PREDICTIVE MODELS FOR PATIENT'S TREATMENT DECISIONS USING PCR AND RFS IN BREAST CANCER

Kevein Samuel Charles
University of Nottingham

Nikhil Raj Ravi Singh
University of Nottingham

Bhavani Pattanam Suresh
University of Nottingham

Afriya Shaffa Seyadu Ali Navas
University of Nottingham

Godwin Ferin Dennis Rajan
University of Nottingham

ABSTRACT

Breast cancer has become one of the most common types of cancer in recent year and tend to have a high mortality rate. Any development for early detection and diagnosis of cancer cells and its type can reduce the death rate and give a good outline for the patients to go further with treatments possible and its outcomes. Here the objective of this review is to predict the *pathological complete response (PCR)* and *relapse-free survival (RFS)* of chemotherapy one of the treatments used for killing the tumorous cells. We have applied machine learning classification algorithms to find PCR and have compared outputs of five different regression models to come to finite solution on RFS value.

Index Terms— Machine learning, Breast cancer detection, chemotherapy effectiveness, Random Forest model, Linear Regression Model, Decision Tree model, SVR model, Neural Network Model.

INTRODUCTION

In this busy modern lifestyle, people around us are affected by one or more types of diseases. Cancer can be illustrated as the rapid and abnormal growth of the cells in certain part which varies in its characteristics and function from a normal cell due to some defects in genetics and epigenetics. Most common types of cancer are Skin cancer, Breast cancer, Lung cancer, Prostate cancer, Colorectal cancer, Melanoma, Bladder cancer etc. among those the most common diagnosed cancer is breast cancer. Breast cancer affected around 300000 people just in US. The estimated death by breast cancer comes around 43700 [1]. Breast cancer is originated through malignant tumor cells, when it growth gets out of control and divides rapidly accumulating to form a lump. The traditional method to predict cancer involves the data of clinical trial and any one of different methods to image the internal such as Breast ultrasound, Diagnostic mammogram, Breast magnetic resonance imaging (MRI), Biopsy. The Machine learning process to predict the cancer involves steps such as pre-processing, feature selection or extraction and classification to find PCR and Regression to find the RFS value. For this prediction model the process starts from the data acquisition. The data used in this consist of a simplified version of clinical outcomes given by the ACRIN 6698/I-SPY2 Breast DWI (ACRIN-6698) [2] and 107MRI-based features. The image-based features were extracted from the tumor region of MRI using a radiomics feature extraction package known as Pyradiomics [3]. The outcome of the machine learning model should be a good overview to suggest the patient if they can go on with the painful chemotherapy and to show its effectiveness and the life expectance after the treatment once the therapy is done.

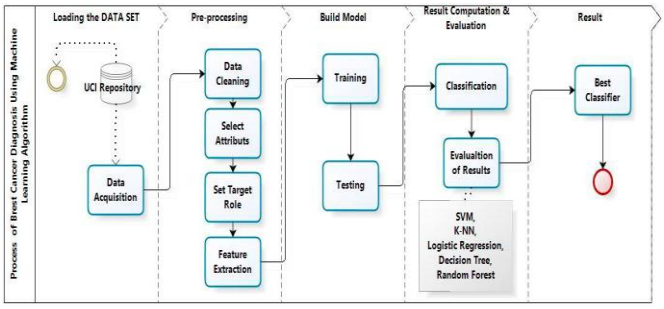

Figure 1: Flow of classification model [4]

I) PREPROCESSING:

1. Mode for filling the empty datasets.

Mode imputation is simple method for filling missing cells with mode which is the most frequently occurring value in the column. The general mathematical expression for a mode is.

$$\text{Mode} = L + (f1 - f0) / (2f1 - f0 - f2) \times h$$

Whereas L = the lower limit of the modal class (the interval with the highest frequency).

$f1$ = is the frequency of the modal class.
$f0$ = is the frequency of the class before the modal class.
$f2$ = is the frequency of the class after the modal class.
h = is the class width.

2. Splitting clinical data and MRI images data

To reduce the outliers and to increase the accuracy rate the predictions the dataset was split into two and Normalization, PCA is done separately for each split and the separated dataset was merged, and Lasso was employed to the merged dataset before fitting it into the model.

3. Normalization

Normalization is a preprocessing technique used to scale and standardize the features of a dataset. The goal of normalization is bring down the values of a data set to a standard range or distribution. This is particularly important for input features that are prone to scaling problem. There are different approaches to normalize the data that are done automatically by the library call from sklearn.preprocessing and the import file import StandardScaler. The several common normalization function each with its own purpose are, Min-Max Scaling (Min-Max Normalization), Z-Score Normalization (Standardization), Robust Scaling Unit, Vector Transformation (Vector Normalization), Log Transformation, Power Transformation.

II) DIMENSIONALITY REDUCTION:

1. PCA (Principal Component Analysis)

PCA is one of the models mainly used for dimensionality reduction. This transforms the data of the original feature set into an uncorrelated feature called the principal components. Some basic terminologies involved in PCA are variance, covariance, standardizing data, covariance matrix, Eigen values and Eigen vectors [5]. The dimensions are reduced to 9 clinical features and 15 MRI features. Selected by visualizing the data and maintaining the variance ratio above 95 percentage.

2. Merging both data clinical and MRI Datasets

After dimensionality reduction process, the datasets are concatenated together using numpy package. This merged dataset is used for the further feature selection process.

3. Lasso (Least Absolute Shrinkage and Selection Operator)

Lasso is a regularization technique used in models to select the best features among the dataset. It is useful when dealing with large datasets. Lasso introduces a penalty term to the ordinary least squares (OLS) loss function, encouraging sparse solutions by shrinking some of the coefficient estimates toward zero.

Lasso function can be represented as,

$$\text{minimize}\left(\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\alpha\sum_{j=1}^{p}|\beta_j|\right)$$

Here,
n is the number of samples.
p is the number of features.
Y_i is the target variable for the i-th sample.
X_{ij} is the j-th feature of the i-th sample.
β_0 is the intercept term.
β_j are the coefficients associated with each feature.
α is the regularization parameter. Increasing α leads to more regularization, resulting in more coefficients being exactly zero.

III) MACHINE LEARNING ALGORITHMS FOR BREAST CANCER PREDICTION:

There are different algorithms available for prediction of cancer using clinical and other image-based datasets. Some of such machine learning models are Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree, K-Nearest Neighbors etc. Classification model was used for predicting pathological complete response (PCR) and for the predicting the relapse-free survival (RFS) the regression algorithms were implemented. Different models were implemented namely, linear regression, decision tree, random forest, support vector machine, Neural network and their performance were evaluated.

PREDICTIVE MODEL FOR RELAPSE-FREE SURVIVAL IN BREAST CANCER PATIENTS: A REGRESSION ANALYSIS

Model Architecture:

Number of hidden layers: The regression model has one hidden layer.

Neurons in the hidden layer: The hidden layer consists of 100 neurons, which indicates the complexity and capacity of this MLP model to capture complicated patterns in the data.

Activation Function: The activation function that has been used in the hidden layer is ReLu(Rectified Linear Unit). ReLu is commonly used in Neural Networks for its ability to observe the non-linearity in the model, which allows it to learn complex relationships in the data.

Solver: The solver which has been used is the Adam optimizer. The chosen architecture with a single hidden layer of 100 neurons and ReLu activation, along with the Adam optimizer, suggests a well-balanced configuration for capturing patterns in the data.

Training Procedure:

Learning rate: The learning rate determines the size of the steps taken during the optimization process. A larger learning rate means larger steps, and vice-versa. The initial learning rate used in the model was 0.001.

Batch size: The training data was processed in batches during each iteration. The batch size represents the number of samples processed together plays an important role in the model training process. For the MLP model used the "batch size = auto", which means the model determines an appropriate batch size based on the size of the training data.

Number of Epochs: The model underwent multiple iterations, commonly referred to as epochs. In the MLP model used to predict the RFS, the "max_iter" is set to 1500. This parameter specifies the maximum number if iterations or epochs.

Hyperparameter Tuning: The process of systematically adjusting the configuration setting of a model to optimize its performance.

Hyperparameter Grid Search: The MLP model used Grid SearchCV to explore various combinations of hyperparameters, such as hidden layer sizes, maximum iterations, and alpha values, to find optimal configuration for the MLP Regressor model.

Best Hyperparameter: The best hyperparameter that was found through the grid search process, was used in the MLP model resulting in a set of value's the maximised the model's performance. The best hyperparameters are mentioned in Table 1.

alpha value	hidden layer sizes	Max_iterations
0.0001	100	1500

Table 1

Reduced Mean Absolute Error: Successfully decreased the mean absolute error to 3.86 from 4.238(before hyperparameter tuning), showcasing a more precise prediction of the target variable and suggesting the effectiveness of the chosen hyperparameter values.

Performance Metrics: The evaluation metrics which were used in evaluating the MLP model were mean squared error, R squared value and mean absolute error.

Mean Squared Error: Recorded a relatively low MSE of 24.39, indication that on average, the squared difference between the predicted and actual values are minimised.

High R squared value: Achieved a high R squared value of 0.973, signifying that approximately 97.3% of the variance in the target variable is explained by the model.

Optimal Mean Absolute Error: Attained a low MAE of 3.86, which makes the MLP model's accurate predictions with an average absolute difference of 3.86 between predicted and actual values.

Precision in Prediction: The relatively low MAE highlights the precision of the MLP model in predicting the target variable, providing confidence in the accuracy of the predicted values for the given dataset.

PREDICTIVE MODEL FOR PATHOLOGICAL COMPLETE RESPONSE IN BREAST CANCER PATIENTS: A CLASSIFICATION ANALYSIS

Model Architecture of Random Forest Classification:

The fundamental building blocks of a Random Forest are decision trees. Each tree is constructed based on a subset of the training data, selected through bootstrapped sampling.

Bootstrapped Sampling: Random Forest employs bootstrapped sampling to create diverse datasets for training each decision tree. This involves randomly selecting samples from the original dataset with replacement. As a result, some instances may be repeated, and others may be left out in each subset.

Feature Selection: At each node of a decision tree, a random subset of features is considered for splitting. The number of features to consider at each split is a hyperparameter that can be tuned. This feature randomization helps to decorrelate the trees and make the ensemble more robust.

Hyperparameters: Random Forest has hyperparameters that can be tuned to control its behavior. Some key hyperparameters include Number of Trees, Maximum Depth, Minimum Samples Split, Maximum Features.

Ensemble Learning: In classification, each tree "votes" for a class, and the class with the most votes becomes the final prediction.

Training Procedure:

Hyperparameter Tuning: Grid Search CV performs a search over all possible combinations of the parameters such as depth, features, number of leaf nodes, estimators and then return the best hyper parameters based on cross-validated performance. The parameter obtained from hyperparameter tuning are listed below.

Max_Depth: The maximum depth of the individual trees in the forest can be found out using this parameter. The grid span value ranges from 0 to 14.

Max_Features: This parameter determines the maximum number of features considered for splitting a node. The grid includes 'sqrt', 'log2', and None. 'sqrt' corresponds to the square root of the total number of features, 'log2' corresponds to the base-2 logarithm of the total number of features, and None means all features are considered.

Max_Leaf-Nodes: This parameter restricts the number of leaf nodes in each tree. The least number of leaf node can be from 0 till the maximum number 14.

N_Estimators: This parameter controls the number of trees in the forest where the grid uses the values of 25, 50, 100, and 150.

Best Hyperparameter The optimal hyperparameter, identified through the grid search process, constitutes a set of values that maximized the model's performance. The discovered best hyperparameters are mentioned in Table 2.

Max_Depth	Max_Features	Max_Leaf-Nodes	N_Estimators
14	log2	13	25

Table 2

Balanced Accuracy Score The evaluation metric consider for this particular model is balanced accuracy score, which ranges from 0 to 1. The model predicted with the balanced accuracy score of 0.90, demonstrating a more accurate prediction of the target variable.

Results:

The models used had the ability to handle complex and non-linear data, and outperformed the other models, which helped in predicting values for PCR and RFS.

From the Figure 2, it is evident that Linear Regression model has the lowest MAE, but due to the simplicity of the model which became a limitation when dealing with complex data, leading to signs of overfitting. Given the data involves complex patters and dependencies that cannot be captured by simpler models, MLP can be more suitable. The dataset comprised of clinical values as well as MRI images which were converted into numerical values, which showed complex non-linear relationships in the data, where MLP outperformed other models.

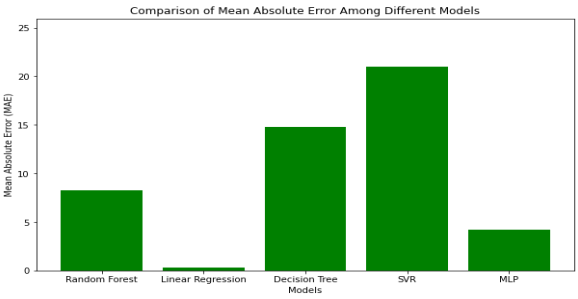

Figure 2 Comparision of Model's MAE

Figure 3 represents the relationship between actual values and the values predicted from the MLP model. The red dashed line serves as a reference line where the data points which have perfect alignment with the line would indicate precise predictions. Deviations of the data points from the red dashed line indicate the variance between the actual and predicted values, highlighting where the model either overpredicts or underpredicts.

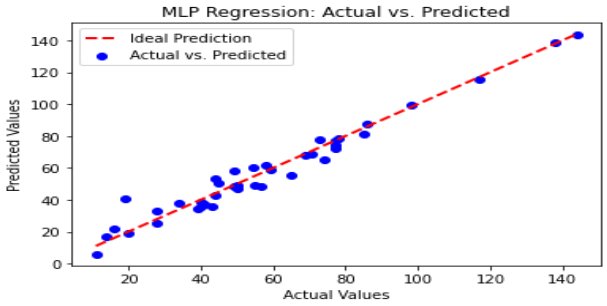

Figure 3 Actual vs Predicted values

Due to the intricate patterns and dependencies present in the data, where simpler models struggled to capture, the Random Forest Classifier proves to be a more fitting choice.

Conclusion:

The most common side effects cancer patients experience is fatigue, nausea and vomiting and hair fall. Additionally, Several chemotherapeutic agents are associated with long-term adverse outcomes, such as persistent neuropathy after treatment with taxanes[6]. Importantly, anthracyclines and HER–2–targeted drugs can lead to cardiomyopathy and congestive heart failure[6]. Therefore, by predicting the accurate RFS and PCR values, the medical professionals can suggest and modify the specificity of treatment individually for each patient, helping them in avoiding unnecessary cancer treatments. By providing accurate predictions of RFS values, the doctors can suggest and adjust the cancer treatment methods according to the RFS

values. Every cancer patient can benefit from this model by allowing them to take specific treatments according to their PCR values and helps in avoiding unnecessary treatments and side effects. Patients can make more informed decisions, by knowing their Relapse Free Survival days after their cancer treatment. With the help of ongoing research and collaboration with medical professionals can lead to the incorporation of additional clinical or MRI image features, which helps in refining the model's accuracy. In conclusion, the model offers hope for advancements in cancer care. The journey toward better cancer care is an ongoing one, and these model serves as a promising catalyst for future breakthroughs.

11. REFERENCES

[1] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians* 2023; 73(1):17-48. Last accessed February 21, 2023. [PubMed Abstract] https://www.cancer.gov/types/common-cancers

[2] Newitt, D. C., Partridge, S. C., Zhang, Z., Gibbs, J., Chenevert, T., Rosen, M., Bolan, P., Marques, H., Romanoff, J., Cimino, L., Joe, B. N., Umphrey, H., Ojeda-Fournier, H., Dogan, B., Oh, K. Y., Abe, H., Drukteinis, J., Esserman, L. J., & Hylton, N. M. (2021). ACRIN 6698/I-SPY2 Breast DWI [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.KK02-6D95

[3] https://pyradiomics.readthedocs.io/en/latest

[4] Flow of classification model https://www.sciencedirect.com/

[5] Premanand S - Assistant Professor Junior, Principal Component Analysis in Machine Learning | PCA in ML, https://www.analyticsvidhya.com/blog/2022/07/principal-component-analysis-beginner-friendly/

[6]American Cancer Society Journal https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21731

Task and Weighting	Data Pre-processing (10%)	Feature Selection (25%)	ML Method Development (25%)	Method Evaluation (10%)	Report Writing (30%)
Kevein Samuel Charles	20	5	35	30	5
Nikil Raj Ravi Singh	20	75	35	30	5
Bhavani Pattanam Suresh	20	10	10	5	15
Godwin Ferin Dennis Rajan	20	5	10	5	70
Afriya Shaffa Seyadu Ali Navas	20	5	10	30	5