Machine Learning in Healthcare

Classification of Mammographic Masses in the UCI Mammographic Masses Dataset

Word Count: 2452

October 27, 2021

Contents

1 Introduction

Breast Cancer, one of the most diagnosed life-threatening cancers for women, affects 1 in 8 women in their lifetime.[1] Detecting breast cancer in the early stages of the disease is a key to increasing survival chance.[2–4] Consequently, large-scale mammography screening programmes have become a priority, with over 42 million exams performed yearly in the US and UK.[5]

Mammograms are annotated by trained radiologists according to the Breast Imaging Reporting and Data System (BI-RADS), a standardised system for assessing breast malignancies on their shape, density and margin. Despite the rise of screening programmes, mammography interpretation remains challenging, with a global shortage of mammography professionals.[6–8] Accuracy varies considerably between experts, producing high false positive and false negative rates, resulting in unnecessary patient anxiety and reducing survival chance respectively.[9–11]

Various data sets have been made available, to encourage development of methods to improve accuracy and speed of diagnosis. The UCI Mammographic Mass data set contains 961 instances of mammographic mass data, with 516 benign and 445 malignant instances collected from full field mammograms from the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006.[12] The data contains the BI-RADS attributes (shape, density, margin) and the patient's age as well as the BI-RAD assessment and goal fields. Table 1 details data attribute specifications. There are missing values from the following predictive fields Age: 5, Shape: 31, Margin: 48 and Density: 76, totalling of 160 missing values from 129 individuals, suggesting these values are generally missing at random.

Age follows a roughly normal distribution. Shape has the highest frequency of data in the irregular category and lowest in the lobular category. Margin data is most distributed in the circumscribed category and least in the microlobulated category. Density has the highest frequency in the low category, with notably few instances in other categories. Classes are not balanced with a higher proportion of individuals with benign masses versus malignant masses, suggesting it will not be very predictive (Figure 1).

Older patients are more likely to have malignant tumours and malignant tumours are most often irregular shaped whereas benign tumours are round or oval, which aligns with the current knowledge (Figure 2,3a). Ill-defined, obscured or spictulated margins have a higher density of malignant masses whereas circumscribed and microlobulated margins are more associated with benign masses (Figure 2,3b).

Variables are not highly correlated with one another. Age is correlated above 40% with Severity, whereas Density is correlated at less than 6% with Severity, likely as most of it falls in the low category (Figure 4). The potential decision boundary can be visualised between high and low margin values, high and low shape values and between young and old ages with older people with more irregular masses with more indistinct margins being more likely to have malignant masses (Figure 5).

Classification of masses using a supervised Decision Tree Classifier (DTC) and AdaBoost Classifier (ABC) is explored, with the ABC achieving the highest overall F1-score.

Table 1: Summary of Data Attributes

Attribute	Description	Category/units	Data type	Predictive
BI-RADS assessment	Standard system to describe mammogram findings and results	1= negative, 2= benign, 3=benign but follow-up, 4= suspicious, 5= highly suggestive of malignant	ordinal	No
Patient's Age	A know risk factor for breast cancer	Years	integer	Yes
Mass Shape	More lobular or irregular shapes indicate potentially malignant masses	round=1 oval=2 lobular=3 irregular=4	nominal	Yes
Mass Density	Higher density indicates potentially malignant masses	high = 1, iso = 2 low = 3, fat-containing = 4	ordinal	Yes
Mass Margin	A poorly defined or spiculated margin is indicative of breast cancer	circumscribed = 1, micro-lobulated = 2, obscured = 3, ill-defined = 4, spiculated = 5	nominal	Yes
Severity	Label/Class Field	benign = 0 or malignant = 1	binominal	N/A

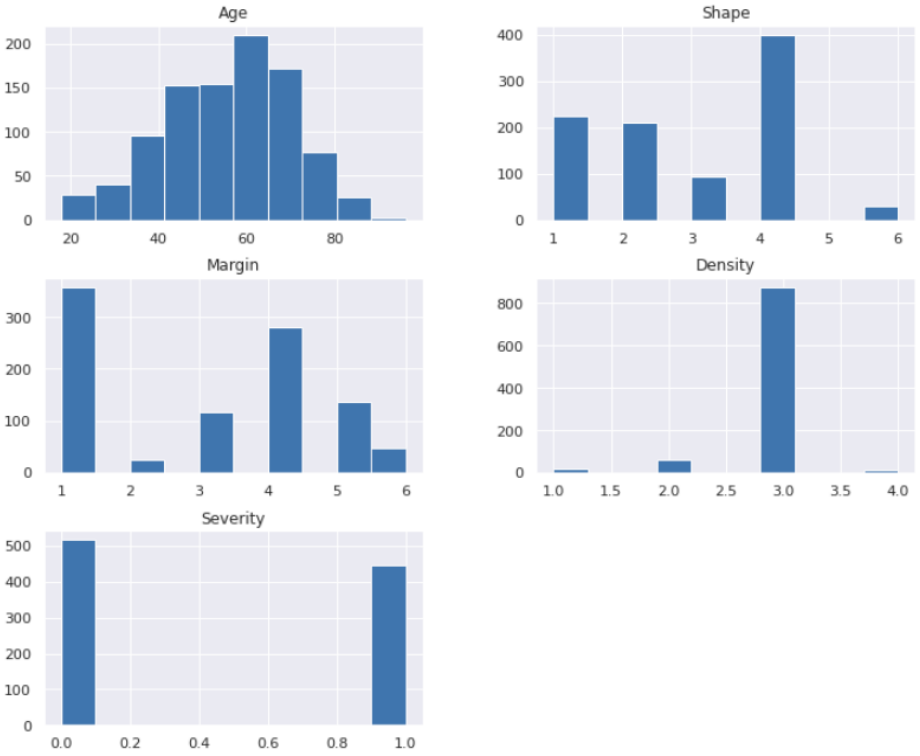


Figure 1: Histogram of Data Set Attributes, where Age: patient's age in years; Shape: round=1 oval=2 lobular=3 irregular=4, unknown= 6; Margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5, unknown= 6; Density: mass density high=1 iso=2 low=3 fat-containing=4; Severity: benign=0 or malignant=1

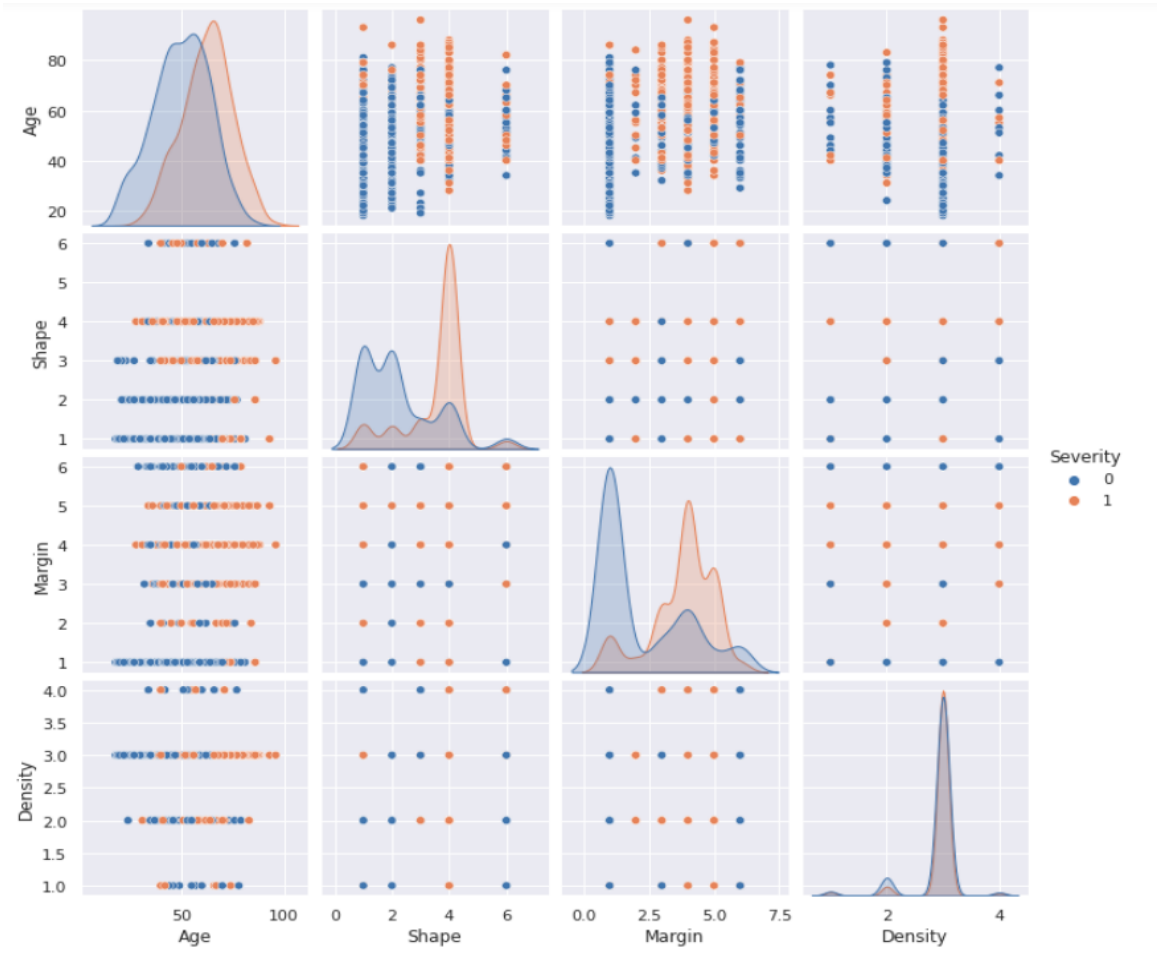
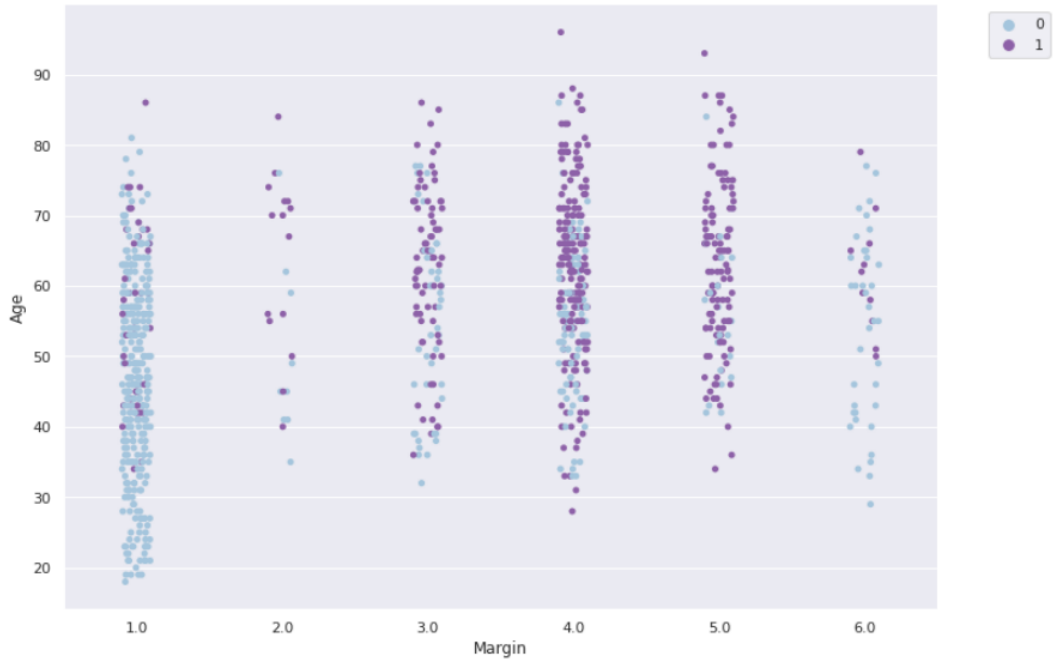


Figure 2: Pairplot of Data Set Attributes, where Age: patient's age in years; Shape: round=1 oval=2 lobular=3 irregular=4, unknown= 6; Margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5, unknown= 6; Density: mass density high=1 iso=2 low=3 fat-containing=4; Severity: benign=0 or malignant=1

(a) Shape



(b) Margin

Figure 3: Scatter plots of Categorical Attributes plotted against Age, where Age: patient's age in years; Shape: round=1 oval=2 lobular=3 irregular=4, unknown= 6; Margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5, unknown= 6; Severity (Class Field): benign=0 or malignant=1

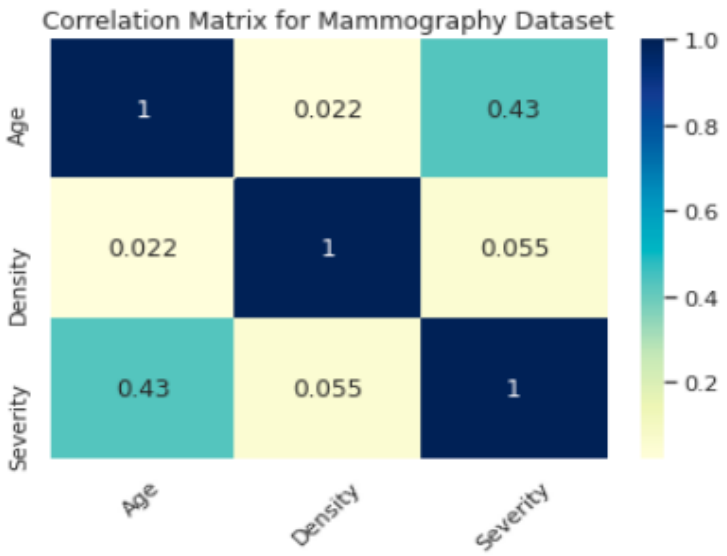


Figure 4: Correlation Matrix of Data Set Attributes, where Age: patient's age in years; Density: high=1 iso=2 low=3 fat-containing=4; Severity (Class Field): benign=0 or malignant=1

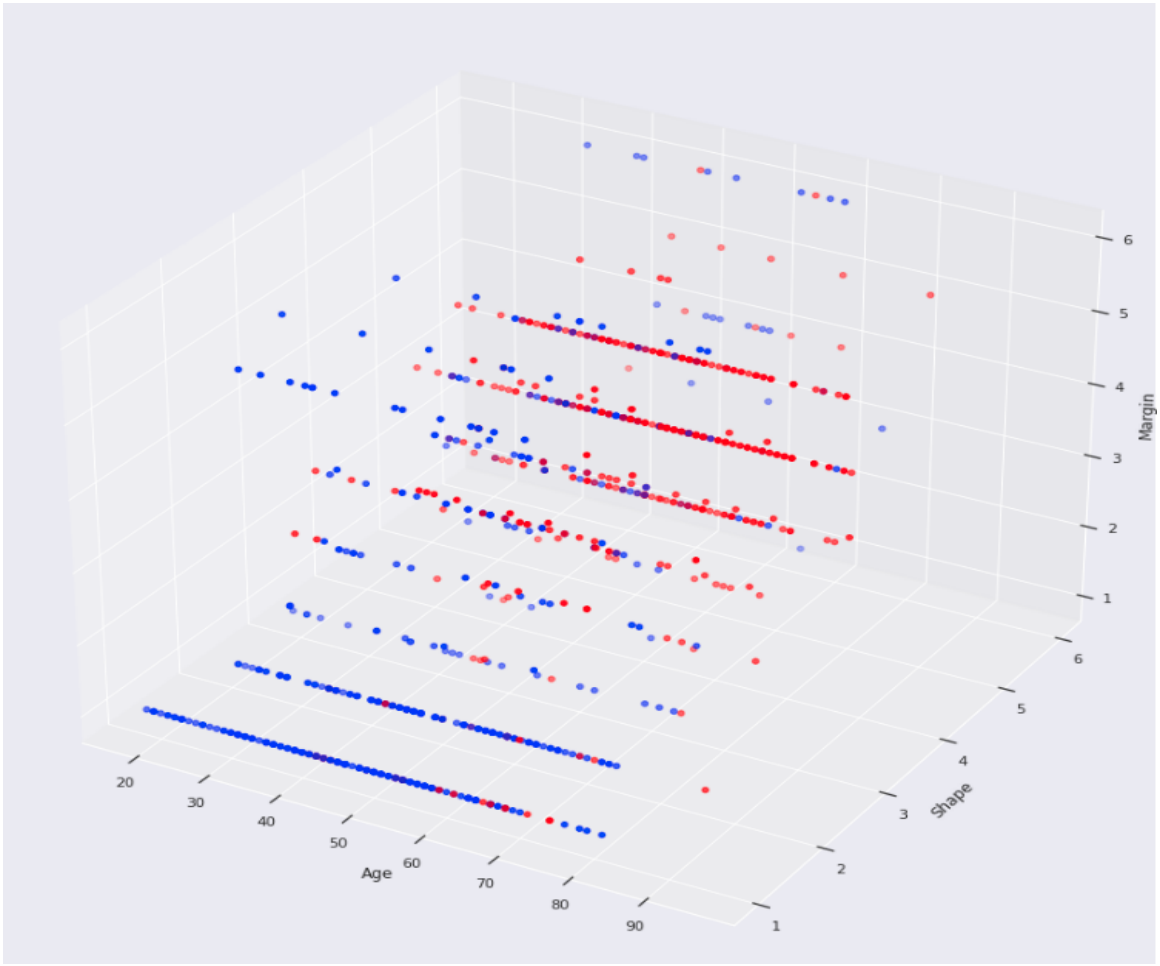


Figure 5: 3D Plot of Data Set Attributes, where Age: patient's age in years; Shape: round=1 oval=2 lobular=3 irregular=4, unknown= 6; Margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5, unknown= 6; Severity (Class Field): benign=0 (blue) or malignant=1 (red)

2 Methodology

2.1 Data Cleaning, Feature Selection and Preprocessing

Missing values are imputed for predictive features, preferable from removing 129 of 961 individuals. For continuous attributes, missing values are replaced with the mean of the corresponding class. For nominal attributes, missing values are replaced by a new unknown category to preserve information. For ordinal attributes, the mode of the attribute for each class is used. Continuous variables are checked for outliers using box plots. Nominal features are converted to objective data-types with each category, encoded into binary. Ordinal data is not encoded, to maintain the numeric importance. The data is split into variables and class with a 70:30 train:test split using the Scikit-Learn Library.[13]

As this dataset is small with minimal class imbalance, resampling was not used. Instead, F1-score and AUC are selected as the main metrics to evaluate model performance, as accuracy can be misleading on imbalanced data sets. Samples are stratified across training and test sets to ensure the same imbalance in both. Analysis of two supervised algorithms is performed on the data set: a Decision Tree Classifier (DTC) and an AdaBoost Classifier (ABC).[13] In each case the training data set is used to train the model and the test data set is used to evaluate model performance.

2.2 Decision Tree Classifier

DTCs output categorical class predictions through sequential binary splits of the data. DTCs are constructed using root, internal and leaf nodes and branches (Figure 6). Each internal node represents a choice between several alternatives and each lead node represents a classification or decision. In the training process, starting at the root node, at each node, each feature is recursively evaluated and the feature that best splits the data is used. Each recursive step selects test condition for an attribute to split the records, $A \leq v$ where v is some number. Proceeding from the top to the bottom of the tree, successive nodes are visited until a leaf node is reached. The maximum depth of the tree is important to control to prevent over-fitting and improve performance on unseen data.

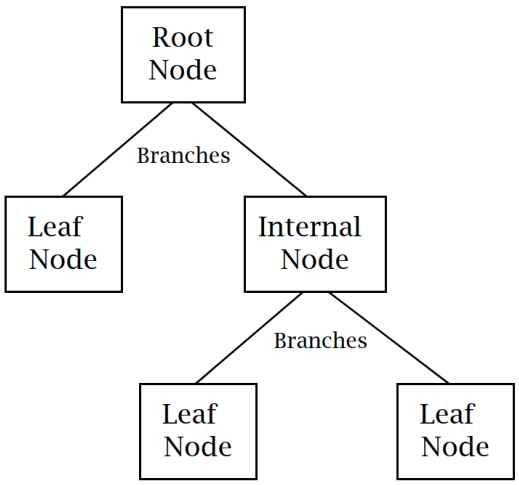


Figure 6: Decision Tree Nodes and Branches

The splitting criteria used to find the best data split are Gini and Entropy. The gini impurity is the frequency at which any element in the dataset will be mislabelled when it is randomly labelled. The optimum split is that which minimises the gini impurity, calculated as follows: $Gini = 1 - \sum_{i=1}^{n} p^2(c_i)$, where $p(c_i)$ is the probability of class c_i at a node.

Entropy is a measure of disorder of the attributes with the target and the optimum split is chosen by the split with the lowest entropy, calculated as follows: $Entropy = 1 - \sum_{i=1}^{n} p(c_i) log_2(p(c_i))$, where $p(c_i)$ is the probability of class c_i at a node.

Scikit-learn uses an optimised version of the CART algorithm /citescikit-learn(Figure 7 shows the psuedocode).

1. Start at the root node
2. For each ordered variable X,
   Convert to an unordered variable X' by grouping its values in the node into a small number of intervals if X is unordered, then set X'=X.
3. Perfrom a chi-squared test of independece of each X' variable versus Y on the data in the node and compute its significance.
4. Choose the variable X* associated with the X' that has the smallest significance probability.
5. Find the split set {X*∈ S^* } that minimises the sum of the Gini indexed and use it to split the node into two child nodes.
6. If a stopping criterion is reached, exit.
   Otherwise apply steps 2-5 to each child node.
7. Prine the tree with the CART method.

Figure 7: CART Pseudocode

2.3 AdaBoost Classifier

The ABC is an adaptive boosting ensemble algorithm fitting a sequence of small DTCs on repeatedly modified versions of data, where each subsequent model attempts to correct the predictions made by the last (Figure 8). The predictions from the DTCs are combined through a weighted majority vote to produce the final prediction. In training, the ABC randomly selects a section of training data and fits a DTC, initialising the weights applied to each training sample as $w_i = 1/N$. The weights applied to each training sample are then modified for subsequent iterations and the weak learner is reapplied to the re-weighted data. Weights are assigned to give the greatest weight to miss-classified findings and increase the chance of these being identified in the next iteration. Thus, across iterations instances that are difficult to predict gain increasing influence.[14]

The learning rate and the number of DTCs can be adjusted in the ABC. There is a trade-off between learning rate and DTCs in that the more DTCs you have, the smaller the learning rate needed. The DTCs underlying the ABC can also be tuned. Scikit-learn employs the SAMME.R(Stagewise Additive Modeling) algorithm[13,15] (Figure 9 shows the psuedocode).

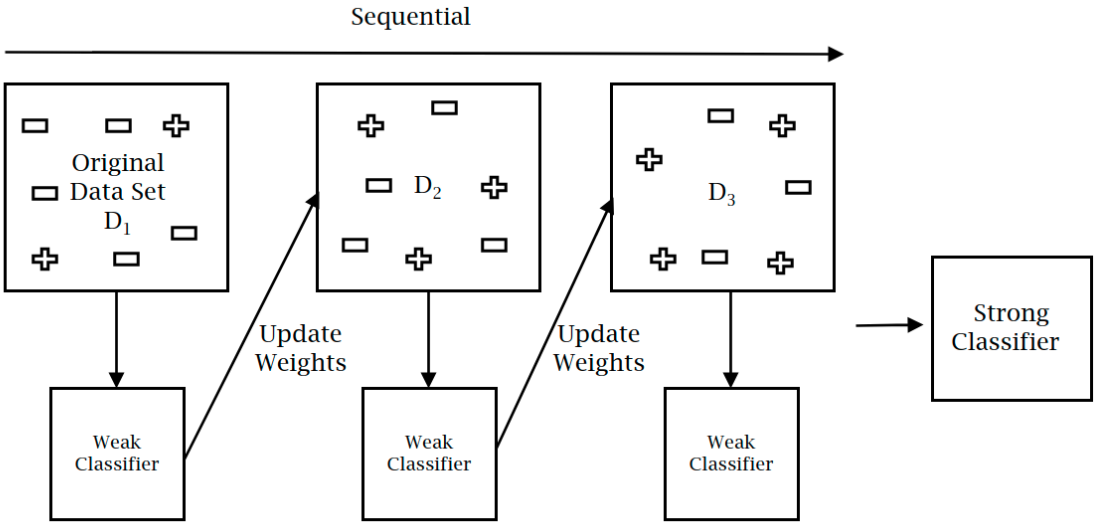


Figure 8: Adaboost Classifier Example

1. *Initialize the observation weights* $w_i = 1/n,\ i = 1, 2, \ldots, n.$
2. *For* $m = 1$ *to* M*:*
   (a) *Fit a classifier* $T^{(m)}(\boldsymbol{x})$ *to the training data using weights* w_i.
   (b) *Obtain the weighted class probability estimates*

$$p_k^{(m)}(\boldsymbol{x}) = \text{Prob}_w(c = k|\boldsymbol{x}),\ k = 1, \ldots, K.$$

   (c) *Set*

$$h_k^{(m)}(\boldsymbol{x}) \leftarrow (K-1)\left(\log p_k^{(m)}(\boldsymbol{x}) - \frac{1}{K}\sum_{k'} \log p_{k'}^{(m)}(\boldsymbol{x})\right),\ k = 1, \ldots, K.$$

   (d) *Set*

$$w_i \leftarrow w_i \cdot \exp\left(-\frac{K-1}{K}\boldsymbol{y}_i^T \log \boldsymbol{p}^{(m)}(\boldsymbol{x}_i)\right),\ i = 1, \ldots, n.$$

   (e) *Re-normalize* w_i.
3. *Output*

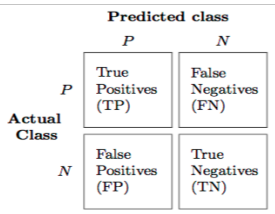
$$C(\boldsymbol{x}) = \arg\max_k \sum_{m=1}^{M} h_k^{(m)}(\boldsymbol{x}).$$

Figure 9: SAMME.R Pseudocode

2.4 Hyperparameter Tuning and Evaluation

Validation curves are plotted to observe accuracy over different values of a hyperparameter of interest, informing range selection for a grid search. This exhaustively considers parameter combinations, to achieve the best model performance.[13] Performance is evaluated using a confusion matrix and associated secondary metrics (Table 2). 10-fold cross validation is applied, which estimates the robustness of a model on unseen data by splitting the data into 10 groups, holding one group back as the test set and training the model on the remaining groups to produce an average overall score.

Table 2: Summary of evaluation metrics

Metric	Formula	Details
Confusion Matrix		True Positives, True Negatives, False Positives and False Negatives of the Target Class
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	The proportion of correct predictions out of all predictions
Precision	$\frac{TP}{TP+FP}$	The proportion of positive identifications which were actually correct
Recall	$\frac{TP}{TP+FN}$	The proportion of actual positives which were correctly identified
F1 Score	$2.\frac{Precision.Recall}{Precision+Recall}$	Harmonic Mean of Precision and Recall
AUC-ROC	Area under the Receiver Operating Characteristic Curve	Trade-off between the true positive rate and false positive rate

Note for each label: True Positives (TP) refers to the number of mammographic masses where the model correctly predicts the positive class; True Negatives (TN) refers to the number of masses where the model correctly predicts the negative class; False Positives (FP) refers to the number of masses where the model incorrectly predicts the positive class; and False Negatives (FN) refers to the number of masses where the model incorrectly predicts the negative class.

3 Results

A simple DTC was initiated with a max depth of 2 and gini splitting criteria. F1-score was 0.743 on the test set and the 10-fold cross validation F1-score was 0.760, suggesting a reasonably stable model. The hyperparameters were optimised by plotting a validation curve to observe the F1-score against maximum depth of the tree. Performance on the validation set does not significantly improve after a maximum depth of four, declining thereafter (Figure 10) suggesting beyond this the DTC starts to overfit to the training data. Grid search was carried out on the training data to confirm the optimum hyperparameters of a maximum depth of four and the splitting criteria was Gini. Using these optimised hyper-parameters, the final model was specified and fitted to the training data. The 10-fold cross validation score improved to 0.794 suggesting a model which will translate better to external data sets. The confusion matrix shows an increase in the number of true positives and true negatives and decrease in the number of false positives and false negatives. The F1-score improved to 0.767 and an AUC of 0.84 indicates this model can distinguish between the malignant and benign cases reasonably well. Visualising the resulting DTC, microlobulated shape is the feature assigned the highest importance and the first split in the data, aligning with initial exploratory plots. Following this a spiculated margin and Age of less than or equal to 59.5 years old, again aligning with original data analysis. Density has low importance and the gini index is high (0.48) when the data is split on density, which fits with density being very similarly distributed for both classes (Figure 12).

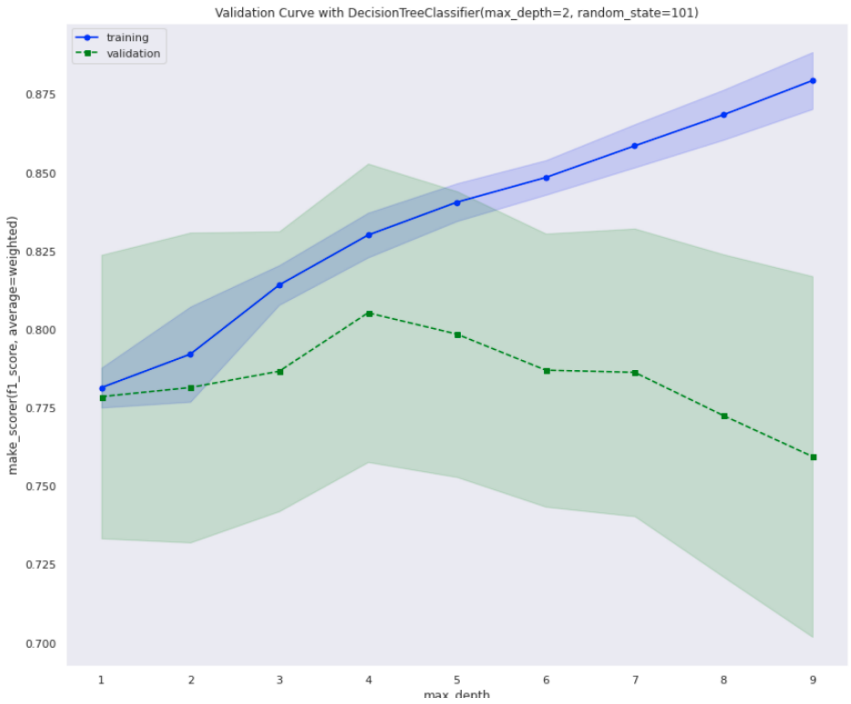
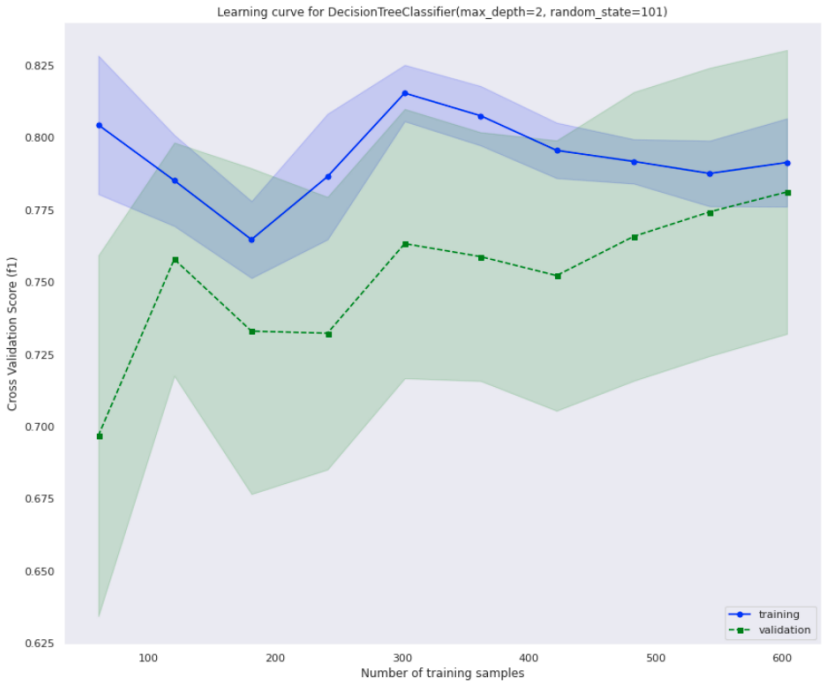
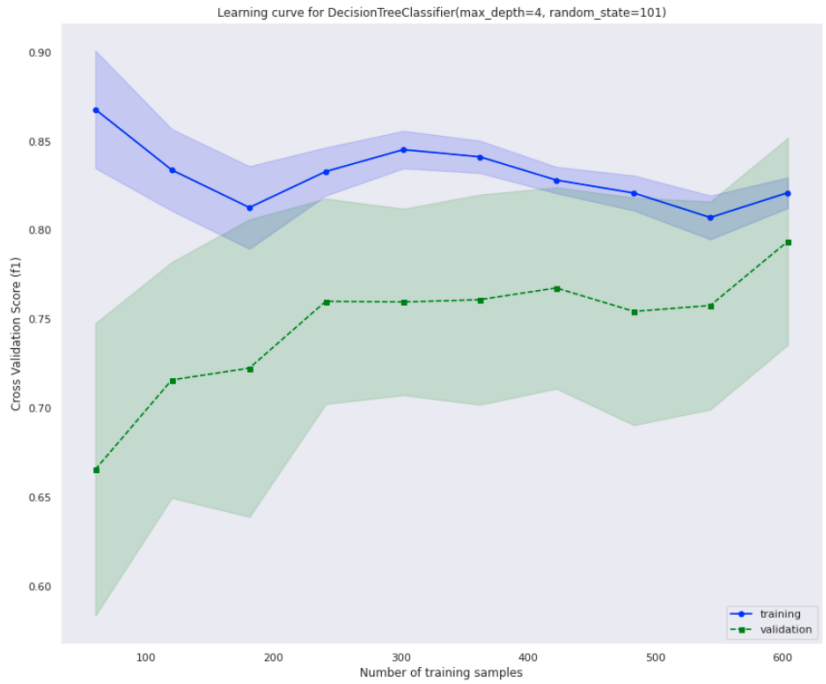


Figure 10: Validation Curve For Decision Tree with Maximum Depth plotted against the Weighted F1 Score

A sensitivity analysis, quantifying the relationship between data set size and model performance, on the original model, shows the training and validation learning curves come together after around 100 samples. This suggests the initial hyperparameters are not creating a stable model. On the final tuned model, we see the training and validation curves moving together gradually as we move to the maximum number of samples for training, suggesting consistent learning and



(a) Original DTC

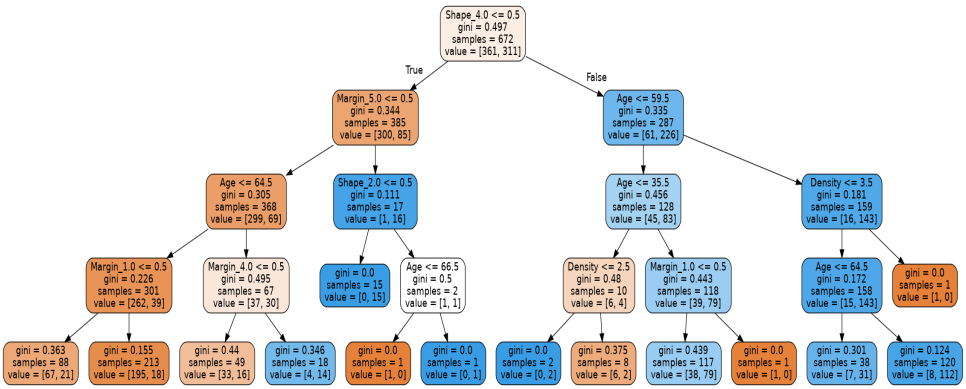


(b) Final DTC

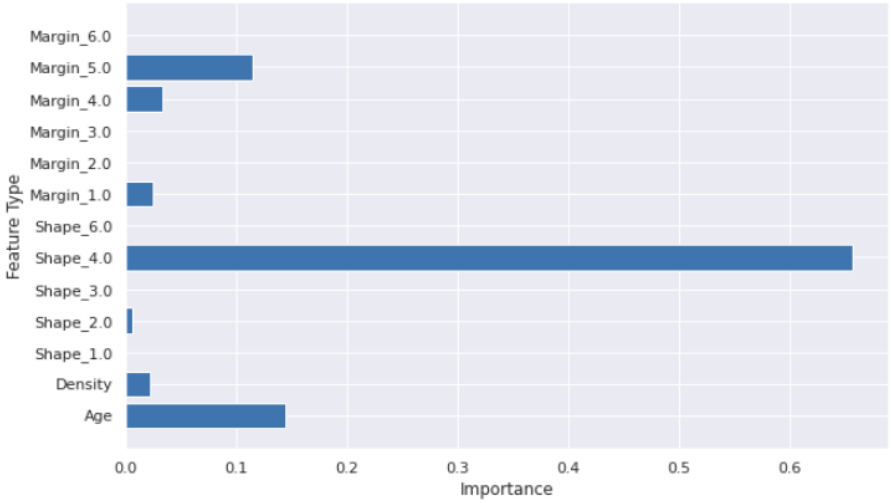
Figure 11: Learning Curves for Decision Tree Classifiers

a good bias-variance trade off by 600 samples (Figure 11).

An ABC was initiated with ten decision trees of 1 depth and learning rate of 0.1 and fitted to the training data. An F1-score of 0.753 on the test data and a 10-fold cross validation f1-score of 0.778, which is a lower performance and stability than the final DTC. The number of estimators



(a) Splits



(b) Feature Importance

Figure 12: Final Decision Tree and Feature Importance

were plotted again the weighted F1-score, which rises rapidly up to 20 estimators, stabilises and declines after 80 (Figure 13). To confirm this a grid search was carried out returning 20 estimators as the optimum. This model was specified and fitted to the training data giving an increase in cross validation f1-score to 0.802, suggesting this model is the most transferrable to external data sets. The confusion matrix shows a higher number of true positives and lower number of true negatives. However, the final F1-score is lower than for the final DTC (0.753). To combat this the ABC was initialised with 6 of the final best DTCs as base estimators and a learning rate of 0.1. This improved the F1-score to 0.780, but slightly decreased the cross validated F1-score to 0.787 so may transfer less well to external data sets.

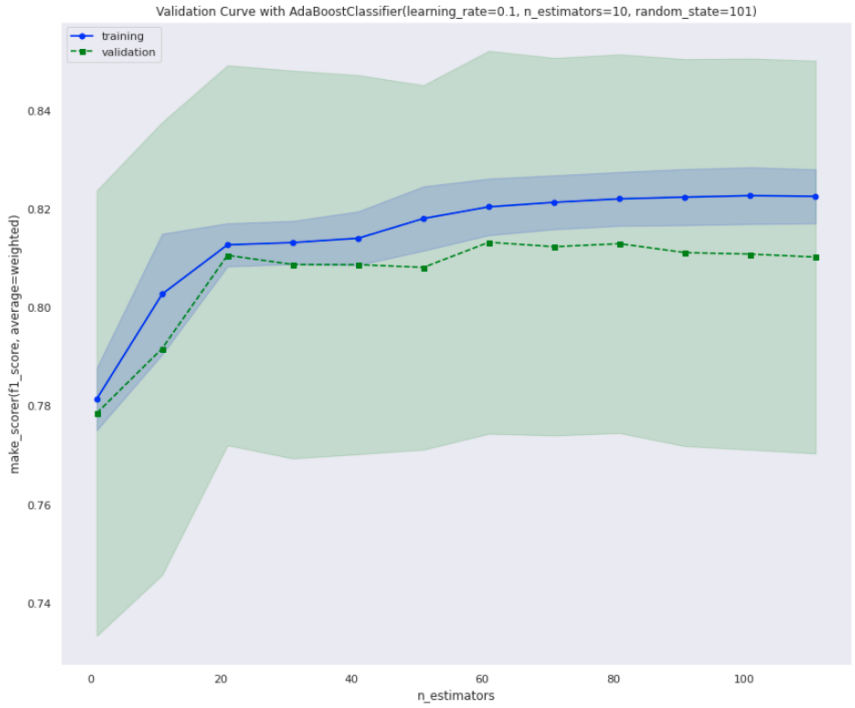
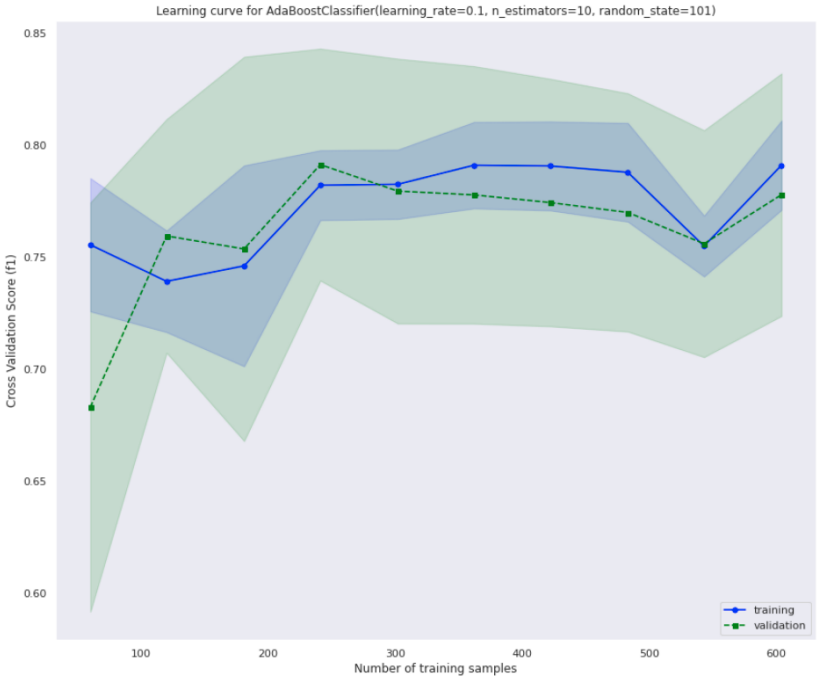


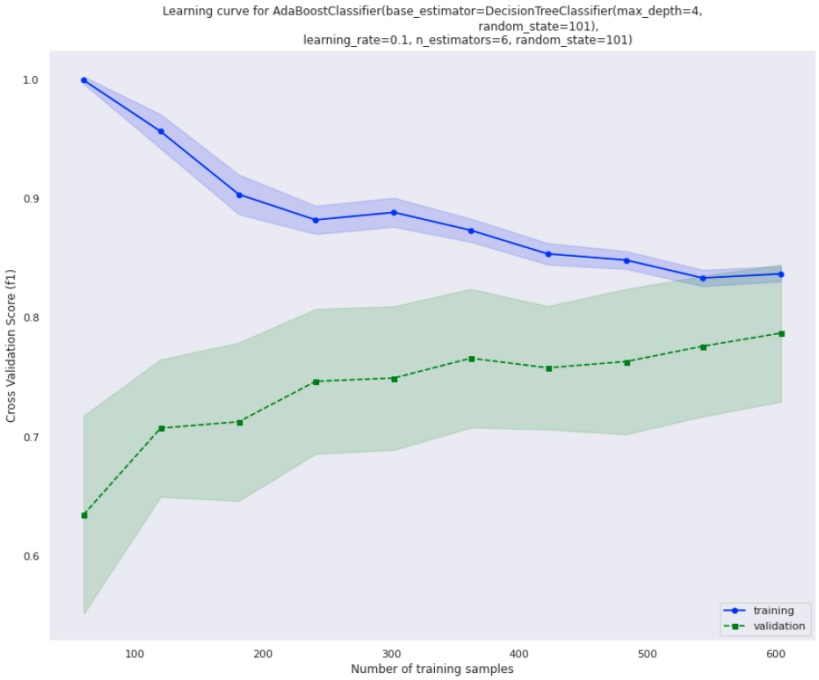
Figure 13: Final ABC Validation Curve

A sensitivity analysis on the original ABC model, shows the training and validation curves start at a low F1-score and come together after only 100 sample. This suggests these initial hyperparameters are not creating a model which improves fit as samples increased and instead the model may be biased. The F1-score peaks at around 230 samples, suggesting this is the point of diminishing returns. On the final tuned ABC model we see the training and validation curves moving together steadily up to the final 600 samples suggesting more consistent learning and a good bias-variance trade-off (Figure 14).

Overall, the final ABC achieved the highest F1-score, Accuracy and Recall but a lower cross validated F1-score meaning it may not transfer so well to external data sets. The results of both classifiers are not drastically different and the AUCs are equal (Figure 15a), likely as this is a small data set with a relatively straight-forward feature space and so boosting algorithms do not offer much advantage (Figure 15b). For malignant tumours the DTC got a higher Precision score suggesting it is more likely to minimise false positives. In practise, missing a malignant case would not be acceptable so we want to maximise Recall, therefore the ABC would be the best option for real-world use.

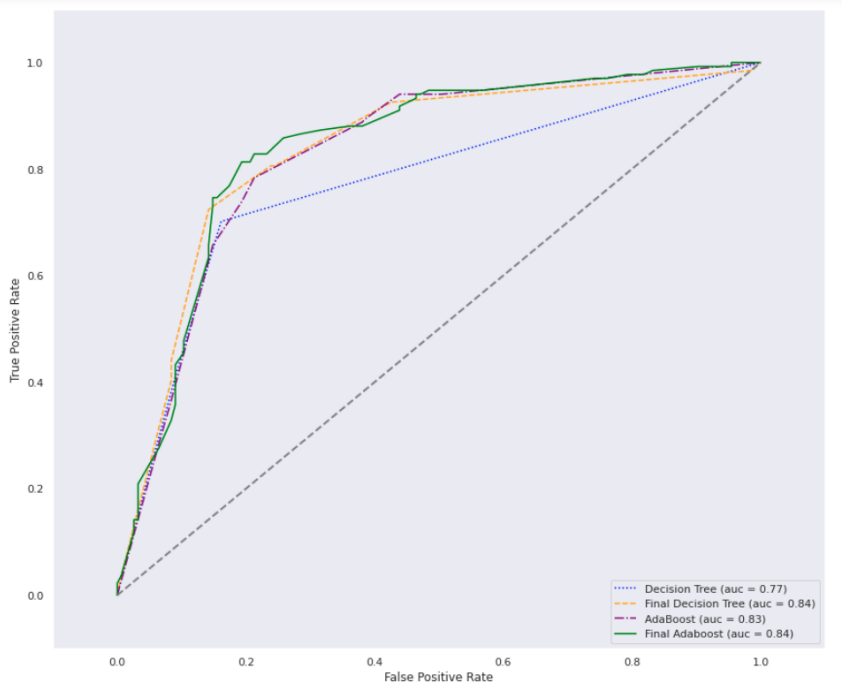


(a) Original ABC



(b) Final ABC

Figure 14: Learning Curves for AdaBoost Classifiers



(a) ROC Plots

Model	Metric	Malignant	Benign	Weighted
Dtree	f1	0.767	0.818	0.795
Ada	f1	0.78	0.815	0.799
Dtree	Precision	0.815	0.782	0.798
Ada	Precision	0.792	0.805	0.799
Dtree	Recall	0.724	0.858	0.796
Ada	Recall	0.769	0.826	0.799

```
Accuracy:
 Dtree  0.796
 Ada  0.799
```

(b) Results Table

Figure 15: Results from Final Decision Tree and AdaBoost Classifiers

4 Discussion and Conclusion

Many studies have applied machine learning techniques to mammography classification on this data se. However, many do not include a rigorous overview of their methodology or comparable evaluation metrics, making conclusive comparisons between study results difficult (Table 3). DTCs previously achieved a very similar AUC to the present study.[16,17] An ABC previously used did not report the AUC but Accuracy is similar but slightly lower than that of their DTC, suggesting they may not have optimised the underlying trees in their ABC which initially affected the present study[17] (Table 3). Support Vector Machines (SVMs) achieved a high AUC, when instances with missing values were removed.[18] Artificial Neural Networks (ANNs) were assessed in multiple studies, achieving relatively high scores ranging from 0.829-0.963.[16,19,20] Interestingly, in a direct comparison one study found dropping missing values gave them higher scores than imputing them.[19] ANNs allow for considerable customisation and so can achieve the highest score on this data set with optimum hyperparameters. However, as this data set is not especially large or complex, very good results can still be achieved with simpler models which are more explainable.

Table 3: Comparison of results with previously literature

Study	**Model**	**Pre-processing**	**Metric**	**Score**
Present Study	DTC	Imputed Missing Values	AUC	0.840
Present Study	ABC	Imputed Missing Values	AUC	0.840
Miao, 2013[20]	ANN	Removed Missing Values	AUC	0.963
Miao, 2015[18]	SVM	Removed Missing Values	AUC	0.936
Huang, 2012[16]	DTC	Removed Missing Values	AUC	0.836
Huang, 2012[16]	ANN	Removed Missing Values	AUC	0.911
Lairenjam, 2010[19]	ANN	Imputed and Removed Missing Values	Accuracy	0.840 when missing values removed, 0.829 when imputed
El Rahman, 2020[17]	DTC	Imputed Age and Removed other Missing Values	AUC/Accuracy	0.810/0.837
El Rahman, 2020[17]	ABC	Imputed Age and Removed other Missing Values	Accuracy	0.826

Note: Different Evaluation Metrics were prioritised in each paper so there was not always a direct comparison that can be made between models. In addition prepossessing steps differed between papers and some did not specify their full methodology.

Based on the literature, removing instances with missing data could improve both models. However, this could potentially remove important information and make these models less generalizable to external data sets. More sophisticated imputation methods which make null values available as a boolean splitting point for DTCs could improve performance, without losing information. Bootstrapping methods to reduce over-fitting and dimensionality reduction to reduce variance, such as Principle Component Analysis, should be tested. Balancing the

data by using a re-sampling technique such as oversampling of the minority class or setting class weights inversely proportional to the class frequencies could also improve performance.

These classifiers could be used to tackle the problems faced by breast cancer screening programmes where interpretation of mammograms remains inconsistent and skilled radiologists are in high demand.[5,9,10] The classifiers could provide a statistical second opinion for a radiologist to help improve consistency and accuracy, especially for marginal cases. However, it is important this is done in a way so as not to encourage deskilling of radiologists. These models would need careful tweaking to balance false negatives and false positives. However, this data set is hand-crafted from images by radiologists carrying out BI-RADs assessments and would need these BI-RADs attributes as inputs to classify a new instance. Thus, this input is vulnerable to subjectivity and the utility is limited. Deep Neural Networks have been used to successfully classify tumours directly from mammography images.[5,21,22] This approach has more potential to fit into the medical workflow, acting as a true second opinion by classifying tumours based on the raw mammogram data and freeing up radiologist time to improve patient care.

5 Appendix

Please refer to the corresponding iPython Notebook.

References

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[2] László Tabár, Bedrich Vitak, Tony Hsiu-Hsi Chen, Amy Ming-Fang Yen, Anders Cohen, Tibor Tot, Sherry Yueh-Hsia Chiu, Sam Li-Sheng Chen, Jean Ching-Yuan Fann, Johan Rosell, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology*, 260(3):658–663, 2011.

[3] Canadian Task Force on Preventive Health Care et al. Recommendations on screening for breast cancer in average-risk women aged 40–74 years. *Cmaj*, 183(17):1991–2001, 2011.

[4] Michael G Marmot, DG Altman, DA Cameron, JA Dewar, SG Thompson, and Maggie Wilcox. The benefits and harms of breast cancer screening: an independent review. *British journal of cancer*, 108(11):2205–2240, 2013.

[5] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[6] Paul Wing and Margaret H Langelier. Workforce shortages in breast imaging: impact on mammography utilization. *American Journal of Roentgenology*, 192(2):370–378, 2009.

[7] Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359, 2017.

[8] Yasuo Nakajima, Kei Yamada, Keiko Imamura, and Kazuko Kobayashi. Radiologist supply and workload: international comparison. *Radiation medicine*, 26(8):455–465, 2008.

[9] Constance D Lehman, Robert F Arao, Brian L Sprague, Janie M Lee, Diana SM Buist, Karla Kerlikowske, Louise M Henderson, Tracy Onega, Anna NA Tosteson, Garth H Rauscher, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology*, 283(1):49–58, 2017.

[10] Joann G Elmore, Sara L Jackson, Linn Abraham, Diana L Miglioretti, Patricia A Carney, Berta M Geller, Bonnie C Yankaskas, Karla Kerlikowske, Tracy Onega, Robert D Rosenberg, et al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*, 253(3):641–651, 2009.

[11] Nehmat Houssami and Kylie Hunter. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer*, 3(1):1–13, 2017.

[12] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[15] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[16] Mei-Ling Huang, Yung-Hsiang Hung, Wen-Ming Lee, Rong-Kwei Li, and Tzu-Hao Wang. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. *Journal of medical systems*, 36(2):407–414, 2012.

[17] Sahar A El_Rahman. Predicting breast cancer survivability based on machine learning and features selection algorithms: a comparative study. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–39, 2020.

[18] Julia H Miao, Kathleen H Miao, and George J Miao. Breast cancer biopsy predictions based on mammographic diagnosis using support vector machine learning. *Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Bioinformatics*, 5(4):1–9, 2015.

[19] Benaki Lairenjam and Siri Krishan Wasan. A note on analysis of mammography data. *Int. J. Open Problems Compt Math*, 3(5), 2010.

[20] Kathleen H Miao, George J Miao, et al. Mammographic diagnosis for breast cancer biopsy predictions using neural network classification model and receiver operating characteristic (roc) curve evaluation. *Journal of Selected Area in Bioinformatics (JBIO)*, 3(9), 2013.

[21] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI: Journal of the National Cancer Institute*, 111(9):916–922, 2019.

[22] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194, 2019.