DISEASE TREATMENT RESPONSE PREDICTION

Sichu Sun, Linxi Ran, Alexander Bull, Qiao Lu, Zehao Li

University of Nottingham

ABSTRACT

Study into varying Lymph node data in MRIs and other clinically measured features in attempt to estimate pathological complete response (PCR) and relapse-free survival time (RFS) should chemotherapy be used. This research explores the application of machine learning methods through disease treatment response prediction. This project defines various types of machine learning methods and compares the performance of each method to discover the best prediction results. The prediction targets are classification task for PCR and regression task for RFS. This study implements classification tasks for PCR and regression tasks for RFS. Eventually, the models get a max accuracy 77.20% and highest balanced classification accuracy of 57.79% for PCR prediction and 20.90 lowest MAE loss for RFS prediction.

Index Terms **—** Machine Learning, Breast Cancer, Multilayer Perceptron, Random Forest

1. INTRODUCTION

1.1. Background

Breast cancer is the most common cause of cancer in women. The increasing public awareness enable people to have earlier diagnosis that is suitable for complete surgical resection and therapies [1].

An in-depth understanding of breast cancer MRI results is crucial to providing better information for both doctors and patients. This comes from finding out whether the chemotherapy on a certain set of data ended up with a PCR (pathological complete response) or at least with a longer RFS (relapse-free survival). This was completed using varying methods of feature selection and ML method selection to attempt to get the best general accuracy possible.

1.2. Aims and Objectives

Our aim in the paper is to draw a way forward for machine learning models in the field of predicting pathological complete response, as well as relapse free survival to see if there are better indicators that the cancer would react well with chemotherapy or not.

2. RELATED WORKS

In this section, we want to talk about the previous research about breast cancer diagnosis by ML/DL, analyse and evaluate their methods.

Aswolinskiy et al. [2] proposed a modular two-step approach to output and encode the specified biomarkers. They use different CNNs for the tissue segmentation and mitoses detection. And then they derived biomarkers from the output to assess their predictive value for PCR. By this creative approach, the tissue segmentation achieved 76% overall accuracy, 90% of tumour and 56% of lymphocytes were correctly predicted. Without the filtering, the mitoses recall was 98% with precision 32%, while with filtering the recall was 64% with precision of 60%. Filtering removed around 77% of the detected mitoses on the NKI slides, 55% on the SCDC slides and 59% on the RUMC slides. However, their results are based on the small cluster dataset and didn't make comparison with the research community. That means their method may not hold the larger dataset with more missing data.

pCR is a presumptive surrogate for disease-free survival in breast cancer patients who have received neoadjuvant chemotherapy (NAC). The accuracy of NAC response prediction by machine learning algorithm and its comparison with human assessments are usually not reported. Huang et al. [3] leveraged the multi-stain histopathologic images, proposed an automatic workflow for breast cancer pCR prediction from pre-NAC biopsies. They proposed a feature extraction pipeline called IMPRESS and it showed good accuracy when predicting response to NAC in breast cancer patients. In HER2+ cohort, IMPRESS achieved better performances ($AUC = 0.8975 \pm 0.0038$) than Pathologists' assessed features ($AUC = 0.7880 \pm 0.0065$) significantly with t-test statistic = 64.59. And in TNBC cohort, IMPRESS achieved slightly better performances ($AUC = 0.7674 \pm 0.0209$) than Pathologists' assessed features ($AUC = 0.7626 \pm 0.0095$) with t-test statistic = 0.94. Although a good AUC score was observed in HER2+ subtype, both models presented inadequate recall value (0.4000). That is the obvious shortcoming.

The methods these two literatures proposed tried to make a better prediction for pCR in different aspects and they both make some achievements significantly. However,

they also have different disadvantages when they were used in practical. But we think they are still valuable for our research and worthy of being referred to.

3. METHOD

3.1. Data Preprocessing

3.1.1. Data Normalization and Handling of Missing Data

When it came to the preprocessing of data, the group decided on using artificial data rather than only the real data despite there not being too many (about 2% of the data samples had missing data). This decision came from the fact that the dataset is already small and so we should be using all that we have access to, in this case these samples with missing data. The averages used were a mix of median and mode depending on the feature. 'PgR', 'HER2', 'LNStatus', 'TrippleNegative' and 'ChemoGrade' had all their missing values calculated using mode whilst 'Proliferation' 'HistologyType' were subsidized with median values.

Further preprocessing included the normalization of all the features using a scalar transform around 0. This transformation was chosen over a logarithmic one due to the fact there is less data loss in the calculation.

3.1.2. Feature Selection

For feature selection 3 different methods were tried of varying complexity to came to different successes over an averaged result.

Firstly, the wrapper method, which is the simplest method of the three. The specific wrapper method used was recursive feature elimination. This method excels at systematically identifying and ranking features based on their individual contribution to the performance metric. By doing this in a simple way, RFE creates a more transparent decision-making system for humans to understand. Sadly, this method can be very computationally expensive which can be overlooked when only run once. Despite this the main drawback is how the method can have vastly different effectiveness depending on the initial feature ranking. In our case this led RFE to be the worst performing feature selection method with an average value up to 0.05 worse than the others.

Secondly, we have an embedded method. This comes about in the form of a random forest model for ranking the features. A positive to this method is the fact it can assess feature importance whilst it is still training. This method is also very adept at managing a wide variety of data types which is good for both the regression and classification problems this paper looks at. In contrast with the wrapper method, this embedded method is far less computationally expensive. Unfortunately, there are cons to this method, such as a bias towards features with more variety of categories, as well as being very abstracted from its answer. Because of the latter, it is hard for a human to understand how the method came to its conclusion of rankings. In our testing this method produced the best results.

3.1.3. Principal Component Analysis

The principal component analysis is used to reduce the features dimensionality in a more non-linear way. PCA is a good method as it can take the original feature space and transform it into a smaller set of uncorrelated components. By doing this we not only achieve a dimensionality reduction (to help fight the curse of dimensionality) but also can identify the most important features for the overall variability of the outcomes. PCA does however, suffer from a similar issue to random forest as there is no interpretability for human beings. Furthermore, PCA assumes there is a linear relationship between features and if this is wrong then the accuracy can drop. Out of the 3 methods PCA landed in the middle when it came to comparison with the previous two.

3.2. Machine Learning Methods

3.2.1. Multilayer Perceptron Network

A Multilayer Perceptron (MLP) is a type of artificial neural network that consists of multiple layers of nodes, or artificial neurons. The training of an MLP involves using a supervised learning algorithm, such as backpropagation, to adjust the weights and biases based on the error between the predicted output and the actual target output. It is versatile and can be applied to various tasks, including regression, classification, and pattern recognition.

3.2.2. Linear Regression (Logistic Regression)

Linear regression is a simplistic machine learning method that works by modelling probability that a given instance belongs to a particular class. The algorithm employs a logistic function to combine the features into a range of 0 to 1 for the probability of the answer belonging to each class. Training of this algorithm consists of optimisation of the weights of each of the features for each class. This model excels at its readability for human beings as it is so simple compared to the modern complex algorithms.

Logistic Regression is used for binary classification tasks, predicting categorical outcomes (e.g., yes/no, 1/0). It uses classification metrics (accuracy, precision, recall, F1 score) appropriate for its use case.

This method's code includes the following steps: model initialization and initial training, cross-validation using Stratified K-Fold, performance metrics calculation and model saving.

Linear Regression is used for predicting continuous outcomes, suitable for regression tasks. It uses MAE for error measurement, fitting its regression context.

This method's code includes the following steps: model training, prediction and evaluation, model saving, cross-validation and residuals analysis.

3.2.3. Support Vector Machine

Support vector machines are powerful algorithms that work well for both classification and regression. SVMs work by finding the best divider through the dimensional space to separate the groups into the different classes. The algorithm tries to maximise the gap between classes whilst also minimizing errors. SVMs are particularly good for high dimensional problems and so depending on how much dimension reduction is done, may have an upper hand on other methods.

3.2.4. Decision Tree

Decision tree models work in an equivalent way to a growing tree, it recursively splits the dataset into subsets depending on the most important feature (the one that divides the set most equally) at each node. Every node may represent a class label or value and so work well with both regression and classification problems. To make the tree, each node splits the data further and further based on the best feature at the node. These models are extremely interpretable as when the model makes a prediction it is just answering questions about the data that slowly work down to find out what the prediction should be.

3.2.5. Random Forest

The model framework works similar to decision trees in that it is essentially made up of a bunch of them together. By randomizing the decision tree creation process in varying ways, you can create a large set of similar but all different tree models. By doing this the random forest model helps to eliminate over fitting as it perturbs the data a lot on its own anyway. The forest then decides on a prediction based on the collective decision of the trees inside of it. This algorithm is great for this kind of data as it works to be a more generalised approach rather than fitting exactly the data and so we can help mitigate the problem of the small train data set by using this approach.

4. EVALUATION

4.1. RFS Prediction Evaluation

4.1.1. MAE

When it came to the mean absolute error (MAE), the best result we were able to produce for the RFS regression problem is 20.90. This means, on average we were off by about 3 weeks with our best MAE performance model (Multi-Layer Perceptron Network with PyTorch) which can be a life changing time for a patient. However, most of our other models would produce an average of over 25 for MAE. However, there were more models besides the best that produced somewhat decent averaged MAE in the context of the group, and that is the SVM and random forest models, both producing an average of roughly 23.

Method	MLPNet	MLPReg	LR	SVM	DT	RF
Average MAE	**20.90**	27.92	26.88	22.92	25.95	23.27

Table 1. Average MAE value across methods

4.2.2. Residual Analysis

In this evaluation method, the x-axis represents predicted values, and the y-axis represents residuals. The closer the points are to being randomly distributed along the horizontal line (y=0), the better the performance of the model. In observing these charts, the model that best fits the data is: Random Forest regressor.

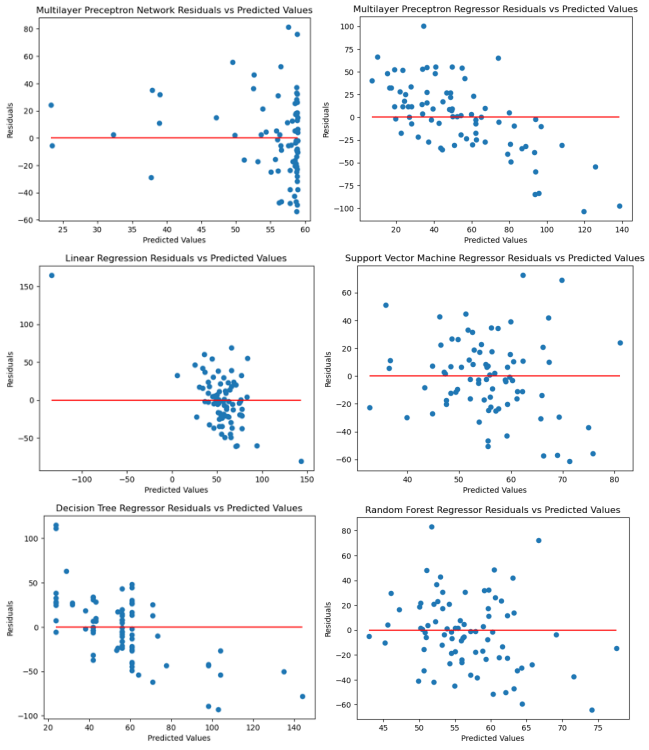

Figure 1. Residuals Analysis results

4.2. PCR Prediction Evaluation

4.2.1. Accuracy and Balanced Accuracy

Accuracy is a very basic metric to use as it only says the amount of time it is right. While this does give a reasonable view of what the model is able to do it can also be quite

misleading if the data set is one sided in representation. Using balanced accuracy is better as it creates an average based on class specific accuracies allowing each class to contribute to the "accuracy" of the model equally rather than one sided.

4.2.2. Precision, Recall, F1 Score

The precision metric is focused on how often a positive result is correct from the model. This means that we focus of making sure if it says yes that that will more likely end up being a yes (with a higher precision value). This metric is less useful when false negatives are less important, however in relation to breast cancer relapse, false negatives are important and so precision is a very good metric for this data.

Recall works in the opposite way to precision, instead of comparing true positives with things predicted positives, recall looks at the model ability to capture all positive instances. It does this by calculating the proportion of true positives among all positives. Sadly, focusing too much on this metric creates a likelihood for false positives which in the context of this work could mean a person going through chemotherapy when it may not help them.

F1 score is the best of both worlds in relation to recall and precision. It provides a balanced metric for the model that takes into consideration both false positives and false negatives. While it may be more useful sometimes to lean more on recall or precision, as both aspects are important here, F1 score is one of the better evaluation metrics for the current research.

Methods	MLPNet	MLPCla	LR
Average Accuracy	**77.20%**	74.02%	75.30%
Balanced Accuracy	52.93%	**57.79%**	54.33%
Average Precision	0.3300	**0.3865**	0.3633
Average Recall	0.1065	**0.2967**	0.17.91
Average F1 score	0.1554	**0.3278**	0.2337

Table 2. Evaluation results for classification

4.2.3. ROC Curve

The shape of the ROC curve depends on various factors, including the performance of the model used, the size and characteristics of the dataset, and the thresholds applied during evaluation. We obtained a jagged ROC curve, but upon further investigation, this is not an error; rather, it reflects an aspect of the model's performance under the current settings. More importantly, the area under the curve (AUC) is a significant indicator. The larger the AUC value, the better the model's performance.

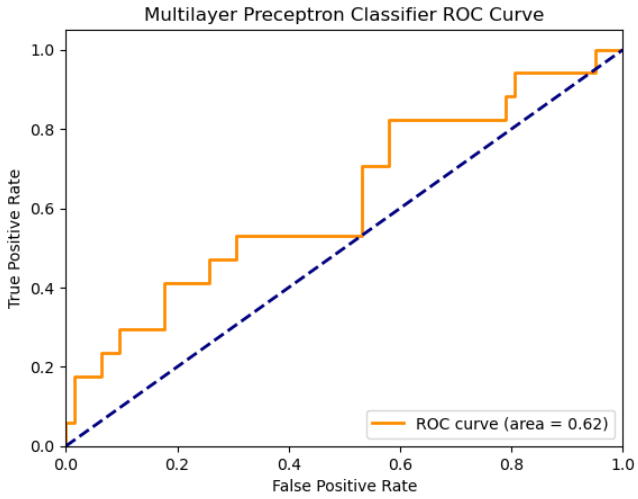

Figure 2. ROC curve for MLP classifier

5. DISCUSSION

Advantages and disadvantages of using Multi-Layer Perceptron classifier and random forest regressor:

Although Multilayer Perceptron Network has the smallest MAE value as 21.41, the residuals analysis figure appears obvious tendency that residuals value increases along with rise of predicted value. The random forest regressor distributes centralized which shows a best performance among residuals analysis but has a lower MAE value.

The MLP classifier has a strong ability to simulate any functions and capture complicated non-linear relationships, it also shows great flexibility through adjusting the number of layers and neurons. However, it needs regularization techniques such as dropout to avoid overfitting.

The random forest regressor is utilized for both feature selection and result prediction, which can automatically select features and can mitigate overfitting problems occurs on single decision tree by integrate various decision trees. Conversely, it is hard to interpret the decision process.

6. CONCLUSIONS

To conclude, we decided on PCR being predicted using Multilayer Perceptron classifier and on RFS using Random Forest regressor. Despite having the lowest "accuracy" when considering balanced accuracy, the MLP classifier had the best value and had the best F1 score when it was used on the classification problem. Regarding the regression problem, the RF was shown to be the best overall out of all the ones we got. Despite this being the most successful method, it should be noted that none of the accuracy values we got were good enough to the point that we would be able to justifiably apply this to a real person and decide whether they should get chemotherapy. To do this far more training data would be required with more test before this happens.

7. REFERENCES

[1] Sharma, Ganesh N., et al. "Various types and management of breast cancer: an overview." *Journal of advanced pharmaceutical technology & research*, 1.2 (2010): 109.

[2] Aswolinskiy, W., Munari, E., Horlings, H.M. et al. "PROACTING: predicting pathological complete response to neoadjuvant chemotherapy in breast cancer from routine diagnostic histopathology biopsies with deep learning." *Breast Cancer Res*, 25, 142 (2023).

[3] Huang, Z., Shao, W., Han, Z. et al. "Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images." *npj Precis. Onc.* **7**, 14 (2023).

Contribution Table

Task and weighting	Data pre-processing	Feature Selection	ML method development	Method Evaluation	Report Writing	Total
Sichu Sun	0%	33.33%	50%	30%	0%	0.238325
Linxi Ran	0%	33.33%	0%	70%	13%	0.193315
Qiao Lu	0%	0%	50%	0%	25.00%	0.2
Alexander Bull	0%	33.33%	0%	0%	37.00%	0.194325
Zehao Li	100%	0%	0%	0%	25.00%	0.175