Machine Learning for Breast Cancer Treatment Response and Survival Prediction

Chaitanya Manem　Rohit Varre　John Gopidi　Dharanija Bantu　Purva Soni

{psxcm4, ppxrv1, psxjg13, psxdb7, psxps10}@nottingham.ac.uk

***Abstract*—Breast cancer affects women more frequently than any other type of cancer across the globe and is one of the main causes of death in females. Chemotherapy is not an effective solution for everyone suffering from breast cancer as only 25 percent of the patients receiving chemotherapy will achieve a complete response. In addition, it leads to various side effects. Thus, predicting a woman's response to chemotherapy prior to the treatment is beneficial. This paper aims to predict the pathological complete response (PCR - Classification) which is the response for chemotherapy and relapse-free survival (RFS - Regression) length of days a woman can survive without getting cancer after chemotherapy and surgery. Our primary goal is to compare and analyze different Machine Learning techniques in order to determine the most suitable method with good prediction accuracy. Weighted Logistic Regression, K Nearest Neighbour, Smote, Random Forest Classifier, Gradient Boosting Method, Neural Network models are explored to predict PCR. Out of all the models, though the accuracy was lower, the recall was highest for the Weighted Logistic Regression method for PCR. Random Forest and Artificial Neural Network models are explored to predict RFS. The best-fitted model for RFS is ANN which has the lowest mean absolute error.**

***Index Terms*—Breast cancer, chemotherapy, machine learning**

I. Introduction

Breast cancer develops tumour in the breast tissues. If the tumour is very large with advanced severe cancer, normally patients undergo chemotherapy, a toxic treatment to shrink the tumour. The success rate for patients going through Chemotherapy is quite low. Moreover, they experience several after-effects including tiredness(fatigue), Anemia, hair loss and others. We developed and compared several approaches based on clinical features and some imaging features to develop prediction models for pathological complete response and Relapse Free Survival among patients with breast cancer.

II. Method

Data includes patients with breast cancer who received chemotherapy. This data comprises 400 records with both clinical and image features (MRI) for each patient. Data driven techniques are used to understand the features and select the features. Different data pre-processing methods are used to make the data suitable for different machine learning models. Also over sampling techniques are used to address the imbalance in the data. Various machine learning models like logistic regression, random forest, Gradient boosting, Neural Networks are used.

A. Data Pre-Processing

Starting with treating the missing values, imputation techniques were used based on the feature type. Continuous features are imputed with mean values and categorical features are imputed with mode values. Outliers were removed using Inter quartile range approach to make sure different statistical tests and linear machine learning models to work better on data. As normality is the common assumption for models and tests used in the machine learning, input features are transformed into normal distribution. Because some of the input features contains both positive and negative numbers yeo-johnson normalisation [1] technique is used. Data is also scaled between the range 0 and 1. This is necessary to work better especially for classification problem.

B. Feature Selection

Significant Features are obtained differently for classification and regression tasks.

For predicting Relapse Free Survival:

- Kendall Rank Correlation [2] is used to find the non-linear relationships between the continuous parameters as the data is not linear with respect to the target variable and the significant features with the cut-off correlation coefficient value of -0.05 and 0.05 is observed on both sides of the scale i.e. -1 to 0 and 0 to +1.
- Pearson Correlation is measured for testing the linear dependency of the different features and the significant features are observed based on the correlation coefficient cut-off of 0.05 for both positive and negative relationship.
- Feature Importance method is used for learning the significant features contributing to the regression output of Random Forest model with sci-kit learn package's feature importance method.

For predicting Pathological Complete Response:

- The parameter Age is tested with a statistical test, a T-test assuming the null hypothesis, that the difference of means is same. With a high p-value between two groups therefore null hypothesis is accepted and the Age is found to be insignificant parameter for PCR prediction.
- The relationship between PCR outcome and different categorical features such as ER, PgR, HER2, TrippleNegative, ChemoGrade, Proliferation, HistologyType, LNStatus and TumourStage are explored graphically and with a

statistical test, Chi square test and the significant features are found to be ER, PgR, HER2, LNStatus with 10% critical value.
- For Continuous Features Point Biserial Correlation [4] is used to select the significance features. Based on the correlation coefficient parameters more than 0.13 are selected.

C. Dimensionality Reduction for Relapse Free Survival

- t-SNE (t-distributed Stochastic Neighbor Embedding) [3] is used to reduce the dimensionality, mapping data of higher dimensions onto low dimensional space. t-SNE is used for the dimensionality reduction when the data is not linearly separable. Unlike Principle Component Analysis it is iterative, i.e. for each time it changes and cannot be applied the same for different data.

For the regression task the feature selection methods and the dimensionality reduction methods wasn't useful to find any significant features as the features have much less correlation coefficient to be at least considered as significant with both Kendall rank correlation and Pearson correlation. Therefore, all the features are considered for training except pathological complete response.

D. Choosing Performance Metrics

- The Mean absolute Error is used as the performance metric for predicting the Regression task outcome which is Relapse Free Survival.
- For classification, as data is imbalanced with majority of data points given for PCR outcome '0', careful examination done between different performance metrics: accuracy, recall, roc_auc_score. Recall metric is identified as the important metric for this case.
- In this type of data case, accuracy might not provide better results, however true positive rate (recall) is a better metric. A model with carefully tuned recall is high likely to spot the patients with good chances of PCR success. However some false positives are expected.
- Some models in this experiment are tuned for high accuracy and some for high recall. Finally, a high recall is achieved than the accuracy using weighted logistic regression.

E. Model Development

For predicting Relapse Free Survival:

- For spotting the non linear relations in the data a tree based method random forest regressor [5]is used. As it found there is not much linear relationship between target variable and features also no existent feature is highly correlated, Neural network with multiple layers and neurons is used to generate high level abstract features.

For predicting Pathological Complete Response:

- Logistic Regression Model is used to observe the classification accuracy.
- A weighted Logistic Regression [6] is used to train the the model in order to penalise the model if it predicts incorrectly.
- K Nearest Neighbour is used as the data sample is small and the new labeled data is expensive to obtain.
- Smote (A Sampling Method) An up sampling method is used to generate new synthetic samples to improve the classification accuracy.
- Random Forest Classifier is used as it goes through different if-else blocks of the decision trees, therefore has a higher probability of getting better accuracy.
- Gradient Boosting Method is used as an ensemble learning technique where it learns from the errors of the Random Forest Classifier.
- Neural Network [8] with sigmoid as an activation function and the binary cross entropy as a performance measure.
- Xtreme Gradient Boosting method [7] is used as an ensemble method to combine the accuracy of many models to better predict the Pathological Complete Response.

F. Method Evaluation

For Relapse Free Interval:

- All features are used to train the Random Forest Model.
- Grid Search with Cross Validation is used to tune the hyper parameters for the Random Forest Model.

The optimal hyper-parameters post Grid Search are found to be:

- bootstrap: True
- max_depth: 5
- max_features: 117
- min_samples_leaf: 20
- n_estimators: 700

For the Artificial Neural Network the hyper-parameter are chosen as follows: All features are used to train the neural network model.

- Number of Layers: 6
- Number of Neurons in a layer:
  - 1st Layer: 240 Neurons
  - 2nd Layer: 480 Neurons
  - 3rd Layer: 480 Neurons, L2 regularisation is used with factor 0.01
  - 4th Layer: 240 Neurons, L2 regularisation is used with factor 0.01
  - 5th Layer: 120 Neurons
  - 6th Layer: 1 Neuron for the the regression output.
- Activation Function: Relu activation function is used across the all the layers of the neural network.
- Metric: Mean Absolute Error is used as performance metric.
- Optimiser: Stochastic Gradient Descent is used as an optimiser for the neural network with the learning rate of 0.1.
- Number of Epochs: 100

TABLE I
MEAN ABSOLUTE ERROR FOR RFS PREDICTION METHODS.

Evaluation Method	Mean Absolute value
Rf	21.49
ANN	17.72

Method Evaluation for Pathological Complete Response:

Method evaluation for extreme Gradient Boosting Classifier method:

- Features selected for logistic regression are ER, PgR, HER2, TrippleNegative, ChemoGrade, Proliferation, HistologyType, LNStatus and TumourStage and continous features based on point biserial correlation.

Hyper-parameters are tuned by using Grid Search with Stratified K-fold cross validation.

The best hyper parameters for the XG Boosting method are:

- n_estimators: 100
- scale_pos_weight: 1
- learning_rate: 0.1
- max_depth: 1
- min_child_weight: 3
- gamma: 0
- subsample: 0.8
- colsample_bynode: 0.7
- colsamle_bylevel: 0.9
- colsample_bytree: 1

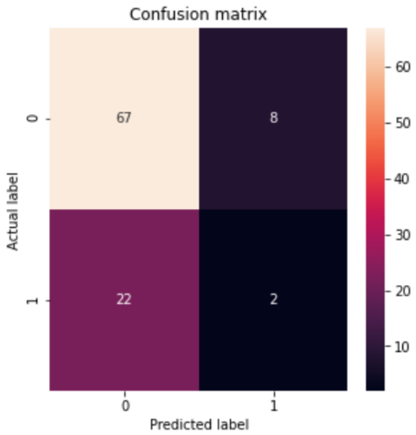

Fig. 1. Extreme Gradient Boosting Confusion Matrix

The performance metrics obtained for XG Boosing method are:Accuracy: 0.69, Precision: 0.20, Recall: 0.08, F1 Score: 0.117

Method Evaluation for Logistic Regression:

Features selected for logistic regression are ER, PgR, HER2, TrippleNegative, ChemoGrade, Proliferation, HistologyType, LNStatus and TumourStage and continous features based on point biserial correlation. Using Grid Search with Cross Validation of 3 folds and recall as a performance metric:

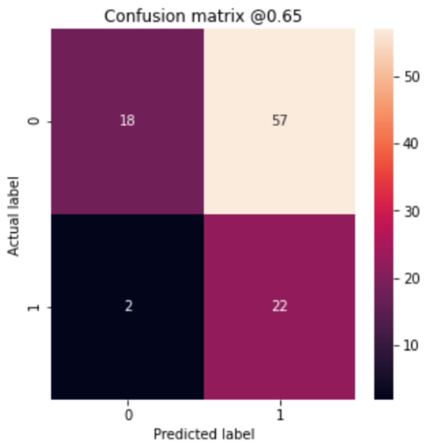

Fig. 2. Weighted Logistic Regression Confusion Matrix

- class_weight: It says about the weights given for each prediction of the logistic regression method for each true and false prediction to penalise when predicted wrong. The optimal weights are found to be: 0:1, 1:100
- threshold: It represents the cut-off probability in the precision vs Recall curve for the optimum accuracy. The optimal threshold is found to be 0.07.
- The precision recall curve is obtained as:

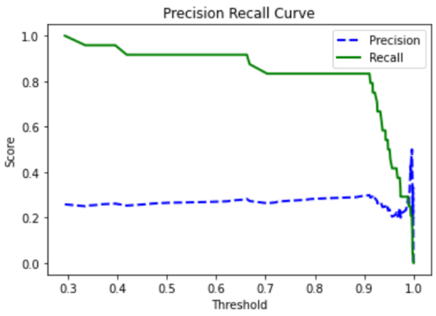

Fig. 3. Weighted Logistic Regression Precision-Recall Curve. The X-axis shows the threshold probability and the y-axis shows the scoring metric of that paticular curve.

- The following are the different metrics obtained, accuracy: 0.40, precision: 0.278, Recall: 0.916, F1 Score: 0.427

Logistic Regression with up sampling SMOTE:

- Features selected for logistic regression are ER, PgR, HER2, TrippleNegative, ChemoGrade, Proliferation, HistologyType, LNStatus and TumourStage and continous features based on point biserial correlation.
- Minority Class is oversampled using Synthetic Minority Oversampling (SMOTE).
- Now the data points are equally split with target label with 448 data points for training.

Grid Search with cross validation is used with 3 folds and recall as a scoring method.

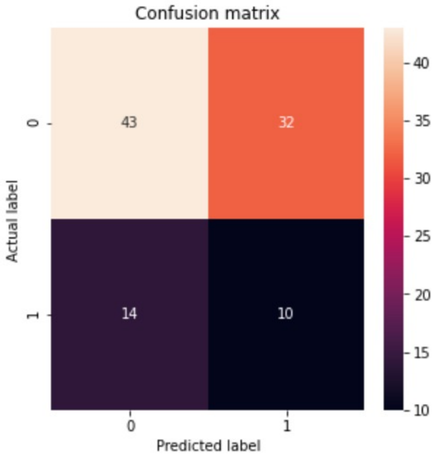

Fig. 4. Confusion Matrix for SMOTE Method

- Accuracy: 0.535, Precision: 0.238 , Recall: 0.417, F1 Score: 0.303

Method Evaluation for Random Forest:

- Features selected for logistic regression are ER, PgR, HER2, TrippleNegative, ChemoGrade, Proliferation, HistologyType, LNStatus and TumourStage and continous features based on point biserial correlation.
- n_estimators’: 70,
- min_weight_fraction_leaf: 0.005,
- min_samples_split: 0.06,
- min_samples_leaf: 1,
- min_impurity_decrease: 0.0,
- max_leaf_nodes: 10,
- max_features: ’auto’,
- max_depth: 8,
- criterion: ’entropy’

The performance metrics for the random forest model are found to be:

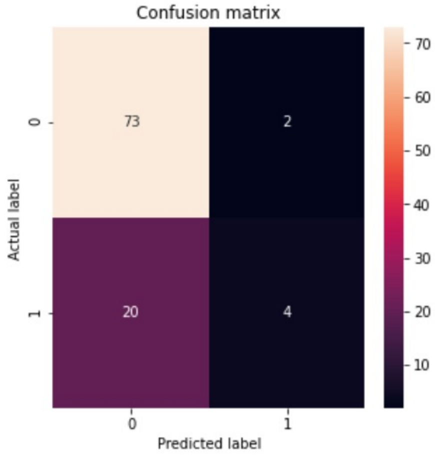

Fig. 5. Confusion Matrix for Random Forest Model

- Accuracy: 0.778, Precision: 0.667, Recall: 0.167, F1 Score: 0.267

For the Artificial Neural Network the hyper-parameter are chosen as follows:

- All features are used to train the neural network model.
- Minority Class is over sampled using Synthetic Minority Oversampling (SMOTE).
- Now the data points are equally split with target label with 448 data points for training.
  - Number of Layers: 9
  - Number of Neurons in a layer:
    - 1st Layer: 170 Neurons
    - 2nd Layer: 240 Neurons, L2 regularisation is used with factor 0.02
    - 3rd Layer: 320 Neurons, L2 regularisation is used with factor 0.025
    - 4th Layer: 480 Neurons
    - 5th Layer: 560 Neurons, L2 regularisation is used with factor 0.025
    - 6th Layer: 240 Neuron
    - 7th Layer: 120 Neurons, L2 regularisation is used with factor 0.02
    - 8th Layer: 24 Neuron for the the regression output.
    - 9th Layer: 1 Neurons for the classification output.
  - Activation Function: Relu activation function is used across the all the layers of the neural network.
  - Metric: Accuracy is used as performance metric.
  - Optimiser: Adam Optimiser is used as an optimiser for the neural network with the learning rate of 0.0001.
  - Number of Epochs: 500
  - Accuracy: 0.77, precision: 0.75, Recall: 0.125

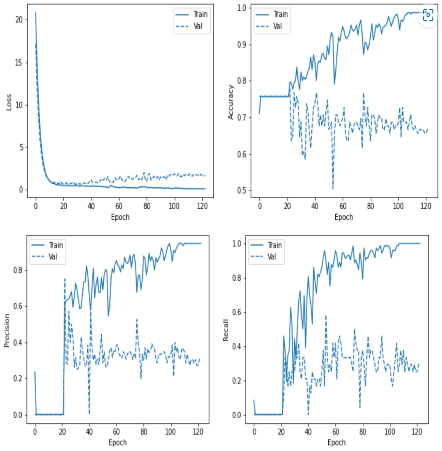

Fig. 6. Performance metrics for the Neural Network

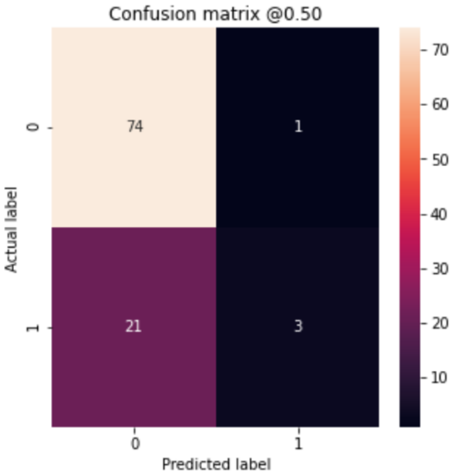

Fig. 7. Confusion matrix for the neural network

TABLE II
METRICS SUMMARY FOR PCR PREDICTION

Method	Accuracy	Precision	Recall
Weighted Logistic Regression	0.40	0.278	0.916
SMOTE	0.535	0.238	0.303
KNN	0.67	0.36	0.45
Random Forest	0.778	0.667	0.267
Gradient Boosting	0.69	0.20	0.08
Deep neural network	0.77	0.65	0.125

G. Method Discussion

Relapse-Free Survival (RFS):

- The feature selection procedure resulted in the use of Kendall, Pearson correlation, and dimensionality reduction techniques, which revealed that there is no significant relationship between features. Implemented ANN with L2 regularizer to decrease the weights of unimportant features. These type of deeper models are prone to overfitting, the L2 regularizer and dropout functioned as anti-overfitting components.
- Dropout has a stronger regularization impact, whereas batch normalization normalizes the output which has no influence on regularization.Used Stochastic gradient descent with L2 as Adam twiks with the weights of the neural network so does the L2.
- In ANN after all the combinations testes these gave the best results:

TABLE III
SUMMARY OF THE NEURAL NETWORK METHOD USED.

Layer (type)	Output Shape	Param
dense$_1$58$(Dense)$	(None, 240)	28320
dropout$_4$0$(Dropout)$	(None, 240)	0
dense$_1$59$(Dense)$	(None, 480)	115680
dense$_1$60$(Dense)$	(None, 480)	230880
dense$_1$61$(Dense)$	(None, 240)	115440
dropout$_4$1$(Dropout)$	(None, 240)	0
dense$_1$62$(Dense)$	(None, 120)	28920
dense$_1$63$(Dense)$	(None, 1)	121

- Total params: 519,361 Trainable params: 519,361 Non-trainable params: 0

Pathological Complete Response (PCR):

- The approach of weighted logistic regression with a threshold set to attain maximum recall (the true positive rate, 95%) increases the percentage of predicting the possibility of a certain patient being effectively treated with PCR.

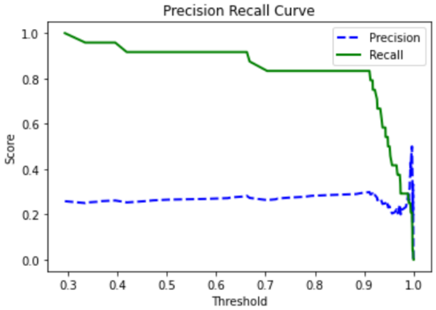

Fig. 8. Precision recall Curve for weighted Logistic Regression

- The advantage of this method is that it decreases the possibility of losing out on a patient who has a higher chance of improving with PCR treatment. While the high true positive rate (recall) is crucial, not every patient should be treated with PCR. The false positive rate of 75% eliminates 25% of false pcr treatment instances, which is slightly disadvantageous given the high noise rate.
- Using weighted logistic regression with threshold tuned for achieving higher recall is a best fit for this type of imbalanced dataset as, predictions cannot be drawn on an accuracy score of 75%, which is low for forecasting critical and crucial outcomes.
- Also the performance of the logistic regression is slightly boosted by the SMOTE technique. This technique does a descent job in balaning the number of examples by generating new minority class data points around the existing minority data points in KNN fashion. One important metric to tune here based on the data is K. In some cases this new generated data might go far from the original data characteristics.
- Another method which showed better performance in accuracy is neural network with L2 regularisation.
- The wide and deep network architecture generated new features from the existing features along with some regularisation worked well for this data, however this has many hyperparameters to tune to find the optimum parameter combination.

H. Conclusion

In conclusion we developed and compared several approaches to modelling PCR and RFS prediction in patients with breast cancer. Considering the highly imbalanced dataset these priliminary models giving promising results. Though the image features not found particularly useful in predicting the target variables some of the categorical variables having high influence. Careful selection of performance metric like precision or ROC_AUC_Score or tuning the weights of the classes might give better results, as next steps either we can find out better features by collaborating with clinical experts or generating some synthetic data by the complex models like GANs to boost the accuracy.

References

[1] Weisberg, Sanford. ”Yeo-Johnson power transformations.” Department of Applied Statistics, University of Minnesota. Retrieved June 1 (2001): 2003.

[2] Stepanov, Alexei. (2015). On the Kendall Correlation Coefficient. 10.48550/ARXIV.1507.01427

[3] van der Maaten, Laurens Hinton, Geoffrey. (2008). Viualizing data using t-SNE. Journal of Machine Learning Research. 9. 2579-2605.

[4] Kornbrot, Diana. ”Point biserial correlation.” Wiley StatsRef: Statistics Reference Online (2014).

[5] Liu, Y., Wang, Y., Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. In: Liu, B., Ma, M., Chang, J. (eds) Information Computing and Applications. ICICA 2012. Lecture Notes in Computer Science, vol 7473. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32

[6] Wilson, Jeffrey Lorenz, Kent. (2015). Weighted Logistic Regression Model. 10.1007/978-3-319-23805-0_5.

[7] Zhang P, Jia Y, Shang Y. Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks. 2022;18(6). doi:10.1177/15501329221106935

[8] Maind, Sonali B., and Priyanka Wankar. ”Research paper on basic of artificial neural network.” International Journal on Recent and Innovation Trends in Computing and Communication 2.1 (2014): 96-100.

Contribution Table:

Task and Weighting(10%)	Data preprocessing(20%)	Feature Selection(30%)	ML method development(10%)	Method Evaluation(10%	Report Writing(30%)
Chaitanya Manem	30%	10%	20%	20%	20%
Varre Rohit	20%	20%	30%	10%	20%
John Paul Reddy Gopidi	20%	30%	10%	20%	20%
Dharanija Bantu	20%	20%	20%	20%	20%
Purva Soni	10%	20%	20%	30%	20%