# Comp 4139 Machine Learning 2024-25

# Machine Learning Techniques for Predicting Breast Cancer Outcomes

Lukshan Sharvaswaran
20154273

Arunavo Dutta
20698791

Shubhankar Shahade
20711210

Yalamanchili Prashanth
20701879

Veerendra Kumar dangeti
20702395

**Abstract**

Breast cancer treatment, particularly chemotherapy, often has varying responses in patients. Prior to treatment, predicting Pathological Complete Response (PCR) and Recurrence-Free Survival (RFS) can improve in patient treatment planning. This study explores advanced machine learning techniques for predicting PCR (classification) and RFS (regression) using a dataset with clinical and MRI-derived features. Various feature selection and machine learning models were implemented. The results were evaluated on balanced classification accuracy for PCR and Mean Absolute Error for RFS. In both models, the Random Forest emerged as the most effective model. This report provides a comprehensive framework for real-world applications by discussing the preprocessing strategies, feature selection methodologies, and hyperparameter tuning.

**Keywords**

Machine Learning, Breast Cancer, Pathological Complete Response, Recurrence-Free Survival, Feature Selection, Random Forest

**Introduction**

Breast Cancer is currently one of the most diagnosed cancers among women across the globe with around 2.3 million new cases reported recently. Breast Cancer stands as a major contributor of cancer related

deaths amongst women, with nearly about 685000 cases of fatality globally.

As in UK, each year approximately 56,000 women are diagnosed with invasive breast cancer. The lifetime risk for women developing cancer is estimated to be 1 out of every 7 women. Even-though advancements in technologies leading to early diagnosis and treatment have significantly brought up positive outcomes backed by the fact that for the past 5 years the survival rate of breast cancer (which hasn't spread beyond the breast tissue) is around 99%.

Thus, Machine Learning techniques particularly classification and regression methods are crucial in advancing breast cancer diagnosis, prognosis as well as treatment planning. Classification method enables to predict categorical outcomes such as differentiating between malignant or benign tumour, identification of potential patients who could receive PCR and chemotherapy and categorize them into potential risk categories. Regression methods enable the prediction of continuous outcomes which are Recurrence Free Survival (RFS), tumour growth rate and potential treatment responses. Integration of clinically measured features like patient demographics with MRI Scans results provide a clinician with deeper insights to tumour biology and behaviour which indeed has enabled in improved early detection and optimized follow-up strategies. For our task, applying classification and regression to predict or categorize data is labelled into various groups on features based on the data[1].

## Methods

### Dataset

Based on the public dataset from The American College of Radiology Imaging Network (I-SPY 2 TRIAL), a simplified dataset is generated. Each patient in this dataset contains 11 clinical features (Age, ER, PgG, HER2, TrippleNegative Status, Chemotherapy Grade, Tumour Proliferation, Histology Type, Lymph node Status, Tumour Stage and Gene) and 107 MRI-based features. The image-based features were extracted from the tumour region of MRIs using a radiomics feature extraction package known as Pyradiomics. In our dataset, we have distinctive ID of 400 patients with the ten characteristics about their age and health and 117 features extracted from the images.

### Feature Selections

Feature selection is very important step in enhancing the model performance and helps us in avoid the overfitting, unnecessary information and improving the accuracy. feature selection helps us in reducing the dataset complexity which helps us in saving the memory usage and becomes easier to handle the data. Irrelevant or redundant features can degrade model performance and increase computational costs. Correlation feature selection helps us in analysing the statistical relationships between features and the target variable. This method is efficient in feature selection Least Absolute Shrinkage and Selection Operator is a regression-based feature selection technique that extends the loss function with an L1 regularization penalty. This has the effect of shrinking some of the feature coefficients exactly to zero and hence selecting only the most important features. In Regression and classification analysis, this paper presents a comparative analysis

of four widely used feature selection and dimensionality reduction methods: Correlation Filter, Tree-Based Filter, Lasso filter, and Principal Component Analysis (PCA),variance thresholds. The aim is to identify the most effective techniques in terms of predictive accuracy and efficiency.

**Model selection and evaluation**

For classification we are using Decision Tree Classifier, Random Forest, Gradient Boosting classifier and MLP classifier. For regression, Predictive models include Linear Regression, Random Forest, Lasso Regression, SVR, and MLP Regressor. Each method was evaluated using k-fold cross-validation (k=5) and MAE as the primary metric. Normalized MAE metrics were also computed to assess the impact relative to target distribution statistics.

**Classification**

**Decision Tree Classification**

Because of the inherent learning algorithm and the classification property, the perceptron, which is the most basic model among the different artificial neural nets, has been influencing and sparking study in the field historically[2]. Decision tree is supervised machine learning algorithm which is easy to interpret. This model gives us the output as a treelike structure. In our task, the parameters we are setting our random state is 42, with maximum depth, minimum samples and minimum samples split are 5,2, respectively the class weights none, balanced.

**MLP classifier**

The multilayer perceptron is a supervised learning algorithm based on the Artificial neural networks(ANN).The parameters we are using are hidden layer size of (50,100), with the maximum iteration of 500 and alpha rate of 0.1.

**Gradient Boosting Classifier**

This algorithm builds the model sequentially to reduce the prediction errors in the previous models. It combined the weaker learners to build a strong predictive model. The Parameters used in this model are after training are estimators: [50, 100, 200],max_depth: [3, 5, 10],min_samples_leaf: [1, 2],random_state: [42].

**Random Forest**

RF is an ensemble machine learning algorithm well-known for its performance on classification problems. It does this by building an ensemble of decision trees chosen at random from the training data[3].This algorithm is powerful learning algorithm. This algorithm builds a numerous decision tree which helps in increasing the accuracy and handling the overfitting. This algorithm performs well both on the small and large datasets. The parameters we are setting for this model are maximum depth value of 10, estimators 50.

**Regression**

The results of the evaluation are summarized in the table below. Random Forest demonstrated the best overall performance with the lowest MAE and RMSE, as well as a positive R2 score. Other models exhibited varying degrees of underperformance, with Linear Regression yielding the poorest results due to potential multicollinearity or inadequate preprocessing.

```
Model Comparison Results:
                Model  MAE (Mean)  RMSE (Mean)  R2 (Mean)
2       Random Forest   20.753202    26.187097   0.053682
3                 SVR   21.112245    26.771807   0.009857
4       MLP Regressor   22.222959    28.067158  -0.100765
1    Lasso Regression   22.711725    29.478284  -0.240195
0   Linear Regression   32.564638    96.777759 -21.033748
```

Table: Regression Model comparison Results

## Discussion

### Classification

For our task the best algorithm for classification is random forest classifier. This model yielded us the better results for our task. The results of this models outperformed other models making us to choose this algorithm from others

### Advantages of choosing Random Forest

Random forest model is best in helping reduce the overfitting. Our algorithm can handle non-linear relationships between the features. It handles the missing data and imbalances with good accuracy. In this algorithm the combining prediction from multiple trees helps us in reducing the noise.

### Disadvantages of choosing Random Forest

The disadvantages of using random forest classification are it requires a lot of time and memory for training and prediction, generally with bigger datasets or with high number of trees. There is a high chance of overfitting if the parameters are not set properly like number of trees or depth of the trees. Noisy dataset can also lead to overfitting. There is a chance that random forest classifier struggles with smaller datasets. Despite these flaws, Random Forest classifier outperforms other models in our task.

### Regression

For regression modelling, the Random Forest achieved the best performance, being robust to the feature selection methods implemented and suitable for datasets with complex interactions. While SVR was close to the Random Forest, it was slightly less effective in capturing variance. MLP Regressor consisted of moderate errors that were most likely due to overfitting. Lasso regression struggled to capture relationships effectively, potentially due to insufficient regularization strength. Lastly, the linear regression model performed poorly due to multicollinearity and high sensitivity to irrelevant features.

## Conclusion

After classifying our task with our dataset with different algorithm and different parameters, random forest has performed better compared to remaining 84.19%.. This study underscores the importance of selecting appropriate feature selection methods based on dataset characteristics. Fore regression model, Random Forest emerged as the most effective predictive model in our experiments. PCA, while computationally efficient, underperformed in predictive accuracy.

## Reference Links

[1]R, G., G, S. Iterative principal component analysis method for improvised classification of breast cancer disease using blood sample analysis. Med Biol Eng Comput 59, 1973–1989 (2021). https://doi.org/10.1007/s11517-021-02405-y

[2]J. Singh and R. Banerjee, "A Study on Single and Multi-layer Perceptron Neural Network," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 35-40, doi: 10.1109/ICCMC.2019.8819775.

[3]Shao, Z.; Ahmad, M.N.; Javed, A. Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface. Remote Sens. 2024, 16, 665. https://doi.org/10.3390/rs16040665

| Task and Weighting | Data pre-processing (10%) | Feature Selection (25%) | ML method Development (25%) | Method Evaluation (10%) | Report Writing (30%) |
|---|---|---|---|---|---|
| Prashanth Yalamanchili | 20% | 20% | 15% | 20% | 20% |
| Lukshan Sharvaswaran | 20% | 20% | 40% | 20% | 20% |
| Shubhankar Shahade | 20% | 20% | 15% | 20% | 20% |
| Arunavo Dutta | 20% | 20% | 15% | 20% | 20% |
| Veerendra Kumar dangeti | 20% | 20% | 15% | 20% | 20% |