Predictive Modelling of Breast Cancer Patient Outcomes Using Machine Learning and Multi-Omic Data Analysis: A Random Forest Approach

Abstract

Background: Breast cancer is a complex and heterogeneous disease with significant implications for women's health globally. The availability of large-scale multi-omic datasets, such as those from The Cancer Genome Atlas (TCGA), provides an opportunity to predict patient outcomes using machine learning algorithms. In this study, the aim was to evaluate the predictive potential of clinical and multi-omic data for breast cancer patient outcomes using Random Forest.

Methods: A subset of the TCGA breast cancer dataset was analyzed using eight different Random Forest models, each utilizing a different set of features. Variable selection was performed using stepwise regression for clinical features and linear models for omics data. The models were evaluated using the Area Under the ROC Curve (AUC) as a performance metric.

Results: The results showed that the predictive performance of the models varied depending on the set of features used. Models utilizing individual omic data types showed poor discriminatory power, while models combining multiple data types exhibited overfitting or limited improvement in predictive performance.

Conclusion: Random Forest, selected for its ability to handle high-dimensional data and robustness against noise, demonstrated limitations in predicting breast cancer outcomes in this study. Further refinement of feature selection and exploration of alternative modelling techniques is suggested. This research contributes to the growing body of literature on machine learning in cancer research and highlights the challenges associated with predicting breast cancer outcomes using Random Forest.

Introduction

Breast cancer is a complex and heterogeneous disease characterized by the abnormal growth of cells in the breast tissue, with significant implications for women's health worldwide. It is the most diagnosed cancer in women globally [1]. Advancements in genomic technology have led to the availability of large-scale multi-omic datasets. The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. Multi-omic approaches, including DNA methylation, gene expression, genetic mutations along with clinical data provide comprehensive insights into the molecular landscape of breast cancer and provide potential opportunities to predict patient outcomes [2].

The aim of this study is to apply machine learning algorithms to analyze the few of the TCGA breast cancer dataset and evaluate the predictive potential of clinical and multi-omic data for patient outcomes for breast cancer. By employing machine learning algorithms, like random forest, the study aims to develop models that predict disease progression, as indicated by the progression-free interval (PFI).

To achieve this aim, eight random forest models has been trained and evaluated, each utilizing a different set of features. By considering various combinations of clinical and omics data, this study will assess the impact of different feature sets on the predictive performance of the models.

The findings of this research hold potential implications for personalized medicine, enabling clinicians to make informed decisions about treatment strategies and patient care. Furthermore, this study contributes to the growing body of literature on the application of machine learning in cancer research, showcasing the utility of integrating clinical and multi-omic data for improved prognostic modeling.

Methods

Dataset Acquisition:

The dataset used in this study has been obtained from The Cancer Genome Atlas Program (TCGA) (https://www.cancer.gov/about-nci/organization/ccg/research/structuralgenomics/tcga). The dataset consists of patient information, including disease outcome - progression-free interval, standard clinical information, and measurements from various omics. The dataset has been extracted as text files and was accessed from the Blue Crystal 4 HPC system from university of Bristol.

Data Processing:

Prior to analysis, the raw dataset requires preprocessing and quality control measures due to its real-world nature. Firstly, a function was created to load tab eliminated dataset into R studio for analyses to be conducted.

Since omics data and clinical data have different characteristics, the variable selection are done differently. Clinical features represent well-established and standardized measurements related to patients' medical history, and clinical assessments. On the other hand, omics data captures thousands or even millions of features.

Variable selection for clinical features has been done using stepwise regression which prunes a list of plausible explanatory variables down to collection of the most useful variables.

For omics data before variable selection, features with all data missing were removed from the dataset as they provide no variability, adds to computational overhead and can negatively impact model performance. After removal of those features, linear models were fit for each feature and p-values were calculated from which the top 25 associations were identified based on the p-values.

Post creating a dataset with only statistically important variates, the dataset has been imputed and standardized. Due to the complexity of multiple imputation (MI), median imputation method has been utilized to handle missing values in this study. While MI can provide more accurate estimates of missing values, it assumes that the data are missing at random (MAR) or missing completely at random (MCAR). In TCGA, the missingness patterns may not conform to these assumptions.

Model Construction and Evaluation:

Random Forest algorithm has been conducted for this study, it is an ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and robustness. In this algorithm, a collection of decision trees is constructed, where each tree is trained on a different subset of the data. At each node of the decision tree, a random subset of features is considered for splitting, reducing the risk of overfitting. The trees are grown independently, making the algorithm highly parallelizable and scalable. During the prediction phase, each tree in the forest independently classifies or predicts the target variable, and the final prediction is obtained by averaging the predictions from all the trees (for regression tasks) or taking a majority vote (for classification tasks) [3].

Random forest here has been evaluated using Area Under the ROC Curve (AUC) which is used as a summary metric, where a higher value indicates better performance. The ROC curve plots the true positive rate against the false positive rate at different classification thresholds which helps understand the trade-off between sensitivity and specificity and provides a measure of the model's power [3].

Software used:

All analyses as been done using R studio version 4.3

For version control Github and Visual studio has been used

Results

From Stepwise regression, the following variables were statistically important: age at diagnosis, progesterone receptor status, estrogen receptor status, lymphocyte infiltration, necrosis percent, and her2 status.

Using these variable, and top 25 features from omics data 8 random forest models were created,

Model	Model features	Training AUC	Testing AUC
1	Clinical features	0.8096	0.49
2	Methylation	0.533	0.5
3	Mirna	0.56	0.5
4	Mrna	0.517	0.5
5	mutation	1	0.5
6	protein	0.511	0.5
7	mRNA and miRNA	1	0.48
8	Clinical and mRNA	0.533	0.49

Model 1: This model uses only clinical features and achieves a training AUC of 0.8096 and a testing AUC of 0.49. It demonstrates good performance on the training set but does not generalize well to the testing set.

Model 2: This model uses methylation features and achieves a training AUC of 0.533 and a testing AUC of 0.5. The performance is similar to random chance, indicating that the model does not effectively discriminate between the classes.

Model 3: This model uses miRNA features and achieves a training AUC of 0.56 and a testing AUC of 0.5. Similar to Model 2, it shows poor discriminatory power.

Model 4: This model uses mRNA features and achieves a training AUC of 0.517 and a testing AUC of 0.5. It also demonstrates weak performance in distinguishing between the classes.

Model 5: This model uses mutation features and achieves a perfect training AUC of 1 and a testing AUC of 0.5. The model appears to overfit the training data as it achieves perfect classification but fails to generalize to the testing set.

Model 6: This model uses protein features and achieves a training AUC of 0.511 and a testing AUC of 0.5. Similar to the previous models, it does not perform well in classifying the data.

Model 7: This model combines mRNA and miRNA features and achieves a perfect training AUC of 1 but a lower testing AUC of 0.48. It indicates overfitting and poor generalization to unseen data.

Model 8: This model combines clinical and mRNA features and achieves a training AUC of 0.533 and a testing AUC of 0.49. Similar to Model 1, it performs reasonably well on the training set but does not generalize effectively to the testing set.

Discussion

The aim of this study was to apply machine learning algorithms, specifically the Random Forest algorithm, to analyze a subset of The Cancer Genome Atlas (TCGA) breast cancer dataset and evaluate the predictive potential of clinical and multi-omic data for patient outcomes. The study examined eight different random forest models, each utilizing a different set of features, to assess the impact of different feature combinations on predictive performance.

The results demonstrate that the predictive performance of the models varied depending on the set of features used. Model 1, which only used clinical features, showed good performance on the training set but did not generalize well to the testing set. Models 2 to 6, which utilized different omic data types individually, showed poor discriminatory power and failed to effectively classify the data. Model 7, which combined mRNA and miRNA features, exhibited overfitting and poor generalization. Model 8, which combined clinical and mRNA features, performed similarly to Model 1, indicating limited improvement in predictive performance.

Random Forest was selected here as it is known for its ability to handle high-dimensional data, which is a characteristic of omics datasets commonly encountered in cancer research. Random Forest's ensemble learning approach, which combines multiple decision trees, can effectively handle high-dimensional data by selecting relevant features and mitigating the risk of overfitting still overfitting can be observed here even after using regularization techniques, such as limiting the depth of decision.

Additionally, Random Forest is robust against noise and outliers in the data which makes a suitable choice for analyzing cancer datasets with potential data quality issues, but it is important to features selected in this study are solely based on numbers without any medical opinion which might have caused poor performance as expert opinion is crucial in these types of data. This suggests the need for further refinement or exploration of alternative feature sets and modeling techniques.

Random Forest algorithm used in this study is just one of many machine learning algorithms available. Exploring alternative algorithms, such as support vector machines or neural networks, could potentially yield improved predictive performance. Each algorithm has its own strengths and weaknesses, and different algorithms may be more suitable for specific types of data or prediction tasks. Therefore, conducting comparative analyses of different algorithms could provide valuable insights into which method is best suited for the given breast cancer dataset.

Overall, this research contributes to the growing body of literature on the application of machine learning in cancer research. The study showcases the utility of integrating clinical and multi-omic data for improved prognostic modeling and provides insights into the challenges associated with predicting breast cancer outcomes using the Random Forest algorithm.

Reference:

[1] World Health Organization (WHO). Breast Cancer: Key Facts. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer]

[2] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.

[3] An introduction to Statistical Learning with Applications in R, Gareth James (Springer)

Additional comments:

Different models can easily be accessed in the same file.