INSTITUTE OF HEALTH INFORMATICS

Graduate Programme in Data Science for Research in Health & Biomedicine

Assessed Coursework Submission

Student candidate number:	SYSD6	
Module:	CHMEGH40: Machine Learning in Healthcare & Biomedicine	
Date due:	Monday, 30th April 2018, 12:00 midday	
Word count: (excluding references, diagrams and appendices)	2,413	
Disability or other medical condition for which UCL has granted special examination arrangements:	Dysgraphia & ADHD	
Formative feedback:	Please address in formative feedback:	
	Please ignore in formative feedback:	

Introduction

Cancer is a multi-faceted disease, where modifications of a cells genome cause it to divide in a manner both beyond the control of the host and to its detriment. Neoplastic growth can be observed in a multitude of different cell types, with further categories of the diesase emerging within these. Generally, most types of cancer share several core charactersitics that are essential for sustained growth and proliferation as well as combatting the response of its host to these (Hanahan and Weinberg, 2011).

Breast cancer is of particular note as it was responsible for over eleven thousand deaths in 2014 in the UK, and the second largest cause of mortality for women with cancer (Jemal et al., 2006; Massat et al., 2016). As with most other forms of cancer, many of these deaths could have been prevented if those patients had their condition identified and thus treated from an earlier stage in it progression (Edge et al., 2009). Survival rates of breast cancer have been steadily increasing golbally, over the past two decades, a phenomenon partially attributred to increased awareness and screening of patients at risk (Allemani et al., 2015). Patients with breast cancer have been shown to have a substantially reduced mortality rate if they have previously been screened for the condition at any time or within three years of their diganosis, with an apprximately 35% and 60% rediction in mortality respectively (Massat et al., 2016).

Medical imaging techniques are commonly used to screen for potentially malignant tumours. One commonly used method for producing these images are sonography, using high frequency sound waves (Lane et al., 2014). Mammography, an imaging technique using X-rays is also widely used to screen for breast cancer. A practioner conducting these tests may then analyse the results to produce a plan of further action.

The Breast Imaging Reporting and Data System (BI-RADS), developed by The American College of Radiologists uses a standardised scoring system, where the medical professional conducting the screening can input categorical variables based on their observations of these images. The attributes assessed in this process include, but are not limited to, the density of the tumour, its observed shape and the margin of tissue surrounding it (Sickles et al., 2013).

The purpose of this investigation is to use the Mammographic Mass Data Set (Elter et al., 2007), hosted publicly on the UCI Machine Learning Repository to train a machine learning algorithm that may outperform the BI-RADS assessment. The features availible for this purpose are those oulined above (see figure 1 for visualisation), with the class lables being binary categories of benign (0) and malignant (1). Within this dataset, there is an uneven split between the benign and malignant classes (53.6% and 46.3% respectively).

The mean age of patients recorded in this dataset is 55.5 years old with a standard deviation of 14.5 years. Patient age is negatively skewed ($P < 0.005$), favouring older patients. The majority of BI-RADS assessments made in this dataset (98.5%) are BI-RADS 3 and above, heavily suggesting that this test was designed to prioritise sensitivity above specifity. The modal category for density is 3 (low density), comprising 90.2% of all observations, for which there is a value present.

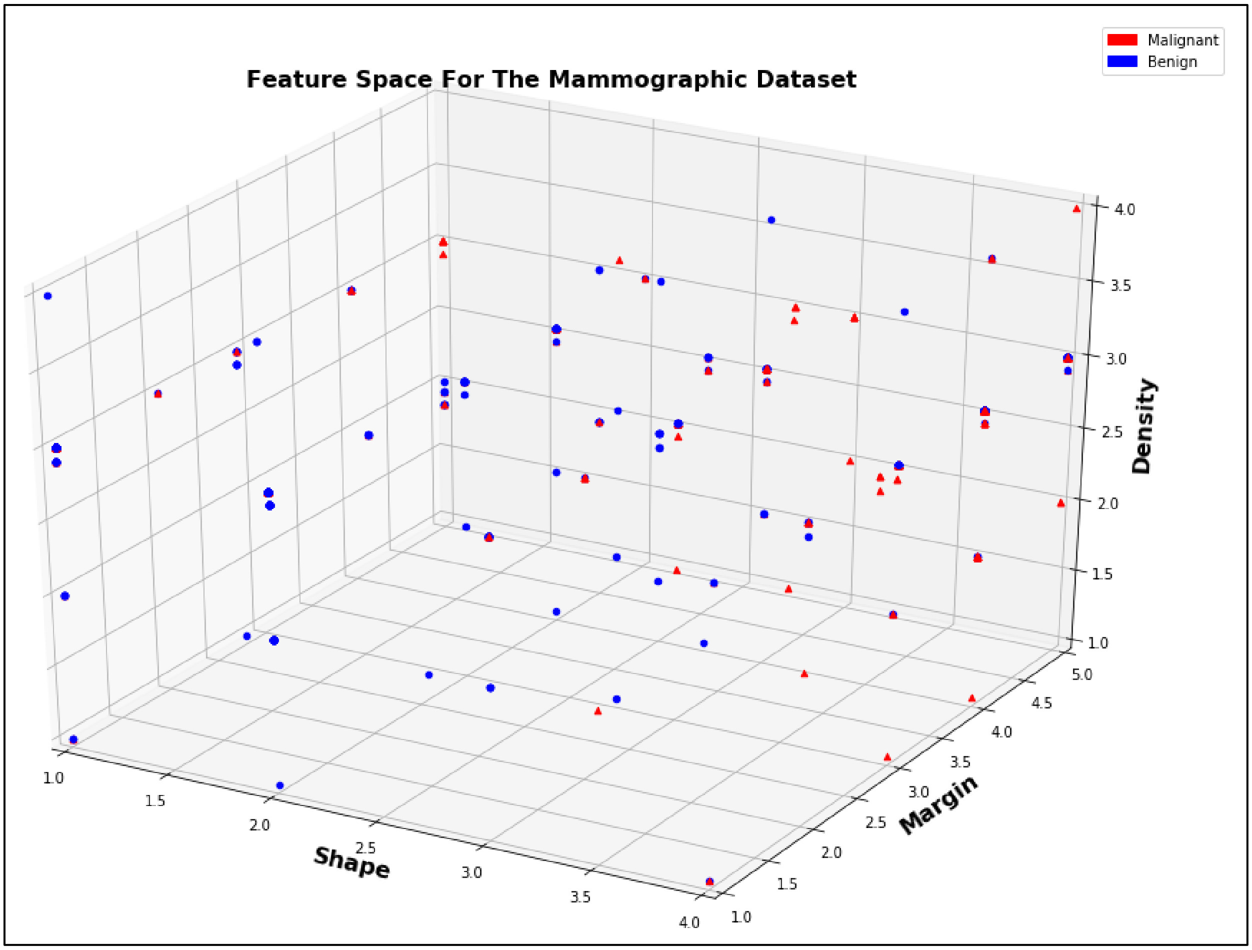

Figure 1 – (Above) Visualisation of feature sapce for the Mammographic Masses Dataset. Blue coloured circles represent obseravations where biopsy has revealed the tumour to be benign, red traingles represent observations for which the tumour was determined to be malignant

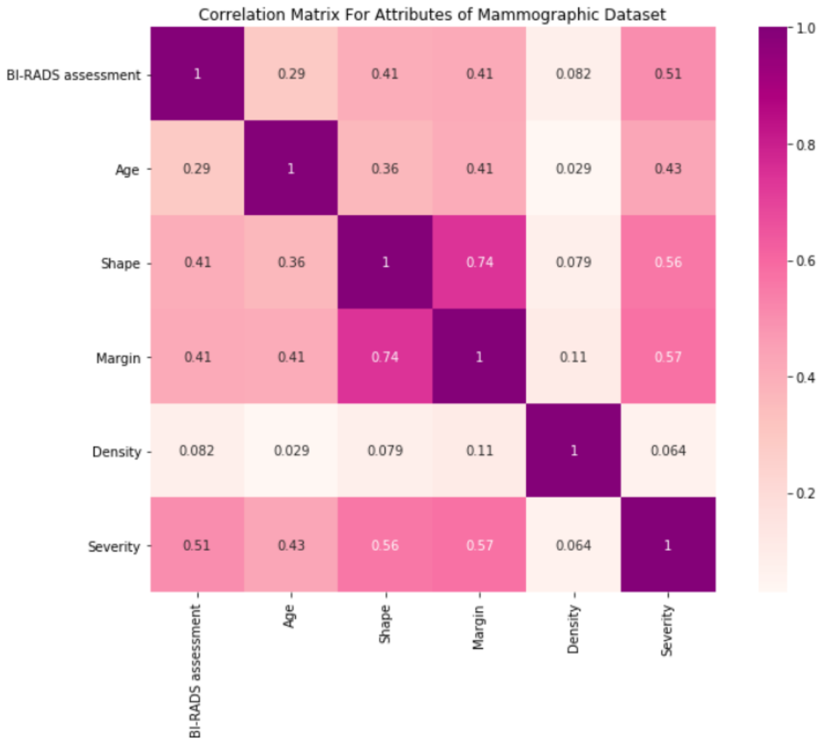

Figure 2 – (Left) Confusion matrix displaying correlation coefficient between all attributes in the dataset. Age, shape and margin show moderate correlations with severity. Tumour density is poorly correlated to all other attributes.

Methodology

Data pre-processing

The data from the mammographic masses dataset needed relatively little pre-processing due to its attributes all being stored as integers, apart from empty values (stored as '?'). Because of this, the mapping of strings to integers or floats was not necessary. Null values (162 in total) were parsed as Not a Number(NaN) by pandas, as the data was read into a data frame.

With this, null values were in a format recognisable by the software. Imputation of null values was considered, however as this would result in the introduction of bias into the classifier due to reducing variance of each feature. Instead, removal of rows containing null values was implemented as a loss of 13.6% of all observations was considered preferable.

Due to the low dimensionality of the dataset, all attributes were used or made available as features, depending on the classifier used. These include: tumour shape, density and margin as well as patient age. Data, which was modelled by the support vector machine classifier, required feature scaling. Normalisation was chosen and was carried out using the minmax scaler function of scikit learn. This scaled the data to values between 0 and 1, and was applied to all values in the feature space. The random forests classifier does not require feature scaling.

Optimisation of Hyper Parameters: Grid Search

For both classifiers, hyperparameter optimisation was partially carried out automatically with the use of the scikit learn GridSearchCV function. This function takes a user-defined grid containing multiple values and tests the classifier with every unique combination of these. This was limited to select hyperparameters as increasing the grid size becomes exponentially more resource intensive. Accuracy was used as a metric of success during the grid search.

For both classifiers, a random state was assigned to a seed based on the date of the code being written. This was done to remove any bias, whilst preserving reproducibility.

First model – Random Forests Classifier

The first model chosen was the scikit-learn random forests classifier, an ensemble model based upon the design of a decision tree. In this model, a (user specified) number of decision trees are created from samples of the training data. Like other methods of bootstrapping, the data from samples are replaced and another decision tree is created from a randomised sample.

This classifier can achieve a lower variance than a standard decision tree without as high a risk of overfitting for doing so. These innate characteristics of the random forests classifier eliminate the need for pruning. This was also preferable as the decision tree classifier module of scikit-learn lacks an in-built pruning function for decision trees.

Each decision tree works by separating the classes based on one feature at a time. Nodes represent features and the leaves emerging from them represent a condition to split the class labels. Nodes that most cleanly split the data occur earlier. The effectiveness of each split, in the model used here, was measured by information gain or entropy as it is referred to in scikit. As entropy is inversely proportional to the purity of data emerging from a leaf, lower entropies are favoured for splitting the data. In the model used, 200 of such decision trees were made and were aggregated to create the final classifier.

A maximum depth of 7 was used to achieve a high accuracy without overfitting and each tree would utilise a maximum of 3 features. The limit of features used was chosen as one of the four features (tumour density) had a substantially weaker correlation to both the other features and class labels. This feature would likely be excluded as it may provide little power to the classifier, whilst increasing complexity and reducing the model's ability to generalise and classify unseen data.

Second model – Support Vector Machine

The support vector machine classifier was selected as the second model for predicting the severity of a breast tumour. Here, observations are plotted as support vectors in the feature space with a hyperplane drawn to best separate points of different class.

A radial bias function (RBF) kernel was selected for the support vector machine. This was done as the data was not entirely linearly separable. Uniform class weights were used for this classifier; however, the benign class could be more heavily weighted to reduce false negatives. The penalty parameter and kernel coefficient were selected using the grid search function. The grid search suggested a penalty parameter of 50. This was reduced to 10, however, to prevent overfitting and came at little cost to accuracy. The suggested kernel coefficient (value of 1) was used.

Model Evaluation

Both models will first be tested using an 80-20, train-testing split and provide an accuracy score and confusion matrix. A more comprehensive method of 10-fold cross validation will be carried out with training and testing occurring within the in-sample training set (80% of the data) previously created. The mean accuracy of the folds will be calculated for evaluation.

The models will then be validated against out-of-sample data (20% testing split). Based on predicted probabilities, a receiver operating characteristic (ROC) curve will be plotted for each fold. This will provide a measurement for the classifiers sensitivity and specificity when classifying unseen data. Mean area under the curve (AUC) scores will be calculated from all the folds of each classifier to give a numeric value to further analyse ROC curves.

Precision recall and F1 scores will also be calculated. Recall is especially important as it gives a measure of the false negative rate.

Results

Overall Performance Comparison

Testing of the above outlined support vector machine (SVM) and random forests (RF) classifiers, using suggested hyperparameters followed by manual optimisation has produced a clear result. The random forests classifier significantly outperformed all implementations of the support vector machine, by most metrics used (Table 1 & Figure 3). There was, however, no significant difference in the overall accuracy of the two classifiers when predicting out-of-sample data, with 10-fold cross-validation. The RF and SVM classifiers achieved a mean accuracy of 0.804 (+/- 0.057) and 0.796 (+/- 0.062) respectively.

Comparison of Sensitivity Between the SVM and RF Classifiers

The RF classifier produced a significantly higher F1 score than the SVM classifier, indicating that it showed greater precision and recall overall. Breaking this metric down into its constituent components shows that the RF classifier had a greater precision and recall then the SVM, however the difference between precision score was not significant. The difference with regards to recall was significant, where the RF achieved a score of 0.943 (+/- 0.012) compared to the SVM score of 0.913 (+/- 0.009) (Table 1).

Precision represents the number of correct positive predictions as a proportion of all positive predictions, effectively being the classifiers true positive rate. The lack of significant difference in this metric between the two classifiers indicates that they share a similar specificity.

Recall, on the other hand, represents the proportion between true positive results and the addition of true positive and false negative results. If a classifier did not falsely predict any negatives, recall would be equal to 0, with an even split indicating that it performs on par with random chance. The significant increase in recall for the RF classifier indicates that it likely has a lower false negative rate when making predictions based off out-of-sample data.

There was a highly significant difference between the area under the ROC curves (AUC) of the RF and SVM classifiers (Table 1). This difference is also clearly visible on the ROC plots (Figure 3), indicating that the RF classifier possesses a superior ability to correctly separate the benign and malignant classes of data.

	SVM		Random Forests	
Metric	Mean	Standard Deviation	Mean	Standard Deviation
AUC	0.865	0.00728	0.940	0.00438
Precision	0.811	0.013	0.823	0.015
Recall	0.913	0.009	0.943	0.012
F1	0.859	0.006	0.879	0.011

Table 1 – Comparison of: AUC (area under reciever operating characteristic curve), precision (specificity), recall (sensitivity) and F1 score (mean of precision and recall), for both the SVM and RF classifiers. The RF classifier significantly outperforms the SVM in terms of AUC and Recall. There is not significant difference between precision scores.

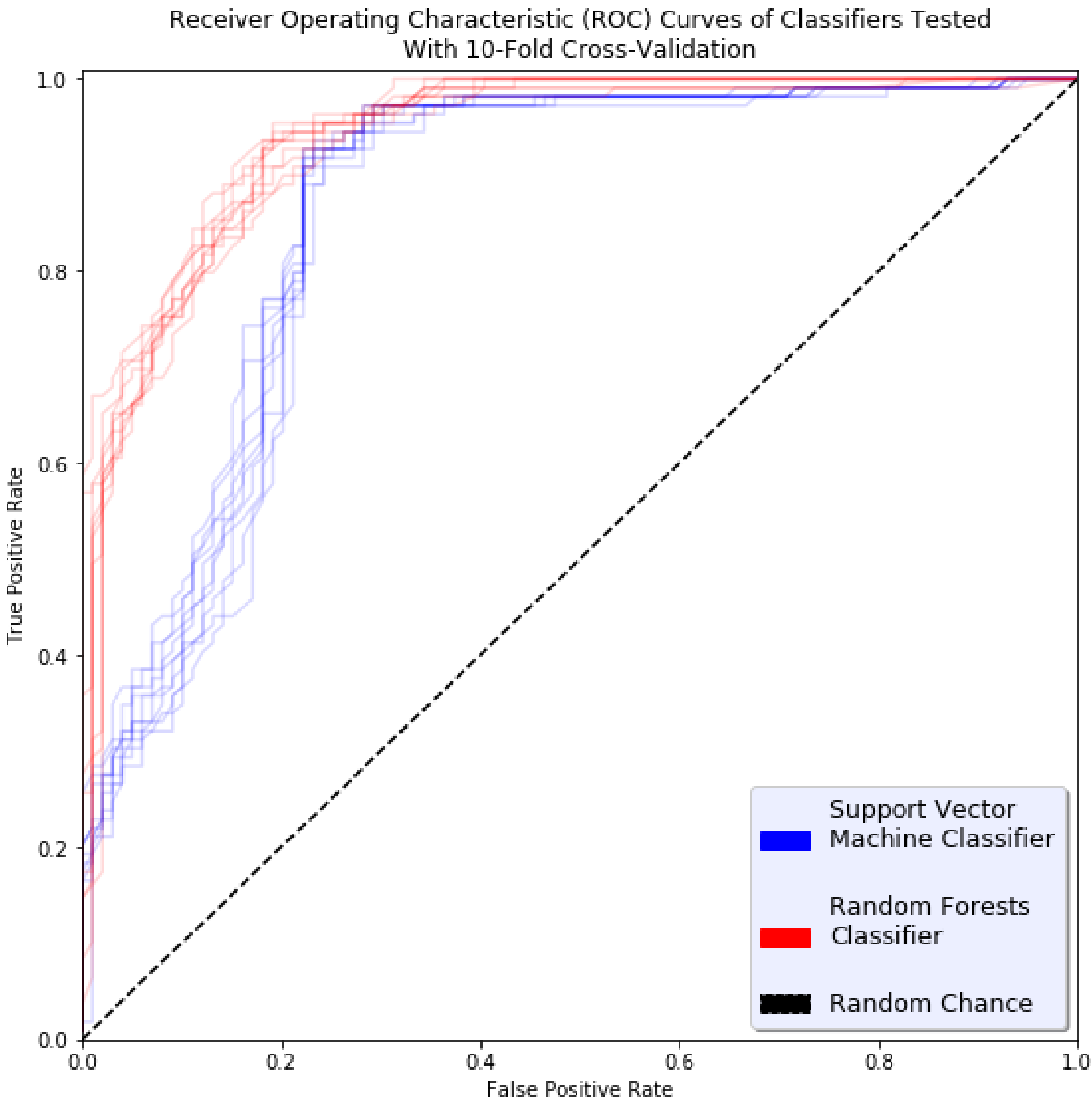

Figure 3 – Receiver Operating Characteristic (ROC) plots for the SVM (blue) and RF (red) classifiers. Present are 10 high transparency ROC curves plotted for each classifier, each representing one of 10 folds from 10 fold cross-validation of the respective classifier. Random chance (y = x) is represented as a dashed black line. Area above y=x and below ROC represents performance greater than what may be explained by random chance. AUC corresponds to total area under ROC curves, used as metric for effectiveness of classifier (see table 1 for breakdown).

The Effect of Hyperparameter Tuning on Performance of The Support Vector Machine Classifier

The SVM was tested, using different values for its penalty parameter (C) (Figure 4). Overall, the optimal penalty parameter for the SVM classifier was found to be 10, with a significantly improved AUC and precision than both values of 0.1 and 1. There was no significant difference between precision scores of C values 1 and 10, though both were significantly higher than C=0.1. This shows that a penalty parameter of 10 likely has a greater ability to separate classes and has a higher specificity than other values for this parameter (Table 2).

Penalty parameters of 50 and 100 were tried. However, these yielded diminishing returns as neither exhibited significantly different scores to C=10 (Results shown in IPython Notebook)

The Effect of Hyperparameter Tuning on Performance of The Random Forests Classifier

Different values were tested for the maximum depth hyperparameter for the RF classifier (Figures 5 & 6). A maximum depth of 7 produced significantly improved scores for AUC and precision, compared to the classifier trained with a maximum depth of 5 (Table 3). Values beyond 5 produced diminishing gains (Figure 6, specific values for maximum depth of 11, 20 and 100 in IPython notebook).

	C = 0.1		**C = 1**		**C = 10**	
Metric	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
AUC	0.81	0.00194	0.841	0.00323	0.865	0.00728
Precision	0.759	0.001	0.761	0.006	0.811	0.013
Recall	0.901	0.006	0.916	0.006	0.913	0.009
F1	0.824	0.002	0.831	0.003	0.859	0.006

Table 2 - Comparison of: AUC (area under reciever operating characteristic curve), precision (specificity), recall (sensitivity) and F1 score (mean of precision and recall), for the SVM classifier with different penalty paramters (C). C=10 significantly outperforms other parameters on all metrics except for recall.

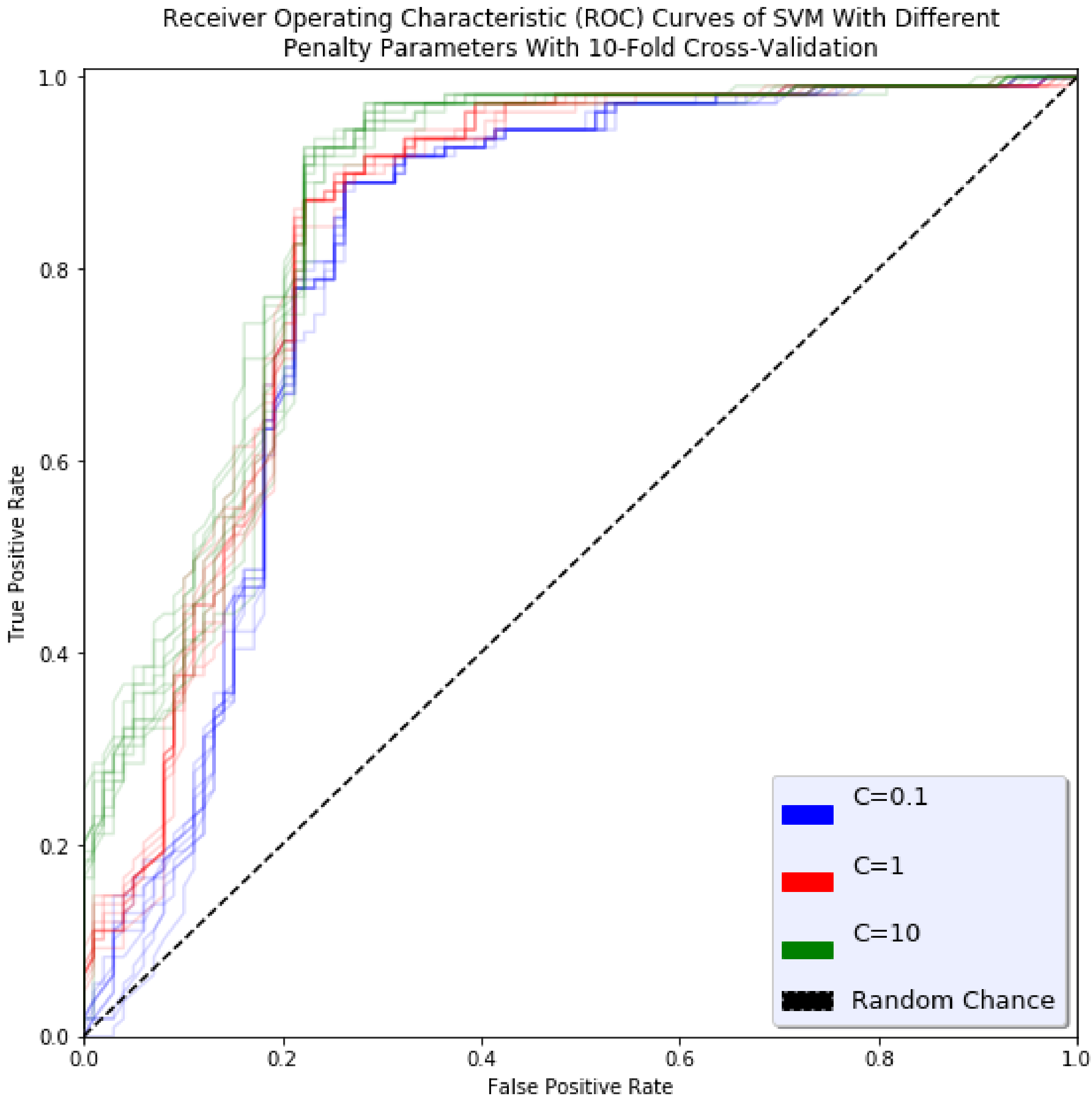

Figure 4 – Receiver Operating Characteristic (ROC) plots for the SVM classifier with different penalty parameters (C) applied. C=0.1 (blue), C=1 (red), C=10 (green). Random chance (y = x) is represented as a dashed black line. Area above y=x and below ROC represents performance greater than what may be explained by random chance. AUC corresponds to total area under ROC curves, used as metric for effectiveness of classifier (see table 2 for breakdown). Higher penalty parameters increase the AUC, up to C=10. Beyond this, increases to C do not result in significant AUC (see IPython notebook for values).

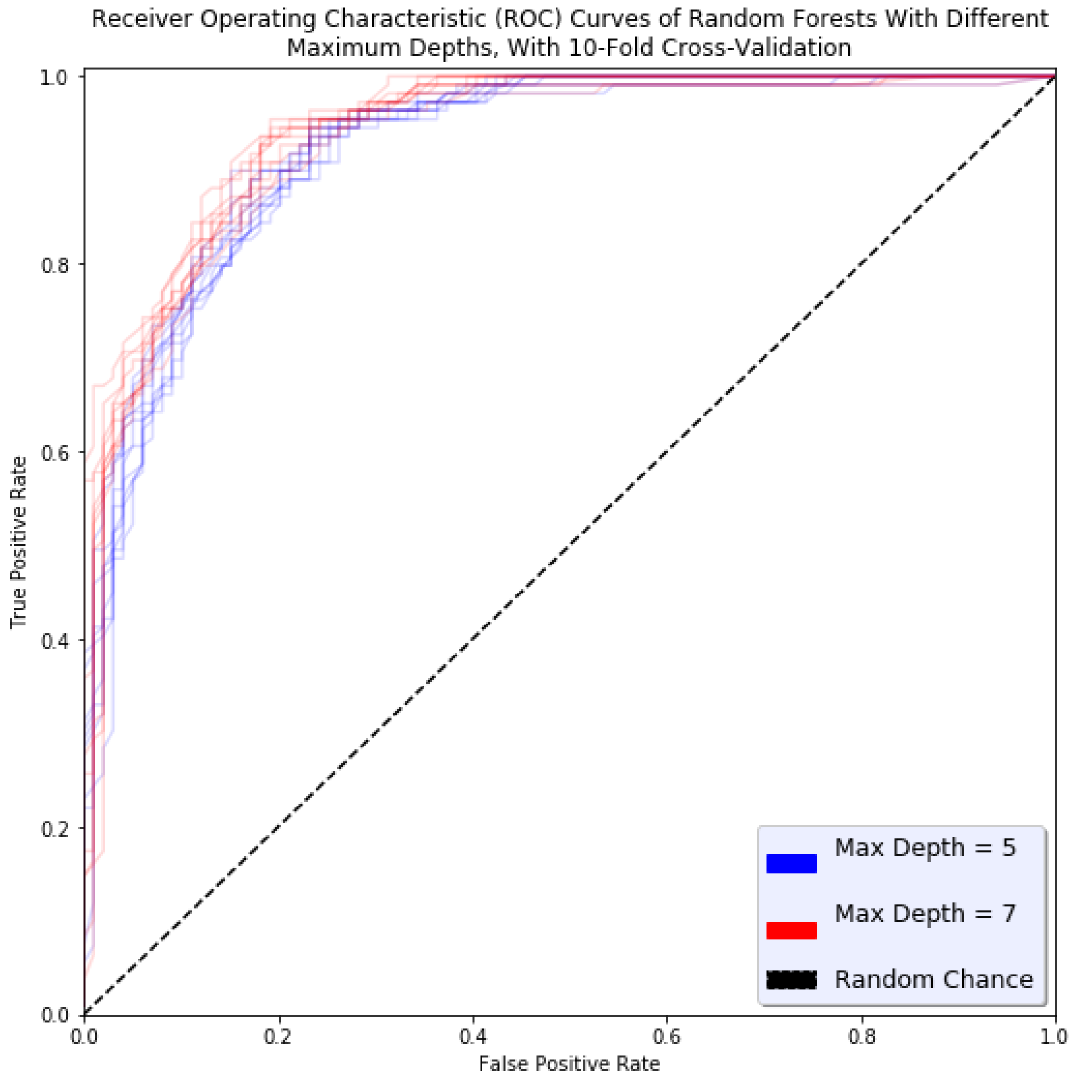

Figure 5 - Receiver Operating Characteristic (ROC) plots for the RF classifier using different maximum depth hyper parameters. Max Depth = 5 (blue), Max Depth = 7 (red). Random chance (y = x) is represented as a dashed black line. Area above y=x and below ROC represents performance greater than what may be explained by random chance. AUC corresponds to total area under ROC curves, used as metric for effectiveness of classifier (see table 3 for breakdown). Greater maximum depth leads to improved AUC scores, up to a maximum depth of 7. Beyond this, increases do not result in significant AUC (see Figure 6 for plot and IPython notebook for values).

	Max Depth = 5		**Max Depth = 7**	
Metric	Mean	Standard Deviation	Mean	Standard Deviation
AUC	0.926	0.00319	0.94	0.00438
Precision	0.795	0.009	0.823	0.015
Recall	0.943	0.011	0.943	0.012
F1	0.863	0.007	0.879	0.011

Table 3 - Comparison of: AUC (area under reciever operating characteristic curve), precision (specificity), recall (sensitivity) and F1 score (mean of precision and recall), for the RF classifier with values for the maximumdepth hyperparameter. A maximum depth of 7 significantly outperforms other parameters on all metrics except for recall.

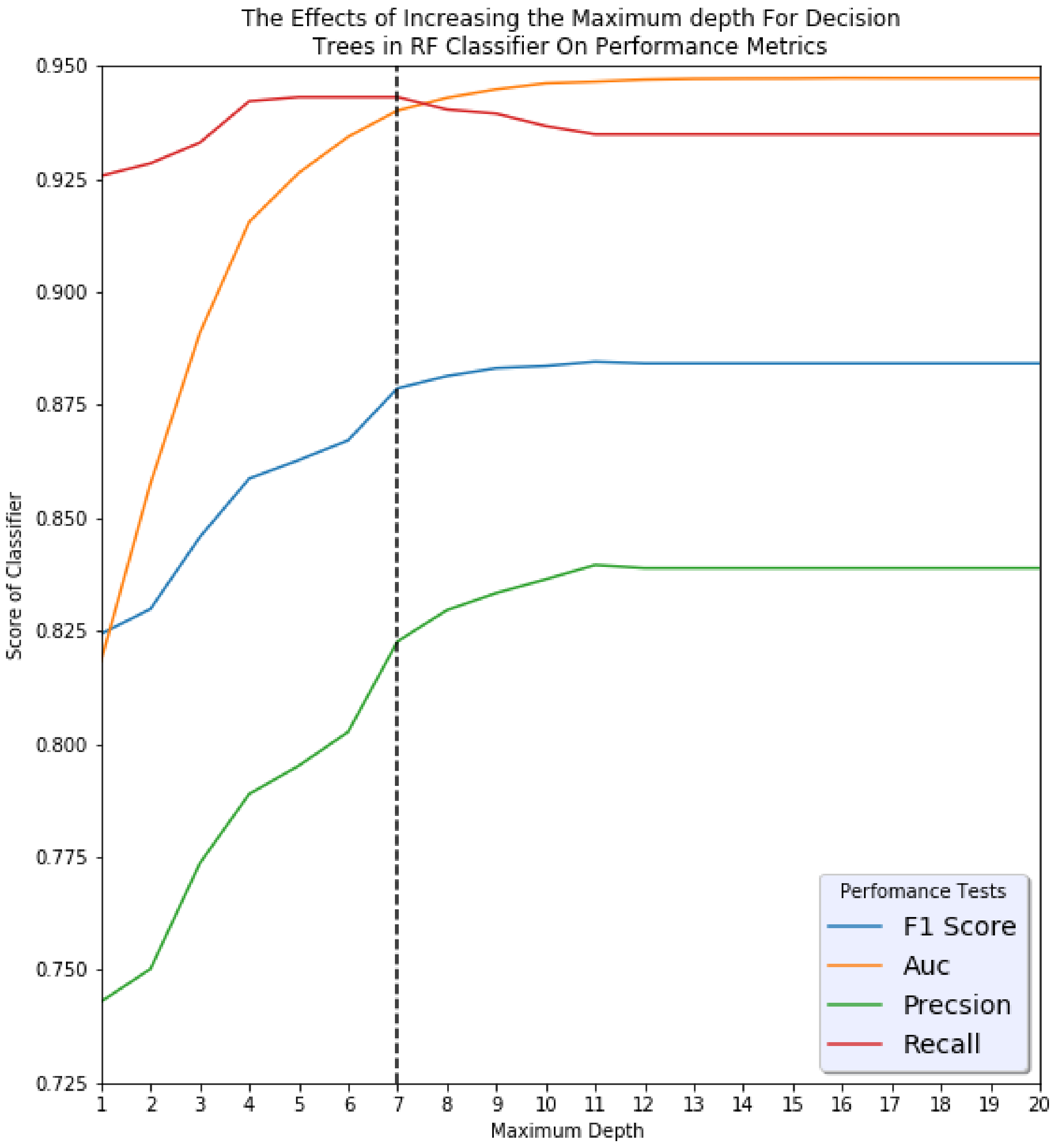

Figure 6 – Plot showing the effect of maximum depths between 1 and 20 for the RF classifier and the effects on the performance metrics: F1 (blue), AUC (orange), Precision (green) & Recall (red). The vertical dashed line (black) represents the final value chosen for this hyperparameter. For values up to between 7 & 12 (depending on metric), increasing maximum depth leads to improved scores for the random forests classifier. Beyond a maximum depth of 7, AUC ceases to improve significantly. Recall ceases to improve above a maximum depth of 4 and begins to decline above a value of 7, though changes throughout are less pronounced than for other metrics. Precision scores cease to increase with maximum depth beyond a value of 11.

Discussion & Conclusion

Of the two classifiers tested in this investigation, the random forests classifier has been shown to be superior to the support vector machine classifier in most of the testing carried out. When compared to the BI-RADS assessment, both classifiers have a far greater precision and accuracy. However, the recall for BI-RADS assessment is near 1. As discussed previously, this is likely due to its designers placing great emphasis on the need to identify malignant tumours at the cost of misclassifying benign tumours, as this would greatly worsen the outcome for those patients.

Very few patients screened with BI-RADS assessment were given a negative result (1.5%) and its accuracy was below that of random chance, bringing into question the purpose of this test beyond prioritising patients with the greatest likelihood of malignancy. As a potential replacement for this assessment, predicted probabilities, outputted by Rf and SVM classifiers could be used instead of a binary classification. This would allow for a more accurate allocation of resources to those most at risk and therefore improving outcomes and survival rates for the group of patients as a whole.

A Naïve Bayes classifier was used by (Bouzghar et al., 2014) to predict malignancy based off of two independent BI-RADS assessments of each patient. Unlike this investigation, leave one out cross validation was used, which may have been more appropriate as study used a smaller sample size of 264. When features of corroborating assessments used, an AUC (A_z) of 0.954 ± 0.016 was attained, not significantly different from the 0.940 AUC of the RF classifier outlined above.

This study also concluded that the predictive power conferred from BI-RADS attributes varied. Asymmetry and echogenicity were two features with notable predictive value that were not present in the mammographic masses dataset. While the latter attribute is dependent on the screening being conducted using sonography, it is unclear why asymmetry was not included as it may have increased the performance of the SVM and RF classifiers. Further investigation may yield improved models if a greater number of attributes are available for training and testing. Moreover, using corroborated BI-RADS attributes, determined by radiologists blind to potentially confounders may help to reduce any bias in the data and improve the models ability to generalise.

In conclusion, the findings of this investigation have shown the Random Forests ensemble classifier to be a powerful model for predicting malignancy in abnormal breast masses, using pre-existing standardised reporting. This model may be used in place of the BI-RADS assessment in the screening of breast cancer, producing more favourable outcomes for patients, both by ensuring those most at risk receive treatment most urgently as well as reassuring patients at lower risk.

References

Allemani, C., Weir, H.K., Carreira, H., Harewood, R., Spika, D., Wang, X.-S., Bannon, F., Ahn, J.V., Johnson, C.J., Bonaventure, A., Marcos-Gragera, R., Stiller, C., Azevedo e Silva, G., Chen, W.-Q., Ogunbiyi, O.J., Rachet, B., Soeberg, M.J., You, H., Matsuda, T., Bielska-Lasota, M., Storm, H., Tucker, T.C., Coleman, M.P., 2015. Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). The Lancet 385, 977–1010. https://doi.org/10.1016/S0140-6736(14)62038-9

Bouzghar, G., Levenback, B.J., Sultan, L.R., Venkatesh, S.S., Cwanger, A., Conant, E.F., Sehgal, C.M., 2014. Bayesian Probability of Malignancy With BI-RADS Sonographic Features. J. Ultrasound Med. 33, 641–648. https://doi.org/10.7863/ultra.33.4.641

Edge, S., Byrd, D., Compton, C., Fritz, A., Greene, F., Trotti, A., 2009. AJCC cancer staging manual, 7th ed. Springer, New York, NY.

Elter, M., Schulz-Wendtland, R., Wittenberg, T., 2007. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Med. Phys. 34, 4164–4172.

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. Cell 144, 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C., Thun, M.J., 2006. Cancer statistics, 2006. CA. Cancer J. Clin. 56, 106–130.

Lane, D.L., Adeyefa, M.M., Yang, W.T., 2014. Role of Sonography for the Locoregional Staging of Breast Cancer. Am. J. Roentgenol. 203, 1132–1141. https://doi.org/10.2214/AJR.13.12311

Massat, N.J., Dibden, A., Parmar, D., Cuzick, J., Sasieni, P.D., Duffy, S.W., 2016. Impact of Screening on Breast Cancer Mortality: The UK Program 20 Years On. Cancer Epidemiol. Biomarkers Prev. 25, 455–462. https://doi.org/10.1158/1055-9965.EPI-15-0803

Sickles, E., D’Orsi, C., Bassett, L., 2013. ACR BI-RADS® Mammography, ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System.