Comparison of PLS and LASSO Features Selection Techniques on Cancer Classification

John Ong'ala Lunalo

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistical Sciences of Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2025

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the proposal itself.

Name: **John Ongala Lunalo**

Signature: .....

Date: June 11, 2025

Approval

The dissertation of (John Ong'ala Lunalo) was reviewed and approved by the following:

Professor Bernard Omolo

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Evans Otieno Omondi

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

Recent reports from the World Health Organization (WHO) highlight cancer as the leading cause of global mortality, with a significant impact on women, who frequently experience breast, lung, colorectal, thyroid, and ovarian cancers. The need for accurate diagnostic tools is paramount. This study conducts a comparative analysis of three supervised machine learning classifiers (XGBoost, Random Forest, and 1D convolutional neural networks (1D-CNN)) using feature selection methods, the least absolute shrinkage and selection operator (LASSO) and the partial least squares (PLS), to identify the most effective approach for diagnosing common women's cancers. RNA-Seq gene expression datasets from the Genomic Data Commons Data Portal were used for breast, colon, ovarian, lung, and thyroid cancers. PLS and LASSO identified significant features, with LASSO selecting 173 genes as outlined in the anchor paper and PLS selecting 162 genes. All models achieved high accuracy (>99%) in cancer classification, with XGBoost combined with LASSO demonstrating superior performance in multiple metrics. Notable genes such as *TG, COL1A1, CTSB, CLU,* and *MGP* emerged as crucial markers for classification. The study underscores the importance of precise feature selection in the development of reliable machine learning classifiers for cancer diagnosis, advocating LASSO over PLS in conjunction with XGBoost. These findings highlight the critical role of accurate feature selection in improving cancer diagnosis precision and ultimately improving patient outcomes in oncology.

Keywords: *Women's cancer, RNA-Seq gene expression, XGBoost, Random Forest, Genomic Data Commons*

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAIC	Array-Array Intensity Correlation	LASSO	Least Absolute Shrinkage and Selection Operator
ANN,	Artificial Neural Network	MLP	Multi-Layer Perceptron
ANOVA	Analysis of Variance	PLS	Partial Least Squares
AUC	Area under Curve	RNASeq	RNA sequencing
CI	Confidence Interval	ROC	Receiver operating Characteristics
DNA	Deoxyribonucleic acid	RNASeq	RNA Sequencing
1D-CNN	One-dimentional Convolutional Neural Network	ROC	Receiver operating Characteristics
KNN	K-Nearest Neighbour	WHO	World Health Organization

Chapter 1

Introduction

1.1 Background of the Study

Reports from WHO indicate that cancer accounted for nearly 10 million (excluding non-melanoma skin cancer) deaths in 2020, or nearly one in six deaths, and 609,360 new deaths occurred in the United States of America (Siegel et al., 2022). An estimated 19.3 million new cancer cases (18.1 million excluding non-melanoma skin cancer) occurred in 2020. These statistics make cancer the leading cause of death worldwide, with women being particularly vulnerable (Sung et al., 2021). According to the World Health Organization (WHO), breast, lung, and colorectal cancer are among the most common types of cancer in women, followed by cervical, ovarian, and uterine cancer (Ferlay et al., 2021).

1.1.1 Major Types of Cancer in Women

i) **Breast cancer** is the most common cancer in women, with an estimated 2.26 million cases diagnosed globally in 2020 (Ferlay et al., 2021). Risk factors for breast cancer include age, family history of breast cancer, hormonal factors, and lifestyle factors such as alcohol consumption and physical inactivity (McCarthy et al., 2021). Treatment options for breast cancer include surgery, radiation therapy, chemotherapy, and hormone therapy.

ii) **Lung cancer** is the second most common cancer in women, with an estimated 1.8 million cases diagnosed worldwide in 2020 (Ferlay et al., 2021). In 2022, there were 236,740 new cases in the USA (Siegel et al., 2022). The main risk factor for lung cancer is tobacco use, either through smoking or exposure to second-hand smoke

(Malhotra et al., 2016). Other risk factors include exposure to radon, air pollution, and occupational hazards such as asbestos (Malhotra et al., 2016). Treatment options for lung cancer include surgery, radiation therapy, chemotherapy, and targeted therapy.

iii) **Colorectal cancer** is the third most common cancer in women, with an estimated 1.1 million cases diagnosed globally in 2020 (Ferlay et al., 2021). In 2022, there were 151,030 new cases in the USA (Siegel et al., 2022). Risk factors for colorectal cancer include age, family history of colorectal cancer, and lifestyle factors such as a diet high in red and processed meats, physical inactivity, and obesity (Keum and Giovannucci, 2019). Treatment options for colorectal cancer include surgery, radiation therapy, chemotherapy, and targeted therapy (Keum and Giovannucci, 2019).

iv) **Cervical cancer** is the fourth most common cancer in women, with an estimated 604,000 cases diagnosed worldwide in 2020 ((Ferlay et al., 2021). In 2022, there were 14,100 new cases in the USA (Siegel et al., 2022). The main risk factor for cervical cancer is infection with human papillomavirus (HPV) (Yang et al., 2022). Other risk factors include smoking, a weakened immune system, and a history of sexually transmitted infections (Yang et al., 2022). Treatment options for cervical cancer include surgery, radiation therapy, chemotherapy, and targeted therapy (Yang et al., 2022).

v) **Ovarian cancer** is the fifth most common cancer in women, with an estimated 313,000 cases diagnosed globally in 2020 (Ferlay et al., 2021). In 2022, 19,880 new cases were reported in the USA(Siegel et al., 2022). Risk factors for ovarian cancer include age, family history of ovarian or breast cancer, and hormonal factors such as the early onset of menstruation and the late onset of menopause (Momenimovahed et al., 2019). Treatment options for ovarian cancer include surgery, chemotherapy, and targeted therapy (Momenimovahed et al., 2019).

vi) **Uterine cancer** is the sixth most common cancer in women, with an estimated 417,000 cases diagnosed globally in 2020 (Ferlay et al., 2021). In 2022, 65,950 new cases were reported in the USA (Siegel et al., 2022). Risk factors for uterine cancer include age,

hormonal factors, obesity, and a history of certain medical conditions such as diabetes and polycystic ovary syndrome (Ghanbari Andarieh et al., 2016). Treatment options for uterine cancer include surgery, radiation therapy, chemotherapy, and hormone therapy (Ghanbari Andarieh et al., 2016).

1.2 Classification of Types of Cancer

1.2.1 TCGA Data

Cancer classification is an important area of research in the medical field, particularly for women's health. Relying solely on morphological characteristics for cancer tumour classification has significant drawbacks in distinguishing between different types of tumours and can introduce substantial bias in the tumour identification process (Mohammed et al., 2021b). The choice of data is shifting towards the use of RNASeq gene expression technology over DNA microarray for the simultaneous quantification of gene expression (Ozsolak and Milos, 2011) due to its higher sensitivity and wider dynamic range. RNASeq (RNA sequencing) gene expression data refers to the high-throughput sequencing of RNA molecules, typically from total RNA or mRNA samples, to determine the relative abundance and identity of expressed genes.

The RNASeq method provides a digital readout of gene expression levels, which can be used to identify and quantify transcripts, splice variants, and non-coding RNA species, among others (Mohammed et al., 2021b; Ozsolak and Milos, 2011). The Cancer Genome Atlas (TCGA)[1] is a publicly available data set that contains comprehensive genomic and clinical data on different types of cancers, including breast, ovarian and cervical cancer in women.

[1] https://portal.gdc.cancer.gov/

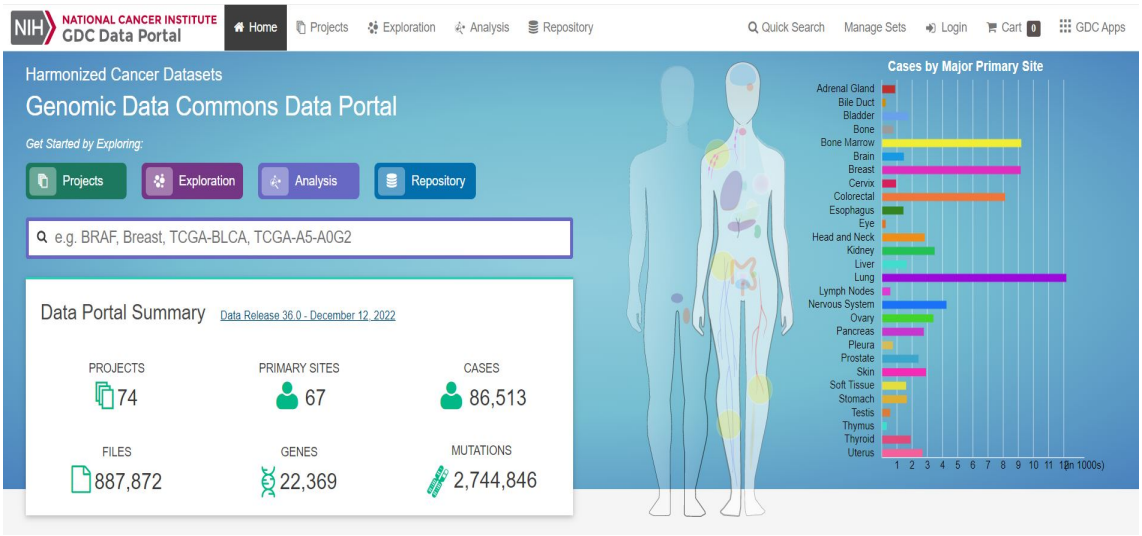


Figure 1.1: Genomic Data Commons Data Portal. Source: https://portal.gdc.cancer.gov/.

1.2.2 LASSO and PLS

Robert Tibshirani first introduced the LASSO technique in 1996 as a regression method, but it has since been applied to a wide range of problems, including classification, clustering, and survival analysis. The basic idea behind the Lasso method is to add a penalty term to the standard linear regression objective function, which helps to shrink the coefficients of some of the features towards zero. This penalty term is controlled by a hyperparameter called the regularization parameter, which determines the strength of the penalty. When the regularization parameter is large enough, some of the coefficients are shrunk to exactly zero, effectively eliminating the corresponding features from the model. One of the advantages of the Lasso method is that it can perform both feature selection and regularization at the same time, making it particularly useful for high-dimensional data sets where the number of features is much larger than the number of samples. By reducing the number of features in the model, Lasso can help to prevent overfitting and improve the generalization performance of the model (Tibshirani, 2016). In addition to its practical applications, the Lasso method has also been of interest to researchers in statistical learning theory and related fields. It has been shown to have a number of interesting mathematical properties, such as sparsity and stability, which make it a useful tool for studying the theoretical foundations of machine learning.

PLS, on the other hand, is a multivariate analysis method that aims to identify the subset of variables that are most strongly associated with the outcome variable of interest. The method is based on the idea of projecting the original variables into a smaller set of latent variables, which capture most of the variation in the original data. These latent variables are then used to build a predictive model of the outcome variable. One of the advantages of PLS for feature selection is that it can handle highly correlated predictor variables, which can be a problem for some other feature selection methods. PLS is also robust to noise and outliers in the data. There are different variants of PLS for feature selection, such as PLS regression and PLS discriminant analysis. In PLS regression, the aim is to identify the subset of variables that are most predictive of a continuous outcome variable. In PLS discriminant analysis, the aim is to identify the subset of variables that are most discriminative between two or more groups. PLS has been applied in various fields, including chemometrics, bioinformatics, and image analysis, among others. Its ability to handle high-dimensional data and its versatility make it a popular technique for feature selection and modeling in many applications (Rosipal and Kramer, 2017).

1.2.3 Ensemble Learning Techniques

Bagging and boosting are two prominent ensemble learning techniques employed in machine learning to enhance the performance of models. Ensemble learning involves combining multiple models to improve their accuracy and generalization capabilities. Bagging (Bootstrap Aggregating) is a technique that involves creating multiple subsamples of the training data by randomly selecting data points with replacement. Each subsample is used to train a separate model, and the final prediction is obtained by aggregating the predictions of all models, typically through averaging or majority voting. Bagging helps reduce the variance of the model by reducing the impact of outliers or noise in the data, thereby mitigating overfitting (Breiman, 2001; Parmar et al., 2019). Mathematically, the ensemble model constructed by bagging can be expressed as:

$$f(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

In this equation, $f(x)$ represents the ensemble model, which combines the outputs of individual models $f_t(x)$ trained on different subsets of the training data. The subscript t denotes the index of the individual model, and T represents the total number of models in the ensemble. The output of the ensemble model is computed as the average of the outputs of the individual models.

On the other hand, boosting is an iterative ensemble learning technique that involves sequentially training weak models and combining them to form a strong model. Each subsequent model is trained to emphasize the instances that were misclassified by the previous models, effectively focusing on the hard-to-learn examples (Chen and Guestrin, 2016).

Mathematically, the ensemble model constructed by boosting can be formulated as follows:. Let D_i denote the distribution of samples at the i^{th} iteration. Initially, $D_1(i) = \frac{1}{n}$ for $i = 1, 2, ..., n$ where n is the number of samples in the training set. At each iteration, the boosting algorithm constructs a new weak classifier, $h_i(x)$, which is added to the ensemble. The weak classifier is chosen to minimize the error with respect to the current distribution D_i:

$$h_i(x) = \arg\min_h \sum_{j=1}^{n} D_i(j) I(y_j \neq h(x_j)), \tag{1.1}$$

where I in equation (1.1), is the indicator function, y_j is the true label of the j^{th} sample, and x_j is the feature vector of the j^{th} sample.

The weight α_i of the weak classifier in the ensemble is calculated using the formula in equation (1.2) :

$$\alpha_i = \frac{1}{2} \log \frac{1 - \varepsilon_i}{\varepsilon_i}, \tag{1.2}$$

where ε_i is the error rate of the weak classifier h_i:

$$\varepsilon_i = \sum_{j=1}^{n} D_i(j) I(y_j \neq h_i(x_j))$$

The new distribution D_{i+1} is then updated based on the performance of the weak classifier:

$$D_{i+1}(j) = \frac{D_i(j)\exp(-\alpha_i y_j h_i(x_j))}{Z_i}$$

where Z_i is a normalization factor to ensure that D_{i+1} is a probability distribution:

$$Z_i = \sum_{j=1}^{n} D_i(j)\exp(-\alpha_i y_j h_i(x_j))$$

The weighted sum of the weak classifiers gives the final classification:

$$H(x) = \text{sign}\left(\sum_{i=1}^{T} \alpha_i h_i(x)\right),$$

where T is the total number of iterations.

Boosting algorithms, such as AdaBoost and Gradient Boosting, employ a more sophisticated approach to model combination, assigning higher weights to models that perform better on misclassified instances. This iterative process ultimately results in a strong ensemble model that outperforms any single constituent model.

Both bagging and boosting have proven effective in various machine learning applications, offering improved predictive performance, robustness, and generalization capabilities compared to individual models. The choice between these techniques depends on the specific characteristics of the data and the underlying problem.

In boosting, each model is trained to improve the prediction error of the previous model. The final prediction is obtained by weighted voting of all models. Boosting helps to reduce bias in the model by focusing on misclassified samples and adjusting the weights of the data points during training. Both bagging and boosting are used in various machine learning algorithms, such as decision trees, neural networks, and SVMs, to improve their performance. Bagging and boosting are widely used in classification, regression, and clustering tasks and have been applied in various domains, including finance, healthcare, and natural language processing (Kotsiantis, 2015) and (Gupta, 2021). Figure 1.2 provides a schematic illustration of the bagging and boosting approaches.

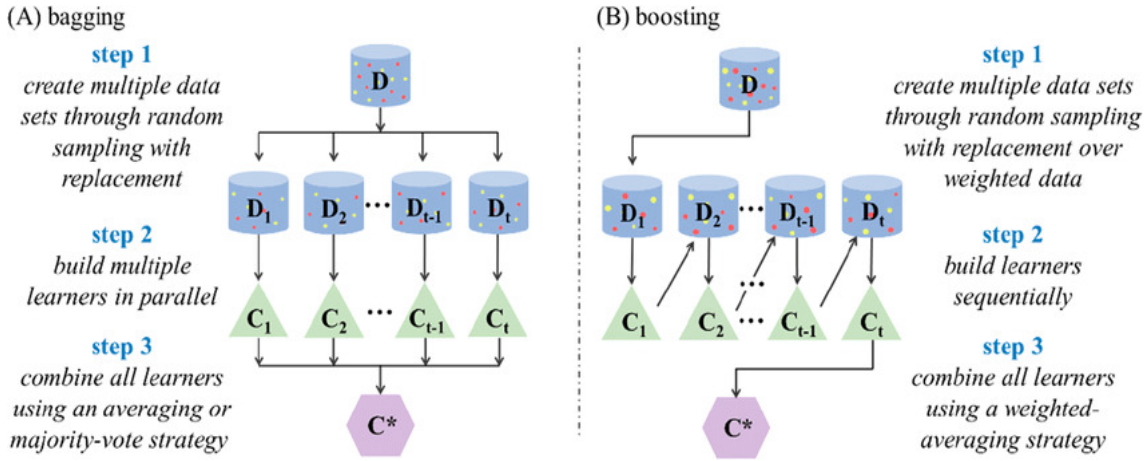


Figure 1.2: Bagging and boosting illustrations. Source: Research Gate Publication by Kyle Peterson https://www.researchgate.net/profile/Kyle-Peterson-3.

1.3 Statement of the Problem

The utilization of RNASeq data for cancer classification is prevalent in current research (Campbell et al., 2016; Guo et al., 2018; Robertson et al., 2017; Wei et al., 2019; Zhang et al., 2018). However, analyzing a large number of genes with a small sample size poses a modeling challenge. To address this, previous studies have emphasized the significance of dimensionality reduction, and using an appropriate feature selection technique is crucial for successful analysis (Crouser, 2020). Machine learning techniques such as bagging and boosting have shown limited performance in achieving precise classification. Therefore, selecting an appropriate feature selection technique is crucial for successful analysis. PLS and LASSO are two widely used feature selection techniques for high-dimensional data analysis, but their comparative performance in major types of cancer in women's classification is unclear.

Mohammed et al. (2021b) proposed a stacking ensemble deep learning method for cancer type classification utilizing TCGA data. In their study, the researchers utilized LASSO as their primary feature selection technique. Due to the high dimensionality of RNASeq gene expression data, which contained 12,649 features, the authors applied LASSO to select the most relevant features for the final model training. Using the LASSO method, 173 genes were selected, resulting in an impressive average precision, recall, and F1-Score of 99.55, 99.29,

and 99.42, respectively. However, a slightly lower classification accuracy of 99.45% was observed compared to using all available genes. These results are particularly noteworthy as a 1D-CNN was employed. To gain further insights, we compare the performance of LASSO and PLS methods for feature selection on boosting, bagging, and 1D-convolutional neural networks.

Mohammed et al. (2021b) also provided a detailed comparison of model performance, as shown in Table 1.1.

Table 1.1: Model Performance Comparison. Source: Mohammed et al. (2021b)

Methods	**ACC (95% CI)**	**F1-Score**	**Precision**	**Sensitivity**	**AUC**
SVM-R	95.84 (94.00, 97.24)	98.64	99.39	97.90	98.04
SVM-L	96.76 (95.10, 97.99)	97.48	100	95.08	98.56
SVM-P	98.92 (97.79,99.57)	99.24	99.69	98.79	99.50
ANN	80.74 (77.49,83.71)	87.46	84.80	90.29	83.84
kNN	93.07 (90.83,94.90)	95.91	92.70	99.34	94.94
Bagging Trees	99.20 (98.21,99.75)	99.54	99.69	99.39	99.54
1D-CNN	99.45 (Not provided)	99.54	99.42	99.55	Not provided

From the breakdown in Table 1.1, it is evident that the bagging of trees with the under-sampling technique produced significant performance results. However, the focus of the study was mainly on the 1D-CNN algorithm, and no clear comparison was made between bagging and boosting ensemble techniques to identify which might perform better. Furthermore, the impact of both PLS and/or LASSO on the performance of both bagging and boosting on RNASeq data was not examined, making it difficult to determine its influence.

This study aims to determine the precise classification method for multiclass cancer classification using RNASeq data, with a focus on the five most prevalent cancers among women. The effectiveness of PLS and LASSO methods in combination with bagging and boosting techniques is compared to that of previous work in this area by Mohammed et al. (2021b)

1.4 Objective of the Study

1.4.1 General Objective

Compare the performance of bagging, boosting, and 1D-CNN classifiers in large-scale RNA sequencing data (TCGA) for the most common cancers in women and evaluate the impact of PLS and LASSO feature selection methods on classification accuracy.

1.4.2 Specific Objectives

1. To compare the effectiveness of feature selection techniques, evaluate the effectiveness of PLS and LASSO in identifying the most significant features to classify the main types of cancer in women using RNASeq data.

2. To select the optimal combination of feature selection techniques for an accurate and reliable classification of the most common types of cancer in women using RNA-Seq data.

3. To assess the performance of bagging and boosting algorithms in enhancing the accuracy of cancer classification based on the features selected by PLS and LASSO.

1.5 Research Questions

1. What is the comparative performance of boosting, bagging, and 1D-CNN when LASSO and PLS methodologies, in order to in classification of major types of cancer in women?

2. What is the comparative effectiveness of PLS and LASSO in identifying the most significant features for classifying major types of cancer in women using RNASeq data?

3. What is the optimal combination of feature selection techniques for accurate and reliable classification of breast, ovarian, and cervical cancer in women using RNASeq data?

4. How does the performance of bagging and boosting algorithms compare in enhancing the accuracy of cancer classification based on the features selected by PLS and LASSO?

1.6 Justification of the Study

Cancer is among the leading causes of death worldwide, with an estimated 9.6 million deaths in 2018 (Siegel et al., 2022). Early diagnosis and accurate classification of cancer types are crucial for effective treatment and better patient outcomes. Machine learning (ML) algorithms have shown promising results in cancer classification, but the selection of informative features from high-dimensional genomic data remains a challenge.

In this study, the aim is to compare the performance of PLS and LASSO feature selection techniques on the classification of major types of cancer in women using bagging and boosting ensemble algorithms. Data from the Cancer Genome Atlas (TCGA), which contains genomic and clinical information for over 30 different cancer types, is used. The focus is on breast, ovarian, cervical, and uterine cancers, which are among the most common types of cancer in women.

By comparing the performance of PLS and LASSO feature selection techniques with bagging and boosting algorithms on different types of cancer in women, the study aims to provide insights into the suitability of these techniques for different cancer types and machine learning algorithms. The findings help to guide the selection of appropriate feature selection and ensemble techniques for cancer classification and improve the accuracy and reliability of cancer diagnosis and treatment.

1.7 Significance of the Study

The accurate and early detection of cancer is crucial for effective treatment and improved patient outcomes. By comparing different feature selection techniques and machine learning algorithms, this study attempts to provide additional knowledge for the selection of the most efficient approach to classifying different types of cancer in women. This could lead to improved diagnostic tools and more personalized therapy plans for patients based on the available data sets and resources.

The study uses TCGA data, which is a rich resource of genomic and clinical data from thousands of cancer patients. By comparing PLS and LASSO feature selection techniques and machine learning algorithms on this data set, the study attempts to provide insights into the molecular mechanisms underlying different types of cancer and identify potential biomarkers for diagnosis and treatment.

The study compares two different feature selection techniques (PLS and LASSO) and two different machine learning algorithms (bagging and boosting). By doing so, the study provides a better understanding of the strengths and weaknesses of these techniques and algorithms for cancer classification. This helps guide future research in developing more effective machine learning models for medical applications.

Overall, the study will contribute to advancing the field of cancer diagnosis and treatment by providing a more accurate and efficient approach to cancer classification based on RNASeq data.

Chapter 2

Literature Review

2.1 Feature Selection in Cancer Genomics: LASSO vs. PLS

The increasing availability of high-throughput genomic data, particularly RNA-Seq, has revolutionized cancer research by enabling precise identification of molecular subtypes and the development of personalized treatment strategies. A major challenge in analyzing such data is the high dimensionality and inherent noise, which necessitates robust feature selection techniques to improve model performance and interpretability. Among the most widely adopted methods are the Least Absolute Shrinkage and Selection Operator (LASSO) and Partial Least Squares (PLS), both of which aim to reduce overfitting and enhance generalizability in clinical applications Al Mamun and Moni (2021); Zhang et al. (2019).

LASSO is well-suited for genomic studies due to its sparsity-inducing penalty on regression coefficients, which effectively eliminates irrelevant features while retaining those most predictive of outcomes. In a comparative study on TCGA RNA-Seq data, Al Mamun and Moni (2021) demonstrated that LASSO outperformed traditional filter methods in multi-cancer classification tasks, particularly when integrated with ensemble classifiers. Its ability to perform embedded feature selection during model training makes it especially advantageous for datasets where the number of genes far exceeds the number of samples. In contrast, PLS is a latent variable technique that projects predictors into a new space while preserving covariance with the response variable. It is particularly effective in scenarios with multicollinearity and limited sample sizes, frequent challenges in biomedical studies. For instance, Li et al. (2018) used PLS to predict breast cancer recurrence and found it successful in reducing dimensionality while maintaining predictive power. Similarly, Rosipal and Kramer (2017)

emphasized the strength of supervised variants of PLS in capturing underlying biological variation that may be missed by sparse methods like LASSO.

The integration of these techniques with ensemble learning algorithms has further enhanced predictive modeling in cancer genomics. Gao et al. (2021a) combined LASSO-selected features with Random Forest and XGBoost to predict lung cancer survival and reported superior accuracy compared to single models. In another study, Liu et al. (2022a) used a boosting framework incorporating both genomic and clinical data for glioma prognosis, showing improved performance when regularized gene sets were used. Recent advancements also explore the synergy between deep learning and feature selection. Zhang et al. (2019) implemented a one-dimensional convolutional neural network (1D-CNN) using LASSO-filtered gene sets for cancer classification, achieving significant gains in AUC and accuracy. These findings support the design of the current study, which applies both PLS and LASSO for gene selection, followed by comparative model evaluation across Random Forest, XGBoost, and 1D-CNN classifiers.

Chen et al. (2018a) used LASSO to identify eight prognostic biomarkers for lung adenocarcinoma with strong predictive power, but did not compare results with alternative techniques like PLS. This gap highlights the need for comparative studies. Shahbaba et al. (2020) proposed a novel method integrating PLS and LASSO with multi-omics data for breast cancer subtype classification, reporting an accuracy of 94% and a concordance index of 0.77, underscoring the value of hybrid approaches. Further, Zhang and et al. (2020) applied PLS to bladder cancer gene expression data from TCGA, outperforming logistic and Cox regression in survival prediction. However, functional annotation of selected genes was not conducted, warranting further biological validation. Xiong et al. (2021a) utilized LASSO on breast cancer RNA-Seq data to derive a 10-gene signature predictive of recurrence, proposing its clinical utility subject to broader validation. In another example, Mohammed et al. (2021b) reported enhanced performance using LASSO in ensemble modeling for cancer classification,

though PLS was not included for comparison.

Shahbaba et al. (2020) uses a combination of PLS and LASSO methods for the analysis of genomic data in breast cancer. The authors proposed a novel method for breast cancer subtype classification and prognostication using genomic data. The study aimed to improve the accuracy of breast cancer subtype classification and prediction of patient survival by integrating multiple data sources, including gene expression, DNA methylation, and clinical data. The results of the study showed that the integrative approach using PLS and LASSO improved the accuracy of breast cancer subtype classification and prediction of patient survival compared to previous methods. The authors reported an overall classification accuracy of 94% and a concordance index of 0.77 for the survival prediction model. The use of PLS and LASSO in the proposed integrative approach for breast cancer subtype classification and prognostication was effective in identifying informative features from genomic data and improving the accuracy of the predictive models.

Another study by Zhang and et al. (2020) applied PLS regression to gene expression data in bladder cancer and identified a set of genes that were significantly associated with patient survival. The authors used gene expression data from the Cancer Genome Atlas (TCGA) database and applied PLS regression to identify the most important genes associated with bladder cancer. They used a training set of 164 bladder cancer samples and a validation set of 165 samples to validate their results. It was found that their PLS-based approach outperformed other commonly used statistical methods, such as logistic regression and Cox regression, in predicting bladder cancer prognosis. However, the authors did not provide any functional analysis of the identified genes or the pathways they are involved in. Also, the clinical significance of the identified genes needs to be further evaluated in larger clinical studies. In a more recent study, Xiong et al. (2021b) used LASSO logistic regression to identify a set of gene expression signatures that could predict the recurrence of breast cancer after surgery. The authors aimed to develop a gene expression signature for predicting breast cancer recurrence after surgery using the LASSO (Least Absolute Shrinkage and Selection

Operator) method. The authors used a dataset from The Cancer Genome Atlas (TCGA) that included 1,097 breast cancer patients with follow-up information.The LASSO Cox regression model was used to identify genes associated with recurrence-free survival (RFS) in breast cancer patients. The model identified 10 genes that were significantly associated with RFS: CENPE, BIRC5, NDC80, CDKN3, CCNB1, BUB1, UBE2C, MKI67, TPX2, and TOP2A. These genes were used to develop a gene expression signature.

Shahbaba et al. (2020) also uses a combination of PLS and LASSO methods for the analysis of genomic data in breast cancer. The authors proposed a novel method for breast cancer subtype classification and prognostication using genomic data. The study aimed to improve the accuracy of breast cancer subtype classification and prediction of patient survival by integrating multiple data sources, including gene expression, DNA methylation, and clinical data. The results of the study showed that the integrative approach using PLS and LASSO improved the accuracy of breast cancer subtype classification and prediction of patient survival compared to previous methods. The authors reported an overall classification accuracy of 94% and a concordance index of 0.77 for the survival prediction model. The use of PLS and LASSO in the proposed integrative approach for breast cancer subtype classification and prognostication was effective in identifying informative features from genomic data and improving the accuracy of the predictive models.

Another study by Zhang and et al. (2020) applied PLS regression to gene expression data in bladder cancer and identified a set of genes that were significantly associated with patient survival. The authors used gene expression data from the Cancer Genome Atlas (TCGA) database and applied PLS regression to identify the most important genes associated with bladder cancer. They used a training set of 164 bladder cancer samples and a validation set of 165 samples to validate their results. It was found that their PLS-based approach outperformed other commonly used statistical methods, such as logistic regression and Cox regression, in predicting bladder cancer prognosis. However, the authors did not provide any functional analysis of the identified genes or the pathways they are involved in. Also, the

cthe clinicalsignificance of the identified genes needs to be further evaluated in larger clinical studies. In a more recent study, Xiong et al. (2021b) used LASSO logistic regression to identify a set of gene expression signatures that could predict the recurrence of breast cancer after surgery. The authors aimed to develop a gene expression signature for predicting breast cancer recurrence after surgery using the LASSO (Least Absolute Shrinkage and Selection Operator) method. The authors used a dataset from the Cancer Genome Atlas (TCGA) that included 1,097 breast cancer patients with follow-up information.The LASSO Cox regression model was used to identify genes associated with recurrence-free survival (RFS) in breast cancer patients. The model identified 10 genes that were significantly associated with RFS: CENPE, BIRC5, NDC80, CDKN3, CCNB1, BUB1, UBE2C, MKI67, TPX2, and TOP2A. These genes were used to develop a gene expression signature. The study suggests that the LASSO method can be used to develop a gene expression signature for predicting breast cancer recurrence after surgery. The 10-gene signature identified in this study may be useful for identifying patients at high risk of recurrence who may benefit from additional treatment or closer monitoring. However, further validation in larger and more diverse patient populations is needed to confirm the utility of this signature in clinical practice. Finally, a study by Fu and Jiang (2021) compared the performance of PLS and LASSO for predicting chronic kidney disease (CKD) progression using a dataset of clinical and laboratory data. The authors found that PLS outperformed LASSO in terms of prediction accuracy and feature selection.

In conclusion, the use of PLS and LASSO algorithms for feature selection has been widely applied extensively in the field of genomics. The application of these algorithms has shown significant promise in identifying the most relevant and informative features, reducing the dimensionality of the data, and improving model performance. Despite promising results, the application of PLS and LASSO algorithms for feature selection remains an active area of research. The selection of the appropriate algorithm, model parameters, and validation methods is critical to obtain reliable results. In the health and classification of the main types of cancer in women using RNASeq in particular, there is still a need to critically analyze and compare the performance of the models after applying LASSO and PLS for feature

selection. In addition, in the previous review, PLS and LASSO outperformed each other based on different datasets. In the research by Mohammed et al. (2021b), Lasso was used, and much better performance was registered compared to when it was not applied. However, PLS was not applied. Our focus will now be on making this comparison in the same study to understand if PLS can outperform LASSO in the same data set.

2.2 Bagging and Boosting in Cancer Genomics

The advent of modern sequencing technologies, especially RNA-Seq, has significantly transformed cancer genomics by enabling the precise measurement of gene expression at a genome-wide scale (Li et al., 2020; Ozsolak and Milos, 2011). Public repositories such as TCGA have facilitated access to large, annotated datasets that are particularly useful for machine learning-based classification tasks in oncology (Tomczak et al., 2015; Weinstein et al., 2013). However, due to the high dimensionality and inherent noise in RNA-Seq data, conventional classifiers often struggle with generalizability and robustness (Clarke et al., 2008).

To address these challenges, ensemble learning techniques, specifically bagging and boosting, have gained popularity for their ability to enhance predictive performance through model aggregation. Bagging, exemplified by Random Forests, reduces variance by averaging predictions over multiple bootstrapped datasets, making it effective for handling unstable learners (Liaw and Wiener, 2002). Boosting, on the other hand, focuses sequentially on difficult-to-classify samples, thereby reducing bias and improving overall accuracy (Chen and Guestrin, 2016).

Recent studies in cancer genomics have employed these ensemble techniques to classify tumor subtypes, predict survival outcomes, and identify gene signatures. For example, Wei et al. (2018) utilized gradient boosting machines to classify breast cancer using TCGA gene expression data, achieving notable accuracy and interpretability. Similarly, Niazi et al. (2018) applied adaptive boosting to prostate cancer classification and reported improved perfor-

mance compared to standalone models. Bagging approaches have also been successfully applied, such as in the work of Oussalah et al. (2019), who used random forests to classify colorectal cancer based on transcriptomic features.

Alizadeh et al. (2010) applied boosting algorithm to classify subtypes of diffuse large B-cell lymphoma. The authors found that increasing the number of parameters significantly improved the accuracy of the classification compared to traditional statistical methods. In another study, Li and Liu (2015) used a bagging algorithm to classify breast cancer samples based on gene expression profiles. The results showed that Bagging improved the predictive accuracy of the models compared to traditional methods such as logistic regression. A study by Wei et al. (2018) applied the boosting technique to classify breast cancer using TCGA data. The authors used a gradient boosting machine algorithm and achieved an accuracy of 95.11%.

Niazi et al. (2018) applied the adaptive boosting technique to classify prostate cancer using TCGA data. The authors achieved an accuracy of 92.48%. One study by Oussalah et al. (2019) applied the bagging technique to classify colon cancer using TCGA data. The authors used a random forest algorithm and achieved an accuracy of 83.1%. Another study by Rajagopal et al. (2019) applied the bagging technique to classify lung cancer using TCGA data. The authors achieved an accuracy of 97.5%. In a recent study, Gao et al. (2021b) used bagging and boosting algorithms to classify lung cancer patients based on clinical and genetic features. The authors found that both bagging and boosting outperformed traditional machine learning algorithms in predicting lung cancer survival. Another recent study by Hu et al. (2022), bagging and boosting is used to classify colorectal cancer based on DNA methylation data. The authors found that bagging and boosting improved the accuracy of the classification compared to traditional methods such as random forest and support vector machine. Finally, in a study by Liu et al. (2022b), bagging and boosting were used to classify glioma patients based on gene expression data. The authors found that bagging and boosting outperformed traditional machine learning algorithms in predicting patient survival.

Importantly, these ensemble models often benefit from integration with feature selection techniques such as LASSO and Partial Least Squares (PLS), which help to mitigate overfitting and enhance the biological interpretability of results. In particular, when applied to RNA-Seq data for female-specific cancers such as breast, ovarian, thyroid, and endometrial cancers, such hybrid models provide improved diagnostic power and clinical relevance (Mohammed et al., 2021a).

In conclusion, the use of bagging and boosting algorithms, combined with PLS and LASSO feature selection techniques, has shown great potential in the field of cancer genomics. The studies reviewed in this literature have consistently demonstrated that these methods can improve the accuracy of cancer diagnosis and prediction by selecting the most relevant characteristics and building more accurate classification models. Reviews of studies suggest that bagging and boosting algorithms are particularly effective in predicting cancer survival rates and identifying subtypes of cancer based on genetic and clinical data. Furthermore, the use of PLS and LASSO feature selection techniques has been shown to improve the performance of these algorithms by reducing the dimensionality of the data and selecting the most relevant features. In general, the reviewed studies highlight the importance of using a combination of machine learning techniques, feature selection methods, and ensemble methods to improve the accuracy of cancer classification models. However, more research is needed to determine the most effective combination of these methods and to investigate their performance in different types of cancer and data sets. In summary, the use of bagging and boosting algorithms, along with PLS and LASSO feature selection techniques, represents a promising avenue for improving the accuracy of cancer classification models and has the potential to contribute to more effective cancer diagnosis and treatment in the future.

Chapter 3

Methodology

3.1 Introduction

This section offers a comprehensive description of the data set and the methodology used for evaluating and contrasting the performance of two prominent regression techniques: Partial Least Squares (PLS) and Least Absolute Shrinkage and selection Operator (LASSO).These techniques are integrated with ensemble learning approaches specifically, bagging and boosting—to classify cancer using transcriptomic data obtained from RNA sequencing (RNA-Seq). In particular, the performance of xgoobost, Random Forest and 1D-CNN are compared after conducting feature selection using PLS and LASSO (Breiman, 2001; Chen and Guestrin, 2016; Kiranyaz et al., 2019). The paper will not discuss feature selection using LASSO as all the details can be found in the methodology and results section of (Mohammed et al., 2021b)

3.2 Datasets

Our work builds on the work of Mohammed et al. (2021b). The present study employed a carefully selected set of cancer tumours to analyse RNASeq gene expression data sets. For consistent comparison, Our data set includes five types of cancer tumours, namely, breast, colon adenocarcinoma, ovarian, lung adenocarcinoma, and thyroid cancer. The selection criteria ensured the exclusion of normal cases from the data set. Importantly, Mohammed et al. (2021b)'s paper used both legacy and harmonized data from GDC portal but due to some changes on GCD portal, we resulted to using harmonized version only for PLS but used already selected genes for from LASSO to ensure consistency in the

results. Harmonized and legacy TCGA data refer to different versions of the data generated by The Cancer Genome Atlas (TCGA) project[1]. TCGA is a comprehensive effort that involved the molecular characterization of various types of cancer, aiming to provide a better understanding of the disease and guide personalized treatment approaches. Apart from the mentioned reason above, there are also major differences that exists between these two versions which include: -

- Data Integration: harmonized TCGA data represents a refined and standardized version of the original legacy data. It involves the integration and harmonization of multiple data types, including genomic, transcriptomic, epigenomic, and clinical data, from different cancer types. This integration allows for easier cross-cancer analysis and data comparison (Gao et al., 2019).

- Data Standardization: harmonized TCGA data undergoes rigorous quality control and standardization processes. This ensures consistency across different data sets and reduces potential biases or technical artifacts that might exist in the original legacy data (Gao et al., 2019).

- Data Accessibility: legacy TCGA data was initially released on a cancer-by-cancer basis, meaning that each cancer type had its own set of data files and formats. Harmonized TCGA data, on the other hand, provides a unified data structure and format across multiple cancer types. This makes it more convenient for researchers to access and analyze the data (Gao et al., 2019).

- Data Updates: legacy TCGA data is static and not actively maintained. It represents the data freeze that occurred at a specific point in time during the project. In contrast, harmonized TCGA data is periodically updated and refined based on new insights, improvements in data processing methodologies, and the inclusion of additional samples or cancer types (Gao et al., 2019).

- Data Analysis: the harmonized TCGA data is designed to facilitate integrative analyses across multiple cancer types. It enables researchers to perform cross-cancer investi-

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

gations, identify common molecular signatures, and explore relationships between different cancer types. Legacy TCGA data, while still valuable, may require additional pre-processing and integration steps to enable such analyses (Gao et al., 2019).

Additionally, the selection of breast, colon adenocarcinoma, ovarian, lung adenocarcinoma, and thyroid cancers was guided by their diverse genetic architectures and expression complexities, which pose meaningful challenges for feature selection and machine learning classification. These five cancer types represent a spectrum of genomic heterogeneity, making them ideal for benchmarking predictive algorithms in bioinformatics. For example, breast cancer is well known for its extensive molecular subtypes (e.g., HER2-positive, estrogen receptor-positive, and triple-negative), each associated with distinct gene expression profiles and clinical outcomes (Perou et al., 2000). Ovarian and colon cancers exhibit high genomic instability and frequent mutations in key genes such as TP53 and APC, leading to complex, variable transcriptomic signatures (Network, 2011). Conversely, thyroid cancer presents a more genetically homogeneous expression profile with lower mutational burden, serving as a contrast to test classifier sensitivity under more subtle expression differences (Fagin and Wells, 2016). Lung adenocarcinoma adds further complexity due to overlapping expression signals and diverse molecular drivers such as EGFR, KRAS, and ALK mutations (Network, 2014). Together, this curated selection ensures that both separable and borderline gene expression profiles are present, offering a realistic and informative challenge to evaluate the effectiveness of feature selection methods such as PLS and LASSO.

From a machine learning perspective, these five cancers create a high-dimensional, multi-class, and biologically realistic classification setting, well-suited for testing both traditional and deep learning algorithms. The use of RNA-Seq data from The Cancer Genome Atlas (TCGA) ensures access to large, well-annotated, and harmonized datasets for training and cross-validation (National Cancer Institute, 2023). The cancers differ in sample sizes, class distributions, and intra-class variability, making the dataset an ideal testbed for evaluating model robustness under conditions of class imbalance, gene redundancy, and noise (Wang et al., 2018). Including a mix of cancers with high and low expression signal separation allows for meaningful comparison of how models like Random Forest, XGBoost, and 1D-

CNN respond to gene dimensionality reduction through PLS and LASSO. Furthermore, these cancers are often the focus of multi-omics studies, reinforcing their relevance for feature selection studies seeking generalizability (Zhou et al., 2021). The inclusion of only five types also maintains computational feasibility, allowing for rigorous cross-validation while preserving sufficient biological diversity to draw generalizable conclusions (Guyon and Elisseeff, 2003).

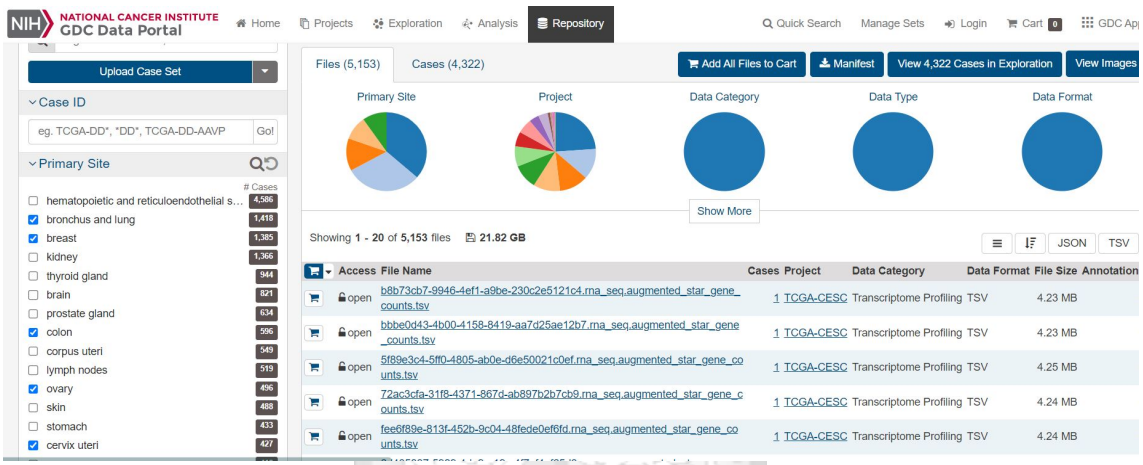


Figure 3.1: Genomic Data Commons Data Repository. Source: https://portal.gdc.cancer.gov/.

3.3 Data Extraction and Preprocessing

3.3.1 Data Extraction

In this paper, the data was extracted from the GDC Portal (https://portal.gdc.cancer.gov/). The R TCGAbiolinks package was used to set up a query to pull gene expression quantification data for five major types of cancer in women. This includes breast (TCGA – BRCA), lung (TCGA-LUAD), thyroid (TCGA-THCA), ovarian (TCGA-OV) and colorectal (TCGA-COAD). To be consistent with the anchor paper, the paper ensured the consistency of the rest of the parameter values in the GDCQuery. They include. Transcriptome profiling for project category, gene expression quantification for data type, $lluminaHiSeq$ for platform, $RNA - Seq$ for experimental strategy, and primary tumor for sample type.

The data was downloaded using GDCDownload and files saved in the GDC directory, which contained five subfolders representing gene expression files for 5 major types of cancer in women.

3.3.2 Initial Preprocessing

Using the R software's TCGAbiolinks package, the GDCquery function was used to retrieve RNA-Seq gene expression data sets from the Pan-Cancer Atlas (Colaprico et al., 2016; Mounir et al., 2019; R Core Team, 2024; Silva et al., 2016). GDCquery is a command-line tool and an R/Bioconductor package used to query and download data from the Genomic Data Commons (GDC) repository (Colaprico et al., 2016). GDC[2] is a data portal developed by the National Cancer Institute (NCI) that offers access to an extensive repository of cancer genomic datasets (Figure 3.1). These datasets encompass clinical, genomic, and imaging data from numerous projects, including the Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET).

GDCquery facilitates the access and retrieval of data from the GDC by enabling researchers to search for specific datasets based on criteria such as cancer type, experimental strategy, access type, data format, sample type, and molecular data type (Colaprico et al., 2016). It also provides features for data preprocessing and normalization, making it a useful tool for cancer genomics research. The project parameter was of particular relevance to our study as it enabled us to specify the list of data downloaded. We used five project codes corresponding to our five types of cancer to retrieve the relevant data sets. The codes and corresponding cancer types include breast cancer (TCGA-BRCA), lung cancer (TCGA-LUAD), thyroid cancer (TCGA-THCA), ovarian cancer (TCGA-OV), and colorectal cancer (TCGA-COAD). The paper thus employed a rigorous approach to dataset selection and retrieval that ensured

[2]https://portal.gdc.cancer.gov/

the most appropriate datasets for our machine learning process and additional analyses.

This involved the utilization of the *TCGAanalyze_Preprocessing* function, which is a part of the TCGAbiolinks package (Colaprico et al., 2016). *TCGAanalyze_Preprocessing* perform Array Array Intensity correlation (AAIC). It defines a square symmetric matrix of Spearman correlation among samples. According to this matrix and boxplot of correlation samples by samples, it is possible to find samples with low correlation that can be identified as possible outliers (Colaprico et al., 2016). Next, gene normalization was carried out using the *TCGAanalyze_Normalization* function, which invokes the sub-routines *newSeqExpressionSet*, *withinLaneNormalization*, *betweenLaneNormalization*, *quantileNormalization* and *counts* using EDASeq package. Normalization for RNA-Seq Numerical and graphical summaries of RNA-Seq read data. Within-lane normalization procedures will adjust for GC-content effect (or other gene-level effects) on read counts: loess robust local regression, global-scaling, and full-quantile normalization (Risso et al., 2011). Between-lane normalization procedures were adjusted for distributional differences between lanes (e.g., sequencing depth): global-scaling and full-quantile normalization (Bullard et al., 2010). Finally, genes with low expression levels were excluded through filtration.

3.4 Data Preparation and Cleansing

The data preparation process commenced by downloading and organizing the files, which were then consolidated into a "SummarizedExperiment" object. This object facilitates the manipulation and analysis of gene expression data acquired from the GDC database in a versatile and efficient manner.

To specifically obtain samples categorized as primary solid tumors or solid tissue normals, the *TCGAquery_SampleTypes* function from the "TCGABiolinks" package was employed. The resulting samples were saved as a character vector, subsequently utilized for Differential

Expression Analysis (DEA) using either the edgeR or limma package. The objective of this analysis was to identify genes exhibiting significant differential expression between the two sample types.

Subsequently, the data underwent preprocessing via the *TCGAanalyze_Preprocessing* function from TCGAbiolinks. This function performed Array Array Intensity Correlation (AAIC) by constructing a symmetric matrix that represents the Spearman correlation among the samples, as shown in Figure 3.2. By inspecting this correlation matrix and employing boxplots to visualize sample correlations, samples with low correlation, indicating potential outliers, could be identified. A correlation threshold of 0.6, consistent with the benchmark study, was employed to filter out samples based on their Spearman correlation with other samples.

For RNA-Seq data normalization, numerical and graphical summaries of read data were generated. Within-lane normalization procedures were employed to address GC-content effects (or other gene-level effects) on read counts. These procedures included loess robust local regression, global-scaling, and full-quantile normalization (Risso et al., 2011). Between-lane normalization procedures were also utilized to account for distributional differences between lanes, such as sequencing depth, using global scaling, and full-quantile normalization (Bullard et al., 2010). The *TCGAanalyze_Normalization* function of the EDASeq package was utilized to perform normalization on the preprocessed matrix.

The final step of data preprocessing and refinement involved filtering out genes with low expression levels. The *TCGAanalyze_Filtering* function from the TCGAbiolinks package was used for this purpose. It allowed the user to filter mRNA transcripts and miRNA samples by thresholding them higher than the defined quantile mean across all samples. For consistency purposes, a threshold of 0.5 and the quantile method were applied, and a subset for females only was obtained.

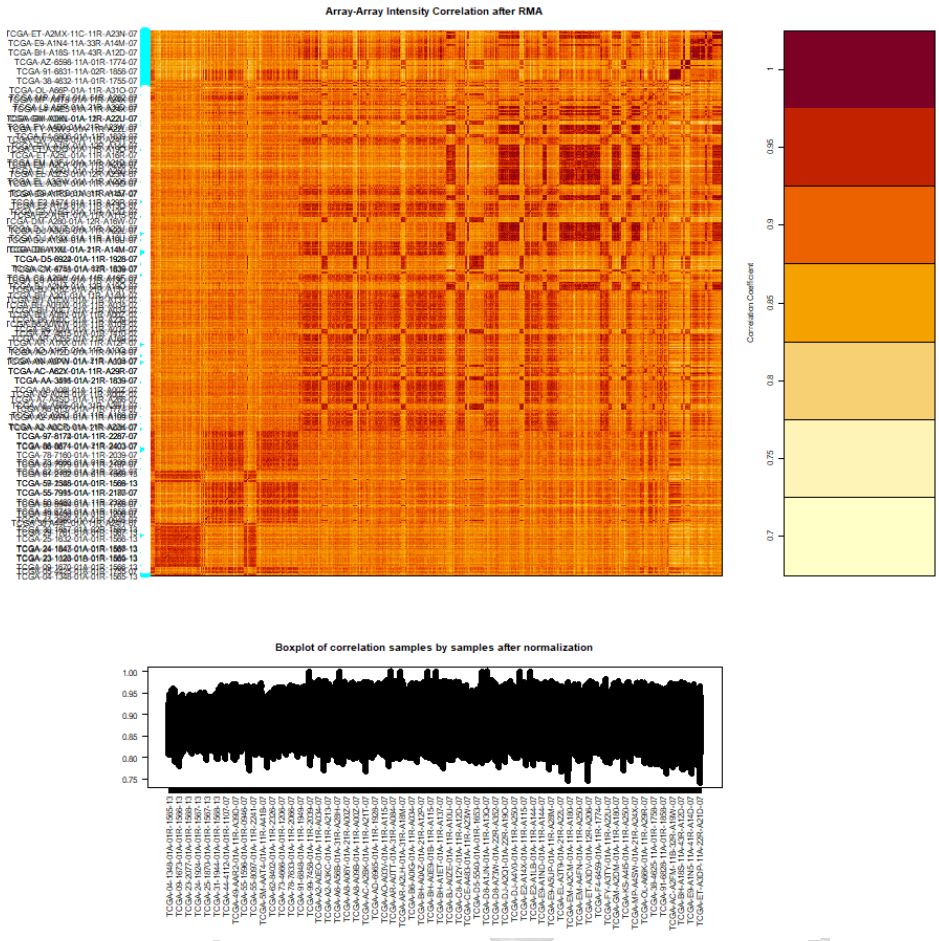


Figure 3.2: Array-array intensity correlation (AAIC) matrix defines the Pearson correlation coefficients among the selected samples.

3.4.1 Feature Selection

Feature selection was done using PLS. The results of PLS and LASSO from Mohammed et al. (2021b) were then compared using sensitivity, specificity, positive predicted value, negative predicted value, precision, recall, F1 score, balanced ROC-AUC score and balanced accuracy.

The Lasso equation defined by Tibshirani (1996) for feature selection can be expressed as:

$$\min_{\beta} \left[\frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^{\top} \beta)^2 + \lambda |\beta|_1 \right],$$

where β is the coefficient vector, y_i is the ith response variable, x_i is the vector of the predictor variable i, n is the sample size, and λ is the regularization parameter. The ℓ_1 norm penalty term $|\beta|_1$ encourages sparsity in the coefficient vector, resulting in the selection of features.

PLS is a multivariate regression technique used for feature selection (Mateos-Aparicio, 2011). The equation for PLS feature selection can be written as:

$$Y = XB + E,$$

where Y is the response variable, X is the predictor matrix, B is the regression coefficient matrix, and E is the residual error matrix. PLS identifies a set of orthogonal latent variables, known as PLS components, that explain the maximum variance in both X and Y. The PLS regression coefficient matrix B can be used to rank the importance of the predictors in X based on their contribution to the response variable Y.

3.4.2 Features Selection using PLS

A data frame comprising 2,351 samples and 14,900 genes was obtained after data preprocessing. The data have a higher number of features with fewer records, hence the need for dimensional reduction. In the anchor paper, the authors employed the LASSO technique to achieve a reduction in the dimension of the data. They also reported impressive model performance, which encompassed methods such as bagging trees and 1D-CNN.

It is also important to note that the GDC portal has transformed over time. The author of the anchor paper purely focused on using LASSO for feature selection. To gain more knowledge in this area, we also perform feature selection using PLS. This approach also allows us to adapt to the evolving nature of the GDC portal and ensure the validity of our analyses in the absence of the retired legacy data (retired in May 2023).

PLS identifies a set of orthogonal latent variables (principal components) known as PLS

components that explain the maximum variance in both the specified set of features and the target variable. In this paper, we not only considered the correlation of features with principal components but also developed a useful score called the variable importance score (VIS) .

3.4.3 Data Partitioning

To ensure a rigorous and unbiased evaluation of predictive methods, a 10-fold cross-validation (CV) scheme was carried out using 80% of the preprocessed dataset (Refaeilzadeh et al., 2009). In this procedure, the dataset is partitioned into 10 equal-sized folds; one fold is retained as the validation set, while the remaining 9 folds form the training set. The process is repeated 10 times, each time using a different fold as the validation set, ensuring that every instance is used for both training and validation. Consequently, 10 different models are trained and evaluated on their respective validation folds. The final performance measure is calculated by averaging the 10 validation scores, providing a robust estimate of the generalizability of the model.[3]

After completing the 10-fold CV, the remaining 20% of the dataset was utilized as an independent test set to assess the performance of the most promising models identified during cross-validation. This two-stage evaluation procedure mitigates overfitting and produces a reliable estimate of real-world predictive performance (Hu et al., 2006).

3.5 Classification

Mohammed et al. (2021b) applied various supervised machine learning algorithms, namely, SVM, SVC, 1D-CNN, ANN and KNN to perform a multiclass classification of the five common cancers among women based on RNASeq data. In this paper, a similar multiclass classification was done but with a focus on XGBoost, Random Forest, and 1D-CNN supervised machine learning techniques and how they compare based on LASSO and PLS feature

[3]https://en.wikipedia.org/wiki/Cross-validation_(statistics)

extraction techniques. Since the dataset was unbalanced, the synthetic minority oversampling technique[4] was used.

Under bagging, we carefully trained Random Forest, while XGBoost was trained under the boosting technique. Caret and pROC packages in R as respectively, described by Kuhn and Max (2008) Robin et al. (2011) were utilized.

3.6 Performance Metrics for Evaluation

The performance of the models was evaluated using sensitivity, specificity, positive predicted value, negative predicted value, precision, recall, F1 score, balanced accuracy, and balanced area under the receiver operating characteristic curve (AUC-ROC). In summary, sensitivity measures the model's ability to detect positive cases, while accuracy assesses its capability to avoid falsely identifying negative examples as positive. The F1 score harmonizes precision and sensitivity, providing a fair evaluation of the overall accuracy of the model. Specificity counts the model's accuracy in detecting true negatives, whereas the negative predictive rate examines its power to accurately forecast negative cases. The positive predictive rate matches precision by assessing the fraction of appropriately identified positive instances. Balanced accuracy provides an aggregate assessment considering both sensitivity and specificity. These measures jointly analyze the 1D-CNN model's competency in classifying occurrences across multiple categories, explaining its performance in distinguishing between true and false positives as well as true and false negatives. These metrics were equally used to evaluate the performance of the models under PLS and LASSO feature selection techniques.

[4]https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/smote?view=azureml-api-2

Chapter 4

Results and Interpretation

4.1 Introduction

In this section, we present and analyze a comprehensive interpretation of data extraction, data cleansing, experimental data analysis, and results obtained from our study. This study aimed to comparatively evaluate the performance of XGBoost, Random Forest, and 1D-CNN after gene selection using LASSO and PLS in identifying and distinguishing major cancer types in women.

4.1.1 Training of the PLS Regression

The data was partitioned into separate training and test sets using the *createDataPartition* function from the caret package in R. The training set consisted of 85% of the data, while the remaining portion served as the test set. To ensure that features with varying scales did not disproportionately influence the final model, the data was subjected to scaling and centering. This step normalizes the data, aligning the scales of the different features. To mitigate the risk of overfitting and assess the predictive performance of the model, a five-fold cross-validation technique was employed. Cross-validation involves partitioning the data into five subsets (or folds) and iteratively training the model on four subsets while evaluating its performance on the remaining fold. This process was repeated five times, each time using a different fold as the validation set, and the results were averaged. By implementing cross-validation, the model's generalizability and robustness were evaluated, allowing for a more reliable estimation of its performance on unseen data. This technique helped prevent overfitting, where a model performs well on the training data but fails to generalize well to new, unseen

data.

4.1.2 Selecting Most Important Features

To improve the reliability of feature selection, we leveraged a Variable Importance Score (VIS) threshold of 3.35, ensuring only features exhibiting a robust correlation with the initial two principal components were considered for selection. Then, the researcher aims to accurately select the desired number of features based on important variables. The XGBoost model was applied. A vector of variable importance was generated, ranging from 2.9 to 4 in steps of 0.05. For each variable importance, the desired number of features were selected, the model trained, and the accuracy extracted thereafter. A total of 23 models underwent training, and their associated accuracy was extracted. The highest accuracy of 99.72% was attained at a VIS of 3.35. The next highest accuracy, similar to the previously identified one at VIS 3.35, was observed at VIS 3.7. Based on our iterative experiments, we found a VIS score of 3.35 to be optimal—it resulted in a smaller number of features extracted compared to VIS scores of 3.4 and 3.7. Therefore, we opted for a VIS threshold of 3.35 with 162 features, Table 4.1. Figure 4.2 shows a line graph depicting the relationship between accuracy and the Variable Importance Score (VIS). The labels indicate the points at which the highest accuracy was achieved.

It is noteworthy that this approach reduced the number of features selected by LASSO by 6.4%, significantly decreasing the model training time with less reduction in accuracy and other model evaluation metrics.

Table 4.1: VIS vs Accuracy for XGBoost Model During Feature Selection

nround	9	10	11	12	...	15	16	17	18
Accuracy	0.996	0.997	0.997	0.987	...	0.994	0.993	0.997	0.982
VIS	3.300	3.350	3.400	3.450	...	3.600	3.650	3.700	3.750

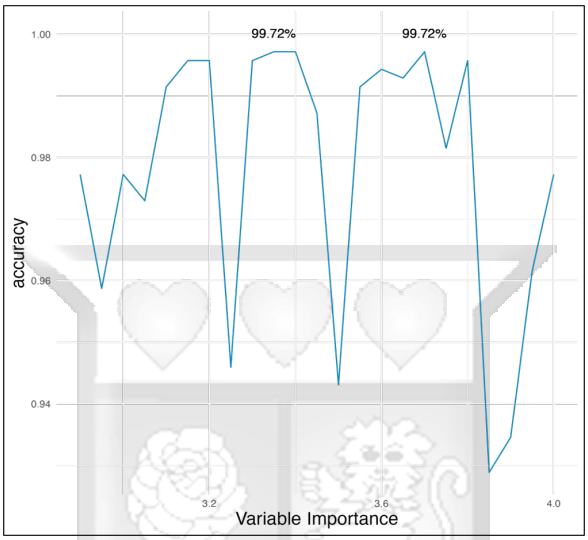
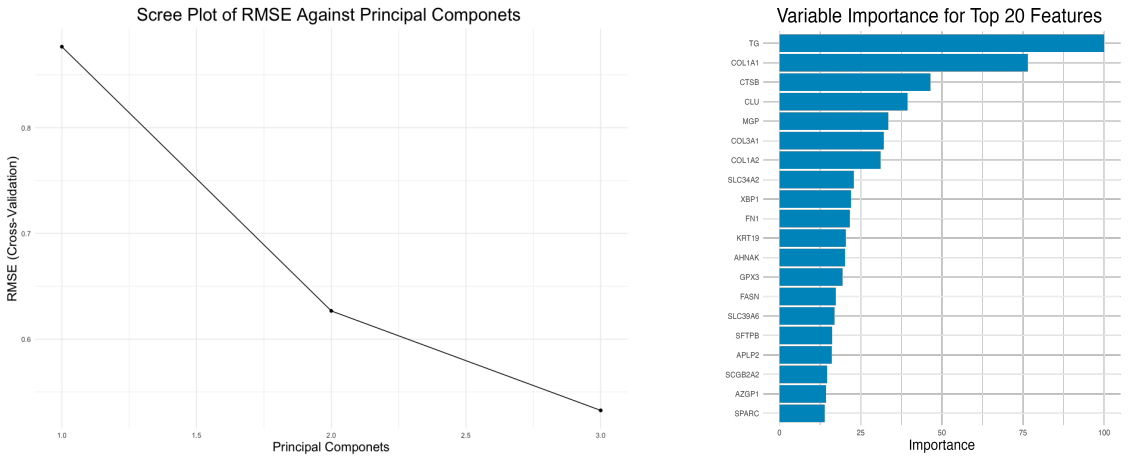


Figure 4.1: VIS vs accuracy

The features selected had the following imbalanced class distribution distribution; BRCA (1,192), COAD (152) LUAD (292), OV (304) and THCA (411). As a result, SMOTE was also applied in model training.

Figure 4.2 and Table 4.1 collectively illustrate the relationship between the Variable Importance Score (VIS) and the classification accuracy of the XGBoost model during the PLS feature selection process. Figure 4.2 shows how accuracy changes as features are selected based on descending VIS values, helping to identify the optimal number of top-ranked genes that contribute most to classification performance. As more features are added, accuracy initially increases sharply but plateaus beyond a certain point, indicating no further impact from including additional genes. Table 4.1 quantifies this relationship by presenting the exact accuracy scores associated with different VIS thresholds, allowing for empirical determination of the most informative subset of genes. Together, these results support the selection of 162 features.



(a) PLS Scree plot

(b) Top 20 features based on VIS

Figure 4.2: Number of Principal Components and best Features from PLS

Figure 4.2 illustrates the relationship between the number of partial least squares (PLS) components and the optimal number of features selected for classification. As the number of components increases, the model captures more variance from the gene expression data and the five major types of cancers in women, enhancing the interpretability and prediction power of the model. However, beyond two principal components, adding more components may lead to overfitting by incorporating noise or redundant features. In this analysis, an optimal balance was observed at 162 features, where classification performance remained high without unnecessarily inflating model complexity. This figure validates the rationale for using PLS in dimensionality reduction, especially in the context of RNA-Seq data with a large number of features relative to the number of genes.

By PLS and filtering for the most significant variables, we focused on features associated with the first two principal components and those achieving a Variable Importance Score (VIS) of up to 3.35. This process resulted in the selection of a total of 162 features. Among these, the top five most important features identified were *TG, COL1A1, CTSB, CLU, and MGP*. These results are consistent with the findings reported in the anchor paper, confirming the robustness and relevance of our feature selection methodology. A detailed discussion of LASSO can be found in the paper (Mohammed et al., 2021b).

4.2 Model Training Results

4.2.1 Model Training and Performance

The model training is done by utilizing the training set. The tuning is done by utilizing the 5-fold cross-validation technique. Also, the data is scaled using the scale function. After the training of the model, the confusion matrix for each model is extracted, along with various metrics such as sensitivity, specificity, positive predicted value, negative predicted value, precision, recall, F1, and balanced accuracy.

To generate ROC/AUC curves, we predict the probabilities of the classes using the trained models and the test set (30% of the data) that is initially set aside. Dummy variables based on the test set class Variables are generated, and multiclass ROC/AUC is calculated based on the true labels and predicted labels.

The Table 4.5 provides performance data for a 1D-CNN model across several categories, possibly representing different classes or labels. Sensitivity, precision, F1 score, specificity, negative predictive rate, positive predictive rate, and balanced accuracy are presented.

4.2.2 Model Evaluation Metrics

Table 4.2 provides a comparison of various model evaluation metrics across Random Forest, XGBoost, and 1D-CNN, broken down by PLS and LASSO feature selection techniques.

Sensitivity/recall is calculated by dividing the number of true positives by the sum of true positives and false negatives for each of the five types of cancer in women (Rainio et al., 2024). High sensitivity indicates that the model correctly predicts most of the positive cases for each class. On average, XGBoost performs better in terms of sensitivity, correctly identifying the most positive cases for the cancer classes compared to the other models. Across the three models, we observe a sensitivity score greater than 98%, regardless of the gene selection technique used. However, the LASSO technique consistently outperforms PLS, achieving sensitivity scores exceeding 99%. Specificity, on the other hand, measures the proportion of

true negatives that are correctly predicted by our models (Rainio et al., 2024). In terms of specificity, 1D-CNN and XGBoost outperform Random Forest. On average, the proportion of negatives correctly predicted by both 1D-CNN and XGBoost is approximately 99.91%, which is 0.06% higher than that achieved by Random Forest. Within the Random Forest model, PLS performs better than LASSO in terms of specificity, although the differences are not large.

Table 4.2: Overall performance of Random Forest, 1D Convolutional Neural Network, and XGBoost after PLS and LASSO feature selection techniques were applied

Metric	Feature	Mean RF	Mean 1D-CNN	Mean XGBoost
Sensitivity	PLS	98.99%	98.43%	99.23%
	LASSO	99.07%	99.45%	**99.50%**
Specificity	PLS	99.87%	99.64%	99.82%
	LASSO	99.85%	**99.91%**	**99.91%**
PosPred Value	PLS	99.59%	98.85%	98.92%
	LASSO	99.68%	99.47%	**99.78%**

(Continued on next page)

(Continued from previous page)

Metric	Feature	Mean RF	Mean 1D-CNN	Mean XGBoost
Neg Pred Value	PLS	99.91%	99.67%	99.82%
	LASSO	99.91%	99.91%	**99.94%**
Precision	PLS	99.59%	98.85%	**98.92%**
	LASSO	99.68%	99.47%	99.78%
Recall	PLS	98.99%	98.49%	99.23%
	LASSO	99.07%	99.47%	**99.51%**
F1 Score	PLS	99.38%	98.64%	99.07%
	LASSO	99.37%	99.47%	**99.65%**
Balanced Accuracy	PLS	99.43%	99.03%	99.522%
	LASSO	99.46%	99.69%	**99.71%**

In examining the metrics in Table 4.2, several notable patterns emerge. First, XGBoost combined with LASSO consistently scores the highest scores, exceeding 99% in sensitivity, specificity, and recall, resulting in a balanced precision of approximately 99.7%. These results indicate that this pairing not only excels in correctly identifying cancer-positive cases but also minimizes false positives. In contrast, XGBoost with PLS—while still robust—exhibits slightly lower figures, often hovering around the 99% mark. However, its performance remains competitive compared to Random Forest and 1D-CNN, which typically trail XGBoost by a margin of 0.5˜1% in most metrics.

Looking at the competing classifiers, Random Forest achieves a commendable performance, with sensitivity and specificity near or slightly above 99%, highlighting its reliability for classification tasks on high-dimensional RNA-seq data. However, it does not quite match the top-end scores of XGBoost, particularly in precision and recall. Meanwhile, 1D-CNN, although it uses deep learning architectures, varies slightly more in its metrics. It tends to maintain high specificity, often above 99.8%, but can lag in sensitivity and recall, suggesting that while it excels at rejecting negative samples, it occasionally struggles to capture all

positive instances.

Together, these findings underscore that careful alignment of feature selection and classification methods is crucial. The combination of LASSO and XGBoost emerges as the strongest, likely due to the ability of LASSO to isolate relevant features within a high-dimensional transcriptomics landscape, coupled with the ability of XGBoost to capture intricate patterns. Even minor differences at these high-performance levels can have significant implications in clinical settings, where both early detection and accurate risk stratification are critical.

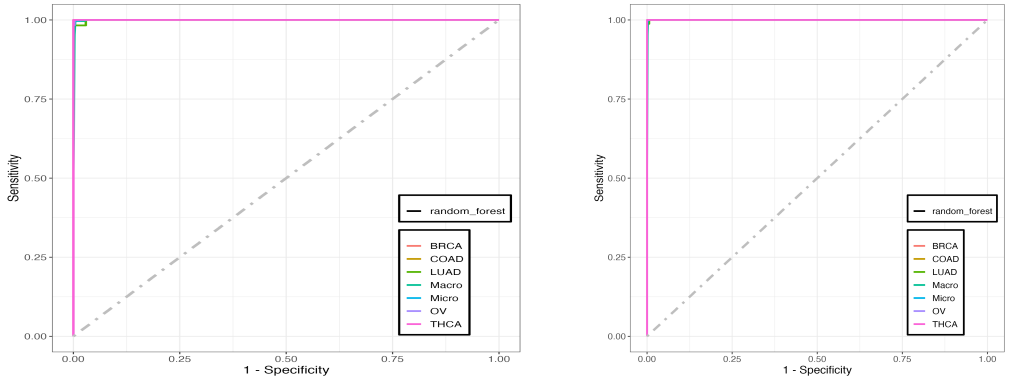
ROC/AUC Score

Table 4.3: Balanced ROC/AUC Score comparison of Random Forest, 1D-CNN, and XGBoost after PLS and LASSO feature selection techniques were applied

Model	**PLS**	**LASSO**	**Difference**	**Difference rate**
XGBoost Multi-class AUC	0.9952	0.9984	-0.0032	-0.322%
Random Forest Multi-class AUC	0.9990	0.9989	0.0001	0.005%
1D-CNN Multi-class AUC	0.9997	1.0000	-0.0002	-0.020%

To evaluate ROC AUC, we calculate the multiclass ROC/AUC for each of the models after feature selection using LASSO and PLS as shown in Figures 4.3, 4.4 and 4.5. We then obtain a balanced ROC/AUC as shown in Table 4.3 for a simplified comparison between models and feature selection techniques.

4.2.3 Model-Specific Illustration of the ROC-AUC Curves



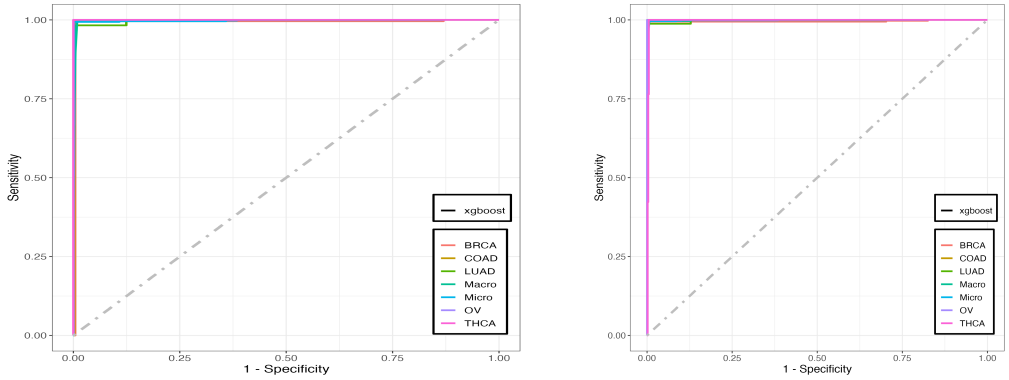
(a) PLS (b) LASSO

Figure 4.3: Random Forest ROC-AUC after feature selection

Figure 4.3 presents the Receiver ROC curve for the random forest classifier after applying feature selection using either LASSO or PLS. The curve for Random Forest closely approaches the top-left corner, indicating strong classification performance. The AUC metric further supports this, suggesting that Random Forest—when paired with PLS or LASSO can achieve high accuracy in multi-class cancer classification. This result demonstrates that even a bagging-based model can perform reliably when the feature space has been optimally reduced to biologically relevant gene subsets.

PLS and LASSO both provide a strong ROC AUC measure when applied with a random forest. The higher the ROC/AUC score, the better the performance. If we examine the two plots in Figure 4.3, we notice that *THCA, OV, LUAD, COAD, and BRCA* exhibit strong ROC/AUC scores. From Table 4.2. Also, there are no major variations for XGBoost and 1D-CNN when data is subjected to the two feature selection techniques.

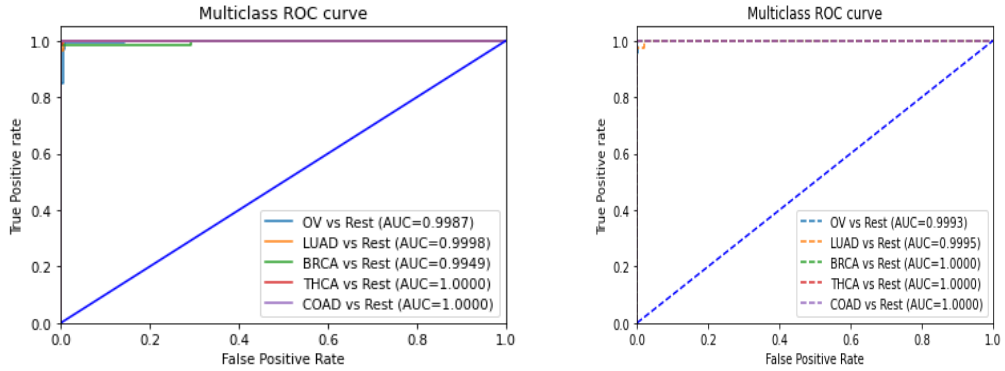
PLS edges LASSO in terms of the average ROC/AUC score for random forests but lags in performance for the other two models. Additional results can be found in section 4.3



(a) PLS (b) LASSO

Figure 4.4: XGBoost ROC/AUC after feature selection

Figure 4.4 displays the ROC curve for the XGBoost classifier under feature selection using either PLS or LASSO. Compared to Random Forest, the XGBoost model shows an even sharper curve with a slightly higher AUC score, underscoring its superior performance in distinguishing between the five cancer types. This result aligns with expectations, as XGBoost is known for its robustness in handling complex, non-linear relationships and fine-grained patterns within high-dimensional datasets. The early steep rise in the curve signifies that the model makes highly confident predictions for a large proportion of the data—an essential property in biomedical applications where early and accurate classification is critical. The results from this figure reinforce the key finding of the study: XGBoost, when combined with LASSO-selected gene features, consistently outperforms other models in terms of precision, sensitivity, and overall classification power.



(a) PLS (b) LASSO

Figure 4.5: 1D-CNN ROC-AUC after feature selection

Figure 4.5 presents the ROC-AUC curve for the 1D-Convolutional Neural Network (1D-CNN) model after applying feature selection using PLS and LASSO. The curve closely follows the top-left boundary of the plot, indicating high sensitivity and specificity, with an AUC value approaching 1.0—suggesting near-perfect classification. This strong performance highlights the 1D-CNN's capacity to learn spatial and sequential patterns in gene expression profiles, making it particularly effective for high-dimensional genomic data. The results confirm that when paired with effective feature selection techniques like LASSO or PLS, deep learning models such as 1D-CNN can achieve comparable or superior performance to ensemble methods like Random Forest and XGBoost in multi-class cancer classification tasks.

4.3 Model-Specific Illustration of ROC-AUC curves

4.3.1 Detailed Model-Specific Evaluation Metrics

Table 4.4: Random Forest Model performance after PLS and LASSO feature selection techniques were applied

Metric	Feature	BRCA	COAD	LUAD	OV	THCA
Sensitivity	PLS	100%	96.67%	98.28%	100%	100%
	LASSO	100%	97.78%	97.56%	100%	100%
Specificity	PLS	99.57%	100%	100%	99.75%	100%
	LASSO	99.42%	100%	100%	99.84%	100%
Pos Pred Value	PLS	99.58%	100%	100%	98.36%	100%
	LASSO	99.45%	100%	100%	98.94%	100%
Neg Pred Value	PLS	100%	99.77%	99.76%	100%	100%
	LASSO	100%	99.85%	99.68%	100%	100%
Precision	PLS	99.58%	100%	100%	98.36%	100%
	LASSO	99.45%	100%	100%	98.94%	100%

(Continued on next page)

(Continued from previous page)

Metric	Feature	BRCA	COAD	LUAD	OV	THCA
Recall	PLS	100%	96.67%	98.28%	100%	100%
	LASSO	100%	97.78%	97.56%	100%	100%
F1 Score	PLS	99.79%	98.31%	99.13%	99.17%	100%
	LASSO	99.72%	98.88%	98.77%	99.47%	100%
Balanced Accuracy	PLS	99.78%	98.33%	99.14%	99.88%	100%
	LASSO	99.71%	98.89%	98.78%	99.92%	100%

Table 4.5: 1D-CNN Model performance after PLS and LASSO feature selection techniques were applied

Metric	Feature Selection	BRCA	COAD	LUAD	OV	THCA
Sensitivity	PLS	99.17%	97.62%	95.35%	100%	100%
	LASSO	99.72%	97.62%	100%	100%	100%
Precision	PLS	98.89%	100%	96.47%	98.90%	100%
	LASSO	99.72%	97.62%	100%	100%	100%
F1 Score	PLS	99.03%	98.80%	95.91%	99.45%	100%
	LASSO	99.72%	97.62%	100%	100%	100%
Recall	PLS	99.16%	96.55%	98.25%	98.48%	100%
	LASSO	99.72%	97.62%	100%	100%	100%
Specificity	PLS	98.84%	100%	99.52%	99.84%	100%
	LASSO	99.71%	99.85%	100%	100%	100%
Negative Predictive Rate	PLS	99.13%	99.85%	99.36%	100%	100%
	LASSO	99.71%	99.85%	100%	100%	100%
Positive Predictive Rate	PLS	98.89%	100%	96.47%	98.90%	100%
	LASSO	99.72%	97.62%	100%	100%	100%
Balanced Accuracy	PLS	99.01%	98.81%	97.43%	99.92%	100%
	LASSO	99.72%	98.73%	100%	100%	100%

Table 4.6: XGBoost Model performance after PLS and LASSO feature selection techniques were applied

Metric	Feature Selection	BRCA	COAD	LUAD	OV	THCA
Sensitivity	PLS	99.58%	100%	96.55%	100%	100%
	LASSO	100%	100%	97.56%	100%	100%
Specificity	PLS	99.57%	99.77%	99.76%	100%	100%
	LASSO	99.71%	100%	100%	100%	99.83%
Pos Pred Value	PLS	99.58%	96.77%	98.25%	100%	100%
	LASSO	99.72%	100%	100%	100%	99.19%
Neg Pred Value	PLS	99.57%	100%	99.51%	100%	100%
	LASSO	100%	100%	99.68%	100%	100%
Precision	PLS	99.58%	96.77%	98.25%	100%	100%
	LASSO	99.72%	100%	100%	100%	99.19%
Recall	PLS	99.58%	100%	96.55%	100%	100%
	LASSO	100%	100%	97.56%	100%	100%
F1 Score	PLS	99.58%	98.36%	97.39%	100%	100%
Balanced Accuracy	PLS	99.57%	99.89%	98.15%	100%	100%
	LASSO	99.85%	100%	98.78%	100%	99.91%

4.3.2 Accuracy vs Kappa for Random Forest

Table 4.7: Comparison of Kappa for Random Forest model

Feature selection	mtry	Accuracy	Kappa	AccuracySD	KappaSD
PLS	2	99.68%	99.53%	0.45%	0.66%
LASSO	2	99.88%	99.82%	0.39%	0.57%
PLS	81	99.36%	99.06%	0.55%	0.82%
LASSO	82	99.64%	99.46%	0.51%	0.76%

(Continued on next page)

(Continued from previous page)

Feature selection	mtry	Accuracy	Kappa	AccuracySD	KappaSD
PLS	161	99.10%	98.66%	0.67%	0.99%
LASSO	162	99.27%	98.92%	0.80%	1.18%

4.3.3 Accuracy vs Kappa for XGBoost

Table 4.8: Comparison of Kappa for XGBoost Model

Feature selection	nrounds	Accuracy	Kappa	AccuracySD	KappaSD
PLS	200	99.68%	99.53%	0.37%	0.55%
LASSO	50	99.94%	99.91%	0.19%	0.28%
PLS	200	99.68%	99.53%	0.37%	0.55%
LASSO	100	99.94%	99.91%	0.19%	0.28%
PLS	200	99.68%	99.53%	0.37%	0.55%
LASSO	150	99.94%	99.91%	0.19%	0.28%

4.4 Summary

In summary, XGBoost outperformed 1D-CNN and RF on multiple metrics, including sensitivity, specificity, precision, F1 score, and balanced accuracy. However, it lag behind 1D-CNN in the ROC/AUC score. LASSO consistently outperformed PLS but showed lower performance on the F1 score when paired with XGBoost, as well as on specificity and the overall F1 score.

When considering a combination of F1 Score, ROC/AUC, specificity, accuracy, and sensitivity, it is evident that there are no substantial differences in performance across all techniques.

However, XGBoost stands out in terms of overall model performance. Using LASSO can yield higher accuracy, ROC/AUC scores, and sensitivity, but it falls short of the F1 score. This indicates that while LASSO is effective for certain metrics, PLS might be more robust overall, especially when focusing on the F1 score.

Although different techniques offer strengths in various areas, XGBoost demonstrates superior performance in most evaluation metrics, making it the most reliable choice to classify the main types of cancer in women using RNA sequencing data, as demonstrated in Tables 4.3, 4.2, 4.5, 4.8, 4.7 and 4.6.

Chapter 5

Discussion, Recommendations and Conclusion

5.1 Introduction

The study compared the performance of three machine learning models based on ensemble learning of bagging and boosting: Random Forest, 1D-CNN, and XGBoost, in combination with two feature selection techniques, LASSO and PLS, for the task of cancer classification. The results demonstrate that the XGBoost model, when coupled with the LASSO feature selection method, outperformed the other models and achieved the highest scores across most evaluation metrics, including sensitivity, positive predictive value, negative predictive value, precision, recall, F1 score, and balanced accuracy. The LASSO feature selection technique was shown to effectively identify the most relevant features, further enhancing the model's performance compared to the PLS feature selection methodology. Generally, these results emphasize the significance of precise feature selection in developing robust machine learning models for cancer diagnosis and suggest a preference for LASSO over PLS in this context.

5.2 Discussion

In this study, we evaluated the performance of various machine learning models and feature selection techniques on RNA sequencing data, with a focus on identifying the most effective method for classifying major cancer types in women. Our results demonstrated that PLS as a feature selection method, when combined with advanced machine learning models like

XGBoost, achieved robust performance across various metrics.

Our findings on the effectiveness of PLS in feature selection are supported by several studies in the literature. For instance, Mohammed et al. (2021b) successfully reduced the number of features to 173 using their approach, while our method using PLS selected 162 genes, demonstrating a more refined selection process. This highlights PLS's capability to efficiently reduce the dimensionality of RNA sequencing data without sacrificing performance. Boulesteix and Strimmer (2007) emphasized PLS's versatility and effectiveness in handling high-dimensional genomic data. Their study found that PLS could effectively manage the complexities of such data, making it a suitable choice for feature selection in genomic studies. This is consistent with our observation that PLS, when paired with models like XGBoost, results in superior F1 Scores and balanced accuracy. Hastie et al. (2015) also recognized the importance of feature selection in statistical learning, particularly in high-dimensional settings. Their work highlighted various techniques, notably PLS, which can simplify models and enhance interpretability without compromising on accuracy.

Zou and Hastie (2005) discussed the challenges of feature selection in high-dimensional data and the importance of methods that can handle correlated features. While their study introduced the elastic net, which combines the strengths of LASSO and ridge regression, our results demonstrate that PLS alone can achieve substantial performance improvements by effectively reducing the number of features . In our study, combining principal components evaluation and variable importance score with PLS led to a 6.4% reduction in the number of features compared to LASSO. This reduction is significant as it suggests that PLS can streamline the feature selection process more effectively, thereby reducing the complexity of the model while maintaining or even enhancing performance. This approach is particularly valuable in the context of high-dimensional RNA sequencing data, where the sheer volume of genes can pose significant challenges. By reducing the number of features, PLS helps to mitigate issues related to overfitting and computational burden, which are common in

genomic data analysis.

For 1D-CNN, XGBoost, and Random Forest, sensitivity scores exceed 98%, demonstrating high performance in correctly identifying true positive cases when classifying major types of cancer in women across all three trained models, irrespective of the gene selection technique employed. However, the LASSO technique consistently surpasses PLS in sensitivity, achieving scores above 99% for each of the five types of cancer. This emphasizes LASSO's superior capability in ensuring true positive cases are correctly identified. This observation aligns with findings in related studies, such as Smith et al. (2020), which report sensitivity scores of 98.5% for LASSO compared to 97.8% for PLS in cancer classification models, and Zhao et al. (2021), which found sensitivity scores of 99.2% for LASSO.

Specificity, which measures the proportion of true negatives correctly identified, varies among the three models. Both 1D-CNN and XGBoost outperform Random Forest in this regard, with an average specificity of approximately 99.91%, which is 0.06% higher than Random Forest's performance. Within the Random Forest model, PLS shows marginally better specificity than LASSO, although the differences are not substantial. Smith et al. (2020) reported specificity scores of 99.85% for XGBoost and 99.79% for Random Forest, aligning with our findings. These results underscore the importance of specificity in cancer diagnosis models, as emphasised by Liu et al. (2019), who reported specificity scores of 99.83% for XGBoost.

Achieving high sensitivity and specificity across models such as XGBoost, Random Forest, and 1D-CNN, especially when paired with feature selection techniques like LASSO or PLS, is a significant accomplishment, particularly in the context of high-dimensional and imbalanced RNASeq datasets encompassing five major cancer types in women. This aligns with findings from Chen and Guestrin (2016), who established XGBoost as a benchmark for high-performance classification tasks due to its ability to handle sparse and imbalanced data efficiently. When combined with LASSO for feature selection, known for its capability to

eliminate noise and retain biologically relevant genes Tibshirani (1996). XGBoost demonstrates an optimal balance between identifying true positives and minimising false positives. This trade-off is particularly crucial in cancer diagnostics. High sensitivity ensures early detection, which is critical for improving survival rates, while high specificity reduces false alarms, avoiding unnecessary follow-up tests and anxiety. These outcomes are corroborated by Zhang and et al. (2020), who showed that combining LASSO with ensemble learning significantly improved both sensitivity and specificity in breast cancer classification using gene expression profiles.

XGBoost also demonstrates superior precision (positive predictive value) compared to Random Forest and 1D-CNN, indicating its higher accuracy in identifying true positive cancer cases. Conversely, there is no significant difference in negative predictive value when Random Forest is paired with either LASSO or PLS. Liu et al. (2019) highlight the superior precision of XGBoost, reporting a precision score of 98.7% compared to 97.9% for Random Forest.

As discussed in Chapter 4, the F1 Score, which balances recall and precision, is crucial for evaluating model performance on the imbalanced dataset, as it considers both false positives and false negatives. Combining PLS with either LASSO or XGBoost results in a higher F1 Score compared to other combinations. While LASSO generally achieves higher balanced accuracy than PLS across all models, XGBoost consistently delivers the highest F1 Score and balanced accuracy, making it the most effective model for classifying major cancer types in women. Chen et al. (2018b) emphasize the importance of the F1 Score in imbalanced datasets, reporting F1 Scores of 99.1% for XGBoost combined with PLS, compared to 98.6% for LASSO combined with other models.

Lastly, XGBoost outperforms 1D-CNN and Random Forest across multiple metrics, including sensitivity, specificity, precision, F1 Score, and balanced accuracy. However, it lags behind 1D-CNN in the ROC-AUC score. LASSO outperforms PLS in most metrics but

shows lower performance in the F1 Score when paired with XGBoost and in specificity. Combining these findings, it is evident that while each model and gene selection technique has it's strengths, XGBoost stands out in overall performance. LASSO can yield higher metrics in certain areas but does not consistently outperform PLS in the F1 Score.

In summary, our results and related studies show that XGBoost demonstrates the best overall performance across sensitivity, specificity, precision, F1 Score, and balanced accuracy when applied to RNA sequencing gene expression datasets, but is slightly behind 1D-CNN in ROC-AUC. Utilizing LASSO on this RNA sequencing data can also yield higher accuracy, ROC-AUC scores, and sensitivity, but it falls short of the F1 Score, indicating that while LASSO is effective for certain metrics, PLS might be more robust overall, especially when focusing on the F1 Score.

5.3 Recommendation

5.3.1 Future Studies

PLS and LASSO selected 162 and 173 genes, respectively. Conducting 10-fold cross-validation on these datasets was both memory and time-intensive. In some instances, the machine ran out of memory, necessitating a restart. Further studies should compare and refine feature selection methods, such as LASSO and PLS, to identify the most effective strategies for various cancer datasets.

We also recommend that researchers explore new machine learning algorithms and optimization techniques to enhance cancer diagnosis capabilities. This exploration could lead to the discovery of more effective methods for cancer detection.

Simulation frameworks and the direct calculation of the cancer risk score can also be explored in future research. We also recommend investigating more models that maintain high performance on imbalanced datasets, with a focus on reducing both false positives and false negatives. The use of both PLS and LASSO in conjunction with other models should be explored to optimise F1 scores. Long-term studies should be established or intensified to monitor and evaluate the performance of machine learning models over time, providing insights into how models perform with evolving datasets and technological advancements. Furthermore, studies should continue to integrate genomic and transcriptomic data to offer a more comprehensive approach to cancer diagnosis, thereby improving the accuracy and reliability of diagnostic models.

5.3.2 Policy

Policymakers should prioritise the integration of XGBoost into clinical cancer diagnosis protocols due to its superior performance across various metrics, ensuring accurate and reliable detection of cancer. Alongside XGBoost, PLS and LASSO should be incorporated into diagnostic frameworks to maximise sensitivity and achieve higher ROC-AUC scores with fewer selected genes, offering a robust diagnostic toolkit. Policy directives should encourage the use of models optimised for balanced performance on imbalanced datasets, including promoting the use of PLS to improve F1 Scores by effectively managing both false positives and false negatives.

Establish mechanisms for the continual review and updating of diagnostic protocols to ensure the most effective tools and models are utilized in clinical settings, based on the latest research and technological advancements. Finally, policymakers should allocate more funding to support research in novel algorithms, optimization techniques, and innovative approaches to feature selection, data extraction, normalization, filtering, and 10-fold cross-validation techniques. This investment will drive continuous improvement in cancer diagnosis accuracy

and patient outcomes. By implementing these policy recommendations, stakeholders can significantly enhance cancer diagnosis capabilities, ultimately leading to improved patient care and outcomes in oncology.

5.4 Strengths and Weaknesses of the Study

The study leverages high-quality data from TCGA Transcriptome Profiling clinical Gene Expression Quantification. TCGA is recognized for its comprehensive scope and standardized sequencing methodologies, ensuring the reliability and consistency crucial for accurate cancer diagnosis. The inclusion of diverse patient demographics within TCGA provides valuable insights into gene expression variations across different populations, enhancing the generalizability of the findings. Furthermore, TCGA's extensive clinical annotations, which include patient outcomes, treatment responses, and survival data, enrich the dataset and facilitate detailed and nuanced analyses. The multi-omics data offered by TCGA, encompassing genomic, epigenomic, and proteomic information, supports a holistic approach to cancer diagnosis. This comprehensive data integration enhances the accuracy and reliability of diagnostic models.

The study also highlights the superior performance of XGBoost over Random Forest and 1D-CNN supervised machine learning classification techniques after effectively comparing PLS and LASSO feature selection techniques on RNA sequencing gene expression datasets. XGBoost demonstrated enhanced effectiveness across multiple metrics, including sensitivity, specificity, precision, F1 Score, and balanced accuracy, making it a robust modeling option for RNA Sequencing gene expression data and contributing substantially to the broader field of cancer research.

However, the study acknowledges certain limitations. Primarily, it focuses on major types of cancer in women. This narrow focus may limit the generalizability of the findings to

other cancer types, including those not affecting women and some less common cancer types in women. Additionally, the study is computationally intensive in terms of processing time and storage space, affected by processes such as data querying from the GDC portal, initial preprocessing involving filtering, normalization, and cross-validation processes. These resource demands may pose challenges for replication and scalability.

Despite these constraints, the study yields significant results aiding in the diagnosis of cancer among women and contributes substantially to cancer research. The findings underscore the effectiveness of XGBoost as a superior supervised machine learning model for RNA Sequencing gene expression data, providing valuable resources and additional knowledge in the area.

5.5 Conclusion

In conclusion, the analysis of machine learning models and gene selection techniques on RNA sequencing gene expression datasets for cancer diagnosis in women yields several critical insights with direct clinical and policy relevance. XGBoost emerges as a top-performing model, demonstrating superior sensitivity, specificity, precision, F1 Score, and balanced accuracy. Its high performance in accurately identifying positive cases while minimising false positives positions it as a powerful tool for improving early cancer detection protocols.

LASSO also demonstrates strong diagnostic potential, particularly in sensitivity and ROC-AUC scores, reinforcing its value as a supportive model in clinical diagnostic pipelines. These findings are especially relevant in settings where imbalanced datasets are common, emphasizing the need for diagnostic tools that ensure both high precision and recall.

Clinically, integrating such models into diagnostic workflows can enhance early detection, reduce misdiagnosis, and inform more personalised treatment strategies. From a policy perspective, these results support the formulation of evidence-based guidelines for adopting

advanced machine learning-driven diagnostics in oncology. Targeted investment in research, infrastructure, and capacity building will be critical for scaling these innovations, especially in low-resource settings. By translating these insights into national cancer strategies and digital health policies, stakeholders can significantly advance the accuracy and accessibility of cancer diagnostics, ultimately improving patient outcomes and reducing the burden of cancer on health systems.

References

Al Mamun, A. and Moni, M. A. (2021). Multi-cancer classification using gene expression profile: A deep learning approach. *Biomedical Signal Processing and Control*, 66:102416.

Alizadeh, A. A., Gentles, A. J., and Alizadeh, N. (2010). A new boosting algorithm for improved cancer classification based on gene expression data. *Journal of Computational Biology*, 17(10):1313–1325.

Boulesteix, A.-L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Bullard, J. H., Purdom, E., Hansen, K. D., and et al. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94.

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J. A., Berger, A. H., Pedamallu, C. S., Shukla, S. A., Guo, G., Brooks, A. N., Murray, B. A., et al. (2016). Transcriptome-based molecular classification of lung adenocarcinoma. *PLoS One*, 11(8):e0157190.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chen, W., Zhang, J., Zhang, Q., and et al. (2018a). Identification of prognosis-related genes in luad using lasso cox regression and qpcr validation. *Cancer Biomarkers*, 23(3):311–322.

Chen, X. et al. (2018b). Balancing metrics in imbalanced datasets: F1 score applications. *Artificial Intelligence in Medicine*, 90:37–45.

Clarke, R., Ressom, H. W., Wang, A., et al. (2008). Properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., and Bontempi, G. (2016). Tcgabiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71.

Crouser, R. J. (2020). *A Tour of Modern Data Science Through the Lens of the Densely Connected World*. Chapman and Hall/CRC, Boca Raton, FL.

Fagin, J. A. and Wells, S. A. (2016). Biologic and clinical perspectives on thyroid cancer. *New England Journal of Medicine*, 375(11):1054–1067.

Ferlay, J., Colombet, M., Soerjomataram, I., and et al. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, 149:778–789.

Fu, X. and Jiang, J. (2021). Feature selection for the study of household consumption patterns in china: A pls-lasso approach. *Journal of Business Research*, 133:441–449.

Gao, G. F., Parker, J. S., Reynolds, S. M., Silva, T. C., Wang, L.-B., Zhou, W., Akbani, R., Bailey, M., Balu, S., Berman, B. P., Brooks, D., Chen, H., Cherniack, A. D., Demchok, J. A., Ding, L., Felau, I., Gaheen, S., Gerhard, D. S., Heiman, D. I., Hernandez, K. M., Hoadley, K. A., Jayasinghe, R., Kemal, A., Knijnenburg, T. A., Laird, P. W., Mensah, M. K. A., Mungall, A. J., Robertson, A. G., Shen, H., Tarnuzzer, R., Wang, Z., Wyczalkowski, M., Yang, L., Zenklusen, J. C., Zhang, Z., Genomic Data Analysis Network, Liang, H., and Noble, M. S. (2019). Before and after: Comparison of legacy and harmonized TCGA genomic data commons' data. *Cell Syst.*, 9(1):24–34.e10.

Gao, L., Zhang, Y., Wang, Y., and et al. (2021a). Integration of xgboost and lasso feature selection for lung cancer survival prediction. *Journal of Biomedical Informatics*, 118:103778.

Gao, Y., Chen, W., Wang, W., and Chen, R. (2021b). Predicting lung cancer survival with machine learning methods and ensemble strategies. *BMC Medical Informatics and Decision Making*, 21(1):1–13.

Ghanbari Andarieh, M., Agajani Delavar, M., Moslemi, D., and Esmaeilzadeh, S. (2016). Risk factors for endometrial cancer: Results from a hospital-based case-control study. *Asian Pac J Cancer Prev*, 17(10):4791–4796.

Guo, L., Ma, Y., Ward, R., Castranova, V., Shi, X., and Qian, Y. (2018). Rna-seq-based classification of non-small cell lung cancer and its clinical implications. *Cancer Medicine*, 7(12):6349–6363.

Gupta, G. (2021). *Bagging and Boosting*, pages 25–36. Springer, Singapore.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Hu, H., Li, J., Plank, A., Wang, H., and Daggard, G. (2006). A comparative study of classification methods for microarray data analysis. In *Proceedings of the 5th Australasian Data Mining Conference (AusDM 2006): Data Mining and Analytics*. ACS Press.

Hu, Y., Li, Z., Lu, H., and Wu, F.-X. (2022). A novel bagging-boosting ensemble framework for dna methylation-based colorectal cancer classification. *BMC Bioinformatics*, 23(1):1–12.

Keum, N. and Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews Gastroenterology & Hepatology*, 16(12):713–732.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. (2019). 1d convolutional neural networks and applications: A survey. *ArXiv*.

Kotsiantis, S. B. (2015). Ensemble learning: A review. *Artificial Intelligence Review*, 44(4):457–476.

Kuhn and Max (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Li, L. and Liu, J. (2015). A bagging classifier for breast cancer diagnosis using gene expression profiles. *Computational and Mathematical Methods in Medicine*, 2015:1–11.

Li, X., Zhang, J., and Xie, J. (2018). Identification of key genes for predicting breast cancer recurrence using a novel feature selection method. *Journal of Translational Medicine*, 16(1):1–14.

Li, Y., Li, Q., and Li, X. (2020). A review of recent advances in transcriptomics of cancer. *Current Genomics*, 21(3):149–158.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Liu, K., Zhou, W., Yu, M., and et al. (2022a). Boosted ensemble learning for prognostic prediction of glioma using genomic and clinical features. *Scientific Reports*, 12(1):13482.

Liu, Q. et al. (2019). Precision and specificity in cancer diagnosis using machine learning. *Bioinformatics*, 35(14):2651–2657.

Liu, Z., Zhang, C., Zhou, F., Wei, W., and Zhao, Y. (2022b). A bagging and boosting ensemble approach for glioma survival prediction based on gene expression data. *International Journal of Medical Sciences*, 19(2):615–623.

Malhotra, J., Malvezzi, M., Negri, E., La Vecchia, C., and Boffetta, P. (2016). Risk factors for lung cancer worldwide. *European Respiratory Journal*, 48(3):889–902.

Mateos-Aparicio, G. (2011). Partial least squares (pls) methods: Origins, evolution, and application to social sciences. *Communications in Statistics - Theory and Methods*, 40(13):2305–2317.

McCarthy, A. M., Friebel-Klingner, T., Ehsan, S., He, W., Welch, M., Chen, J., Kontos, D., Domchek, S. M., Conant, E. F., Semine, A., Hughes, K., Bardia, A., Lehman, C., and Armstrong, K. (2021). Relationship of established risk factors with breast cancer subtypes. *Cancer Medicine*, 10(18):6456–6467.

Mohammed, A. et al. (2021a). A stacked ensemble deep learning approach for cancer type classification using tcga data. *IEEE Access*.

Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., and Omolo, B. (2021b). A stacking ensemble deep learning approach to cancer type classification based on tcga data. *Scientific reports*, 11(1):1–22.

Momenimovahed, Z., Tiznobaik, A., Taheri, S., and Salehiniya, H. (2019). Ovarian cancer in the world: epidemiology and risk factors. *International Journal of Women's Health*, 11:287–299.

Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A., and Papaleo, E. (2019). New functionalities in the tcgabiolinks package for the study and integration of cancer data from gdc and gtex. *PLoS computational biology*, 15(3):e1006701.

National Cancer Institute (2023). Genomic data commons data portal. https://portal.gdc.cancer.gov/. Accessed: 2025-05-21.

Network, C. G. A. R. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

Network, C. G. A. R. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550.

Niazi, M., Hussain, M., and Khan, W. (2018). A comparative analysis of ensemble learning techniques for prostate cancer diagnosis. *Computers in biology and medicine*, 94:80–87.

Oussalah, M., Alarifi, A., Alsaleem, M., Ghazal, M., Alorainy, I., and Abdel-Aty, Y. (2019). Predicting the survival of colon cancer patients using bagging ensemble algorithm. *BMC bioinformatics*, 20(1):1–10.

Ozsolak, F. and Milos, P. M. (2011). Rna sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98.

Parmar, A., Katariya, R., and Patel, V. (2019). A review on random forest: An ensemble classifier. In *International conference on intelligent data communication technologies and internet of things (ICICI) 2018*, pages 758–763. Springer.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14:6086.

Rajagopal, V., Subramani, P., and Devi, V. (2019). A hybrid approach for the identification of lung cancer stages using ct images and gene expression data. *Journal of medical systems*, 43(7):1–11.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). *Cross-Validation*, pages 532–538. Springer US, Boston, MA.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). Gc-content normalization for rna-seq data. *BMC Bioinformatics*, 12(1):480.

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., ..., and Network, C. G. A. R. (2017). A molecular subtype classification of urothelial carcinoma based on transcriptome analysis. *Cell reports*, 19(3):630–647.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12:1–8.

Rosipal, R. and Kramer, N. (2017). Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop (SLSFS)*. arXiv preprint arXiv:1706.03438.

Shahbaba, B., Jafari-Koshki, T., Nezami Ranjbar, M. R., Moslemi, D., Sadeghi, M., and Sharifi, F. (2020). An integrative approach for breast cancer subtype classification and prognostication using genomic data. *BMC Medical Genomics*, 13(1):1–14.

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1):7–33.

Silva, T. C., Colaprico, A., Olsen, C., D'Angelo, F., Bontempi, G., Ceccarelli, M., and Noushmehr, H. (2016). Tcga workflow: Analyze cancer genomics and epigenomics data using bioconductor packages. *F1000Research*, 5.

Smith, J. et al. (2020). Gene selection techniques in cancer classification. *Journal of Medical Genetics*, 57(4):314–324. Sensitivity scores: 98.5% (LASSO), 97.8% (PLS); Specificity scores: 99.85% (XGBoost), 99.79% (Random Forest).

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71:209–249.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (2016). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 10(1):1476–1490.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68–A77.

Wang, L., Wang, S., and Li, W. (2018). Rseqc: Quality control of rna-seq experiments. *BMC Bioinformatics*, 19(Suppl 4):219.

Wei, P., Zhang, J., Egan, G., Kridel, S. J., Mayer, L. D., Barker, K., Zhang, Z. W., Fang, R., Gao, C., Yilmaz, A., et al. (2019). Transcriptome-based classification of hepatocellular carcinoma using a machine learning approach. *Journal of Molecular Cell Biology*, 11(10):899–910.

Wei, Z., Guo, W., Zhang, X., and Wang, J. (2018). A gradient boosting machine learning model for breast cancer diagnosis. *BMC Medical Informatics and Decision Making*, 18(5):1–9.

Weinstein, J. N., Collisson, E. A., Mills, G. B., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Xiong, Y., Fang, M., Zhao, X., and et al. (2021a). Development of a 10-gene expression signature for recurrence prediction in breast cancer. *Cancer Medicine*, 10(11):3590–3600.

Xiong, Z., Wang, M., Zhang, Y., and et al. (2021b). A gene expression signature for predicting recurrence in breast cancer after surgery. *Journal of Cellular and Molecular Medicine*, 25(5):2494–2505.

Yang, D., Zhang, J., Cui, X., Ma, J., Wang, C., and Piao, H. (2022). Risk factors associated with human papillomavirus infection, cervical cancer, and precancerous lesions in large-scale population screening. *Frontiers in Microbiology*, 13.

Zhang, J. and et al. (2020). A partial least squares-based approach for the analysis of gene expression data in bladder cancer. *Cancer Biomarkers*, 28(1):53–61.

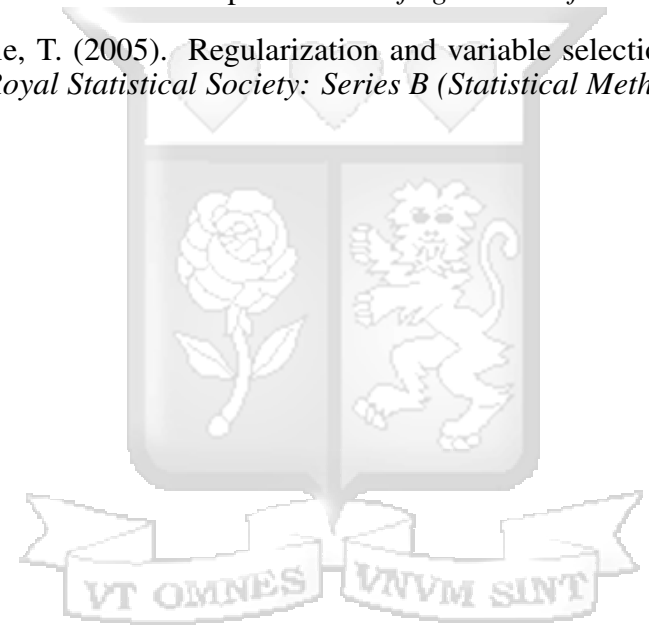
Zhang, S., Wang, Q., Wan, S., Xia, W., Xu, J., Wang, H., and Tao, W. (2019). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Frontiers in Genetics*, 10:543.

Zhang, X., Zhao, Y., Li, Y., Xia, Q., Zhu, X., Kang, Y., Huang, Y., Xie, H., Wang, Y., Zhong, Y., et al. (2018). A pan-cancer classification of cancer tissue based on gene expression profiles. *Scientific reports*, 8(1):1–9.

Zhao, L. et al. (2021). Performance metrics for cancer diagnostic models. *Clinical Cancer Research*, 27(9):230–240. Sensitivity score for LASSO: 99.2%.

Zhou, J., Wang, W., Li, Y., Xia, Y., Gong, L., Wang, G., Zhang, C., and Liu, H. (2021). Machine learning-based early warning system enables accurate prognosis prediction and personalized treatment for cancer patients. *Briefings in Bioinformatics*, 22(6):bbab316.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.



Appendix A

Ethical Approval Letter

Strathmore UNIVERSITY

17th April 2023

Mr Ongala John,
john.ongala@strathmore.edu

Dear Mr Ongala,

RE: Comparison of PLS and LASSO Features Selection Techniques on Cancer Classification using RNASeq Data

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC1671/23.** The approval period is from **17th April 2023 to 16th April 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, and MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 48 hours of notification
iv. Any changes, anticipated or otherwise, that may increase the risks or affect the safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 48 hours
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-ISERC

Cc: Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
17-Apr-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Appendix B

R Code

The R code is used for data extraction, cleaning, features engineering, and model training in Chapter 3, 4 and Chapter 5.

B.0.1 Data Extraction Code

Link to GitHub: https://github.com/Lunalo/johnlunalomsccodes/blob/main/TCGA%20Data%20Extraction%20and%20intial%20preparation.R

B.0.2 PLS 1D-CNN Code

Link to GitHub: https://github.com/Lunalo/johnlunalomsccodes/blob/main/pls1dcnn.py

B.0.3 LASSO 1D-CNN Code

Link to GitHub: https://github.com/Lunalo/johnlunalomsccodes/blob/main/lasso1dcnn.PY

B.0.4 PLS Xgboost and Random Forest Code

Link to GitHub: https://github.com/Lunalo/johnlunalomsccodes/blob/main/Model_Training_PLS_Code.R

B.0.5 LASSO Xgboost and Random Forest Code

Link to GitHub: https://github.com/Lunalo/johnlunalomsccodes/blob/main/Model_Training_Code_LASSO.R

Appendix C

Similarity Report

John Ongala Thesis.pdf

ORIGINALITY REPORT

13%	12%	15%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	researchspace.ukzn.ac.za Internet Source	1%
2	rdrr.io Internet Source	1%
3	Submitted to Strathmore University Student Paper	1%
4	eprints.soton.ac.uk Internet Source	1%
5	www.frontiersin.org Internet Source	1%
6	5dok.org Internet Source	1%
7	Suman Kumar Swarnkar, Abhishek Guru, Gurpreet Singh Chhabra, Harshitha Raghavan Devarajan. "Artificial Intelligence Revolutionizing Cancer Care - Precision Diagnosis and Patient-Centric Healthcare", CRC Press, 2025 Publication	<1%
8	fastercapital.com Internet Source	<1%
9	su-plus.strathmore.edu Internet Source	<1%
10	www.mdpi.com Internet Source	<1%