A Machine Learning Lab Project Report on

“Classifiers to Recognize Tumor State”

Submitted by

Chinam Srihari Akash

14

130905202

Semester 7 Section B

Arushi Gupta

27

130905372

Semester 7 Section B

Poorva Gupta

19

130905210

Semester 7 Section D

Department of Computer Science & Engineering



**1. Abstract**

Breast Cancer has become the common cause of death among women. Due to long hours invested in manual diagnosis and lesser diagnostic system available emphasize the development of automated diagnosis for early diagnosis of the disease. Our aim is to classify whether the breast cancer is benign or malignant. To achieve this, we have used machine learning techniques such as Principal Component Analysis, Logistic Regression and KNN. These techniques are coded in Python. We have compared the accuracies of different techniques and observed the results. We found Logistic Regression most suited for predictive analysis and KNN performed best for our overall methodology.

**2. Introduction**

Breast Cancer is the most common type of cancer World Wide and is the leading cause of death among women. The most effective way to reduce breast cancer deaths is by detecting it earlier. This is possible by performing various tests like MRI, mammogram, ultrasound and biopsy. Breast Cancer refers to uncontrolled growth of cells in the breast tissue. If these cells are not stopped or controlled, then it might cause an adverse effect on the whole body. Breast cancer can occur in men too having a higher mortality rate.

Diagnosis of breast cancer is done by classifying the tumor. Tumors can be either benign or malignant but only latter is the cancer. Malignant tumors are more cancerous than the benign. Unfortunately, not all physicians are expert in distinguishing between the benign and malignant tumors. So we need a proper and reliable diagnostic system that can detect the malignant tumor. Thus there is a need of proper parameter and feature selection so that the error rate and cost is minimized.

Using PCA the number of dimensions in the data set is reduced to two so that it can be plotted and visually represented. We have implemented K- nearest neighbor where a similarity criterion is Euclidean distance. Logistic regression is also implemented for solving the problem of over fitting the data. Based on the observed accuracy values certain conclusions are made.

**3. Survey**

The discovery of the survival rate or survivability of a certain disease is possible by extracting the knowledge from the data related to that disease. The data source is taken from kaggle and includes attributes like:

1) ID number

2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and worst or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

**4. Gaps**

* We faced a problem with the accuracy as the number of tuples we had were 569, from which we used a meagre 25 tuples for testing. As the number of cases we had access to were restricted, the models that we created did not perform as expected
* We needed more attributes such as age of the person and number of pregnancies undergone to make a better analysis of a tumor

**5. Objectives**

* The main objective of this project is to create a model that could be trained on given samples of training data and later recognize other samples and classify them under their required category
* Another objective was to compare the designed model to other models that could be suitable for the given dataset, to see which model fared better

**6. Model**

* We initially based our model on the K Nearest Neighbor approach to classify an unseen case as Benign or Malignant
* Once we completed this model, we designed another model based on the Logical Regression approach, after which we compared the results of both models

**7. Methodology**

* To implement the first model, we applied PCA on the dataset to reduce it to a two dimensional set
* Once applied, we then segregated our dataset into three parts – the first part contained the training set with 544 cases of both benign and malignant tumors, the second part contained the first test set containing 10 cases of malignant tumors and the third part contained the second test set containing 15 cases of benign tumors
* We then tested the latter two parts on the training set using the lazy learner KNN
* In the other model, we first applied PCA on the dataset bringing it down to two dimensions and then formulated a hypothesis (h(x)) that would be trained to create a partition between the cases of benign and malignant tumors
* We started the process of training the model with a random value set for theta and then trained the hypothesis using gradient descent
* Once the gradient descent reached a minima, we used the hypothesis to predict the state of a tumor based on the test cases

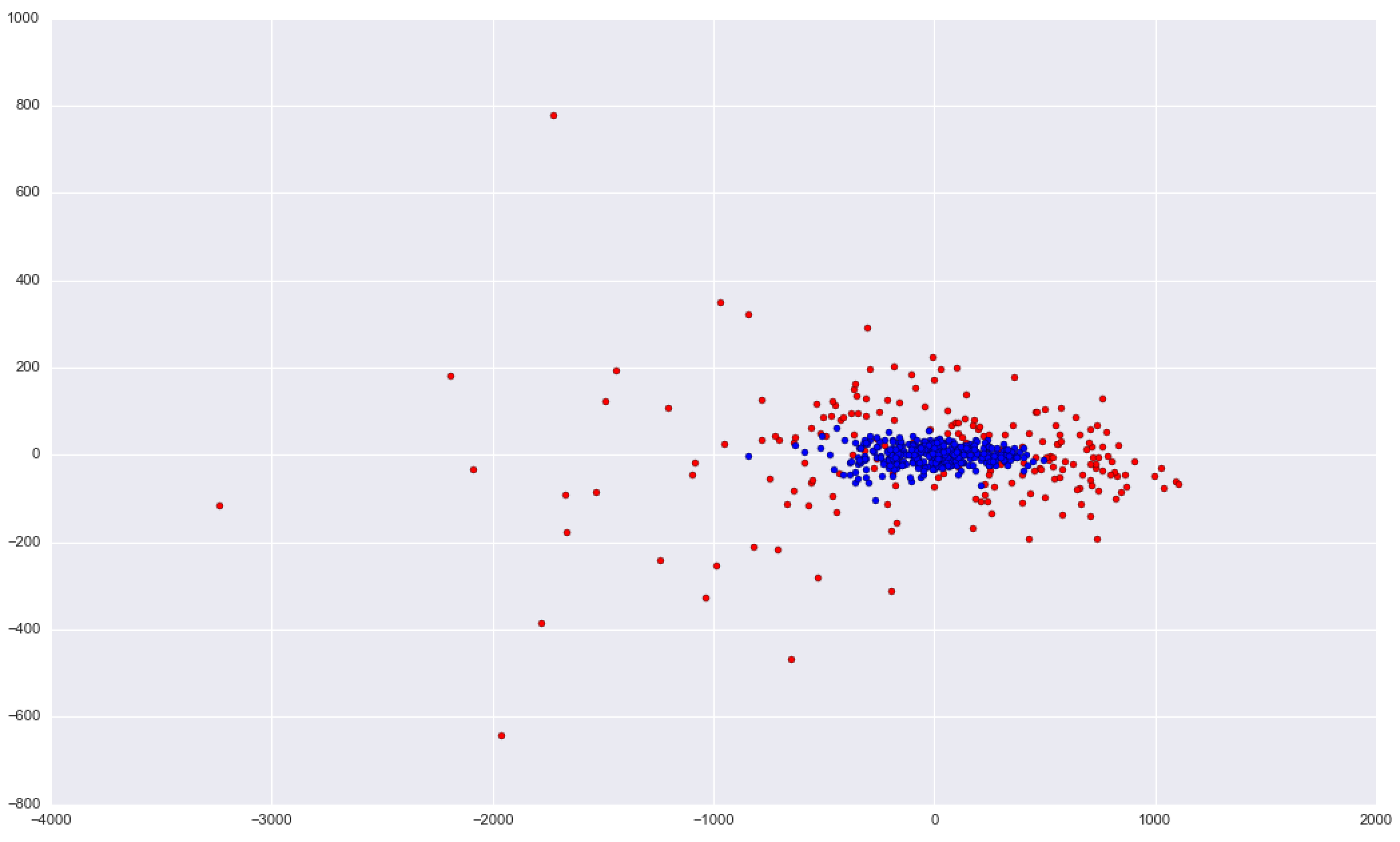
****

Figure 1 – 2D scatter plot of data after applying PCA

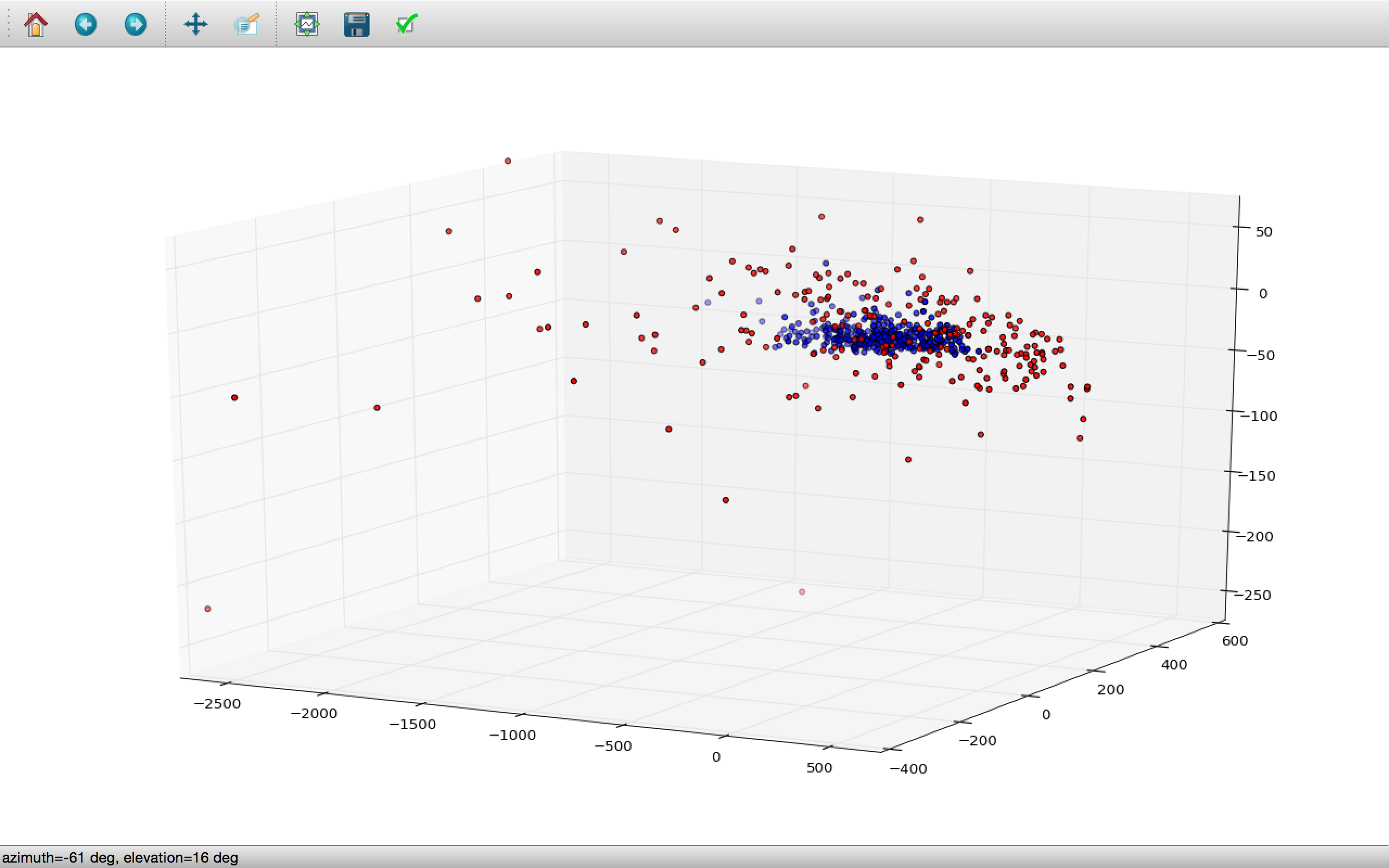


Figure 2 – 3D scatter plot of dataset after applying PCA

**8. Results**

* Using the model based on KNN, we got an accuracy of 60% on the malignant cases where the value of ‘k’ was set to 5 and we got an accuracy of 100% on benign cases with k value as 2
* Using the model based on Logistic Regression, we achieve an accuracy of 80% on cases with malignant tumor, while we only achieved an accuracy of 47% on cases with benign tumor

**9. Conclusion**

* We concluded that we require more data to improve the accuracy of the classifier which in cases of higher accuracy can be used for public healthcare
* We also realized that a combination of these two classifiers can improve the accuracy of performance
* The KNN classifier has a better specificity while the classifier based on logistic regression has a better recall

**10. Future Scope**

* We can create an Ensemble of Classifiers given that another classifier is designed based on a different methodology
* Once these three classifiers are created, we can apply all of them on the same case and take a majority voting, which will improve the accuracy as each classifier has its own pros and cons