A Comparative Study of Machine Learning Approach for Predicting Breast Cancer Progression using Multi-Omic and Clinical Data from the Cancer Genome Atlas Database

StduentNo : 2332635

June 7, 2023

Abstract

The research aims to predict the progression of breast cancer using clinical information and multi-omic data. To compare different Machine Learning models built on combinations of different omics and clinical features. (8 models built and trained on 5 ML approaches, KNN, SVM, Random Forest, Logistic Regression, and Elastic Net). The Cancer Genome Atlas Program database will be used. The research will employ median imputations to handle missing data and feature selection for each model will be done using the Boruta feature selection method. The models will be trained using the training dataset, and the predictive power of each model will be obtained by testing using the testing dataset. Each machine learning approach will be compared on the basis of prediction accuracy to find the best approach. This research will contribute to the existing knowledge of breast cancer progression. The findings could be used to develop personalised treatment plans for breast cancer patients and a more accurate prediction of cancer progression.

1 Introduction

Breast cancer is one of the most common cancers and the second leading cause of cancer deaths among women worldwide [1]. It is a disease in which cells in the breast grow out of control [2]. Although surgery is effective in removing the tumour containing the cancer cells, the complex nature of cancer progression is challenging [3]. Understanding the factors that may influence cancer recurrence post-treatment is crucial for improving treatment outcomes [4].

This study employs supervised learning techniques to investigate the progression of breast cancer using clinical information and multi-omics data obtained from patients as a part of The Cancer Genome Atlas Program. The study will focus on comparing the accuracy and predictive power of different machine learning approaches used for prognosis. The Cancer Genome Atlas Program database contains data of over 20,000 primary cancer and matched normal samples spanning 33 cancer types collected from the year 2006 onwards [5]. The cancer samples considered by TCGA included those with poor prognosis and primary, untreated tumours with an available source of matched normal tissue or blood sample [5]. This study uses a subset of breast cancer data from the original study group.

In this study, eight combinations of datasets are developed with different omics and clinical features. The omics measurements include gene expression, microRNA expression, protein abundance, DNA methylation, and genetic mutation counts. Each combination of datasets is pre-processed, imputed to handle missingness, cantered and scaled for better prediction, and relevant features are sorted using feature selection techniques. The data is further transformed using principal component analysis to handle the remaining multicollinearity in the data. The research employs median imputations to handle missing data. Feature selection for the models is done using the Boruta feature selection method which is a wrapper method built around the random forest classification algorithm [6]. Each dataset will be used to train and build machine learning models using k-nearest neighbour, support vector machine, Naïve Bayes, elastic net model and random forest approaches. The trained models were used for prediction on the testing dataset. The principles followed here are supervised learning, training the models on labelled data, and evaluating the model performance using an unseen dataset.

After training, the best modelling algorithm is selected for each dataset, based on the pre-processing steps involved and complexity of data. The fitted model can be further used for the prognosis of breast cancer recurrence on additional data. This research will provide valuable insights into the influence of multi-omics expressions and clinical information on the recurrence of breast cancer. This would further help clinicians and researchers in developing personalised treatment plans and conducting further research, improving patient outcomes in breast cancer patients.

2 Methods

2.1 Overview

The database used in this study is The Cancer Genome Atlas Program database. It consists of six text files with data on each patient's clinical information and omics measurements, including gene expression, microRNA expression, protein abundance, DNA methylation, and genetic mutation counts. The target variable relevant to our study is progression (pfi) and is of type, binary. It indicates whether the cancer reappeared or not post-treatment. The clinical information includes cancer risk factors such as age, family history, previous diagnosis, tissue density, hormone therapy, obesity, height, alcohol consumption status, radiotherapy status and the target variable, progression status. This information was arranged in the clinical dataset with the rows representing each patient sample and columns representing risk factors. Each omics contains patient samples arranged in columns and omics features in rows. All the omics measurements are made relative to the matched normal sample.

The analysis is done in R studio using R programming language. The required R packages are installed and loaded for analysis, imputation, feature selection, model building and visualisation of the data at different points in the research. The training and testing folders containing the dataset are downloaded from HPC Blue Crystal phase-4, University of Bristol, UK.

2.2 Data Loading

As the first step, data paths for training and testing directories were read and corresponding relevant data is loaded into the working environment. Slicing, grep and lapply functions are performed for this purpose.

2.3 Data Pre-processing

After the training and testing data directories were read and data was loaded, exploratory analysis was performed on each dataset to understand them better. This helped in understanding the structure, summary, and sample datapoints. This also helped in exploring the missingness in the data and in identifying relationships between different features within and between datasets.

2.3.1 Data Transformation

The features in each relevant text files in the training and testing directory is converted to numeric type for prediction accuracy and better feature representation. This is done using as.numeric(), after converting each into categories using as.factor(). Features that were coded with null values of type character were respecified as type null and missing. The distinct values were explored using unique() function from dplyr package, applied to each feature. The target variable, progression is extracted and stored as a separate variable, outcome and converted to type factor using as.factor(). This is then converted to Yes/No values ideal for classification.

2.3.2 Handling Missing Data

The presence of missing data is explored using miss_var_summary() from visdat package. The presence of outliers is also detected by the Turkey method using quartiles and inter-quartile distance. Considering the outliers and skewness in the data, median imputation is used to handle missingness in the data. Median imputation replaces missing values in a dataset with the median value of the non-missing values for each feature.

2.3.3 Standardising Data

The data is centered and scaled using preProcess() function from caret package. This centring and scaling is then applied to the dataset using predict() function. This is done to bring all the features to a similar scale and to remove biases in the data.

2.3.4 Boruta Feature Selection

Boruta feature selection method is based on the random forest algorithm. It identifies relevant features in a dataset by comparing them to randomly generated ”shadow” features. Shadow features are copies of the original features that are randomly permuted to destroy any potential relationship with the target variable. Gini impurity metrics are used to measure the importance of features. Boruta() function from the Boruta package is used for this. The getSelectedAttributes() function is fed with the Boruta output to get the selected features. The dataset is then updated with the selected features as the next step towards predictive modelling.

2.3.5 Principle component analysis (PCA)

Principle component analysis is a dimensionality reduction technique used to identify patterns and structures in data. The correlation among the variables is reduced by transforming the original variables into a new set of uncorrelated variables called principal components. The prcomp() function from the stats package was used for this. This new dataset formed by the principle components will further represent the dataset in model training and testing.

2.4 Machine Learning

Supervised machine learning is a type of machine learning where the algorithm learns from labelled training data to make predictions or classifications on unseen data. The learning is done on the training dataset and the testing dataset acts as the unseen data. The target variable acts as labels for model training and as the values predicted on the unseen data.

Each dataset in the training set is trained on 5 machine-learning strategies. These trained models are used to predict the target variable in the testing datasets. Depending on the nature of the dataset, prediction accuracy and other factors, the best strategy for predictive modelling is chosen.

2.4.1 10-fold cross validation

Cross-validation is a resampling technique used in machine learning. The available data is partitioned into 10 subsets, called folds. The model is trained and evaluated on different combinations of these folds. The createFolds() function from the caret package is used for this. The output of the cross-validation function will be specified in the control parameters in model training.

2.4.2 Trian controls

Common control parameters for all the models were defined using trainControl() from the caret package. The summary function is defined as twoClassSummary for handling the binary nature of the target variable. The classProbs is set to true, as the model is asked to calculate the class probabilities in addition to the predicting class labels. The verboseIter is set to false, indicating to not display the progress of cross-validation. The savePredictions is set to true, indicating to save of the predictions made by the model during the cross-validation process. The method and number indicate that k-fold cross-validation will be performed with 10 folds.

2.4.3 Elastic Net Model

The Elastic Net model is a linear regression model that combines the L1 (Lasso) and L2 regularization (Ridge) techniques. It is used for variable selection and regularization in the presence of high-dimensional data. It addresses some limitations of the Lasso and Ridge models by including both penalties.

The Elastic net model was trained using glmnet method, representing logistic regression, belonging to family binomial. The grid of hyperparameters to be searched during the model tuning process is specified with an alpha value of 0:1, indicating values between 0 and 1 (inclusive) and a lambda value of 0:1/10, indicating 10 equally spaced values between 0 and 1. This grid adds the elastic net model characteristics to the logistic regression.

2.4.4 Random Forest

The Random Forest is an ensemble machine-learning algorithm used for both classification and regression tasks. It makes use of decision trees for predictive modelling, by combining multiple decision trees to create powerful models.

The Random Forest model was trained with the ranger method. The grid of hyperparameters to be searched during the model tuning process is specified with expand.grid(), to get a combination of all the parameters specified, mtry, splitrule, and min.nide.size during tuning.

2.4.5 K-Nearest Neighbors (KNN)

K- Nearest Neighbors is a non-parametric classification algorithm used for both binary and multi-class classification tasks. It is based on the principle that similar data points tend to have similar labels. In KNN, the class of a new data point is determined by the majority class of its K nearest neighbours.

The k-nearest neighbour model was trained with knn method. The tuneLength parameter, set to 20 indicates the k-value, the number of neighbours.

2.4.6 Support Vector Machine (SVM)

A support Vector Machine is a supervised machine learning algorithm used for both classification and regression tasks. It uses the concept of an optimal hyperplane that maximizes the margin between different classes or fits the data points with minimal error in regression.

The Support Vector Machine (SVM) model was trained with the svmRadial method. The tuneLength parameter, set to 10 indicates the number of values to be explored for the cost and sigma parameters during the model tuning process.

2.4.7 Naïve Bayes

Naive Bayes is a probabilistic machine learning algorithm. It is based on Bayes' theorem and assumes independence among the predictors. It is commonly used for classification tasks and is known for its simplicity and speed. The Naïve Bayes model was trained with the naive_bayes method.

2.5 Model selection

All trained models will be added and saved in a list, model_list. The resamples() function from the caret package will be used to calculate the resampling results from the list of models using cross-validation. The summary() from the caret package will be used to view the summary of the sampling results, such as the mean and standard deviation of the performance metrics across different models. A visual representation of the performance metrics is done using box-and-whisker and dot plots for each model in the list on the sample plot.

2.6 Prediction

The trained models were used to make predictions on the testing dataset. The predict() function from the caret package was used for this. The predictions from each of the 5 models were stored in a list. The training and testing datasets were pre-processed in the same ways. The model list was unlisted and each model performance metric such as accuracy and lower and upper limits of the 95% confidence interval were extracted and added to a new data frame. The performance is compared and visualised across the models in the form of a table. Depending on the nature of the dataset, prediction accuracy and other factors, the best predictive model was chosen and the confusion matrix is further explored and concluded. The confusionMatrix() from the caret package is used for this.

2.7 Reproducibility

Git is a version control approach for managing source code and tracking changes in software development projects. This project is version-controlled using git and uploaded to the remote git repository. The files in this project are organised in the form of an **executable research compendium** for reproducibility. **Pipelining** is achieved using conda snakemake.

3 Results

3.1 Directories and Datasets

The training directory consists of 7 data files and 3 annotation files, and the testing directory consists of 7 data files. All the files are in text format. The annotation files are used to link and produce comparable omics characteristics. The target variable is progression is found in the clinical file.

3.2 Model 1 (Clinical features)

In the clinical dataset, the data is arranged in rows representing each patient sample and columns representing risk factors. The dataset containing clinical information in the training and testing directory is pre-processed and the training set is fitted on different machine learning strategies. No missingness is observed in the dataset. The following plot gives the ROC characteristics of each model fitting.

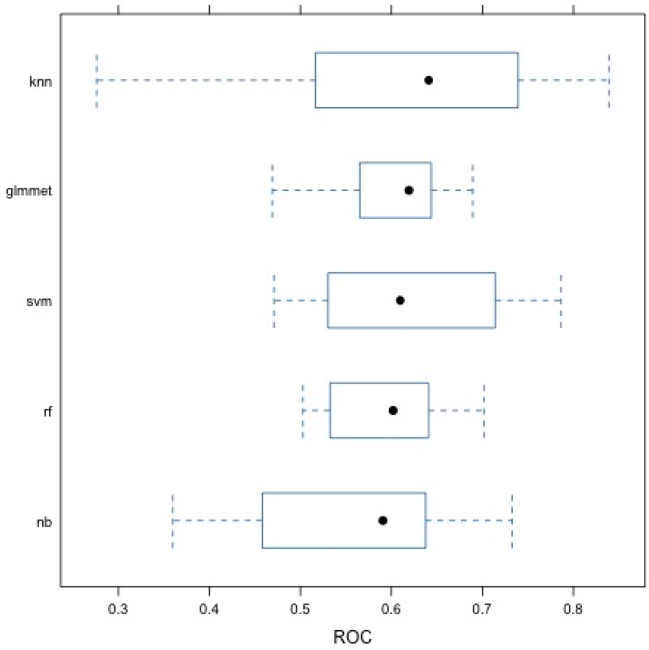

Figure 1: Comparison of machine learning algorithms for model clinical features

The KNN followed by glmmet and svm appears to be the best model according to the AUC plot. Due to the presence of outliers and considering the number of predictors and nature of the outcome, the random forest model is chosen for prediction. The model accuracy is 0.942 and 95% confidence interval ranges from 0.922 to 0.958.

Table 1: Model Accuracy

Model	Accuracy	95% CI
rf	0.942	0.922 - 0.958

3.3 Model 2 (microRNA expression measurements)

The dataset containing microRNA expression measurements in the training and testing directory is pre-processed and the training set is fitted on different machine learning strategies. No missingness is observed in the dataset. The following plot gives the ROC characteristics of each model fitting.

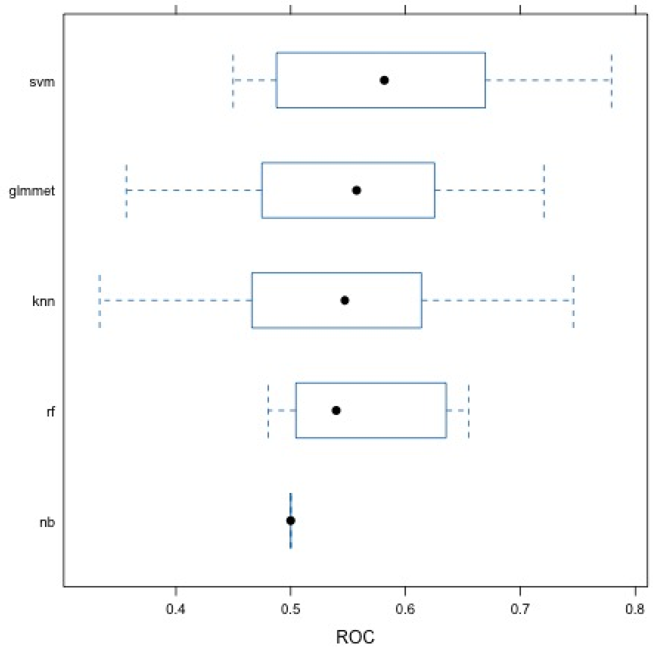

Figure 2: Comparison of machine learning algorithms for model on microRNA measurements

The Support Vector Machine (SVM) appears to be the best model according to the AUC plot. It also shows the presence of severe bias in the model due to the presence of outliers. As a result, the random forest model will be chosen for prediction. The model accuracy is 0.874 and 95% confidence interval ranges from 0.847 to 0.897.

Table 2: Model Accuracy

Model	Accuracy	95% CI
rf	0.874	0.847 - 0.897

3.4 Model 3 (gene expression measurements)

The dataset containing gene expression measurements in the training and testing directory is pre-processed and the training set is fitted on different machine learning strategies. No missingness is observed in the dataset. The following plot gives the ROC characteristics of each model fitting.

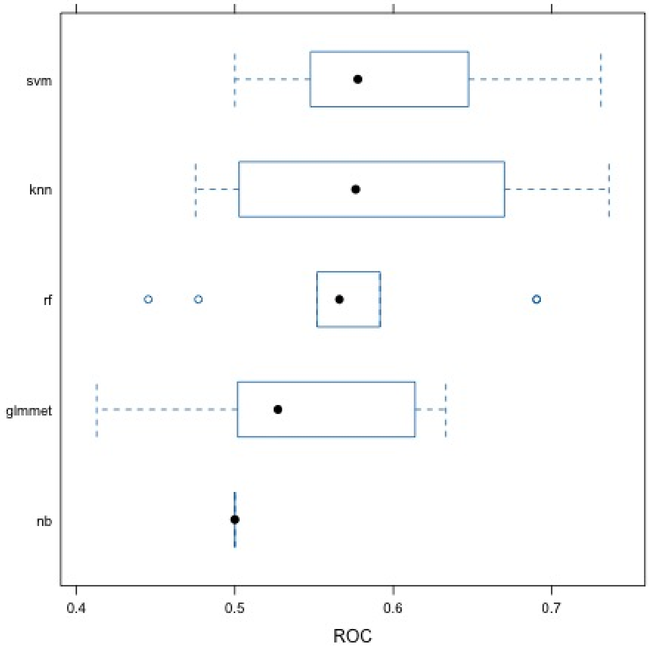

Figure 3: omparison of machine learning algorithms for model on gene expression measurements

The Support Vector Machine (SVM) appears to be the best model according to the AUC plot. Due to the presence of outliers and considering the number of predictors and the nature of the outcome, the random forest model is chosen for prediction. The model accuracy is 0.875 and 95% confidence interval ranges from 0.849 to 0.898.

Table 3: Model Accuracy

Model	Accuracy	95% CI
rf	0.875	0.849 - 0.898

3.5 Model 4 (DNA methylation measurements)

The dataset containing gene expression measurements in the training and testing directory is pre-processed and the training set is fitted on different machine learning strategies. No missingness is observed in the dataset. The following plot gives the ROC characteristics of each model fitting.

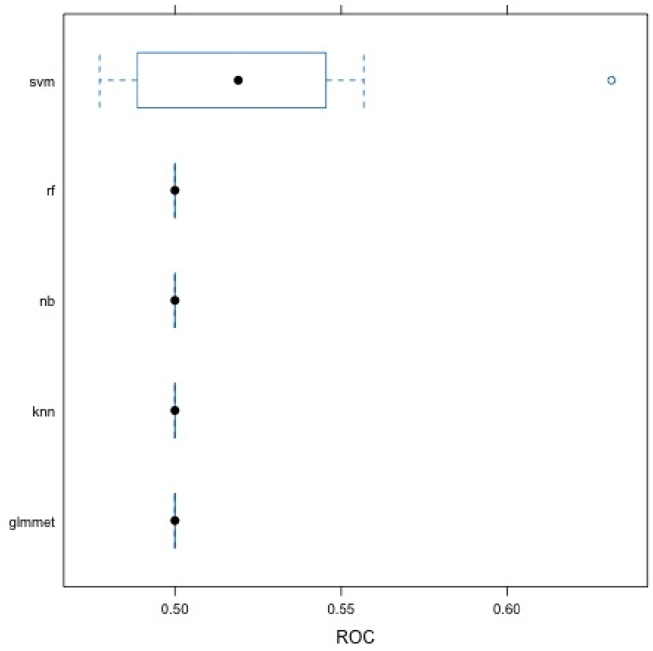

Figure 4: omparison of machine learning algorithms for model on DNA methylation measurements

The Support Vector Machine (SVM) appears to be the best model according to the AUC plot. Due to the presence of outliers and considering the number of predictors and the nature of the outcome, the random forest model is chosen for prediction. The model accuracy is 0.865 and 95% confidence interval ranges from 0.833 to 0.894.

Table 4: Model Accuracy

Model	Accuracy	95% CI
rf	0.865	0.833 - 0.894

3.6 Model 5 (Protein abundance measurements)

The dataset containing protein abundance measurements in the training and testing directory is preprocessed and the training set is fitted on different machine learning strategies. No missingness is observed in the dataset. The following plot gives the ROC characteristics of each model fitting.

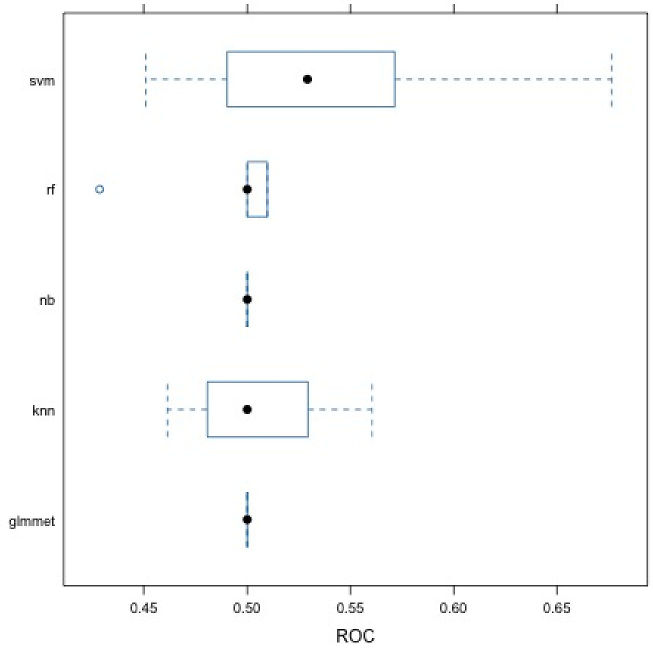

Figure 5: omparison of machine learning algorithms for model on protein abundance measurements

The Support Vector Machine (SVM) appears to be the best model according to the AUC plot. Due to the presence of outliers and considering the number of predictors and the nature of the outcome, the random forest model is chosen for prediction. The model accuracy is 0.883 and 95% confidence interval ranges from 0.854 to 0.908.

Table 5: Model Accuracy

Model	Accuracy	95% CI
rf	0.883	0.854 - 0.908

3.7 Model 6 (Genetic mutation counts)

The dataset containing genetic mutation counts in the training and testing directory is pre-processed and the training set is fitted on different machine learning strategies. No missingness is observed in the dataset. The following plot gives the ROC characteristics of each model fitting.

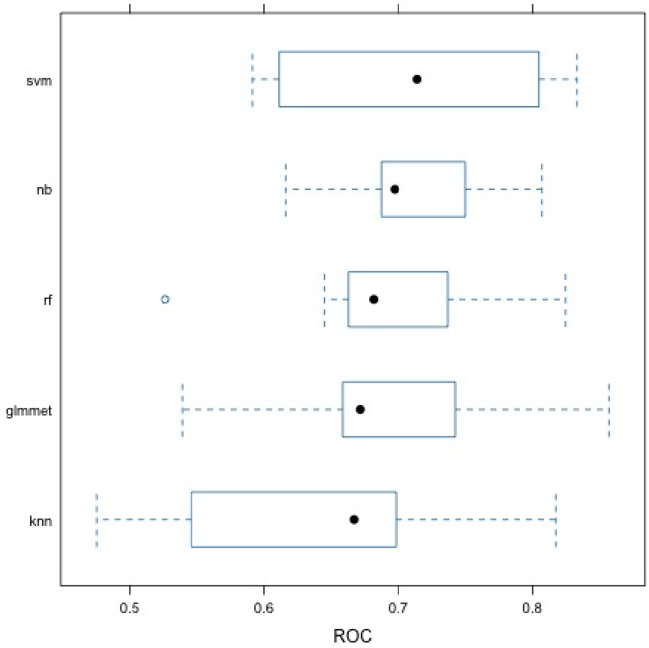

Figure 6: omparison of machine learning algorithms for model on mutation measurements

The Support Vector Machine (SVM) appears to be the best model according to the AUC plot. Due to the presence of outliers and considering the number of predictors and the nature of the outcome, the random forest model is chosen for prediction. The model accuracy is 0.903 and 95% confidence interval ranges from 0.878 to 0.925.

Table 6: Model Accuracy

Model	Accuracy	95% CI
rf	0.903	0.878 - 0.925

3.8 Model 7 (Gene expression measurements and Protein abundance measurements)

The dataset containing gene expression measurements and Protein abundance measurements in the training and testing directory is pre-processed and the training set is fitted on different machine learning strategies. No missingness is observed in the dataset. The following plot gives the ROC characteristics of each model fitting.

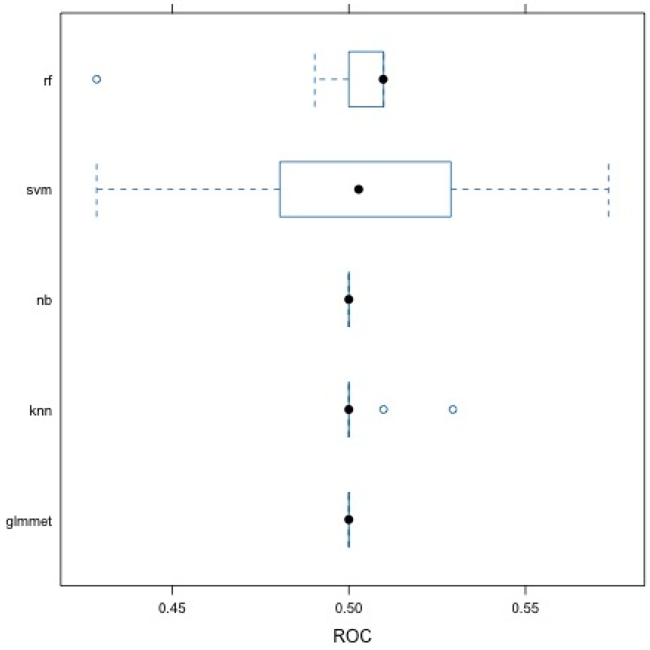

Figure 7: omparison of machine learning algorithms for model 7

The Random Forest appears to be the best model according to the AUC plot and will be used for prediction. The model accuracy is 0.998 and 95% confidence interval ranges from 0.99 to 1.

Table 7: Model Accuracy

Model	Accuracy	95% CI
rf	0.998	0.99 - 1

3.9 Model 8 (Gene expression measurements and Protein abundance measurements combined with clinical information)

The dataset containing gene expression measurements and Protein abundance measurements combined with clinical information in the training and testing directory is pre-processed and the training set is fitted on different machine learning strategies. Missingness observed in the dataset is handled using median imputation. The following plot gives the ROC characteristics of each model fitting.

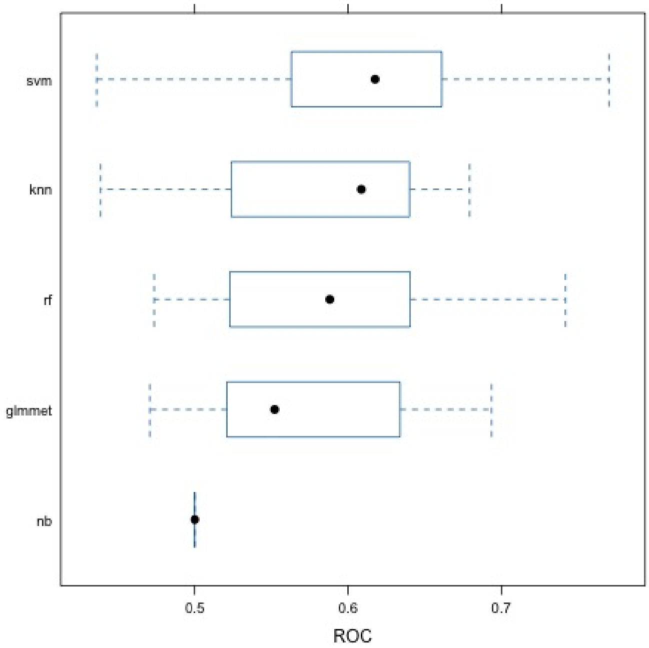

Figure 8: omparison of machine learning algorithms for model 8

The Random Forest appears to be the best model according to the AUC plot and will be used for prediction. The model accuracy is 0.883 and 95% confidence interval ranges from 0.854 to 0.908.

Table 8: Model Accuracy

Model	Accuracy	95% CI
rf	0.883	0.854 - 0.908

4 Discussion

This study showed the feasibility of using supervised machine learning techniques to predict breast cancer progression using clinical information and multi-omics data. The choice of the appropriate machine learning algorithm depends on the research question, the nature of the data, and the ability to handle high-dimensional data. For the 8 models, we considered svm, knn, random forest, naïve bayes and elastic net machine learning algorithms and the best turned out to be random forest for all the models. Even though the models varied in their predictive power, each provided unique insights into the progression of cancer. Thus, the Random Forest algorithm is used for making predictions on the testing dataset.

Biological data often appears messy and has many limitations. This study provides a promising direction for future research. The study aimed to investigate the association between various clinical and multi-omics features on recurrence in patients with breast cancer. This provided valuable insights for clinicians and researchers, aiding them in developing personalised treatment plans and conducting further research. Further research of alternate machine learning techniques for similar studies can be conducted.

5 References

[1] Sun, Yi-Sheng, Zhao Zhao, Zhang-Nv Yang, Fang Xu, Hang-Jing Lu, Zhi-Yong Zhu, Wen Shi, Jianmin Jiang, Ping-Ping Yao, and Han-Ping Zhu. 2017. “Risk Factors and Preventions of Breast Cancer.” International Journal of Biological Sciences 13 (11): 1387–97. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5715522/

[2] https://www.cdc.gov/cancer/breast/basic$_info/what - is - breast - cancer$

[3] Sopik, Victoria, and Steven A. Narod. 2018. “The Relationship between Tumour Size, Nodal Status and Distant Metastases: On the Origins of Breast Cancer.” Breast Cancer Research and Treatment 170 (3): 647–56. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6022519/

[4] Wu, Mengjuan, Ting Zhao, Qian Zhang, Tao Zhang, Lei Wang, and Gang Sun. 2022. “Prognostic Analysis of Breast Cancer in Xinjiang Based on Cox Proportional Hazards Model and Two-Step Cluster Method.” Frontiers in Oncology 12: 1044945. https://doi.org/10.3389/fonc.2022.1044945

[5] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

[6] Kursa, Miron B., and Witold R. Rudnicki. 2010. “Feature Selection with the Boruta Package.” Journal of Statistical Software 36 (September): 1–13. https://www.researchgate.net/publication/280138095$_Feature_Selection_with_Boruta_Package$