Graphical user interface, text

Description automatically generated

# Coversheet for submission of MSc coursework

* DO NOT WRITE YOUR NAME ON YOUR WORK. Instead, please write your student number on this coversheet.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Student No. | 2 | 3 | 9 | 8 | 7 | 3 | 6 |

|  |  |
| --- | --- |
| Date Submitted: | 2024.3.20 |
| Word Count: | 1460 |

PLAGIARSIM DECLARATION:

|  |
| --- |
| This assignment is entirely my own work, and it adheres to the University of Bristol’s policy on plagiarism and academic integrity.  Quotations from secondary literature are indicated by the use of inverted commas around ALL such quotations AND by citations in the text or notes to the author concerned. ALL primary and secondary literature used in this piece of work is indicated in the bibliography placed at the end, and dependence upon ANY source used is indicated at the appropriate point in the text.   * I confirm that no sources have been used other than those stated. * I confirm that I have not used artificial intelligence or chatbot software to create any writing or content in this assessment * I confirm that I have not written this work in another language and translated it into English using translation tools * I confirm that I have not used grammar checkers that suggest rewrites   **I understand that plagiarism, collusion, and cheating constitute misconduct and may result in disciplinary action being taken.** |

Comparative Analysis of Clinical and Genomic Data for Predicting Breast Cancer Recurrence: Evaluating Machine Learning Models

**Abstraction:**

**Background:**

Breast cancer is one of the most common malignant tumors with less than 50% survival rate after recurrence. Effective recurrence prediction is essential to improve patient survival.

**Study aim:**

1. To compare the predictive efficacy of using only clinical data versus combining it with genetic data for breast cancer recurrence;

2. To evaluate and compare the performance of various machine learning models, including XGBoost, Elastic Net, Decision Tree, Random Forest, and LASSO, in predicting breast cancer recurrence.

**Method**:

The study utilized different combinations of datasets including clinical data, mRNA data, miRNA data, protein expression data, and genetic mutation data. Further, the study compared the effectiveness of XGBoost, Elastic Net, Decision Tree, Random Forest, and LASSO using clinical plus genetic mutation data.

**Result and conclusion:**

By comparing the predictive abilities of different datasets, I found genetic mutation data significantly enhanced prediction accuracy. The results demonstrate that the XGBoost model performed optimally in handling complex genetic relationships, providing the highest prediction accuracy.

1. Introduction

Breast cancer is a very common malignant tumor today, with patients accounting for more than 36 percent of oncology patients, and is one of the leading causes of female mortality(Liew, 2021). The incidence of breast cancer is increasing year by year. Once cancer has recurred, the survival rate is less than 50%, making it important to predict cancer recurrence after surgery(Smolarz, 2022).

Polygenic factors and clinical information, which includes cancer grade, degree of histological malignancy, hormonal receptors, and so on play an important role in predicting breast cancer outcomes. Therefore it is necessary and possible to use clinical information as well as genetic data to make predictions about the prognosis of the disease.

With the development of artificial intelligence, researchers tend to use various machine learning models to predict cancer relapse. At the same time, using omic data to establish a prediction system is also feasible with the increased capacity of computers to store data and build models.

Alzu’bi(2021) used clinical data to predict disease recurrence by using bagging, logistic model, SVM, KNN, MLP, PART, and OneR, then found that the bagging and OneR performed better than other models. In Behravan’s research(2020), they used a machine learning approach to identify risk factors for breast cancer by combining genetics with family history and estrogen metabolism. In particular, XGBoost was used to evaluate the importance of features in this model. In another article, Das(2015) used ENCAPP which is an elastic-net-based approach to predict the prognosis of different cancers. In this model, Elastic Net was used to help choose features. Thus, combining genetic data with clinical data can lead to more accurate predictions and provide better advice to doctors in clinical decisions to reduce disease recurrence.

Genetic testing is expensive, ranging from $2,000-$5,000 per individual, so the use of genetic testing in clinical diagnosis is worth investigating the question. The research aims to figure out the question of whether adding omic data in a relapse prediction system will improve the prediction accuracy.

As a result, it is necessary to build up machine learning models with and without omic data. However, the outcome variable is binary, which simplifies the complex nature of the disease. The feature selection method may not include all relevant features, and a very complex model will make it hard to interpret results.

1. Methods
   1. Dataset and data preprocessing

This project contains datasets from The Cancer Genome Atlas Program (TCGA) including clinical data and omic data for 1085 patients. The clinical dataset has patients’ disease records including age at diagnosis, race, disease stage, and tumor category along with a binary variable that records the tumor progress. In the genomic dataset, I acquired data on mRNA and microRNA expression levels, DNA methylation, protein abundance, and genetic mutations.

Within these complex datasets, there is much missing data. So I will conduct median imputation to address these gaps which can preserve the overall distribution of the data. Also, I will remove some outliers to prevent them from skewing the data. Character variables will be converted into factorial variables to facilitate their use in subsequent models. For omic datasets, I will implement Z-score normalization to standardize each variable, ensuring comparability between different types of omic data and making sure each feature has an equal contribution to the model.

* 1. Feature Extraction

I will use Elastic Net to extract features, which is a linear regression model combining L1 and L2 regularization to constrain the complexity of the model and prevent overfitting. It selects features by adjusting two parameters, and , where controls the relative impact of L1 and L2 regularization terms, and controls the strength of regularization. Overall, it is suitable for datasets with multicollinearity or more features than samples. And with this algorithm, I will select clinic features and omic features.

* 1. Model Training

In this part, I will use several classifications to predict the tumor reappearance including XGBoost, Elastic Net, Lasso, Random Forest, Decision tree. The training set will be 75% of the dataset and the testing set will be 25%. The division strategy will be applied through all of the datasets.

Firstly, I will train models with clinic data using selected clinic features. Then assess each model’s performance by evaluation metrics. Secondly, repeat the pipeline with omic data and omic features. Finally, combine clinic features and omic features, and train classifiers on the combined dataset. Then evaluate the models and compare the performance against models based on clinical dataset. The aim is to evaluate the prediction ability between clinical information and omic information and calculate the exact extent to which omic data improves predictions.

* 1. Model Evaluation

The evaluation matrix will contain accuracy, precision, recall, ROC curve, and AUC curve.

1. Result
   1. Application results for clinical data and different genetic data

In order to evaluate the effectiveness of combining clinical data with genetic data, I used the XGBoost model to compare the performance of using only clinical data with using clinical data combined with different genetic data. I conducted prediction models with many different datasets. The effect of each dataset was assessed by key metrics including AUC, acc, sensitivity, specificity, precision, recall, and fbeta.

Table 1 below is the result of comparing different datasets. And Figure 1 below shows the combination of ROC curves generated by the model for different datasets, which can visualize the effect of adding genetic data on the prediction ability very effectively. We can see that the difference between using only clinic data and clinic plus omic data is not very large, and even adding all omic data will test the predictive power down. Of all the genetic data, genetic mutation improves the predictive ability the most.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **auc** | **acc** | **sensitivity** | **specificity** | **precision** | **recall** | **fbeta** |
| **clinic** | 0.60 | 0.85 | 0.96 | 0.15 | 0.88 | 0.96 | 0.92 |
| **clinic & mrna** | 0.54 | 0.87 | 1 | 0.023 | 0.87 | 1 | 0.93 |
| **clinic & protein** | 0.61 | 0.86 | 0.99 | 0 | 0.87 | 0.99 | 0.93 |
| **clinic & mutation** | 0.66 | 0.86 | 0.97 | 0.17 | 0.88 | 0.97 | 0.92 |
| **clinic & mirna** | 0.62 | 0.86 | 0.99 | 0.009 | 0.86 | 0.99 | 0.93 |
| **clinic & all above** | 0.57 | 0.87 | 0.99 | 0.011 | 0.87 | 0.99 | 0.93 |

Table 1

A graph of a curve

Description automatically generated with medium confidence

Figure 1

* 1. Application results for different prediction models

In this section I compared XgBoost, Elastic Net, Decision Tree, Random Forest, and LASSO in the best dataset: clinical data with mutation data. Table 2 below shows the results of different models. And Figure 2 below shows the combination of ROC curves generated by different models, which can visualize the prediction ability very effectively.

We found that the xgboost model performs best, the decision tree and random forest perform slightly worse, and elastic net and LASSO are not suitable for genetic data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **auc** | **acc** | **sensitivity** | **specificity** | **precision** | **recall** | **fbeta** |
| **Xgboost** | 0.66 | 0.86 | 0.97 | 0.17 | 0.88 | 0.97 | 0.92 |
| **Elasric Net** | 0.5 | 0.87 | 1 | 0 | 0.87 | 1 | 0.93 |
| **Decision Tree** | 0.61 | 0.86 | 0.96 | 0.16 | 0.88 | 0.96 | 0.92 |
| **Random Forest** | 0.62 | 0.86 | 0.99 | 0 | 0.87 | 0.99 | 0.93 |
| **LASSO** | 0.5 | 0.87 | 1 | 0 | 0.87 | 1 | 0.93 |

Table 2

A graph of a variety of blue lines

Description automatically generated with medium confidence

Figure 2

1. Discussion
   1. Model preference

When we evaluate the impact of different kinds of genetic data’s predicting ability, we can conclude that only adding genetic mutation in the model can significantly improve the predicting performance. The possible reason is that genetic mutation is the key factor in breast cancer recurrence.

Overall, the using of genetic data did not provide a significant improvement in the prediction of breast cancer prognosis, and the use of genetic testing is not optimal in today's context where genetic testing is more expensive.

When we compare the different machine learning models, the XGBoost model performed optimally. This indicates that when dealing with complex relationships between clinical and genetic data, XGBoost algorithms can account for the nonlinear interactions between features, which is very important when combining clinical and omic data.

In contrast, Elastic Net and LASSO models cannot fully consider the interactions and correlation structures between genes, resulting in their limited effects in handling high-dimensional feature space.

* 1. Strength and limitation

In this research, the main strength is the systematic comparison of the effects of different types of genetic data (protein expression, genetic mutation, mRNA, and miRNA) on the prognostic prediction of breast cancer. This research also compared a variety of machine learning models, which can provide a clear direction for future algorithm exploration.

The limitation of this study is the limited understanding of the biological mechanisms behind different kinds of genetic data, which requires further studies to elucidate.

Reference:

1. Liew XY, Hameed N, Clos J. An Investigation of XGBoost-based Algorithm for Breast Cancer Classification. Machine Learning with Applications. 2021 Sep;100154.

2. Smolarz B, Nowak AZ, Romanowicz H. Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). Cancers [Internet]. 2022 May 23;14(10):2569. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9139759/>

3. Alzu’bi A, Najadat H, Doulat W, Al-Shari O, Zhou L. Predicting the Recurrence of Breast Cancer Using Machine Learning Algorithms. Multimedia Tools and Applications. 2021 Jan 18;

4. Behravan H, Hartikainen JM, Tengström M, Kosma V, Mannermaa A. Predicting Breast Cancer Risk Using Interacting Genetic and Demographic Factors and Machine Learning. Scientific Reports. 2020 Jul 6;10(1).

5. Das J, Gayvert KM, Bunea F, Wegkamp MH, Yu H. ENCAPP: elastic-net-based Prognosis Prediction and Biomarker Discovery for Human Cancers. BMC Genomics [Internet]. 2015 Apr 3 [cited 2022 Nov 10];16:263. Available from: <https://pubmed.ncbi.nlm.nih.gov/25887568/>