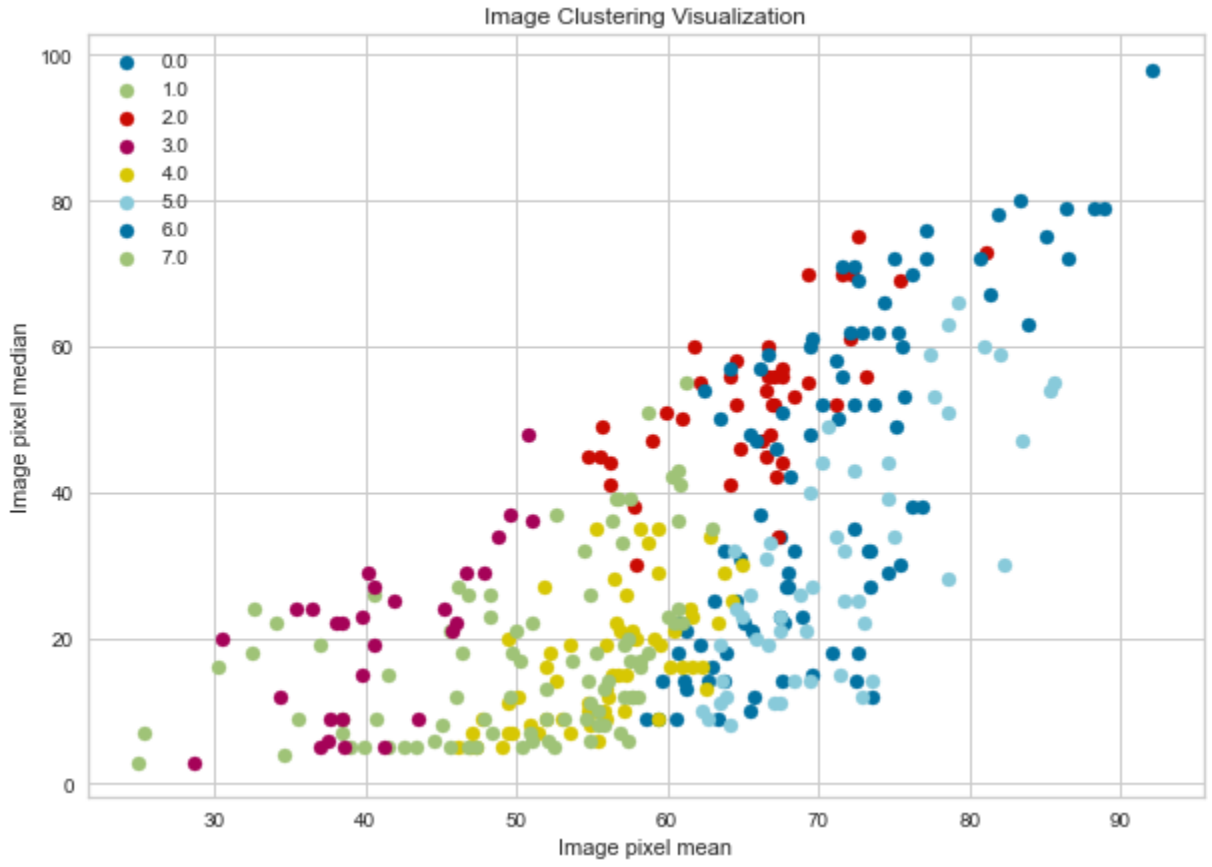
**Roll Number:** 19I-1865

Ali Kamal
19I-1865

# Clustering and Sampling

K-means algorithm was used for clustering the images. Optimal value of k was determined through the K-Elbow method (which, in our case, returned 8 as the optimal value of K). These 8 clusters were now treated as the ground truth, and we shall be using these labels for further use.

Against each of our 8 clusters, 70% of images were taken as our sample. So, in the end, we ended up with equal distribution of images with respect to clusters, as we are choosing 70% of the total images against each cluster.



Ali Kamal
19I-1865

# Image Segmentation

**Original Image**

**Segmented Image**

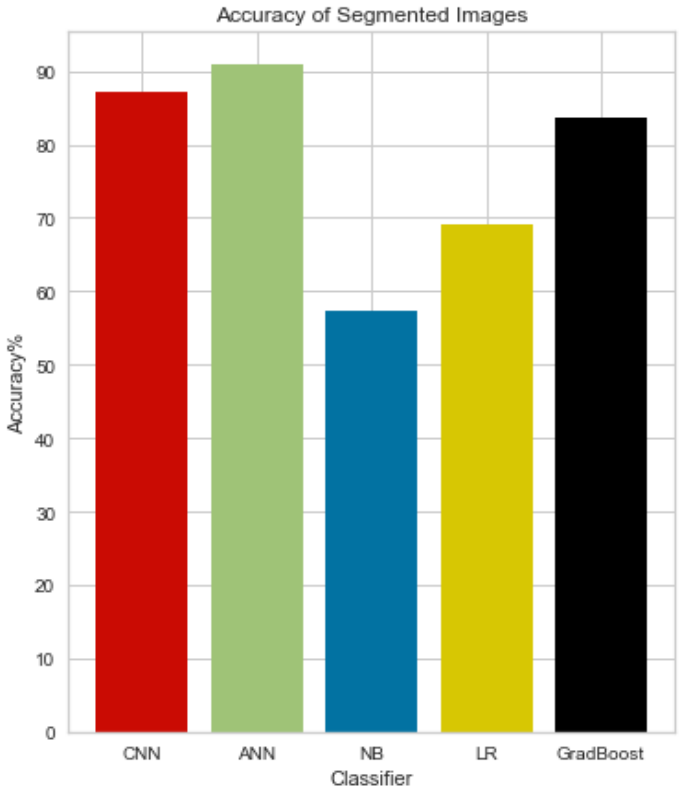
# Comparison of Models for Segmented and Non-Segmented Images

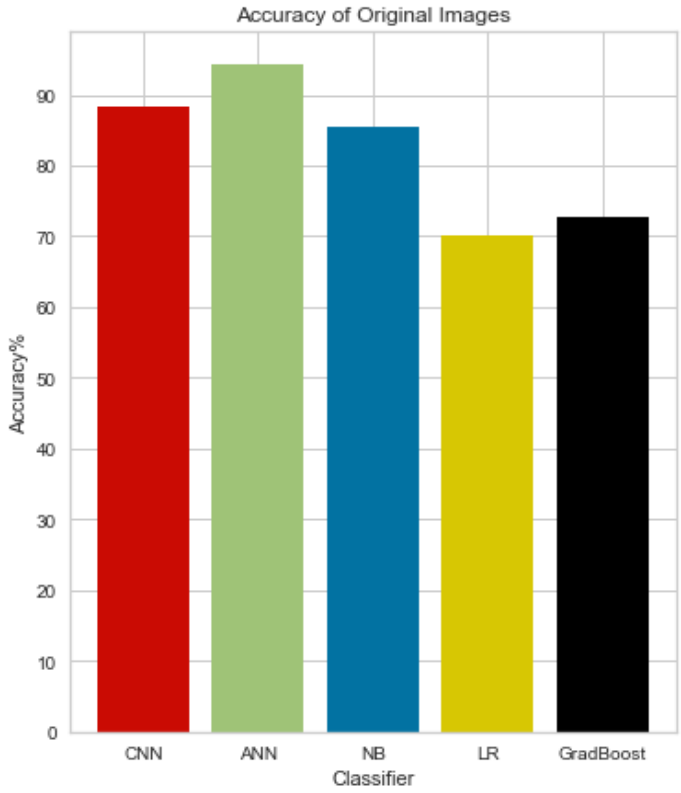
## Accuracy:

| | Segmented Images | | Non-Segmented Images | |
|---|---|---|---|---|
| | Accuracy(%) | Time(ms) | Accuracy(%) | Time(ms) |
| CNN | 87 | 178 | 88 | 232 |
| ANN | 90 | 137 | 94 | 231 |
| Naiive Bayes | 57 | 0.02 | 85 | 0.02 |
| Logistic Regression | 69 | 0.46 | 70 | 0.4 |
| GradBoostingCLF | 83 | 17 | 72 | 25 |

Deep Learning models perform better on non-segmented images, where they give the better accuracy. However, the deep learning models are quite a bit faster on segmented images, as they were able to run on 1000 Epochs in less time than on 10 Epochs for Non-Segmented Images.

Ali Kamal
19I-1865



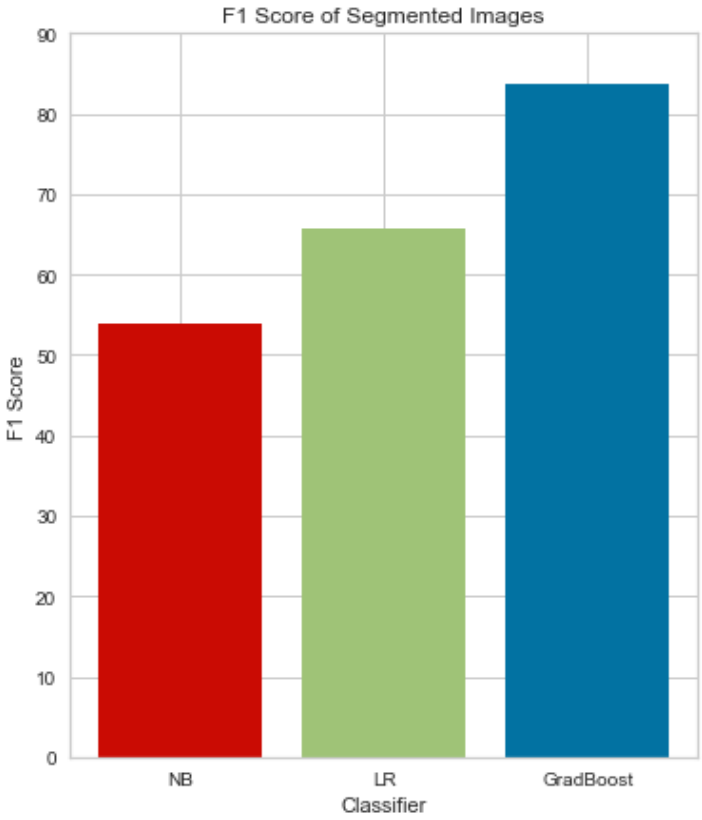

Gradient boosting classifier performed significantly better on segmented images, as seen by the above figures. All of the models were substantially faster on segmented images.

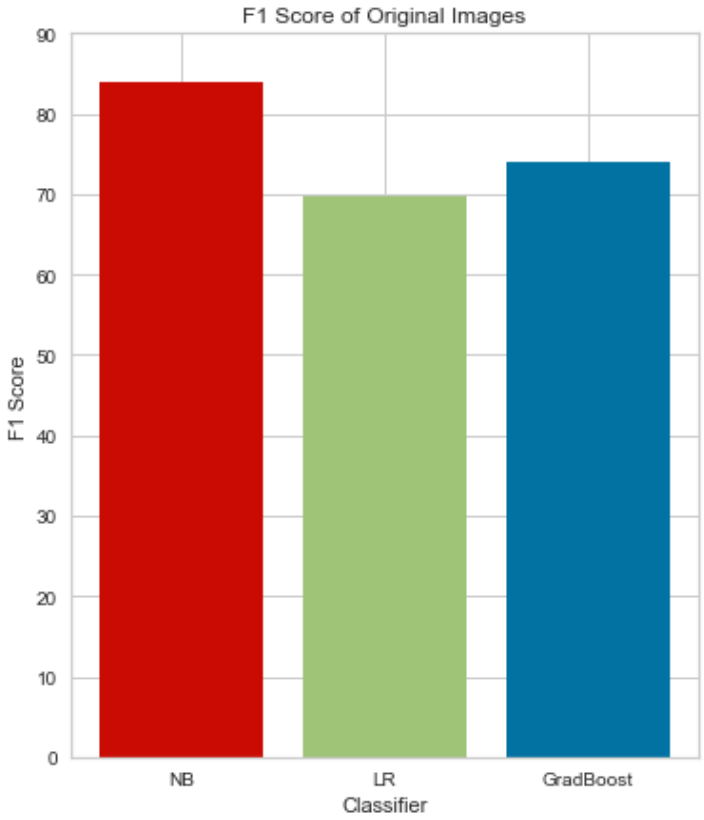
## F1 Score:

| | Segmented Images | Non-Segmented Images |
|---|---|---|
| | F1 Score | F1 Score |
| Naiive Bayes | 53 | 83 |
| Logistic Regression | 65 | 69 |
| GradBoostingCLF | 83 | 74 |

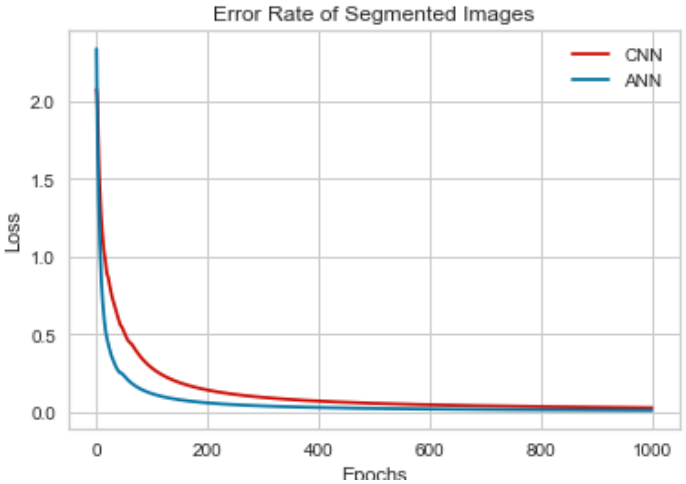
Classical ML models tended to perform better on Non-segmented images, in terms of their respective F1 scores, with the exception of Gradient Boosting Classifier, which performed quite a bit better on Segmented Images.

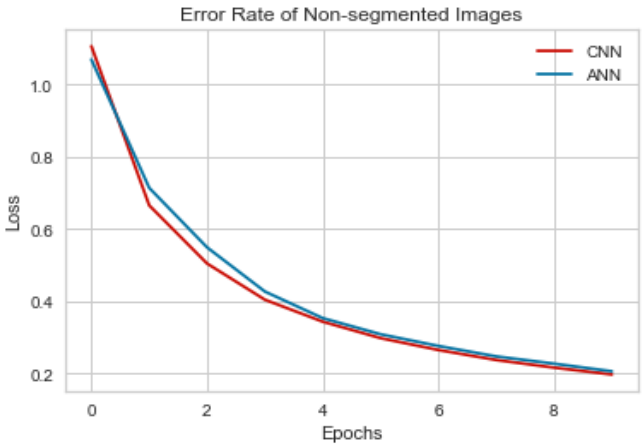
Ali Kamal
19I-1865




## Error Rate:




In terms of the error rate of Deep Learning models, ANN converges faster than CNN on segmented images, while CNN converges faster than ANN on non-segmented images. Both, however, provide very good accuracy overall, as show in the above graphs.