Machine Learning for Disease Treatment Response Prediction

1st Nai-Jui Yeh
School of Computer Science
University of Nottingham
Nottingham, UK
psxny2@nottingham.ac.uk

2nd Chenyang Jin
School of Computer Science
University of Nottingham
Nottingham, UK
psxcj9@nottingham.ac.uk

3rd Raonak Shukla
School of Computer Science
University of Nottingham
Nottingham, UK
psxrs15@nottingham.ac.uk

4th Kumar Harsh
School of Computer Science
University of Nottingham
Nottingham, UK
psxkh2@nottingham.ac.uk

5th Palak Bajaj
School of Computer Science
University of Nottingham
Nottingham, UK
psxpb13@nottingham.ac.uk

***Abstract*—Breast Cancer is the most frequent kind of cancer a large number of women are affected by it. Any development in early prediction can significantly affect the treatment aspect and survival of the patient. The main objective of this paper is to use different machine learning algorithms like Logistic Regression, Gradient Boost, and Random Forest to classify patients and estimate their survival time. The steps taken in the study are data preparation, feature selection, dimensionality reduction, tuning of the base model, and ensemble learning. For classification, our ensemble model got a balanced classification accuracy of 0.681 and for regression a mean absolute error for relapse-free survival time of around 12.637 weeks.**

***Index Terms*—Machine Learning, Feature Selection, Dimensionality reduction, Ensemble Learning, Classification, Regression.**

I. Introduction

Chemotherapy is a commonly used treatment to reduce tumor size before surgery. However, its effectiveness varies among patients, and it can lead to serious side effects. To provide the most appropriate treatment for patients, we developed a machine learning system to predict pathological complete response (PCR) and relapse-free survival (RFS). This information will then be utilized to determine the suitability of chemotherapy for each patient.

II. Method

A. Feature Selection and Dimentionality Reduction

1) Filter Method: We employed Chi2 [1], ANOVA [2], and correlation analyses [3] to conduct statistical analysis between various pairs of features. Chi2 was applied to categorical-to-categorical feature pairs, ANOVA to categorical-to-continuous pairs, and correlation analysis to continuous-to-continuous pairs. Subsequently, f-statistics and p-values were calculated for each statistical analysis, providing insights into the degree of correlation between the features.

By setting the p-value to 0.05, 19 features were chosen for the regression task, including 1 clinical (categorical) feature and 18 image (continuous) features. For the classification task, 42 features were selected, comprising 4 clinical features and 38 image features.

2) Embedded Method: Random forest [4] is used for feature selection. In this method, each tree of the random forest can calculate the importance of a feature according to its ability to increase the pureness of the leaves. The higher the increment in leaves purity, the higher the importance of the feature. Once we have the importance of each feature, we perform feature selection using a procedure called Recursive Feature Elimination. Using this method, 39 image features are selected and one non-image feature 'age' is selected for regression.

3) PCA: PCA (Principal Component Analysis) [5] is a dimensionality reduction technique that's beneficial when dealing with high-dimensional data. The dataset contains 107 MRI features, which might lead to the curse of dimensionality. PCA helps in reducing these dimensions by creating linear combinations of the original features. This reduction aids in computational efficiency and helps mitigating issues like overfitting. Then there is standardizing the data after PCA transformation. This ensures that the transformed features are on the same scale. The output of pca.explained_variance_ratio_ shows variance explained by each principal component. These values indicate how much information each principal component retains from the original data. It helps in deciding the number of components to keep while still preserving a significant portion of the data's variance. For instance, PCA_1 shows 32.6%, followed by PCA_2 with 14.5%.

B. Ensemble Learning

Three ensemble learning techniques [6], namely Voting, Bagging, and Boosting, are employed to combine multiple machine learning models, aiming to achieve better performance and robustness.

1) Voting: Voting is a technique that aggregates the results from multiple models trained with the same dataset. There are two different aggregation methods. One is to take the average result of all models, known as soft voting. The other involves taking the majority result from all candidates, referred to as hard voting. Soft voting can be used in both classification and regression tasks, whereas hard voting is applicable only in classification tasks.

2) Bagging: Bagging is similar to voting, with all models trained using different sample data selected by bootstrapping. Moreover, models in bagging are identical, while voting combines different models. The results from each model are averaged to form the final result.

3) Boosting: The aim for boosting is to create a strong learner from multiple weak learners. Instead of training identical models individually, it accumulates the importance of challenging observations that previous models struggled with. This process compels the current model to work more diligently on the most difficult data encountered so far.

C. Resampling

Upon analyzing the original dataset, we observed an imbalance in the output for classification. The negative result, represented by '0,' consists of 316 data points, while the positive result, represented by '1,' consists of 84 data points. Building a machine learning model to predict accuracy with such an imbalanced dataset may yield ineffective results, as the model could be biased towards the majority class, potentially misclassifying the minority class.

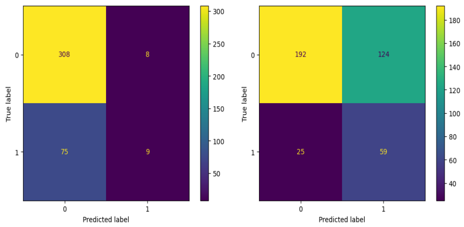
Resampling data is a commonly preferred approach to address imbalanced datasets, with two main methods: 1) undersampling and 2) oversampling. To tackle the imbalance issue, we are exploring two methods: the Synthetic Minority Oversampling Technique (SMOTE) and random undersampling.

1) Synthetic Minority Oversampling Technique(SMOTE): SMOTE [7] is an oversampling technique that generates synthetic samples for the minority class, effectively minimizing the impact of imbalanced data.

2) Random under-sampling: Random under-sampling involves randomly deleting samples from the majority class to balance the overall data distribution.

To evaluate the effectiveness of these methods,we utilize both the F1 score and balanced accuracy to compare the results between the default imbalanced dataset and the resampled dataset. we apply logistic regression model for the experiment. With the default imbalanced dataset, the F1 score is 0.11, and the balanced accuracy is 0.52. Subsequently, using SMOTE and random under-sampling improves the F1 score to 0.45 and the balanced accuracy is 0.65, indicating that this approach is better for training the model on the imbalanced dataset.

In addition, we apply a confusion matrix in Fig1 to visualize the differences between the default imbalanced dataset and the resampled dataset. The dataset after resampling demonstrates a more balanced distribution compared to the default imbalanced dataset.



(a) original dataset (b) after resampling

Fig. 1. confusion matrix compareing the distriution of original data and after resampling

III. Evaluation

We evaluated our methods using a simplified public dataset obtained from The American College of Radiology Imaging Network[1]. The dataset comprises 117 features, including 10 clinical (categorical) features and 107 image (continuous) features, across 400 entries. There are two target columns, namely PCR and RFS. PCR serves as the binary categorical target for a categorical task, while RFS is the numeric target for a regression task.

For the regression task, our methods were assessed through 10-fold cross-validation, involving training with 360 data points and validating with 40 data points. As described in Section II-C, the classification task involves unbalanced labels, which are addressed by employing resampling methods—specifically, SMOTE and random undersampling. The dataset size is increased to 568, with 284 data points for each label. Similar to the regression task, we applied 10-fold cross-validation (511 training data and 57 validating data) to evaluate the classification methods. Regression methods are evaluated by Mean Absolute Error (MAE), and classification methods are assessed using balanced classification accuracy.

In the following sections, three experiments were conducted. Firstly, various feature selection and dimensionality reduction methods (as mentioned in Section II-A), along with their combinations, were compared to address our data. After determining the most effective approach for reducing features, we implemented multiple machine learning models using ensemble learning methods mentioned in Section II-B. Following a thorough evaluation of each model, the one with the highest accuracy (or MAE for regression) was selected as our final model.

A. Feature Selection and Dimentionality Reduction

The aim of this section is to identify the optimal mechanism for feature reduction. A total of five methods were experimented with and compared against the baseline, where no feature reduction was applied. The five feature reduction methods considered are as follows:

- Filter Method (see Section II-A1)
- Embedded Method (see Section II-A2)

[1] I-SPY 2 TRIAL: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=50135447

- PCA (see Section II-A3)
- PCA with Filter Method

For PCA with Filter methods, PCA is conducted before applying the filter method.

All the mentioned methods, including the baseline, were examined using models with default settings. The models employed for both regression and classification tasks are listed below:

- **Regression**: Linear Regression [8], LASSO [9], Random Forest [4], Bayesian Ridge [10], ElasticNet [9], Gradient Boosting [11], Stochastic Gradient Descent [12], Support Vector Machine [13], and Ridge Regression [14]
- **Classification**: Logistic Regression [15], Support Vector Machine [13], Naive Bayes [16], XGBoost [17], Decision Tree [18], K-nearest-neighbor [19], Random Forest [4], Gradient Boosting [11], Artificial Neural Network [20]

The results for the regression (RFS) and categorical (PCR) tasks are presented in Table I and II respectively. Table I is sorted in ascending order by MAE for the top 5 models, while Table II is sorted in descending order by balanced classification accuracy for the top 5 models. The best result for each task is highlighted in bold.

For the regression task, the Embedded Method demonstrated the highest performance, achieving an MAE of 21.110 for the top 5 models and 21.060 for the top 3 models (see Table I). Its top 5 models include Random Forest [4], Gradient Boosting [11], ElasticNet [9], Stochastic Gradient Descent [12], and Lasso [9]. These models were trained with 39 continuous features, including age and other image features. Based on these results, the Embedded Method, specifically employing Random Forest feature selection, will be utilized in the experiment in Section III-B.

As shown in Table II, the best-performing method for classification task is PCA with Filter Method, achieving a balanced accuracy of up to 0.627 for the top 5 models and 0.638 for the top 3 models. The top 5 models chosen are Naive Bayes [16], Logistic Regression [15], K-nearest-neighbor [19], Gradient Boosting [11], and Support Vector Machine [13]. They were trained on 6 selected features, namely HER2, PgR, ER, Principal Component 1, Proliferation, and LNStatus. Consequently, PCA with Filter Method will be utilized in the experiment in Section III-C.

TABLE I
EVALUATION OF TOP 5 AND TOP 3 REGRESSION MODELS' MEAN ACCURACY FOR EACH FEATURE REDUCTION METHODS

Method	Top 5 MAE	Top 3 MAE
Embedded Method	**21.110**	**21.060**
PCA with Filter Method	21.159	21.140
Filter Method	21.204	21.106
Baseline	21.306	21.125
PCA	21.331	21.237

TABLE II
EVALUATION OF TOP 5 AND TOP 3 CLASSIFICATION MODELS' MEAN BALANCED ACCURACY FOR EACH FEATURE REDUCTION METHODS

Method	Top 5 Acc.	Top 3 Acc.
PCA with Filter Method	**0.627**	**0.638**
PCA	0.617	0.636
Filter Method	0.605	0.623
Embedded Method	0.561	0.565
Baseline	0.535	0.535

TABLE III
PARAMETER SETTINGS FOR TOP 3 REGRESSION MODELS

Model	Hyper parameters
Random Forest Regressor	ccp alpha: 0.01, criterion: friedman mse, max depth: 20
Gradient Boost Regressor	alpha:0.2, ccp alpha:0.01, loss: squared error, criterion: squared error, learning rate: 0.02,
Lasso	alpha:0.1

B. Relapse-Free Survival Estimation

During the evaluation of baseline models, it is evident that the Random Forest Regressor performed the best and Lasso being the last. After applying ensemble method, the mean absolute error of Random Forest Regressor is the least indicating it to be the top model to predict RFS followed by Gradient Boost Regressor. While the MAE for models like Ridge Regressor, ElasticNet is high. All the models were tuned with the [Table III] hyper parameters in order to achieve a better MAE. After bagging, only few models performed better. Following are the models with top 5 MAE value:

- Tuned Random Forest Regressor after bagging
- Tuned Gradient Boost Regressor
- Tuned Gradient Boost Regressor after bagging
- Lasso after bagging
- Random Forest after bagging

In addition, soft voting has been chosen for the final model by combining the top 5 models. Following are those top 5 models after soft voting:

- Tuned Random Forest Regressor after bagging
- Tuned Gradient Boost Regressor
- Tuned Gradient Boost Regressor after bagging
- Random Forest after bagging
- Lasso after bagging

TABLE IV
EVALUATION OF THE MEAN ABSOLUTE ERROR FOR BOTH TUNED MODELS AND TUNED MODELS WITH BAGGING

Model	Tuned	Bagging
Random Forest Regressor	20.973 ±1.651	20.563 ±1.817
Gradient Boost Regressor	20.675 ±2.002	20.783 ±1.751
Lasso	20.886 ±1.812	20.861 ±1.336

C. Pathological Complete Response Classification

Upon the evaluation of the various classification models, Logistic Regression, SVM, Artificial Neural Network, Decision Tree, Gradient Boosting, KNN, Random Forest, and Naive

TABLE V

Evaluation of the Mean Absolute Error among the best model without voting, and the result of soft voting

model	Mean Absolute Error
best model (Random Forest)	20.563 ±1.817
Soft voting	12.637

Bayes, the top-performing ones were selected based on their balanced classification accuracy. These included Logistic Regression, Random Forest, XGBoost, Artificial Neural Network, and Naive Bayes. GridSearch was used to search for the best hyperparameters, optimizing parameters like the number of estimators, learning rate, regularization parameters, etc.

Subsquently, the Bagging technique was employed on Logistic Regression, Random Forest, XGBoost, Artificial Neural Network, and Naive Bayes models to potentially improve their performance. To further enhance model performance, the two types of ensemble voting methods: hard voting and soft voting were employed.

After applying SMOTE and Under-sampling, different models with best scores were selected along with their best suited hyper parameters (see Table VI). After comprehensive experimentation and evaluation of various ensemble methods, soft voting emerged as the optimal classification model for the dataset.

TABLE VI

Parameter Settings for Top 5 Classification Models

Model	Hyper parameters
Logistic Regression	C: 1, penalty: l1, solver: liblinear, tol: 0.000001
Random Forest	max depth: 9, min samples leaf: 10, n estimators: 10
XGB	booster: gblinear, learning rate: 0.01, n estimators: 100
MLP	alpha: 0.05, hidden layer sizes: (50, 50,50), learning rate: constant, solver: adam
Naive Bayes	var smoothing: 1e-08

After tuning parameters for all the models, we use the top five models as mentioned in Table VII. Subsequently, we apply bagging to each of these models. Table VII represent the balanced accuracy for both tuned models and the tuned models with bagging.

TABLE VII

Evaluation of the balanced accuracy for both tuned models and tuned models with bagging

Model	Tuned	Bagging
Logistic regression	0.690 ±0.095	0.663 ±0.098
Random forest	0.675 ±0.090	0.616 ±0.073
XGBoost	0.669 ±0.082	0.652 ±0.092
Artificial Neural Network	0.662 ±0.077	0.665 ±0.110
Naïve Bayes Gaussian	0.658 ±0.092	0.647 ±0.094

After bagging, we can see most of the models do not perform better than before. With the working principle of bagging methods, we choose different data from the whole dataset by bootstrapping so that some data maybe choose repeatedly, or else some data maybe not be chosen. Then we select the top 5 models based on balanced accuracy, the top 5 models are:

- Logistic regression
- Random forest
- XGBoost
- Artificial Neural Network after bagging
- Logistic regression after bagging

We apply soft voting and hard voting for the above 5 models. The best model among the tuned models are displayed in Table VIII, and the result of soft voting and hard voting. We compare these outputs and then choose soft voting as our final model. The reason is that soft voting considers the probability scores of each model for every class, subsequently calculating a weighted average of these probabilities to formulate the final prediction. By combining the top 5 models, soft voting can generalize better and enhance the overall performance.

TABLE VIII

Evaluation of the balanced accuracy among the best model without voting, and the result of soft voting and hard voting

model	balanced accuracy
best model (Logistic regrsssion)	0.690 ±0.095
Soft voting	0.687
Hard voting	0.681

IV. Conclusions

In this study, a comprehensive analysis of feature selection, dimensionality reduction, ensemble learning, and resampling techniques was conducted to enhance the performance of machine learning models for both regression and classification tasks. In regression tasks, Random Forest feature selection along with soft voting outperformed other methods, achieving the lowest Mean Absolute Error (MAE) of 12.637. For the classification task, Soft Voting, combining predictions from multiple models, demonstrated the highest balanced accuracy of 0.687.

As future scope, further exploration of hyper parameter tuning for ensemble methods and finding the impact of different algorithms within ensemble structures. Exploration of additional resampling techniques and evaluation of their effectiveness in addressing class imbalance.

References

[1] Mary L McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.

[2] Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.

[3] Mordecai Ezekiel. Methods of correlation analysis. 1930.

[4] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[5] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical methods*, 6(9):2812–2831, 2014.

[6] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[8] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[9] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[10] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[11] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[12] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

[13] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[14] Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[15] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[16] Harry Zhang. The optimality of naive bayes. *Aa*, 1(2):3, 2004.

[17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[18] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[19] Jigang Wang, Predrag Neskovic, and Leon N Cooper. Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters*, 28(2):207–213, 2007.

[20] Imad A Basheer and Maha Hajmeer. Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*, 43(1):3–31, 2000.