

Report

End to End ML Project on FaceMask Detection

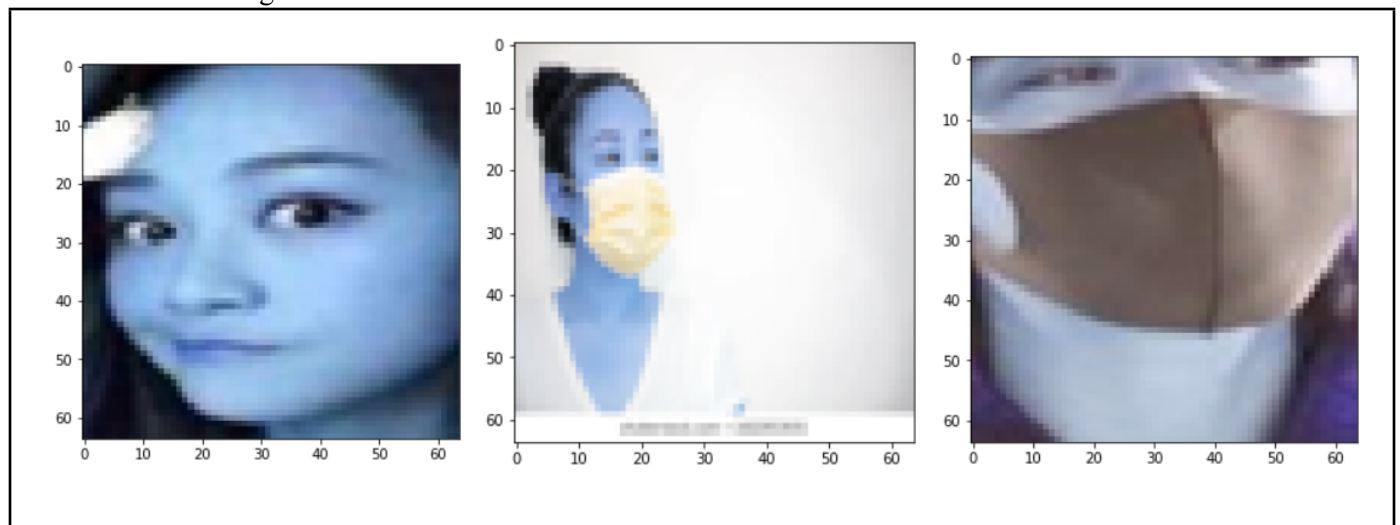
Team Members - Nirbhay Sharma , Mayank raj

Under the guidance of - Dr. Richa Singh

1) Data Preprocessing

Data is given in the form of images from that the data is transferred to one folder where both of the classes images are there then on those images we reshape the images to (64 , 64) using PIL module and then the images are converted into pixel values using cv2 module in python then the whole dataset is converted into dataframe and stored as data_64.csv in the system and later used in the project .

Visualization of images



Analysis :

In all there are 2 classes 1 (for the person who wear masks) and 0 (who does not wear masks)

2) PCA results

Since the data is huge with train data size as (4600 , 12288) so pca is necessary to reduce some dimension that are not that relevant to the data and we can reduce the dimension of the data so we applied pca with 0.95% of variance to be reserved

after reducing, the dimension was (4600, 214) so we got pretty good reduction here

3) Models used and analysis performed

For the model part 4 models are used (Support Vector machines , MLP classifier , KNearest Neighbors ,Random forest classifier)

SVM

- Svm was applied on the train set and the results are shown below (accuracy = 98.19 %)
- Roc curve is also shown below so the reason why we choose this is that generally svm's performs good in image dataset so we found this a strong candidate for this classification and it performs good as well

MLP

- Application of MLP classifier on train and test dataset results into these observations (accuracy = 96.72 %) we also choose this because neural networks are strong contenders in the field of classification and it performs good here as well

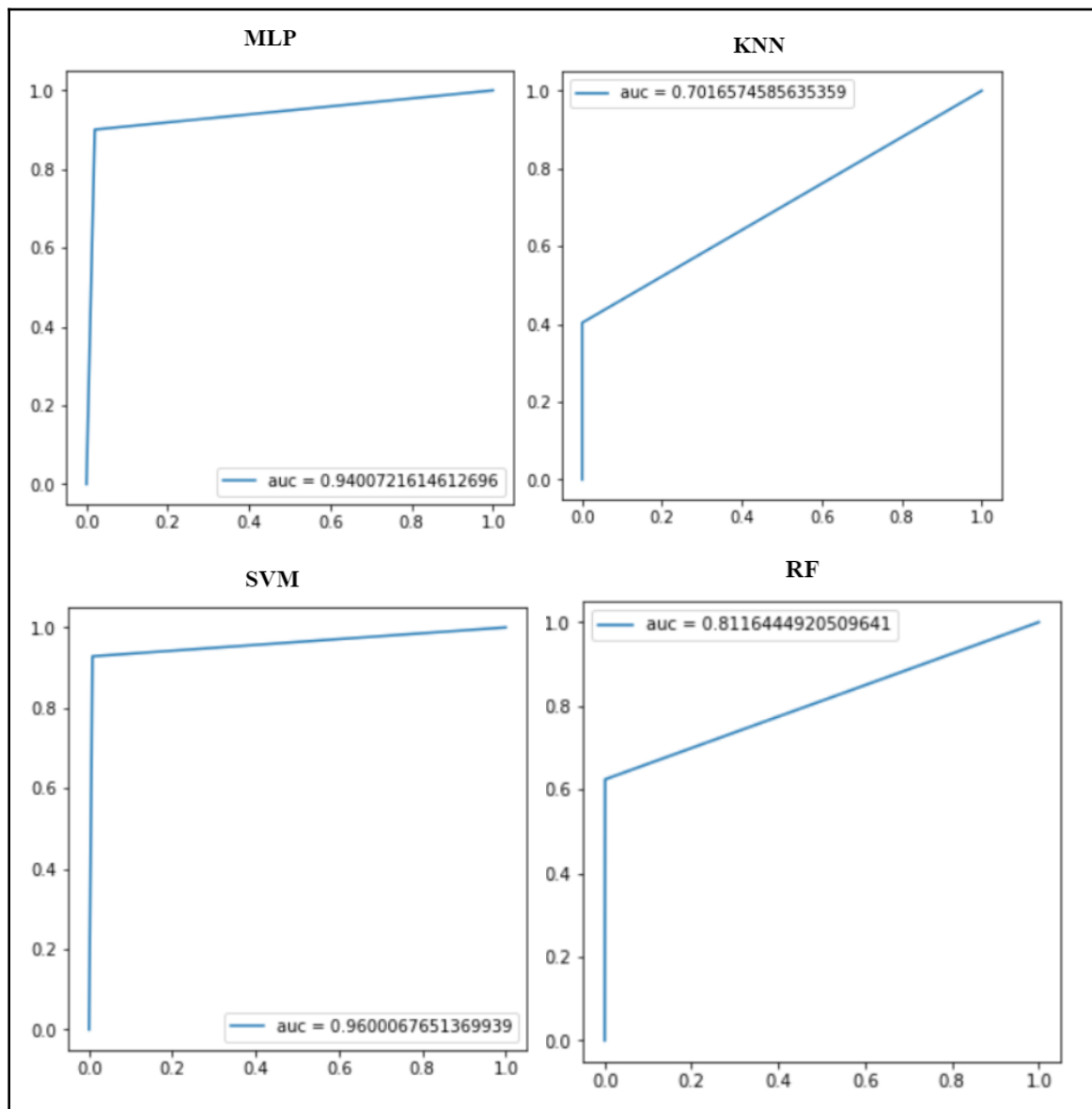
KNN

- Accuracy for KNN is 90.69% we choose KNN because it is simple and easy to implement and it is also a famous algorithm for classification

RandomForest

- Accuracy for Random Forest comes out to be 94.05% we used this because in industries also after CNN and neural networks second privilege is Given to Random forest Classifier as they are strong classifier

Roc for all



4) Cross validation for all the models

Boxplots are drawn for the cross validation of different models

SVM

Cv scores = array([0.9795479 , 0.98062433, 0.98062433, 0.98170075, 0.9762931])

MLP Classifier

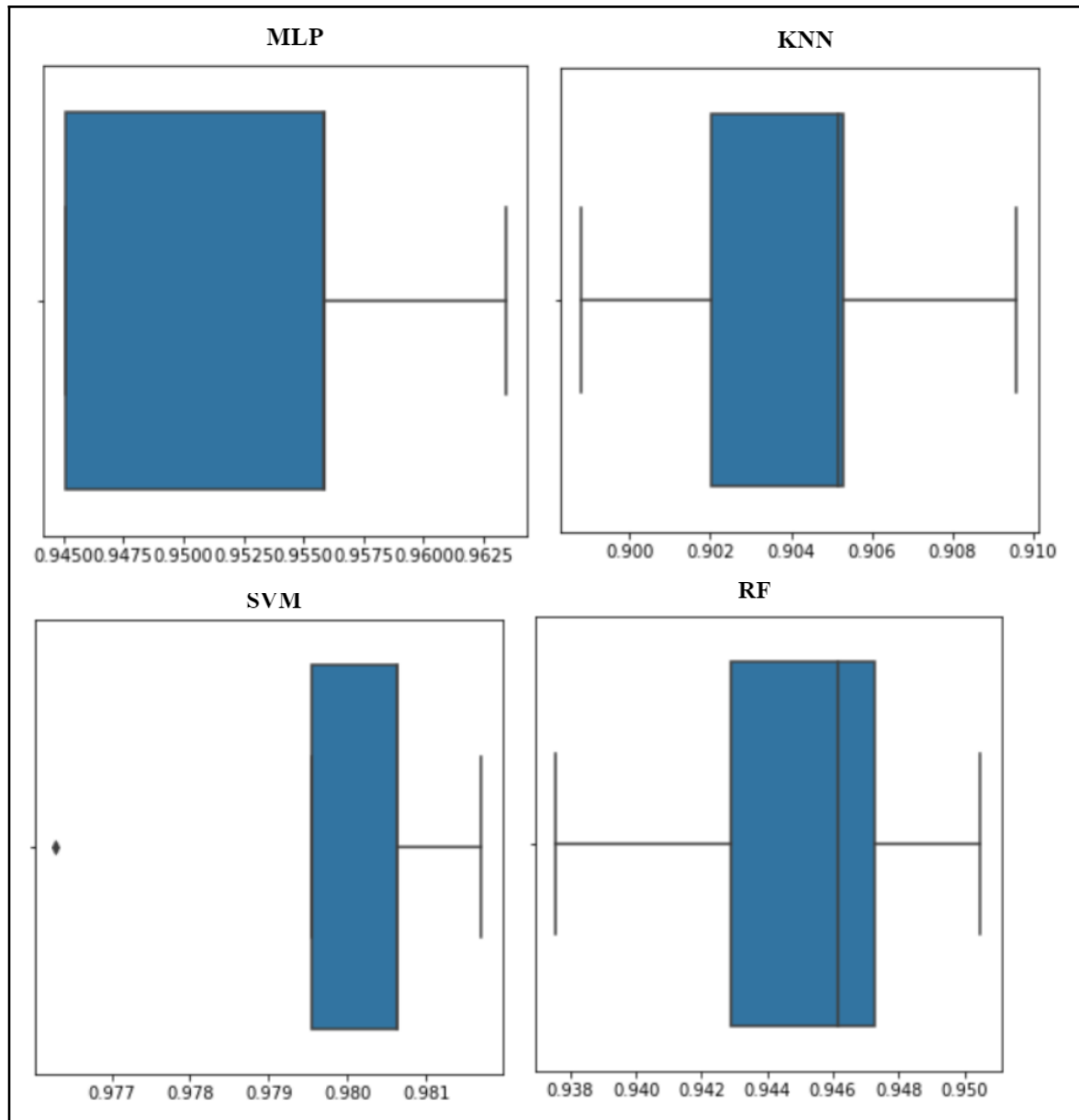
Cv scores = array([0.94510226, 0.96340151, 0.95586652, 0.94510226, 0.95581897])

KNN

Cv scores = array([0.90527449, 0.89881593, 0.90958019, 0.90204521, 0.90517241])

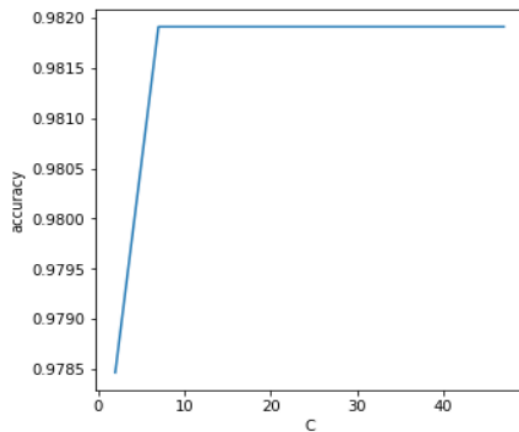
RandomForestClassifier

Cv scores = array([0.95048439, 0.94725511, 0.94617869, 0.93756728, 0.94288793])

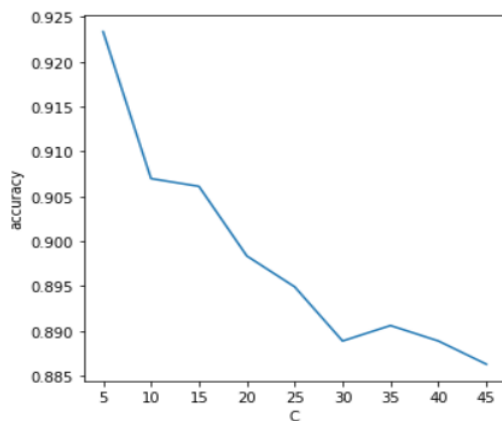


5) Some more findings from the experiments

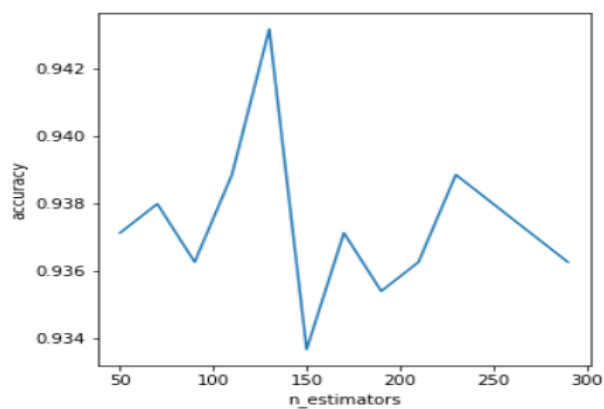
For svm a plot for variation of C with accuracy is shown below we can easily see that if C increases then the testing accuracy is reducing as we know that when C is more then the model overfits and hence the testing accuracy reduces the optimum value of C as seen from this plot is around 5



For KNN classifier the variation of accuracy with the variation of K is shown below it is clear that as value of N_neighbors increases then accuracy increases and then it is decreasing as the neighbours increases



For RandomForest the variation of accuracy with n_estimators is shown below



6) Comparison of all the model

On the basis of box plots:

From the box plots we can easily get a clear idea that svm is performing better than all of the classifiers since median of SVM is greater than all these like for SVM median is 98 % for MLP median is somewhere around 95% for KNN the median is around 90.57% and for RandomForest the median is 94.6% hence they are performing in the order of SVM > MLP > RandomForest > KNN

Reasons for difference in results

Why svm gives good results :

- SVM works relatively well when there is a clear margin of separation between classes. SVM is more effective in high dimensional spaces and since it is a binary classification margin between classes are well defined and due to which SVM works better than other classifiers.

Why MLP gives good results but not better than SVM

- SVMs are based on the minimization of the structural risk, whereas MLP classifiers implement empirical risk minimization. So, SVMs are efficient and generate near the best classification as they obtain the optimum separating surface which has good performance on previously unseen data points.

Why RandomForest does not performs better than SVM and MLP

- Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model but it may also overfit sometimes which may degrade its performance sometimes.

Why KNN performs less compared to other models

- Since KNN is a lazy classifier and the most simplest among all of these models so it is not performing well in comparison to other models

7) Chosen approach, hypothesis, expectation

- Chosen approach is end to end machine learning pipeline which involves process like preprocessing then train test split then applying the model , fitting the model and then testing the model using cross validation and changing different parameters of the model and plotting their curves and at last some of the visualization curves like roc curves are drawn in order to have a clear idea of model .
- The model should be able to correctly classify the new data point which it has not encountered before into whether it is masked or unmasked.

8) Experiments result and conclusion

After the experiments the conclusion is that svm is performing better than all other classifiers for the given image datasets and our main focus is on that in this Covid time our model should predict the person without masks more correctly than to identify persons with masks correctly.

9) References :

- a) Sklearn documentation for classifiers
- b) Pillow module documentation
- c) Open cv documentation
- d) Os module documentation

Who did what -

Nirbhay Sharma (B19CSE114) - KNN(and analysis in report) , MLP (and analysis in report) ,
Preprocessing (Image extraction and conversion)

Mayank Raj (B19CSE053) - SVM(classifier and analysis in report),Random Forest(Classifier and
analysis in report),PCA(dimensionality reduction), Preprocessing(conversion and dimensionality
reduction)