STATS780 Final Project:
Prediction of Cancer Recurrence in Breast Cancer Survivors

John Heisey 400151293

Abstract

This report applies a variety of statistical methods onto a dataset of breast cancer survivors containing information relevant to their cancer and treatment process, and classified based on whether they have experienced a recurrence in their cancer or not. Logistic regression, bagging, random forest, boosting, neural network, and VSCC classification methods were applied to the data, with comparisons made between the prediction accuracy results. The random forest model proved to have the lowest false negative percentage ($\sim 40\%$) while possessing a prediction accuracy comparable to the other methods ($\sim 75\%$). The results can be used to help predict the likelihood of cancer recurrence for a patient.

1 Introduction

Breast cancer is the most commonly diagnosed cancer among women, as well as the second leading cause of cancer death [1]. While it is not strictly women that are diagnosed with breast cancer, there is data specific to women that can be used to help predict likelihood of a breast cancer diagnosis, as well as prognosis and the likelihood of a recurrence in patients already diagnosed. This report applies a variety of classification methods onto a dataset containing information pertaining to female patients diagnosed with breast cancer. this report has the intended goals of determining the most accurate classification methods to use, and obtaining useful information to help predict likelihood of cancer recurrence in the patient.

The dataset used for this report is the *Breast Cancer Data* set from the University Medical Centre Institute of Oncology in Yugoslavia [2]. The data is comprised of 286 observations of 9 input variables of varying type and one classification output variable. The input variables themselves all relate to the patient's history with breast cancer, including information such as the location size and type of the tumour, the treatment process, and biological information of the patient. The output variable is a binary classification stating whether they have or have not had a recurrence of the cancer (201/85 split of no recurrence/recurrence classification, respectively). The first linear attribute is the patients age at the time of diagnosis, divided into decades (e.g. age 30-39,40-49, etc.), ranging from ages 20 to 79, and normally distributed with the average patient being diagnosed at $\sim$ 47 years old and a standard deviation of $\sim$ 10 years (*Fig. 1*). The other variable not directly related to the cancer is a categorical variable which indicates if the patient has begun menopause before the age of 40, after the age of 40, or is premenopausal. The majority of the observations fall into the "menopause after 40" or "premenopausal" categories as it is unlikely to begin menopause before 40 - however it is worth looking into potential correlations between an early menopause and other factors.

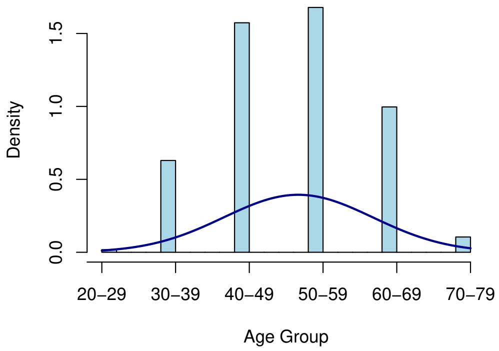

Figure 1: Density plot of patient ages ($\mu = 46.6$yrs, $\sigma = \pm 10.1$yrs)

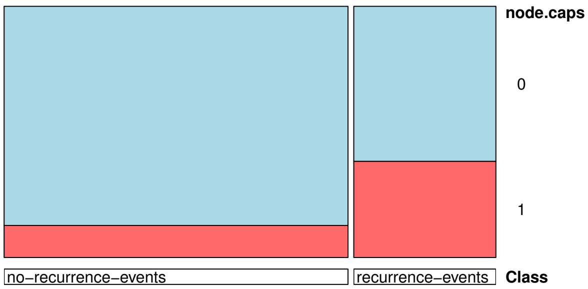

Figure 2: Double-decker plot of incidence of recurrence vs. incidence of node penetration (0/1 ↔ no/yes)

The next four variables pertain to the cancerous tumour itself. The first two are linear variables measuring the size of the tumour in millimetres, quantized similar to age in groups of $5mm$ (5-$9mm$, 10-$14mm$, etc.) and ranging from under $1mm$ to $54mm$. The next variable being the number of lymph nodes containing cancerous cells (again quantized in groups of 3 nodes) ranging from 0 to 26 nodes (*Fig. 3*). Related to the number of involved lymph nodes is the third (binary) variable stating whether or not the cancer has penetrated through the lymph node capsule, allowing the spread of the cancer to other organs in the body (metastasis). These three variables are all closely related to the fourth variable, the degree of malignancy. The degree of malignancy classifies the cancer as grade 1, 2, or 3 - with 1 being the least aggressive (highest survival rate) and 3 being the most aggressive (lowest survival rate) [3]. The three previous variables all contribute to the classification of the cancer - for instance a positive observation of node capsule penetration indicates a higher grade cancer (no observations of a grade 1 cancer with node capsule penetration) as well as a higher rate of recurrence (*Fig. 2*). While the degree of malignancy could essentially be considered a rudimentary classification based off of the previous variables, it will be important when constructing a generalized linear model to determine if the degree of malignancy predictor alone is comparable to combinations of the other linear variables.

The next two variables give information on the physical location of the cancer on the patient's breasts, including a binary variable indicating whether the cancer is located on

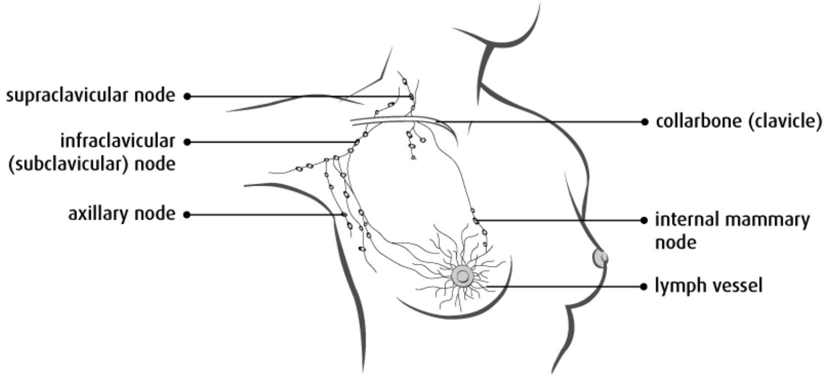

Figure 3: Common involved lymph nodes in close proximity to the breast (*source: cancer.ca*)

the left or right breast, and a categorical variable stating which "quadrant" on the breast the cancer is located (upper left, lower left, upper right, lower right and central). Cancer located on the left breast is slightly more common (the current theory is that this is likely due to the left breast being slightly larger on average, thus having more tissue for cancer to potentially develop [4]), and overwhelmingly common on in the upper or lower left quadrant of either breast (this is believed to be due to the greater amount of breast tissue located in these quadrants [5]).

The final input variable is a binary variable indicating whether or not the patient underwent radiation treatment as part of their treatment process. This is a very nuanced variable, since the decision to undergo radiation treatment is highly dependent on the tumour size relative to the patient's breast. Generally if the tumour is small relative to the breast, a lumpectomy is performed, followed by radiation treatment with the purpose of eliminating any cancer cells potentially missed by the lumpectomy. If the cancer is large or quite invasive, a mastectomy is usually performed, which often does not require radiation therapy unless the cancer was located particularly close to the chest wall. As invasive cancers are more likely to be grade 3, this variable might prove to be a better predictor for lower grades of cancer.

The output variable classifies each patient by whether or not they have had a recurrence of their cancer. If a patient does have cancer recurrence it can be incredibly difficult for them, and the people closest to them, physically, emotionally, and financially. Because of this, there is great importance put towards predicting the likelihood of a patient experiencing cancer recurrence. If predicted, they can take preventative measures to avoid recurrence, or even prepare for it if the likelihood of recurrence is high. The ability to predict likelihood of recurrence is perhaps more important to the medical industry itself, where logistics regarding treatment can be tailored to a patient depending on their likelihood of recurrence in order to give them the best treatment possible.

The intended goal of this project is to address the problem of accurately predicting the likelihood of cancer recurrence in a patient for the reasons stated above. This will be done by implementing a variety of statistical methods using the `R` programming language onto the dataset. Methods including logistic regression and classification tree analysis, along with the less discussed methods neural network classification and variable selection for classification will be utilized and compared using a supervised approach to determine prediction accuracy.

2 Methods

2.1 Logistic Regression

Logistic regression is a method which is applied to extract a useful model from a dataset of continuous input variables, and a binary output variable. The general form of the model can be expressed as:

$$Y = \mathbf{E}[Y|X] + \epsilon = \pi(X) + \epsilon, \quad \pi(X) = \frac{exp\{\beta_0 + \beta_1 x\}}{1 + exp\{\beta_0 + \beta_1 x\}}$$

where the error ϵ follows a distribution with a mean 0 and variance $\pi(X)[1 - \pi(X)]$. Since the dataset possesses multiple linear input variables (age, number of involved nodes, tumour size, and degree of malignancy), and a binary output recurrence variable, this is a well suited method for the dataset - as long as the parameters are restricted to the linear data. A generalized linear model (GLM) was applied to the linear data and the χ^2 value analyzed. The model was then reduced by removing the predictor variables with the largest p values and compared to the full model using deviance calculations. Additionally, odds-ratios were utilized to create some intuition on what biological parameters predict a cancer recurrence.

2.2 Classification Tree Analysis

Classification tree analysis is a method which is applied to datasets where the response variable is categorical in nature (in this case two categories), and whose performance can be potentially improved by bagging, boosting, and random forest methods. Supposing there are G classes in a given dataset ($G = 2$ for this dataset), classification can be performed via discriminant analysis, where the probability that a data vector $\mathbf{x}$ is in a class g follows the formula (quadratic version):

$$\mathbb{P}[g|\mathbf{x}] = \frac{\pi_g p(\mathbf{x}|g)}{\sum_{h=1}^{G} \pi_h p(\mathbf{x}|h)} = \frac{\pi_g \phi(\mathbf{x}|\mu_g, \mathbf{\Sigma}_g)}{\sum_{h=1}^{G} \pi_h \phi(\mathbf{x}|\mu_h, \mathbf{\Sigma}_h)}$$

where π_g is the probability of an observation belonging to class $g \in G$. From this, an observation $\mathbf{x}$ will be assigned to class g if $\mathbb{P}[g|\mathbf{x}] > \mathbb{P}[h|\mathbf{x}] \ \forall \ h \neq g$. This method allows the use of linear variables in the dataset as well as the categorical variables for classification. As mentioned above, the bootstrap aggregating (bagging) method was then applied by generating M bootstrap ensembles, obtaining a majority vote from the M samples, and measuring success based on the misclassification rate. This method can be extended to the random forest method by decorrelating the M ensemble trees, where for each split a random predictor subset approximately equal to the square root of the number of total predictors is used instead. This is done to avoid a small subset of predictors dominating each of the M ensemble splitters [6]. Lastly the boosting method was applied, where instead of M trees being grown independently they are now grown sequentially depending on the most recent residual response (for boosting, having a large M is not necessarily useful or practical). Additional parameters shrinkage λ and split number d are utilized in boosting, which need to be weighted appropriately relative to one another to produce useful results. The use of these extended methods could prove important in ensuring both the categorical and linear data can contribute to the recurrence prediction accuracy.

2.3 Neural Network Classification

Neural network (NN) classification is a relatively new method developed separately both in the fields of statistics and artificial intelligence [7]. Given a set of K classes $\mathbf{Y} = Y_1, \ldots, Y_K$ that are associated with a dataset, hidden units $\mathbf{Z} = Z_1, \ldots, Z_M$ which are each some linear combination of the input variables in the data.

Consequently, the classes $\mathbf{Y}$ are modeled as a function of linear combinations of Z_m expressed by [7]:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T \mathbf{X}), \quad m = 1, \ldots, M$$

$$T_k = \beta_{0k} + \beta_k^T \mathbf{Z}, \quad k = 1, \ldots, K$$

$$f_k(\mathbf{X}) = g_k(\mathbf{T}), \quad k = 1, \ldots, K$$

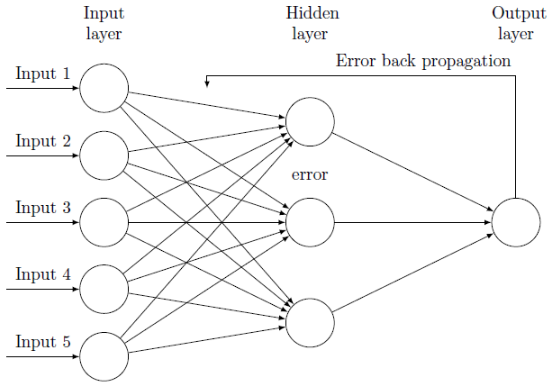

Figure 4: Graphical presentation of a NN with back propagation

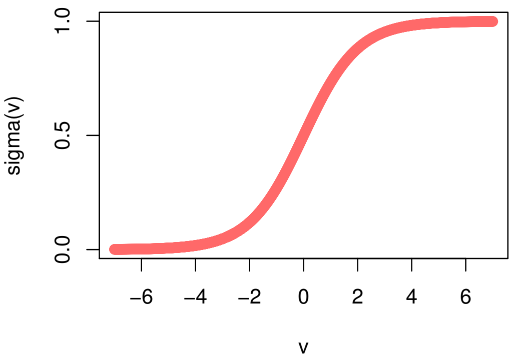

Figure 5: Sigmoid activation function (activation threshold $\in [0,1]$)

where α and β are weight parameters. NN's can be used for either regression or classification problems, and apply a nonlinear classification approach which differs it from the two previous methods. The standard single hidden layer backpropagation network was the method used for the classification of the data (*Fig. 4*), along with a sigmoid activation function $\sigma(\nu) = \frac{1}{1+e^{-\nu}}$ (*Fig. 5*) whose activation threshold can be tuned to increase predictive performance.

2.4 Variable Selection for Classification

Variable selection for clustering and classification (VSCC) is a form of dimension reduction similar to principal component analysis or factor analysis. The VSCC technique searches for a subset of variables that minimizes within-group variance while simultaneously maximizing between-group variance [8]. While the concept of variable selection and reduction is also used in other methods such as logistic regression, this technique and it's associated R package effectively automates the process in an efficient way. Comparisons in prediction accuracy between the logistic regression method and VSCC technique should generate useful insight towards what variables are most important when determining the likelihood of cancer recurrence.

2.5 Comparison Method

As the intended goal of the project is to predict likelihood of cancer recurrence for patients, all methods were compared primarily on their ability to accurately predict cancer recurrence for a given patient. Alongside this, secondary factors specific to each method were considered. These include such things as: amount of data required, type of data required, and results specific to the method. Since the dataset is unequal in terms of classified groups ($\sim 2/5$ recurrence/no recurrence class ratio) the use of stratified random sampling for the labelled/unlabelled splits was used in an attempt to reduce sampling error issues in the prediction accuracy comparisons.

3 Results and Discussion

3.1 Logistic Regression

The logistic regression approach was first applied to the data - namely the linear variables “age”, “tumour size”, “number of involved nodes””, and “degree of malignancy”, using the `glm()` function in `R`. Upon overview of the GLM summary (*Fig. 6*), certain predictor variables stand out; particularly a high correlation between having cancer recurrence and the predictor variables “cancer grade”, “number of involved nodes”, and “tumour size” due to their very small p values ($p \ll 0.01$). Conversely, the “age” predictor variable appears to have little effect on the model based on it’s large p value, and was removed to compare the resultant model with the previous model. Upon removing the “age” predictor, the residual deviance for the model goes from 345.90 to 308.98. Operating under the null hypothesis that the new reduced model is no better than the original, the difference between the residual deviances can be used to calculate the non-central Poisson χ^2 value using the `pchisq()` function in `R` and test the null hypothesis. The χ^2 result and corresponding p value were calculated to be on the order of $1-\mathcal{O}(10^{-8})$ and $\mathcal{O}(10^{-8})$, respectively. These results strongly suggest a rejection of the null hypothesis (i.e. the new reduced model without the “age” predictor variable was more effective).

Figure 6: Table Summary of GLM

Coefficients	Estimate	Std. Error	z value	Pr($>\|z\|$)
Age	-0.1542	0.1289	-1.196	0.232
Tumour Size	0.03741	0.01284	2.913	0.00358
Num. Inv. Nodes	0.16898	0.04106	4.116	3.86e-5
Cancer Grade	0.9622	0.1988	4.841	1.29e-6

Figure 7: Predictor Odds Ratios

Input Variable	Δ Quantity	Odds Increase
Tumour Size	+5*mm*	1.21
Num. Inv. Nodes	+4 *nodes*	1.97
Deg. Malignance	+1 *deg.*	46.9

The new, reduced, model was now tested to determine whether or not the degree of malignancy - which is somewhat dependent on tumour size and involved node number predictor variables - benefited at all from still having these predictors incorporated in the model. The first removed was the "tumour size" variable, which actually increased the residual deviance from 308.98 to 311.97, which subsequently ruled out the new model. Similarly, the "number of involved nodes" predictor was removed and again the residual deviance was increased to 316.87, ruling out the model without node involvement as well. This suggests that "degree of malignancy", although determined through factors including tumour size and number of involved nodes, does not produce a better model without them. It is worth noting that the accepted reduced model possessed a reduced Akaike information criterion (AIC), and all rejected models possessed an increased AIC, further reinforcing the acceptance and rejection decisions outlined. The correlation between the predictor variables that were deemed important and the probability of cancer recurrence can also be analyzed using the odds ratio e^{β_1}, where β_1 is estimated for each predictor variable from the GLM (*Fig. 7 & Fig. 8 for associated logit plot*).

The odds ratio calculations suggest a 1.21 increase in the likelihood of recurrence if a tumour is $5mm$ larger, as well as an almost twofold increase with 4 addition involved lymph nodes. The most striking however is the almost 47-fold increase when the degree of malignancy is increased a level. This result should be considered with some skepticism due to the existence of only 3 degrees of malignancy, as well as no observed cancer recurrences for degree 1 malignancy.

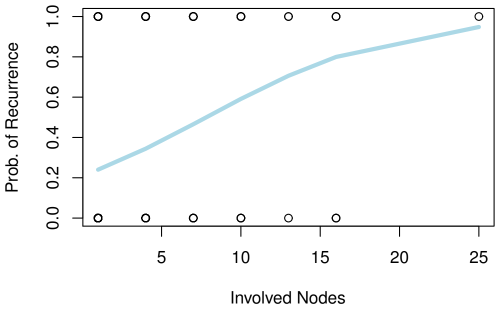

Figure 8: Logit plot of involved node number vs. probability of cancer recurrence

The predictive accuracy of the model was then tested using ten different 75/25 stratified randomly sampled labelled/unlabelled splits (split using `R`'s `caTools` package). The results proved less than optimal, with the mean prediction accuracy of the model being $\sim 72.1\%$ with a standard deviation of $\sim \pm 2\%$. While the model predicted the "no recurrence" class to a high accuracy (on average $\sim 94\%$ true negative), the main concern of the predictive model lied in the high rate of false negatives - only correctly predicting patients in the "recurrence" class $\sim 19\%$ of the time (81% false negative). This model as it stands should not be used, since accurately predicting a patient to have a cancer recurrence is a much higher priority than predicting no recurrence. For this predictive model to become useful as a solution to the presented problem it's sensitivity needs to be raised substantially.

3.2 Classification Tree Analysis

The prediction accuracy outlook is more promising for this particular kind of analysis, since all input variables available are now available to aid in prediction. The dataset was first analyzed using a general classification tree with no additional methods or pruning (*Fig. 9*) using `R`'s `rpart` package (note that a regression tree analysis is not applicable as the response variable is categorical in nature). Two resultant splitters arose as the best predictor variables for recurrence classification: the degree of malignancy, and the number of invasive nodes. A quick scatterplot analysis (*Fig. 10*) shows that the classes recurrence/no recurrence do indeed show a degree of clustering according to these properties.

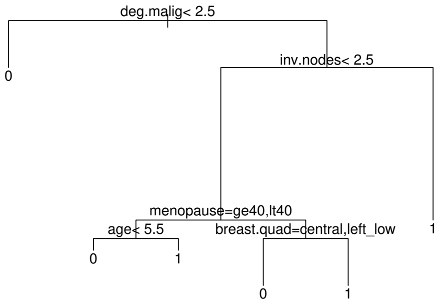

Figure 9: General recurrence classification tree

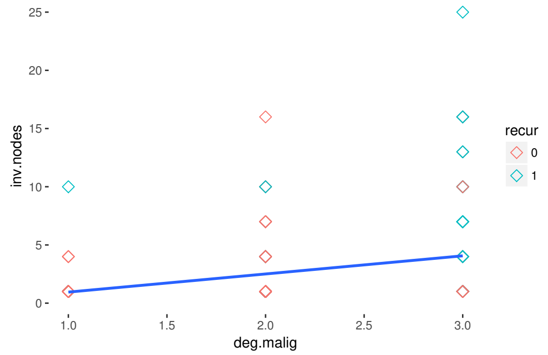

Figure 10: Deg. of malignancy vs. involved nodes with best fit line, coloured by recurrence class

While the "degree of malignancy" and the "number of invasive nodes" predictors were used in the previous linear regression model, the general classification tree has also incorporated the categorical variables regarding the patient's stage of menopause and the breast quadrant location of the cancer for classification. Using the same training/testing splits outlined earlier, the general model classified at $\sim 70.3\%$ accuracy (*Fig. 11*), which was in fact a lower accuracy than the logistic regression approach (false negative rate was roughly equal as well). The bagging method was then applied using 500 separate ensembles ($M = 500$). The bagging analysis produced predictor variables similar to the top predictors from the general model (*Fig. 12*) while indicating that the next best predictor variables for mean decrease in accuracy (MDA)/total node impurity decrease (Gini) are the tumour size and breast quadrant predictors.

Figure 11: Method Comparison

Method	Pred. Acc.	C. Disagree
General	0.703	0.297
Bagging	0.708	0.292
R. Forest	0.751	0.249
Boosting	0.943	0.057

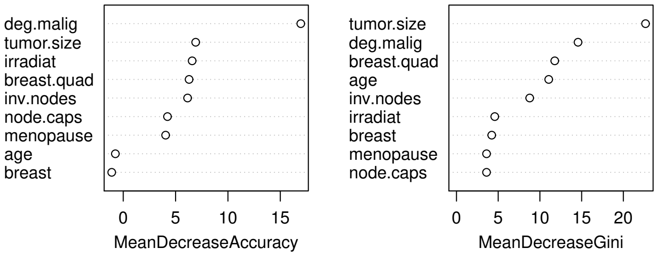

Figure 12: Bagging MDA & Gini Plots

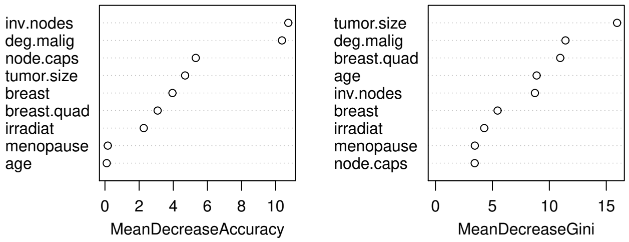

Figure 13: Random Forest MDA & Gini Plots

Using the same ten training/testing splits, the bagging model classified at $\sim$ 70.8% accuracy, which is roughly equal to the general method. The major difference however is that the bagging method had a much lower false negative rate (from $\sim$ 81% in the general model to $\sim$ 50% in the bagging model). Since reducing the false negative rate is very important in the context of the data and the problem at hand, this is a step in the right direction. The random forest method was then applied, where 3 randomly drawn predictors were used for each split in each tree. The random forest model classified at $\sim$ 75.1% accuracy while still maintaining the same reduced false negative rate from the bagging model (this essentially means the false positive rate was reduced compared to the bagging model). Upon overview of the random forest MDA/Gini plots, "number of involved nodes" moved from rank 6 to rank 1 for MDA, while the Gini plot remained virtually the same (*Fig. 13*). While this is an important jump in accuracy, the "number of involved nodes" data for a patient is obtained through a surgical procedure (usually during the lumpectomy/mastectomy to remove the cancer), which may render this data hard to acquire for hospitals with smaller budgets or large surgical time constraints. The final extension of the classification tree applied was the boosting method beginning with parameters $\lambda = 0.001$, $d = 4$, and $M = 500$. These parameters initially produced a classification accuracy of $\sim$ 76.5%, which at first glance appears to be the best model so far. Unfortunately, the false negative rate was on par with the linear regression model, leaving the model ineffective. Utilizing the optimal boosting iteration finder function `gbm.perf()`, it was found that the optimal number of iterations was likely higher than M, indicating the shrinkage parameter λ should be increased. Increasing λ threefold to

0.003 and reducing $M = 273$ returned results (*Fig. 14*) approximately equal to the previous parameter model at $\sim 75.7\%$ classification accuracy and a roughly equivalent false negative rate. Upon overview of the relative influence plot (*Fig. 15*), the degree of malignancy has once again risen to rank 1, with the number of involved nodes moving from rank 1 to 2. Comparing the prediction accuracies and false negative rates of all 4 classification methods, the random forest method appears to have the lowest false negative rate, along with a comparable prediction accuracy. This makes the random forest method the most useful model for recurrence prediction so far.

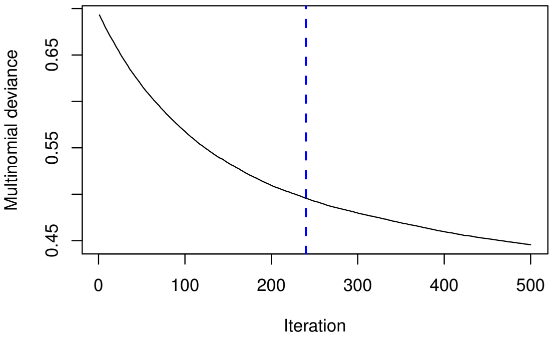

Figure 14: Boosting optimal iteration plot

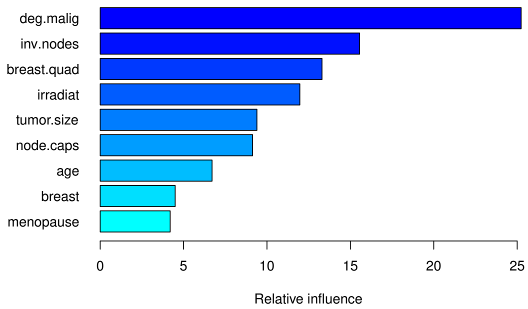

Figure 15: Boosting relative influence

3.3 Neural Network Classification

To apply neural network classification onto a dataset this small (< 300 observations) one must be very wary of the possibilities of overfitting. This issue can be mitigated by reducing the number of nodes, reducing the number of hidden layers, and increasing the the size of the training set. While increasing the training set size is not an option, and the neural network applied is already single layer, focus can be put on the node size as well as decay rate to achieve a useful model. After scaling the data, a single hidden layer neural network with back propagation was applied to the data. Initial parameters were set to `maxit`$= 100$, `decay`$= 0.25$, and `size`$= 6$ (*Fig. 16*), along with a sigmoid activation function. After fitting the model to the training sets, the model was tested and achieved a classification accuracy $\sim 76.1\%$, along with a false negative rate at $\sim 60\%$. The predictive accuracy is again similar to the other models implemented so far, but the false negative rate ranks above logistic regression, boosting, bagging, and general classification models. Increase and reduction of the size, decay, and maximum iterations parameters was undergone in an attempt to increase

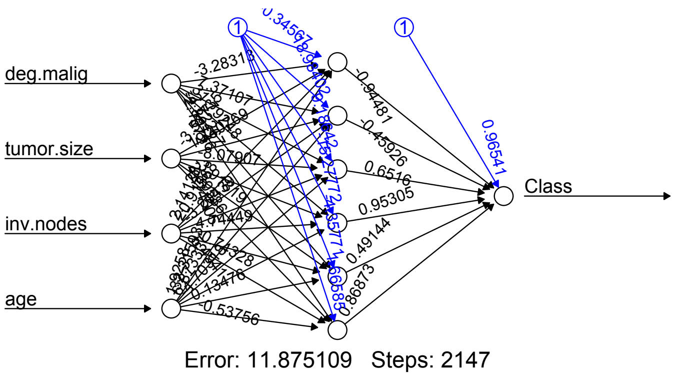

Figure 16: Single layer NN with 6 nodes applied to the linear variables

the prediction accuracy and decrease the false negative rate - however the initial parameters yielded the best results. Although this model was still below the random forest approach in terms of false negatives, it could prove more useful for larger future breast cancer datasets. If the breast cancer dataset acquired more observations, or was combined with another dataset with equal patient attributes, the neural network model could likely become the best predictive model.

3.4 VSCC

Figure 17: VSCC suggestion parameters

Parameter	VSCC Suggestion
Variables	4
Relation	3
BIC	-2770.83
Model	EEV
Family	MClust
Num. Groups	2

In a sense VSCC is an automated, much more efficient, version of the manual variable selection undergone in section 3.1, coupled with a model based approach. While VSCC is generally used on datasets of high dimensionality[8], it will still be beneficial to compare the results with the manual variable reduction from 3.1. Subsetting and scaling the dataset's linear variables (can apply binary variables in a linear format here as well), the `vscc()` function was applied to the dataset as a whole with zero guidance. The initial results proved unsatisfactory, due to the VSCC summary recommending the use of all 6 available input variables and suggesting 9 groups. This was easily fixed by providing a "group number suggestion" in the associated function parameter between 1 : 3, for which it correctly suggested two groups. To ensure this was not due to the limited suggestion options, the function parameter was increased to 1 : 5, for which two groups was again correctly suggested. The VSCC model suggested 4 variables (*Fig. 17*), with the top suggestions being the degree of malignancy, whether or not the patient received radiation treatment, node capsule penetration, and tumour size. While the results are similar to the manually obtained logistic regression results, both node capsule penetration and the instance of radiation treatment were suggested instead. These new variable suggestions are very interesting due to them having the most intuitive correlation with a cancer recurrence (particularly node capsule penetration, which suggests a higher degree of malignancy and thus a higher rate of recurrence). Upon prediction testing of the model, the results possessed a false positive rate of $\sim 50\%$ which is on par with the best predictive models so far. However the model will likely not be feasible due to its very low true negative rate $\sim 22\%$ which classifies a large majority of the negative population as candidates for cancer recurrence. This case of an "overly sensitive" classification model would simply indicate that too many patients will have recurrence, likely leading to a lot of wasted resources on patients who will likely not have a recurrence.

4 Conclusions

The proposed problem of accurately predicting a patient's likelihood of cancer recurrence have been addressed by a plethora of different statistical methods using a variety of dataset parameters. All models produced a prediction accuracy rate in the low-to-mid 70th percentile accuracy range (*Fig. 18*), which is relatively accurate. Comparing with other's attempts to classify the data accurately [9], it appears $\sim 77\%$ is the current highest prediction accuracy, which is comparable to these results. Relating the prediction accuracy to the data, is it highly likely than cancer recurrence in a patient is inherently quite random and hard to predict given the data available. In the context of the problem, these accuracies are likely not worth any considerable logistical or infrastructural change within the healthcare

Figure 18: Complete comparison of methods
(** low false neg. rate is offset by high false pos. rate*)

Method	Pred. Accuracy	False Neg.
Logistic Regression	$72.1 \pm 2\%$	$\sim 80\%$
Gen. Classification	$70.3 \pm 2.5\%$	$\sim 80\%$
Bagging	$70.8 \pm 3.4\%$	$\sim 50\%$
Random Forest	$75.1 \pm 3.1\%$	$\sim 40\%$
Boosting	$75.7 \pm 4.1\%$	$\sim 80\%$
Neural Network	$76.2 \pm 3.4\%$	$\sim 60\%$
VSCC	$71.1 \pm 2\%$	$\sim 50\%^{*}$

system since the problem deals with patient livelihoods, and thus demands high accuracy. Along with a requirement for high accuracy, the other parameter continuously referenced throughout this report is the percentage of false negative classifications. The need for the false negative percentage to be low is especially important in the context of this problem, since telling a patient that *is* likely to have cancer recurrence that they are unlikely to have recurrence leaves them unprepared for probable recurrence. Unfortunately, the majority of models used produced false negative percentages far too high to warrant any legitimate use in the oncological field. Regardless, the random forest model was shown to have the lowest false negative rate at $\sim 40\%$ (*Fig. 19*), along with a high prediction accuracy in comparison to the other models. If this model was treated as an auxiliary predictor in conjunction with a more precise model, there is potentially a use within the oncological field for it.

Aside from the primary results, secondary analysis showed every method had it's respective pros and cons. The logistic regression method - while possessing a poor false negative percentage - produced a set of odd's ratios which are potentially useful for a healthcare professional to make a more informed prognosis for a patient. The classification methods were able to utilize all the attributes in the dataset, which makes the method more adaptable to a range of similar datasets. The neural network approach, although similar in performance to the other models, could potentially increase in predictive accuracy if more data were available to train on. The final method applied (VSCC) was useful to compare with the other models to prioritize variables and determine the number of attributes needed.

Figure 19: Random forest test set classification

	Pred. No Rec.	Pred. Rec.
No Rec.	42	7
Rec.	8	12

To reiterate, the random forest model proved to be the most effective in solving the presented problem - although further analysis and attempts to reduce the false negative rate should be done to make the model better suited for field use.

References

[1] *Breast Cancer Statistics* Cancer Research UK,
`cancerresearchuk.org/health-professional/cancer-statistics`

[2] *Breast Cancer Data*, Zwitter & Soklic, OpenML
`https://www.openml.org/d/13`

[3] *Historical grading and prognosis in breast cancer*, Bloom & Richardson
British Journal of Cancer, 1957

[4] *Breast size, handedness and breast cancer risk.*, Hsieh & Trichopolous
Harvard School of Public Health, 1991

[5] *Why is carcinoma of the breast more frequent in the upper outer quadrant? A case series based on needle core biopsy diagnoses.*, Lee AH
Nottingham City Hospital, 2005

[6] *Ensemble Methods*, Zhou
CRC Press 2012

[7] *The Elements of Statistical Learning*, Hastie et. al
Springer 2008

[8] *Variable selection for clustering and classification*, Andrews & McNicholas
Journal of Classification, 2014

[9] *Supervised Classification on Breast Cancer Data*, Vanschoren et. al, OpenML
`https://www.openml.org/t/13`