Question 1.3

(a) The training time for SVM on the Training Set: 34.952703488000225 Seconds
Prediction Time for SVM on Validation Set is 2.3632573140002933 Seconds
Prediction Time for SVM on Test Set is 1.762775120999322 Seconds

Training Time for Random Forest on Train Set is 4.163597747000495 Second
Prediction Time for Random Forest on Validation Set is 0.14906643900030758 Seconds
Prediction Time for Random Forest on Test Set is 0.11664836300042225 Seconds

It can be concluded that SVM takes longer to train and validate as compared to Random Forest. In addition, the time taken to predict the labels on the test set is higher than that of Random Forest by quite a margin.

The above values are shown under Question 1.2 with code.

(b) On comparison of the accuracy for various SVM kernel and parameters it can be concluded that gamma=’auto’ linear kernel, gamma=’scale’ linear kernel and gamma=’scale’ polynomial kernel gives best accuracy of 94.737%. As Polynomial kernel takes more computation time so we chose linear kernel with gamma = ’auto’.

The comparison of various parameters from python notebook (Under Question 1.2 Line no In[10]) was:

1. with gamma = 'auto', kernel = 'rbf', decision_function_shape = 'ovo', probability = True accuracy is 89.474%
2. with gamma = 'scale', kernel = 'rbf', decision_function_shape = 'ovo', probability = True accuracy is 93.421%
3. with gamma = 'scale', kernel = 'linear', decision_function_shape = 'ovo', probability = True accuracy is 94.737%
4. with gamma = 'auto', kernel = 'linear', decision_function_shape = 'ovo', probability = True accuracy is 94.737%
5. with gamma = 'auto', kernel = 'poly', decision_function_shape = 'ovo', probability = True accuracy is 44.737%
6. with gamma = 'scale', kernel = 'poly', decision_function_shape = 'ovo', probability = True accuracy is 94.737%
7. with gamma = 'auto', kernel = 'sigmoid', decision_function_shape = 'ovo', probability = True accuracy is 77.632%
8. with gamma = 'scale', kernel = 'sigmoid', decision_function_shape = 'ovo', probability = True accuracy is 44.737%

From the code in Python notebook it can be seen for Random Forests (under Question 1.2 line no in [17]) that:

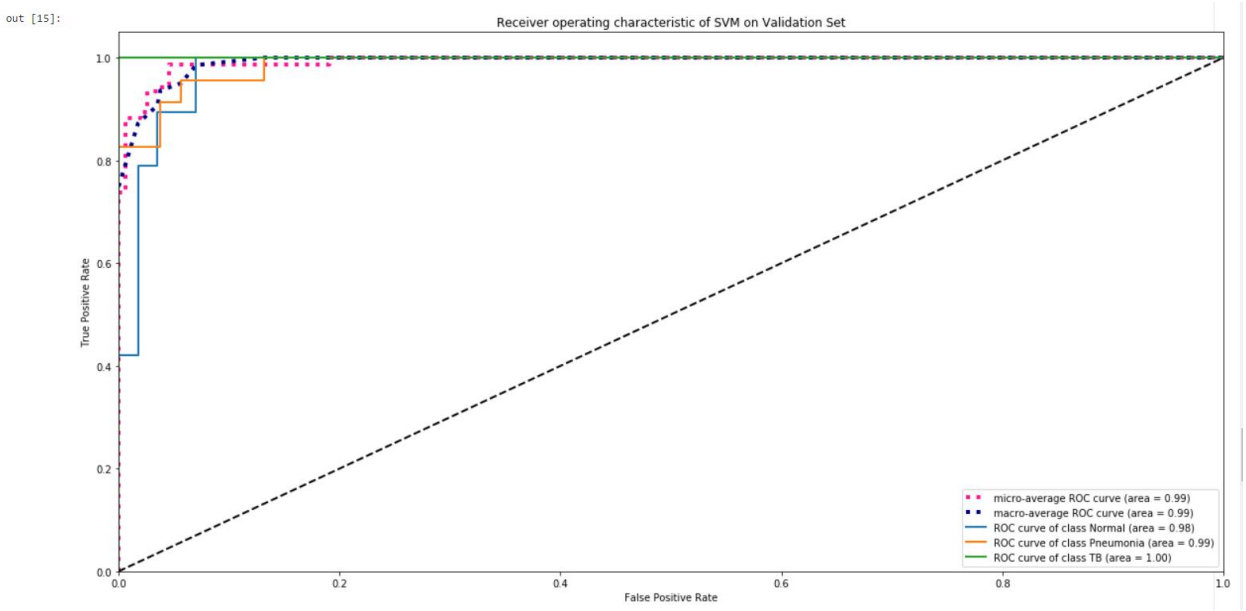
1. With criterion = "entropy", max_depth= None, bootstrap = True, random_state= 42, n_jobs = -1 the accuracy score is 92.105% because changing max_depth has no effect.
2. With criterion = "gini", max_depth= None, bootstrap = True, random_state= 42, n_jobs = -1 the accuracy score is 90.789%

In the case of Random Forest the best accuracy score obtained is 92.105% with With criterion = "entropy", max_depth= None, bootstrap = True, random_state= 42, n_jobs = -1

On comparison between SVM and Random Forest, the best accuracy is of SVM with above defined parameters that is 94.737%.

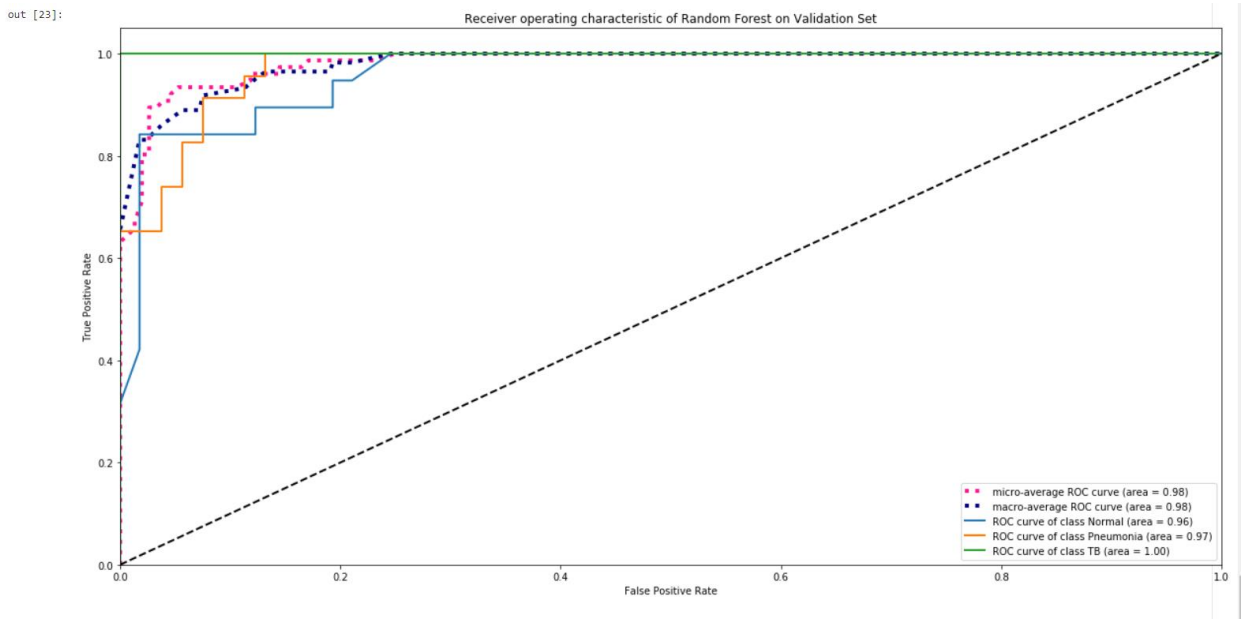
Hence, SVM is performing better than Random Forest because SVM uses support vectors which is the closest points to the boundary in each class

(c) For SVM the ROC curve is shown below:



It can be seen from above figure that the Area under the curve (AUC) for Normal class is 0.98, Pneumonia class is 0.99 and TB class is 1.00. The micro-average and macro-average of both AUC are 0.99.

For Random Forest the ROC curve is shown below:



It can be seen from above figure that the Area under the curve (AUC) for Normal class is 0.96, Pneumonia class is 0.97 and TB class is 1.00. The micro-average and macro-average both AUC are 0.98.

Comparison among the classifiers: It can be seen that the Area under the curve (AUC) is better for both Normal and Pneumonia class in SVM as compared to RF. However, the AUC for TB in both cases is 1.00 which means that both the algorithms are perfectly classifying the TB class.

However, in case of Normal and Pneumonia class there is some misclassification of the labels which results in the less value of AUC. This misclassification is more for RF as compared to SVM for these classes.

It is also evident from the accuracy score that SVM outperforms RF which is verified by the ROC curves for each of the classifier.

(d) The best performing model in our case is SVM so we have filled up the predicted labels for SVM in the excel sheet ‘Testing Labels.xlsx’.