ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

An Ensemble Method for Cancer Classification and Identification of Cancer-Specific Genes from Genomic Data

A thesis submitted in partial fulfillment of the requirements
for the degree of B. Sc. in Software Engineering

Authors

Siana Rizwan, 180042105

Farzana Tabassum, 180042119

Sabrina Islam, 180042122

Co-Supervisor

Tasnim Ahmed

Lecturer

Supervisor

Tareque Mohmud Chowdhury

Assistant Professor

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

A Subsidiary Organ of the Organization of Islamic Cooperation (OIC)

Dhaka, Bangladesh

May 20, 2023

Academic Year: 2021-2022

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out under the supervision of Tareque Mohmud Chowdhury, Assistant Professor and co-supervision of Tasnim Ahmed, Lecturer of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Siana Rizwan

_____

Student ID - 180042105

Farzana Tabassum

_____

Student ID - 18042119

Sabrina Islam

_____

Student ID - 180042122

Approved By:

Supervisor:

Tareque Mohmud Chowdhury
Assistant Professor
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT), OIC

Co-Supervisor:

Tasnim Ahmed
Lecturer
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT), OIC

Acknowledgement

We would like to express our grateful appreciation for **Tareque Mohmud Chowdhury**, Assistant Professor, Department of Computer Science & Engineering, IUT for being our adviser and mentor. His motivation, suggestions, and insights for this research have been invaluable. Without his support and proper guidance, this research would never have been possible. His valuable opinion, time, and input were provided throughout the thesis work, from the first phase of thesis topics introduction, subject selection, proposing algorithm, and modification till the project implementation and finalization which helped us properly do our thesis work. We are really grateful to him.

We are also grateful to **Tasnim Ahmed**, Lecturer, Department of Computer Science & Engineering, IUT for his valuable inspection and suggestions on our proposal of Cancer Gene Classification.

Dedicated this work fully to our cherished friends, family, and teachers, who have been the cornerstones of our paths with their unwavering support and motivation

Abstract

Classifying cancer using gene expression can be an important tool for understanding the specific characteristics of a patient's cancer and for guiding the most appropriate treatment approach. By identifying the specific genes that are involved in the development and progression of a particular cancer, it may be possible to tailor treatment to target those genes and improve outcomes for the patient. In addition, by understanding the genetic makeup of a patient's cancer, it may be possible to identify clinical trials or targeted therapies that may be more effective for that patient. Here, in our study, we worked with the TCGA Pan-Cancer dataset where we used the RNA-seq data for analyzing the gene expressions. The dataset comprises 33 types of cancer. Our study mainly focuses on implementing an explainable AI-based panCancer classification approach using gene expression analysis. The goal is to accurately detect the type of cancer in individuals within a short time. We employed seven classifier algorithms- Logistic Regression, SVM, XGBoost, Random Forest, MLP, 1-D CNN, and TabNet. To enhance the performance of the models, we utilized feature selection techniques such as Lasso, SelectFromModel, Select-K-Best, and ElasticNet. SelectFromModel with 500 features yielded the best performance. We applied ensemble methods of probability averaging and max voting, with probability averaging achieving the highest accuracy of 96.60%. Validation of the selected features' contribution and comparison with gene sets from DESeq2 analysis confirmed their significance and relevance. This approach provides insights into cancer-specific molecular mechanisms and pathways. Overall, our study demonstrates the effectiveness of feature selection in reducing dimensionality while maintaining predictive power and biological relevance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Cancer is one of the most common causes of death in the whole world, causing the deaths of nearly 10 million people in 2020 alone [1]. Between 30 and 50 percent of cancer cases can be prevented by quickly diagnosing them and implementing existing evidence-based prevention strategies [1]. There are several ways to detect cancer early, including screenings such as mammograms for breast cancer and colonoscopies for colon cancer, as well as regular check-ups and self-exams to identify any unusual changes in the body [2]. It's important for individuals to be aware of their own bodies and to report any unusual symptoms to their healthcare provider for further evaluation. Many cancer patients have a high chance of recovery if they are diagnosed early and treated appropriately. Sometimes, a surgeon must know the abnormal cell condition of a patient in the middle of an operation. That's why it is vital to diagnose cancer very quickly so that treatment can start as soon as possible.

The domain of computer science most relevant to our research is Bioinformatics. Bioinformatics is a field of study that combines biology, computer science, and information technology to analyze and interpret biological data. It is a multidisciplinary field that involves the use of computational techniques and tools to analyze large amounts of biological data, such as DNA sequences, protein structures, and gene expression patterns. The goal of bioinformatics is to understand the underlying biology of living organisms and to use this knowledge to improve human health and advance scientific research. Bioinformatics has applications in a wide range of areas, including genomics, proteomics, drug discovery, and systems biology. It is an important field of study that is helping to drive many of the advances in modern biology and medicine.

A significant portion of Bioinformatics research is focused on cancer diagnosis.

By using computational techniques and tools, bioinformatics can help analyze large amounts of biological data, such as gene expression data, and identify patterns and trends that may be useful for cancer classification and diagnosis [3]. Bioinformatics can also be used to predict the response of cancer cells to different treatments, which can help guide the selection of the most appropriate treatment for a particular patient. In addition, bioinformatics can be used to identify potential targets for new cancer therapies, such as specific genes or proteins that are over or under-expressed in cancer cells [4]. Overall, bioinformatics has the potential to significantly improve our understanding of cancer and help improve patient outcomes.

1.2 Understanding Cancer

Cancer is a multifaceted disease that is characterized by the uncontrolled development and division of aberrant cells [5]. Benign tumors are growths that are not cancerous and do not spread to other regions of the body, whereas malignant tumors are cancerous growths that can spread to other parts of the body [6]. Hematologic cancers are those that affect organs and tissues that are responsible for blood formation, such as bone marrow and the lymphatic system. Solid tumor cancers, on the other hand, are those that grow tumors in organs other than those that are responsible for blood formation [7]. Hematologic malignancies involve aberrant blood cell formation, which has an impact on the body's ability to fight infections and regulate bleeding. On the other hand, cancers of solid tumors can affect any organ and vary in how aggressively they behave [7]. The mutation of genes is a necessary step in the progression of cancer, which can be caused by a variety of causes including consumption of nicotine and alcohol, genetics, food, physical activity, infections, radiation exposure, and chemical exposure. It is a difficult process because there are various factors that can contribute to cancer to determine which genes are responsible for each type of cancer [8]. Ongoing research is aimed at improving our understanding of the genetic and environmental impacts of cancer in order to develop techniques for cancer prevention, diagnosis, and therapy.

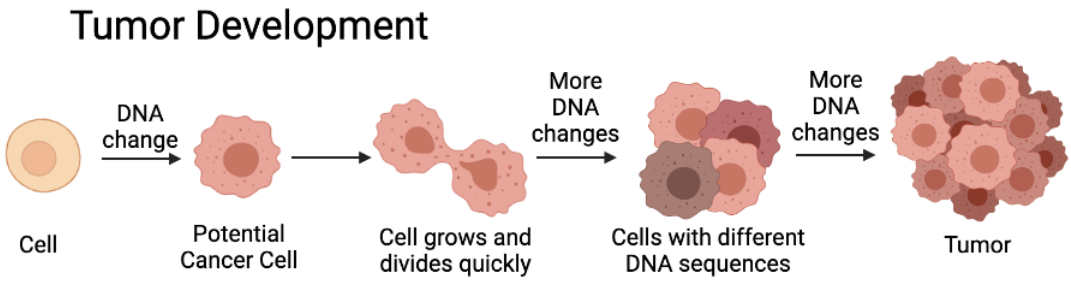


Figure 1.1: Tumor cell development

1.3 Gene Expressions for Cancer Identification

Gene expression analysis is a useful approach for classifying cancer and identifying the specific type of cancer that an individual may have [9]. By analyzing the expression levels of certain genes, it is possible to determine the likelihood that an individual has a particular type of cancer and to differentiate between different types of cancer [9].

There are several different approaches that can be used for gene expression analysis, including microarray analysis and RNA sequencing [10]. Microarray analysis involves hybridizing labeled RNA samples to a microarray chip, which contains probes for many different genes [10, 11]. This allows researchers to measure the expression levels of multiple genes simultaneously. RNA sequencing, on the other hand, involves sequencing the RNA molecules in a sample and analyzing the sequence data to determine the expression levels of different genes [10, 11].

1.4 Motivation and Scope

For proper diagnosis, prognosis, and treatment planning, cancer must be properly classified due to the disease's complexity and heterogeneity [12]. Time, erroneous results accuracy (cases of false positives or false negatives), invasiveness, and high prices all contribute to the shortcomings of conventional cancer diagnostic approaches such as intrusive biopsies and imaging studies (X-Rays, CT

scans, ultrasounds, MRIs, and PET scans). In light of these difficulties, it is clear that new methods are required, ones that are less invasive, more cost-effective, and faster at yielding reliable results. Subjective criteria are also frequently used in conventional cancer classification systems, which can lead to discrepancies and accuracy issues [13]. And so, there is a rising interest in creating rigorous and objective methods for cancer classification in light of recent developments in genetic technologies and computational methodologies. Machine learning and data-driven methods can help researchers find previously unseen connections inside massive cancer datasets, leading to more precise diagnoses. These developments have the potential to dramatically alter cancer diagnosis and customized treatment, leading to better overall patient outcomes [14].

On the other hand, since circulating miRNAs may be identified directly from biological fluids such as blood, urine, saliva, and pleural fluid and are less intrusive than the invasive techniques currently employed for cancer detection, the use of these molecules as potential biomarkers is being discussed [12]. mRNA, or messenger RNA, is a type of RNA that carries genetic information from DNA to the ribosome, where it is used to synthesize proteins [15]. In normal cells, mRNA plays a crucial role in the production of proteins that are necessary for the proper functioning of the cell. In cancer cells, however, the production of mRNA and the synthesis of proteins can be altered, leading to the development and growth of cancer [16]. For example, certain genes may become overactive or underactive in cancer cells, leading to the production of abnormal proteins or the suppression of proteins that normally help to regulate cell growth and division.

Researchers are studying the role of mRNA in cancer to better understand the molecular changes that occur in cancer cells and to identify potential targets for cancer diagnosis and treatment [17]. For example, some cancer therapies are designed to target the mRNA of specific cancer-related genes in order to inhibit the production of abnormal proteins or to restore the function of proteins that help regulate cell growth.

That is why, we propose a pipeline for cancer classification procedure for analyzing the gene expressions beforehand for ensuring the reduction of time, early

diagnosis, and avoidance of invasiveness. By analyzing gene expression data, it may be possible to identify patterns and characteristics that are associated with specific types of cancer, which could help doctors accurately diagnose and classify cancer in its early stages. This could lead to earlier treatment and a higher likelihood of successful treatment outcomes. In addition, we will investigate the identification of key genes responsible for each type of cancer using explainable AI systems, which are designed to provide transparency and accountability in their decision-making processes. By analyzing the learned patterns and importance of features derived from our trained models, we can identify the genes that substantially contribute to classification accuracy. This analysis will allow the identification of key genes associated with specific types of cancer, casting light on the molecular mechanisms and pathways underlying tumor development and progression.

1.5 Problem Statement

Based on the aforementioned discussions, this research aims to propose a pipeline to classify 33 types of cancer with high accuracy and identify cancer-specific important gene sets. And so, the objectives of this research are -

- Evaluating the performance of different classifier models on both the raw dataset and the normalized dataset with and without integrating feature selection techniques.
- Integrating different ensembling approaches to improve accuracy.
- Determining the feature contribution for each sample by applying Explainable machine learning models.
- Extracting the list of globally significant genes and patient-specific gene sets for each cancer type.
- Validating the gene sets by statistical analysis.

1.6 Research Challenges

Both the process of identifying cancer based on gene expression data and the process of locating important genes include several challenges. It is of the utmost importance to preserve data quality while also ensuring that heterogeneity in preprocessing and measurement platforms is well managed. When working with high-dimensional feature spaces, it is essential to have effective procedures for feature selection. It is necessary to resolve the class imbalance to prevent the creation of models that are prejudiced. It is crucial to construct models that can be interpreted and that can offer insights into significant genes and pathways. The important steps of generalization and validation of several datasets are often overlooked. It is necessary to carry out biological validation and functional research on the genes that have been discovered. To successfully control model complexity and prevent overfitting, it is vital to avoid overfitting. Both the variable nature of cancer and the accurate classification of its various subtypes present several obstacles to researchers. Lastly, various environmental setup-related concerns were also engaged in the experiment, such as limitations on time and resources, such as when experimenting.

1.7 Organization of thesis

The subsequent sections of the dissertation are structured in the following manner. Section 2 discusses the background and motivation for cancer classification using RNA-seq gene expression. The work also provides an analysis of the current literature pertaining to the subject matter and its respective limitations. A detailed description of the dataset used in this study is included in section 3. In Section 4, the proposed framework is presented, which demonstrates the ability to accurately classify 33 distinct types of cancers and identify gene sets specific to each cancer type. Section 5 of the paper undertakes an analysis of the performance of the proposed architecture and conducts a comparative evaluation with existing state-of-the-art architectures in the context of a cancer classification task. The aforementioned method detects both cancer-specific and patient-

specific gene sets and subsequently conducts statistical validation on the cancer-specific gene set. Section 6 of this paper serves as a conclusion to our discourse and offers guidance for potential areas of future investigation.

Chapter 2

Background Study

2.1 RNA (Ribo Nucleic Acid)

RNA (Ribo Nucleic Acid) is a molecule that is essential in many biological activities, such as gene expression, protein synthesis, and cellular function regulation [18]. It is made up of nucleotide-building units and is physically similar to DNA (Deoxyribo Nucleic Acid). However, RNA is normally single-stranded and includes sugar ribose rather than deoxyribose [19]. RNA has many implications in molecular biology and genetics. Several aspects of the contribution of RNA molecules were explored before digging in depth. They are briefly listed below -

Central Dogma of Molecular Biology: The central dogma analyzes and explains the transfer of genetic information from DNA to RNA to protein. It offers the essential concept required for comprehending the role that RNA plays in the process of gene expression [20].

Transcription: A template molecule of DNA is used in the process of RNA transcription, which results in the synthesis of an RNA molecule. It entails moving genetic information that is encoded in DNA to RNA, more specifically to RNA that is known as messenger RNA (mRNA). The information contained in DNA is utilized in this process, which is an essential stage of gene expression. Functional RNA molecules are produced as a result [21].

RNA Processing and Modifications: RNA processing converts the initial RNA transcript into a mature, functioning RNA molecule. Capping, splicing, polyadenylation, and RNA editing are examples. These mechanisms regulate gene expression and RNA stability [22].

RNA Interference and Gene Regulation: RNA interference (RNAi) is the biological process that regulates gene expression. It entails using tiny RNA

molecules to limit gene expression by targeting and degrading messenger RNA (mRNA) or interfering with translation [23].

2.2 Types of RNA regulating Gene Expression

There are several types of RNA molecules involved in gene expression and regulation within cells. Some important ones that we have explored through our study are discussed below -

2.2.1 mRNA (messenger RNA)

mRNA is a molecule of single-stranded RNA that transports genetic information from DNA to the ribosomes for protein synthesis. It functions as a link between DNA and protein synthesis. Before being translated into proteins, mRNA endures numerous modifications after being transcribed from specific genes [23].

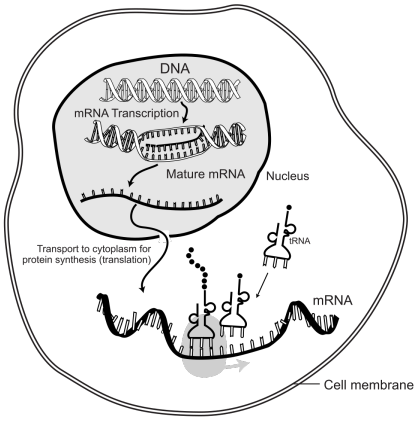


Figure 2.1: mRNA (messenger RNA)

2.2.2 miRNA (microRNA)

miRNAs are minuscule, approximately 22 nucleotide-long noncoding RNA molecules. By binding to target mRNA molecules, they perform a key role in post-transcriptional gene regulation. This binding can lead to mRNA degradation or translational repression, thus affecting gene expression. miRNAs play multiple roles in development, cellular processes, and disease [24].

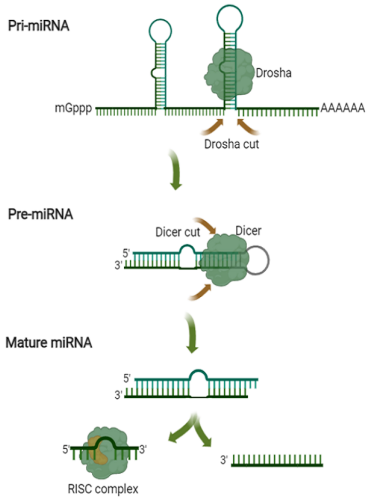


Figure 2.2: miRNA (microRNA)

2.2.3 lncRNA (long non-coding RNA)

lncRNAs are a heterogeneous group of RNA molecules longer than 200 nucleotides that do not encode proteins. They have emerged as crucial regulators of gene expression and participate in a variety of biological processes. The interaction of lncRNAs with DNA, RNA, and proteins influences chromatin structure, transcriptional regulation, and post-transcriptional processes [25].

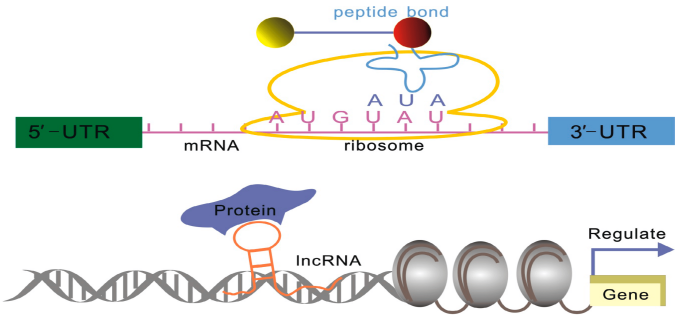


Figure 2.3: lncRNA (long non-coding RNA)

2.2.4 DNA methylation

DNA methylation is an epigenetic modification in which a methyl group is added to the DNA molecule. It is essential for the regulation of gene expression and is involved in numerous biological processes, such as development, genomic imprinting, and X-chromosome inactivation. DNA methylation patterns can be sta-

ble and heritable, influencing gene activity without modifying the DNA sequence [26].

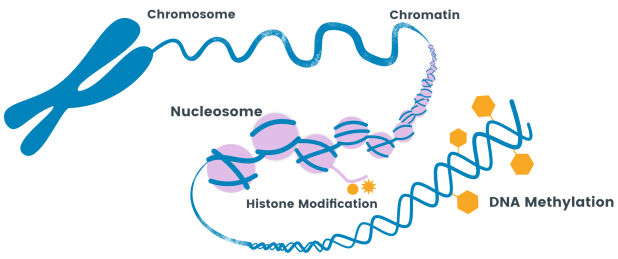


Figure 2.4: DNA methylation

There are several other types of RNA molecules involved in regulating gene expression, like rRNA (ribosomal RNA), tRNA (transfer RNA), siRNA (small interfering RNA), snRNA (small nuclear RNA), and many more. But they have not been studied in detail because we won't need them for conducting our research work, but they can be explored in the future if required.

2.3 RNA in Cancer Classification

Methods based on RNA have made significant contributions to the classification of cancer by revealing important information about tumor heterogeneity, prognosis, and treatment response. A few examples of RNA's contribution to cancer classification are discussed below in detail.

2.3.1 Gene Expression Profiling

The approach of gene expression profiling Figure 2.5 involves assessing the quantities of RNA transcripts from various genes in a cell, tissue, or organism [27]. This comprehensive perspective of gene activity and regulation assists researchers in better understanding biological processes, disease states, and treatment responses. To gather gene expression data, procedures such as microarrays or RNA sequencing are frequently used. This information can then be used to

detect differentially expressed genes, characterize gene expression patterns, categorize samples, and infer underlying biological pathways. Gene expression profiling has had a dramatic impact on cancer research. Researchers have identified unique molecular subtypes of many tumors utilizing RNA-based technologies such as microarrays and RNA sequencing. This classification based on gene expression patterns has greatly improved our understanding of tumor biology and paved the way for individualized treatment approaches. The identification of dysregulated genes, prediction of prognosis, and creation of targeted and personalized cancer treatments have all been made possible by gene expression profiling.

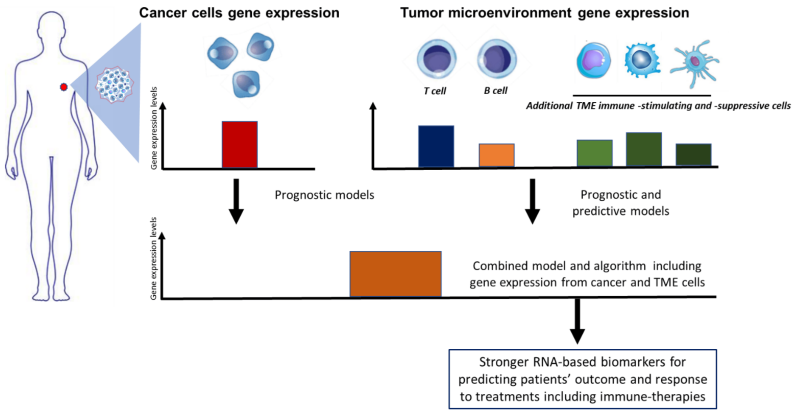


Figure 2.5: Gene expression profiling in early breast cancer [28]

2.3.2 Non-coding RNA Signatures

Non-coding RNA signatures Figure 2.6 are distinct patterns or variations in the expression levels of non-coding RNAs, such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), that have been seen in numerous disorders, including cancer [29]. These signals have diagnostic and prognostic value, providing information about illness presence, progression, and therapy response. They add to our understanding of disease genesis by revealing molecular pathways. Furthermore, non-coding RNA profiles provide prospective therapeutic targets for personalized medicine and targeted therapeutics in cancer research.

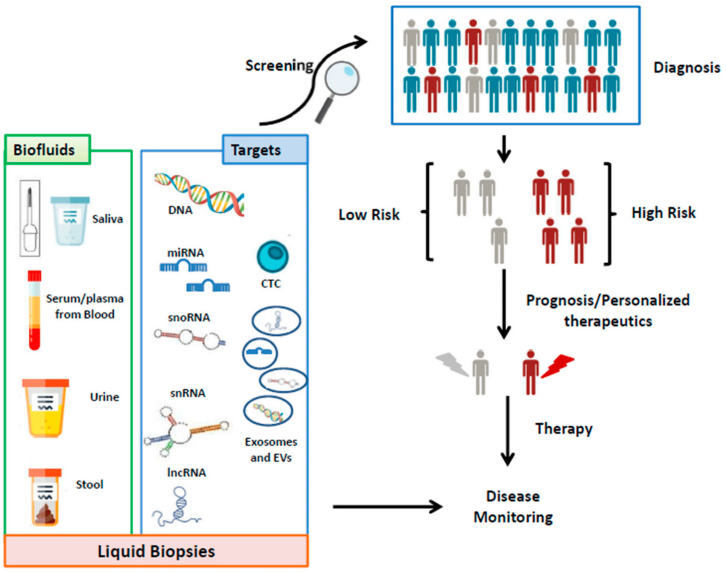


Figure 2.6: Noncoding RNAs in extracellular fluids as cancer biomarkers [30]

2.3.3 Fusion Gene Detection

RNA-based methods have been critical in detecting fusion genes Figure 2.7, which are produced as a result of chromosomal rearrangements and have the ability to induce oncogenesis. Fusion gene identification approaches, such as RNA sequencing (RNA-seq), have proven very useful in the categorization and diagnosis of specific tumors, such as pediatric sarcomas and hematological malignancies. Identifying and characterizing hybrid genes using molecular approaches such as RNA sequencing and fluorescence in native hybridization is part of the detection process. These approaches aid in the detection of fusion transcripts, provide insights into disease causes, and direct therapy options. Fusion gene discovery helps with cancer categorization, diagnosis, and understanding of distinct tumor subtypes linked to fusion genes, which can lead to the development of targeted therapeutics [31].

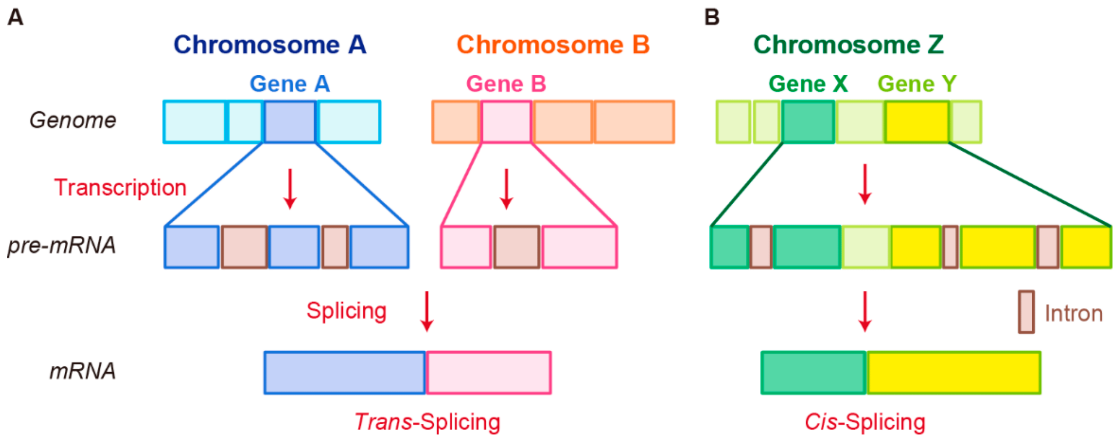


Figure 2.7: Fusion gene detection [32]

2.4 Literature Review

Due to the high dimensionality of the data, it is essential to explicitly specify characteristics while classifying cancer. And so based on the success rate in finding key characteristics and accessibility to data from high throughput equipment, machine learning-based [33] and deep learning-based approaches have recently become more prominent in the categorization of cancer cells using different types of datasets composed of single-omics like - mRNA data, miRNA data, lncRNA data, DNA-methylation data or combination of different datatypes (multi-omics), etc.

Since our work was based on gene expression analysis of mRNA data, we first tried to go through some works relevant to it. Amrane et al. [34] used the Breast Cancer Dataset (BCD) to determine whether a tumor is benign or malignant using two classifiers: the Naive Bayesian Classifier and K-Nearest Neighbor (KNN). The ID property was removed from the dataset, which included 11 other properties. After preprocessing, 683 samples remained. Compared to NB, which had an accuracy of 0.961932, KNN had a greater accuracy of 0.975109. And for computing the final prediction, in the KNN algorithm, new instances were compared to the training set to determine the final prediction, while the NB algorithm involved dividing the dataset into testing and training sets, calculating the mean and standard deviation of each feature and class, and measuring the probability

of each feature and class. But their limitation was that although KNN did better if the dataset was larger, KNN would have dropped out of the top spot due to the computation's time complexity.

Another experiment performed by Sara Tarek [35], proposed an ensemble system for cancer classification using gene expression data, addressing the drawbacks of enhancing result accuracy, covering more cancer types, and mitigating the effect of over-fitting. They employed three benchmark cancer datasets (Leukemia, Colon, and Breast cancer datasets) and preprocessed the data using feature selection algorithms, normalization, and logarithmic transformation. Five base classifiers using the 3-NN algorithm and other gene selection techniques made up the proposed ensemble system. In order to post-process the data, they additionally employed majority voting and error estimation techniques. Using different measures including ROC, AUC, and BCI, the performance of the classification was evaluated. Reducing ensemble error, BCI, and AUC, the proposed approach exhibits notable improvements in performance parameters for all three malignancies.

Podolsky et al. [36] processed four publicly available lung cancer data sets, including those from the Dana-Farber Cancer Institute, the University of Michigan, the University of Toronto, and Brigham and Women's Hospital. These datasets were analyzed using seven machine learning methods, including k-NN, Naive Bayes, SVM, and the C4.5 decision tree. AUC values were used to gauge these algorithms' efficacy. The findings demonstrated that various algorithms worked best for various datasets, with k-NN and SVM typically providing higher AUC values. On the University of Toronto dataset, the C4.5 decision tree outperformed all other algorithms, whereas, on the University of Michigan dataset, C4.5 was the exception. The dataset from the Dana-Farber Cancer Institute revealed that k-NN had the greatest averaged AUC while Naive Bayes had the lowest.

Experiments based on deep learning-based approaches have also been conducted on mRNA datasets. Lyu et al. [37] applied a deep learning-based approach to tumor type classification using normalized-level-3 RNA-Seq gene expression data

from 33 tumor types in the Pan-Cancer Atlas. To exclude noisy and unimportant genes, the data were log-transformed and filtered. The filtered data were normalized and molded into a 102x102 picture. Three convolutional layers with max-pooling and batch-normalization layers were employed in the classification model, which was then followed by three fully connected layers with drop-out layers in between. The workflow is presented in Figure 2.8. With two pairings of misclassification for the READ and CHOL samples, the model has an accuracy of 95.59% after being trained using 10-fold cross-validation. Significant genes were found for six different cancer types using KEGG pathway analysis. Because of the smaller sample size, CHOL was incorrectly classified as LIHC.

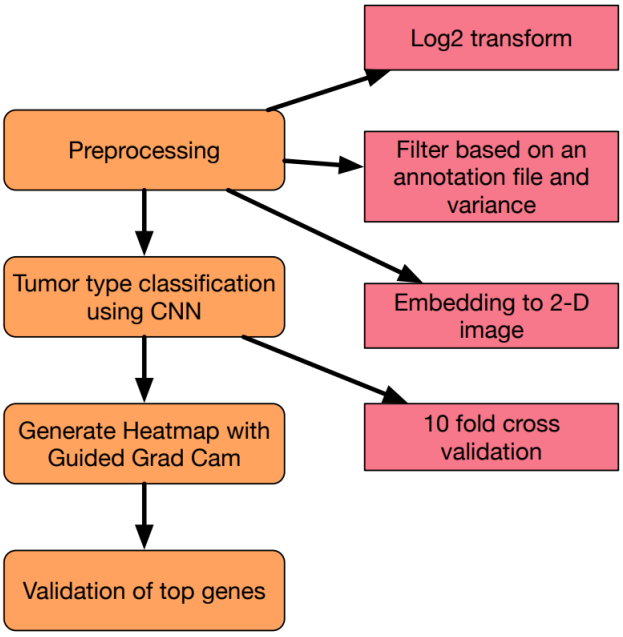


Figure 2.8: Workflow proposed by Lyu et al. [37]

Joseph M. de Guia et al. [38] in their experiment with TCGA RNAseq data containing 33 types of cancer, used normalized and log-transformed gene samples to create 2D images with a size of 102×102 pixels. After preprocessing, they used a convolutional neural network (CNN) with three hidden layers and an output layer to classify the images into 33 different cancer types. Tenfold cross-validation was used to validate the model after it had been trained using

two gradient history technique algorithms. Additionally, the scientists developed a heatmap to show the scores of each gene in the classification job and utilized guided backpropagation and Grad-CAM to analyze the coarse localization map for the relevant genes. With the notable exception of CHOL and READ, the accuracy of the model was higher than 95.6% for all cancer types. A gene functional categorization method was used to assess 400 possible biomarker indicators and annotate the genes according to how similar their functions were. Using criteria like p-values for correlation with the major gene pathways and their associated biomarkers, the functional analysis results were then compared with the pertinent pathways of the cohort cancer types.

Apart from the above-mentioned works based on mRNA gene expressions, we also tried to go through experiments relevant to the dataset (TCGA PanCancer dataset) that we used for our work. Li et al. [39] worked with 31 types of tumor data collected from TCGA. RNA-seq expression of 9096 tumor samples was used in the analysis. They have applied k-Nearest Neighbour (k-NN) [40] for classification, log2-transform for data normalization, and genetic algorithms for gene selection. They ran a classification experiment using those data sets after choosing 20 genes.

Pan-cancer analysis is becoming more popular as researchers use improved sequencing technology and resources such as The Cancer Genome Atlas (TCGA) to discover critical determinants in cancer formation.

Hsu et al. [41] used TCGA RNA-sequencing data to classify 33 kinds of cancer Figure 2.9. The accuracy, training time, precision, recall, and F1-score of five machine learning methods were evaluated: decision tree, k-nearest neighbor, linear support vector machine, polynomial support vector machine, and artificial neural network. The linear support vector machine (SVM) outperformed the other methods, with the greatest accuracy rate of 95.8%. The research also emphasizes the significance of advanced data pre-processing approaches in improving the model's performance. Overall, this study highlights the benefit of utilizing machine learning and RNA-sequencing data for pan-cancer categorization, adding to a better understanding of the disease.

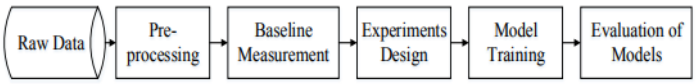
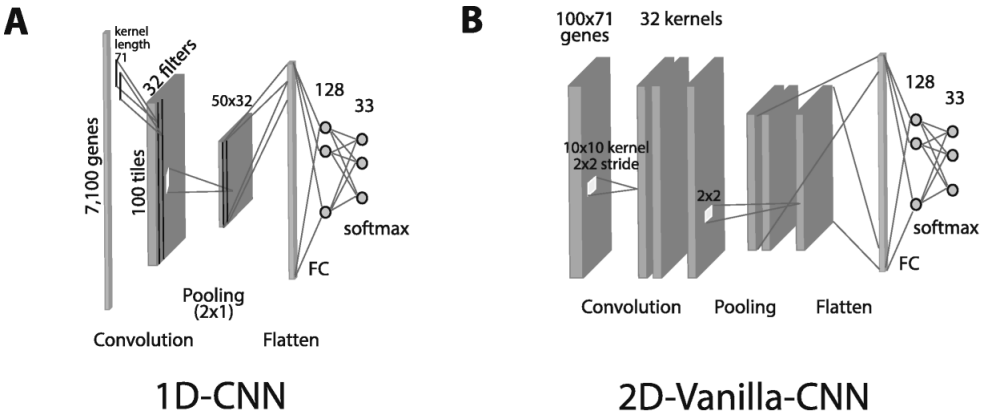


Figure 2.9: Cancer type prediction and classification based on RNA- sequencing data [41]

Mostavi et al. [42] along with the other authors, introduced three Convolutional Neural Network (CNN) models for the classification of tumor and non-tumor samples using gene expression data. In 34 classes (33 tumors and 1 normal), the models 1D-CNN, 2D-Vanilla-CNN, and 2D-Hybrid-CNN obtain high prediction accuracies (93.9–95.0%). Figure 2.10 portrays the architectures of these three CNN models. A saliency technique is used to further analyze the 1D-CNN model, revealing 2090 cancer indicators that exhibit agreement with the cancer kinds they represent. Notably, popular indicators like GATA3 and ESR1 have been discovered in breast cancer. The 1D-CNN model is further extended for subtypes of breast cancer prediction, with an average accuracy of 88.42% across 5 subtypes. Independent of tissue-of-origin effects, the authors contend that their models provide precise cancer detection and shed light on the biological significance of cancer marker genes. These replicas have lightweight hyperparameters, enabling easy adaptation for future clinical applications.



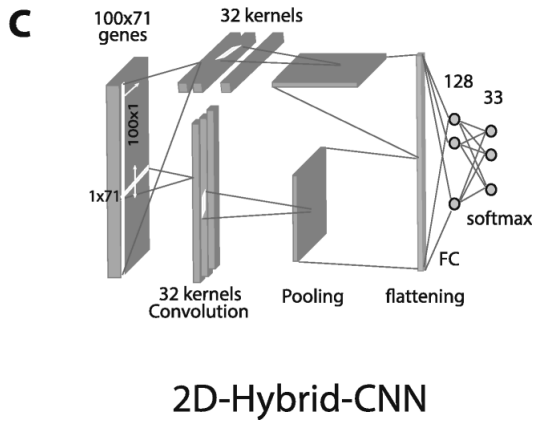


Figure 2.10: Illustration of three CNN models proposed by Mostavi et al. [42]

Laplante et al. [43] suggest using a deep neural network classifier as a means of determining the location of malignancies throughout the body. They were able to achieve an astounding 96.9% accuracy in the classification of tumors across 20 anatomical sites by using 27 TCGA miRNA stem-loop cohorts. They begin by preprocessing the expression data, then classify the data utilizing a neural network with six hidden layers, and lastly, they train the model utilizing the ADAM optimizer and the Categorical Cross-entropy loss function. Their method indicates the feasibility of employing data from miRNA stem-loop sequencing for precise tumor localization. The F1 score as a whole is 96.88%, and the majority of classes have achieved a score of 90% or higher, with the exception of cervical and endocervical cancers, which are frequently misclassified as uterus tumors due to their proximity to the uterus. These results highlighted the effectiveness of miRNA stem-loop sequencing data in oncology inference tasks. However, they did not differentiate between the various cohorts, nor did they confirm the transferability of their model by conducting experiments using data from a different source.

Mahin et al. [44] introduced PanClassif, a technique to improve the performance of a variety of machine learning classifiers while using only a small number of efficient genes to identify cancer from RNA-seq data. The Cancer Genome Atlas (TCGA), which has 8287 cancer samples and 680 normal samples, was used to collect 22 different types of cancer samples. To deal with data noise, PanClassif used k-Nearest Neighbour (k-NN) smoothing to smooth the samples.

Then, using an Anova-based test, effective genes were chosen. The oversampling technique SMOTE was used to balance the train data. Six classifiers—SVM (linear kernel), SVM (RBF kernel), Random Forest(RF) [45], Neural Network, k-Nearest Neighbor(KNN), and Adaboost [46] algorithm—were used to assess the performance of the technique. Figure 2 depicts the mechanism of their proposed method.

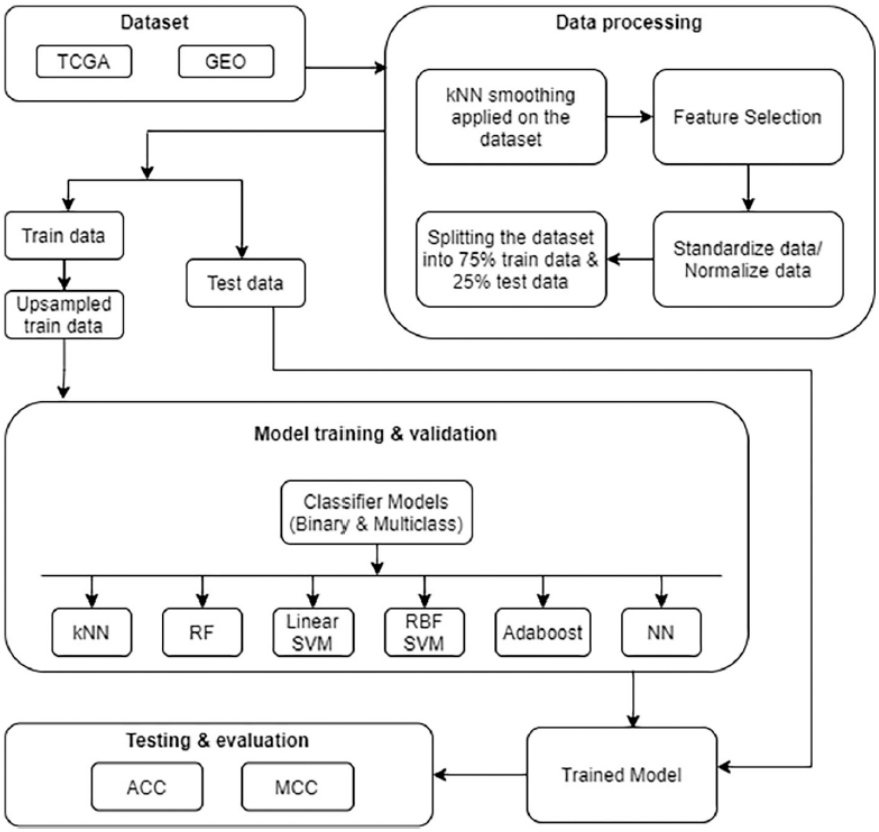


Figure 2.11: Complete workflow of PanClassif [44]

The top genes from the TCGA cancer datasets were then subjected to gene set enrichment analysis, and the majority of the genes chosen showed an association with cancer and tumor-related ailments. Additionally, they tested the efficacy of their suggested technique using single-cell RNA-Seq datasets from the Gene Expression Omnibus (GEO). Among the 6 classifiers, RF performed best considering both datasets and on the TCGA dataset, KNN performed better. Using five distinct datasets with various numbers of genes chosen, figure 3 shows a spider plot for the ACC score for the multi-class classification task generated using

TCGA data. However, they did not provide any short list of features for each cancer classification. For example, the gene that is responsible for skin cancer is certainly not responsible for lung cancer.

The process of identifying a restricted set of genes that hold significance for classification has been the focus of several research endeavors. Several authors have addressed the task of identifying genetic markers for diseases by utilizing feature selection techniques in machine learning algorithms [47, 48, 49, 50]. Methods for filter selection place characteristics in order of classification effectiveness in order to find the best ones. According to mutual knowledge, Pavithra et al. [48] established the filters, while Guyon et al. [51] arranged them in accordance with the weights of a recursive Support Vector Machine (SVM) trained for classification.

Lopez-Rincon et al. [52] proposed an ensemble approach to extract 100 miRNA for multi-class cancer classification.

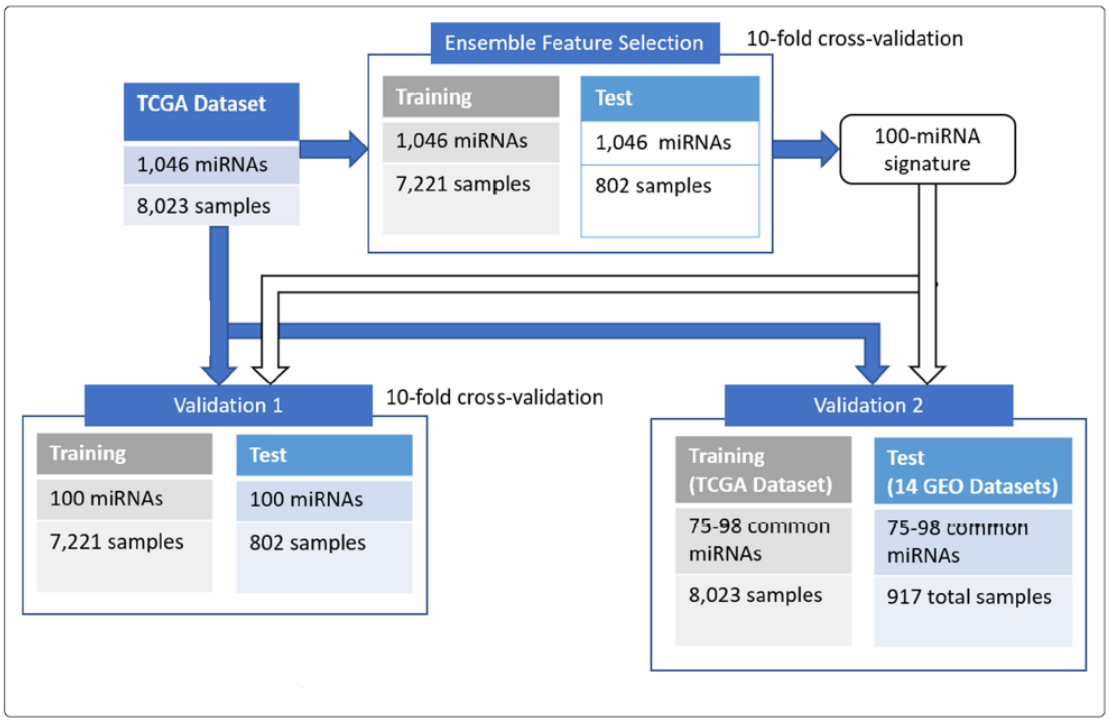


Figure 2.12: Summary of the different datasets and their use in the experiments [52]

A subset of The Cancer Genome Atlas dataset (TCGA) encompassing 8023 cases, 28 distinct kinds of cancer, and 1046 unique stem-loop miRNA expressions served as the basis for their ensemble feature selection methodology. In order to distinguish between various types of cancer and normal tissues, they implemented 8 different classifiers (Bagging, Gradient Boosting, Random Forest, Ridge, SGD, SVC, Logistic Regression, Passive Aggressive) along with 10-fold cross-validation. Based on them they also proposed an ensemble approach to identify the minimal miRNA entities. They examined 14 GEO datasets from 5 different platforms and cancer types to corroborate their findings Figure 2.12. Additionally, they examined the 777 BRCA samples in the TCGA dataset, which were divided into 5 distinct subtypes (Triple-negative/basal-like, Luminal A, Luminal B, HER2-enriched, and Normal-like). They used the 100-miRNA signature in five different tests (BRCA subtype in TCGA, BRCA subtype in GEO datasets, tumor type classification, tumor tissue versus normal tissue, and GEO datasets) Table 2.1. According to their findings, Logistic Regression performed best across all experiments, and Ridge had the worst accuracy. They then conducted a bibliographical meta-analysis, which confirmed that 77 out of the 100 miRNAs in the signature are found in lists of circulating miRNAs utilized in cancer research, either in stem-loop or mature-sequence form. They also demonstrated that hsa-miR-21, which was produced from their suggested output, appeared to be the miRNA that was overexpressed most frequently across all kinds of tumors. However, they only provided a certain number of features for all kinds of cancer. Additionally, global feature attribution and person-specific feature attribution have not been discussed.

Lastly, we have also gone through some other works relevant to ours to get a deeper insight into the most commonly used models, feature extraction techniques, and much more relevant to our field of interest. A database from the UCI Repository that contains 801 samples and 20,531 characteristics that are specific to 5 forms of cancer (breast, kidney, colon, lung, and prostate) has been successfully used to test the Grouping Genetic Algorithm (GGA) by García-Díaz et al. [53]. A few candidate classifiers ($<$50 from the total of 20,531) with an average

Classifier	TCGA	TT vs NT	GEO	TCGA (Subtype)	GEO (Subtype)	Global
Gradient Boosting	0.9359	0.9846	0.6697	0.9725	0.8909	0.8907
Random Forest	0.9324	0.9839	0.8085	0.9725	0.8634	0.9121
Logistic Regression	0.9237	0.9799	0.9351	0.9647	0.8476	0.9302
Passive Aggressive	0.8831	0.9606	0.8678	0.9556	0.8197	0.8974
SGD	0.9035	0.9767	0.9393	0.949	0.8145	0.9166
SVC	0.9154	0.9791	0.7724	0.9451	0.8355	0.8895
Ridge	0.8305	0.947	0.8867	0.9503	0.83	0.8889
Bagging	0.911	0.9812	0.7682	0.9555	0.907	0.9046

Table 2.1: Comparison of the 8 classifiers, for the different experiments with the 100-miRNA signature as demonstrated by Lopez-Rincon et al. [52]

accuracy of 98.81% are chosen by the GGA from a total of 20,531 attributes. The potential discrepancy in the solution space exploration by the suggested approach is a limitation. The selection of several tens of genes from 20,531 genes from the enormous number of characteristics implies a vast search field. Also, they did not work with 33 different types of cancer, as the dataset is highly imbalanced and may create a bias towards a larger sample class.

Lai et al. [54] investigated the feasibility of developing an accurate approach of prognostic stratification for patients with non-small cell lung cancer (NSCLC). In order to make an accurate prediction of overall survival, they used a deep neural network (DNN) that took into account both gene expression data and clinical information. Patients were initially classified into biomarker-positive and biomarker-negative subgroups based on their responses to a panel of seven well-established NSCLC biomarkers. The authors then found eight more novel prognostic gene biomarkers by employing a systems biology approach to the research. Along with clinical data, these 15 biomarkers were incorporated into a bimodal learning DNN to predict the 5-year survival status of NSCLC patients with a high degree of accuracy (AUC-ROC of 0.8163, accuracy of 75.44%). This predictive model has the ability to guide decisions regarding individualized therapy and contribute to the evolution of precision medicine in NSCLC. It achieves this promise through the utilization of deep learning. However, they worked with only one cancer type and did not perform a variety of cohort analyses.

Zhang et al. [55] integrated an unsupervised feature learning framework for the purpose of recognizing a variety of traits based on gene expression profiles. In order to obtain features, the system utilizes a combination of principal component analysis (PCA) and an autoencoder neural network. For the purpose of predicting clinical outcomes in breast cancer, an ensemble classifier known as PCA-AE-Ada is created. This classifier is based on the AdaBoost algorithm. Comparisons are made between the proposed method and several other gene signature-based algorithms, one of which is a baseline method known as PCA-Ada. The results of the experiments show that the suggested strategy, which makes use of deep learning techniques, is superior to other approaches in terms of the AUC-ROC, the maximum classification accuracy (MCC), and other evaluation criteria across a variety of different breast cancer datasets. The purpose of this study is to improve the accuracy of cancer prognostic predictions using a supervised classifier learning mechanism that incorporates feature selection, feature extraction, and deep learning approaches. However, there are still obstacles to overcome in understanding the intricate structure of the deep learning model and enhancing the generalization capacity with more datasets that are accessible to the public. In our study, we have tried to overcome the shortcomings mentioned above.

Chapter 3

Dataset

The Cancer Genome Atlas (TCGA) [56] is a collaborative initiative aimed at expediting our comprehension of the molecular underpinnings of cancer via the utilization of genome analysis techniques, such as extensive genome sequencing. This initiative is comprehensive and well-coordinated. The TCGA dataset includes a wide variety of genomic data, including gene expression data generated using RNA-Seq.

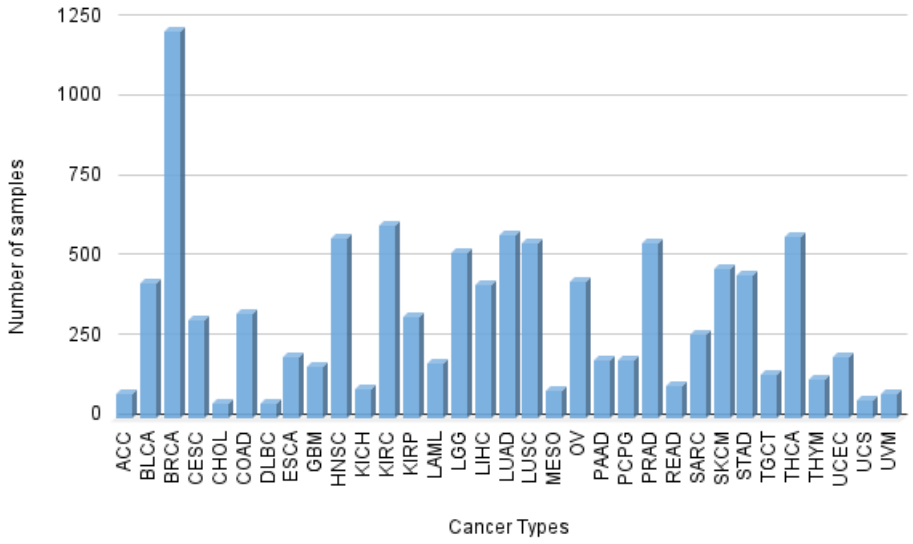


Figure 3.1: Number of samples in each cancer type

RNA-Seq is a technique for profiling the expression levels of RNA molecules in a sample. One common way to quantitate gene expression levels from RNA-Seq data is by using FPKM (Fragments Per Kilobase of transcript per Million mapped reads) [57, 58]. FPKM is a measure of the abundance of a transcript in the sample, normalized for differences in the total number of reads produced by the sequencer, and for the length of the transcript. TOIL (Transcriptome-Oriented Incremental Learning) is a software tool that performs reference-based transcriptome assembly and quantification using RNA-Seq data. It is based on

RSEM (RNA-Seq by Expectation Maximization) [59], which is a software package that estimates the abundance of transcripts in an RNA-Seq sample by using the Expectation Maximization algorithm.

In summary, TCGA gene expression data generated using RNA-Seq can be quantified using the FPKM metric, and tools like TOIL and RSEM can be used to analyze and interpret this data. And combinedly, the TOIL RSEM fpkm data is based on mRNA sequencing data. It estimates the expression level of genes based on the sequencing reads of mRNA transcripts

The data was downloaded from the TCGA Data Portal, on September 1, 2016. There are a total of 10535 patient samples along with their 60,499 features. We take into account the file's protein-coding values and eliminate all of the features when an item doesn't adhere to the study protocol. It reduces the total number of features to 19,238. Furthermore, we download the corresponding phenotype data and map each cancer type with each tumor tissue sample. Figure 3.1 presents the total number of samples in each cancer type.

Chapter 4

Proposed Methodology

4.1 Overview

In this research, we propose a novel pipeline for classifying 33 distinct cancer types. The proposed architecture utilizes SelectFromModel [60] to select 500 features from a total of 19,239 features. We classified the cancer categories using seven classification models, including logistic regression [61], support vector machine (SVM) [62], Extreme Gradient Boosting (XGBoost) [63], random forest [45], multi-layer perception (MLP) [64], 1-D CNN [65, 66], and TabNet [67]. To enhance the performance of our classification models, we combined the top three performing models, logistic regression, SVM, and XGBoost, using the Probability averaging [68] and Max voting [69] ensemble method.

SHAP (SHapley Additive Explanations) [70] was utilized to analyze the performance of the models and determine the significance of the features in predicting cancer types. SHAP is a potent instrument that permits the calculation of both local and global feature importance values for the various models. SHAP was applied to logistic regression, SVM, XGBoost, random forest, MLP, and 1d CNN models, while TabNet's feature importance values were utilized for the TabNet model.

To validate our findings, we performed DEG [71] analysis and used DESeq2 [72] to identify globally significant genes for each specific malignancy. To further validate the significance of the features identified by our models, we compared the genes obtained from our classification model with the DEG analysis gene set.

The results indicate that the proposed architecture obtains a high degree of classification precision for the 33 distinct cancer types. In addition, our SHAP analysis revealed the most significant characteristics of each cancer type and shed light on the decision-making processes of the models. The overview of the proposed methodology is shown in Figure 4.1.

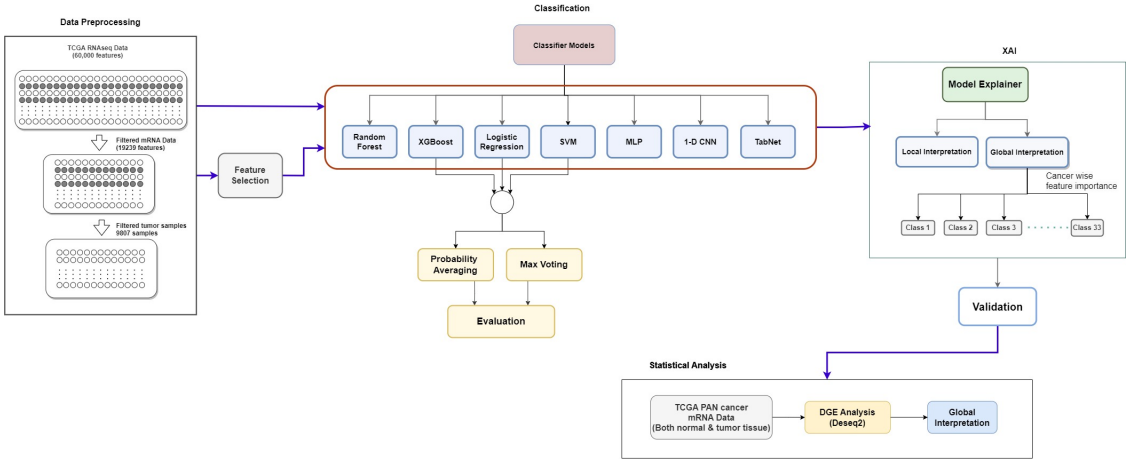


Figure 4.1: Proposed architecture

4.2 Data Preprocessing

4.2.1 Filtration of Features

The TCGA dataset, comprising nearly 60,499 features, has been used in this study. In order to facilitate interpretability and ease of analysis, our examination was restricted solely to mRNA characteristics. The process involved the mapping of ensemble IDs to their respective gene symbols, thereby enabling the utilization of gene common names instead of ensemble IDs for reference purposes. Furthermore, it was essential to eliminate characteristics apart from the mRNA data in order to streamline the dataset and reduce its dimensionality. Following the exclusion of non-mRNA features, our dataset comprised a total of 19,235 mRNA features. This process of identifying only the pertinent characteristics enabled us to concentrate our analysis on the most informative genes and enhance the precision of our findings. Figure 4.2 represents the overall procedure of data preprocessing.

4.2.2 Discarding Normal Tissue Samples

Our study centered on the identification of genes specific to cancer through the analysis of a dataset comprising 10,535 samples of tumor and normal tissues across 33 distinct cancer types. In order to ascertain the exclusivity of cancer-specific genes, the normal tissue samples were eliminated from the dataset, and

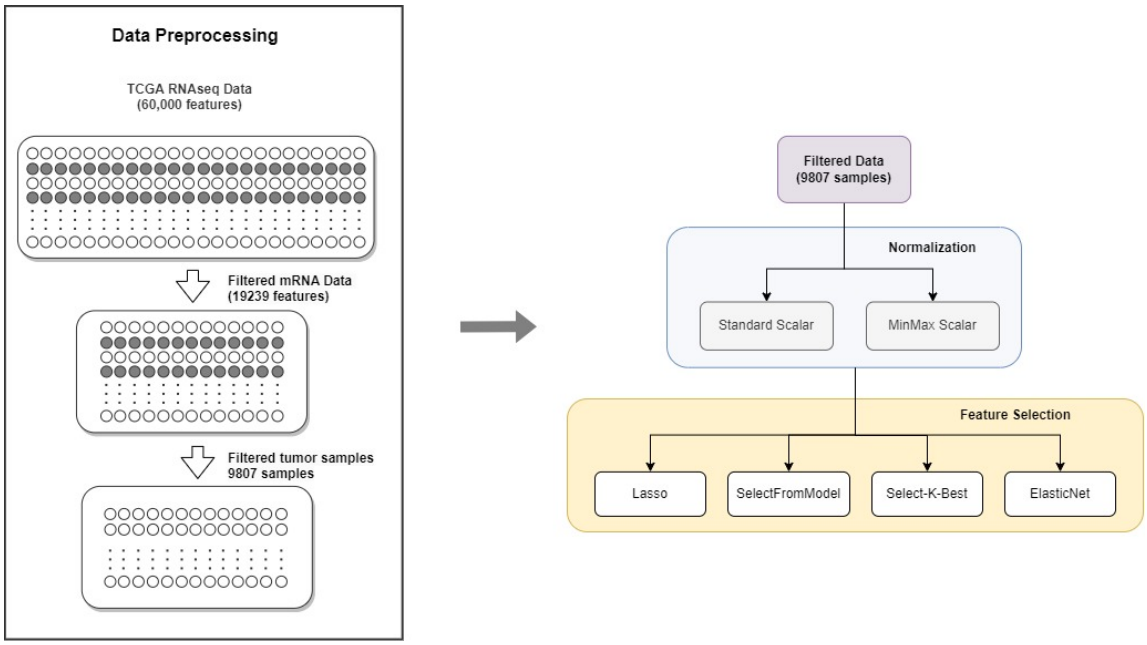


Figure 4.2: Data pre-processing

solely the tumor tissue samples for each cancer type were retained. It was imperative to mitigate any potential impact that the existence of non-cancerous tissues may exert on the genetic profile of particular malignancies. The objective of our study was to isolate genes that exhibited predominant expression in tumor tissues, while being absent or minimally expressed in normal tissue samples. Excluding normal tissue samples from the analysis aided in mitigating the possibility of encountering erroneous positive or negative outcomes. The total number of samples obtained after excluding all normal tissue samples was 9807. The sample size for each cancer type was sufficiently large and representative, enabling the identification of cancer-specific genes with a high level of confidence.

4.2.3 Normalization

This study involved the utilization of two distinct data normalization methodologies, specifically the Min-Max Scaler [73, 74, 75] and Standard Scaler [76, 77, 78], for the purpose of preprocessing the dataset prior to classification. The objective was to assess the impact of these techniques on the performance of the classification models.

The technique of normalization holds significant importance in the fields of ma-

chine learning and data analysis [41, 79]. It guarantees that the features are standardized to a common scale, which is a crucial prerequisite for numerous algorithms to function efficiently. In the event that the input data features exhibit dissimilar scales or significant differences in magnitudes, it may result in certain weights within the model being updated at a much faster rate than others. This can impede the convergence process, leading to a slow convergence or even a complete failure of the model to converge. Data normalization, which involves scaling and shifting the input data to have a similar range of values, can help mitigate this issue. The process of normalization is utilized to mitigate the impact of feature scale on the performance of a machine-learning model. This is especially crucial when working with data that exhibits varying units or orders of magnitude. The normalization of data has the potential to enhance the efficiency and convergence of specific algorithms.

The normalization technique known as the min-max scaler involves scaling data to a predetermined range, often spanning from 0 to 1 (Equation 1). This is achieved by performing two operations on the data: subtracting the minimum value and dividing by the range of the data. The utilization of a straightforward normalization technique can prove to be advantageous when dealing with data that possesses a predetermined range and remains unaffected by outliers. In certain scenarios, the utilization of a particular scale may be imperative for the algorithm to effectively process the data, rendering it advantageous. The min-max scaler technique is susceptible to outliers and may lead to a reduction of data integrity in the tails of the data distribution.

$$X_{std} = \frac{X - X.\min(\text{axis} = 0)}{X.\max(\text{axis} = 0) - X.\min(\text{axis} = 0)}$$
$$X_{scaled} = X_{std} * (\max - \min) + \min \tag{1}$$

here min, max = feature range

Conversely, the standard scaler is a normalization technique that converts the data to possess a mean of zero and a variance of one through the process of mean

subtraction and division by the standard deviation (Equation 2). This normalization technique exhibits greater resilience in handling data with unknown ranges and is comparatively less susceptible to the influence of outliers.

$$X_{scaled} = (x - u)/std \tag{2}$$

In the context of training samples, the variable u represents the mean, which is set to zero if $with\_mean = False$. Similarly, the variable std represents the standard deviation, which is set to one if $with\_std = False$.

4.2.4 Feature Selection

The process of Feature Selection is a statistical technique utilized to decrease the dimensionality of data by selecting relevant features and disregarding irrelevant ones within a given dataset [80]. It holds significant importance in the classification of cancer genes, as it aids in the reduction of data dimensionality and the identification of pertinent genes that are linked to cancer [52, 81, 82, 60]. Through the implementation of diverse feature selection methodologies, it is feasible to identify the most important genes that hold the utmost significance in the classification of the 33 distinct categories of cancer. The reduction of dataset complexity can enhance the precision and efficacy of classification models [83]. In general, the process of selecting features can be categorized into three types: Filter methods, Wrapper Methods, and Embedded methods [84].

The filter method is a frequently employed technique in the preprocessing of data. The present approach integrates ranking methodologies with primary criteria and employs sorting methodologies for variable selection. The wrapped approach involves identifying features that are appropriate for the machine learning algorithm employed. When employing a machine learning algorithm, the filter method is utilized prior to the wrapped method. The wrapped method is then employed during the machine learning process until an appropriate feature is identified.

Embedded methods refer to a class of techniques that preserve the iterations of

the model training process and selectively extract the most significant features that contribute to the training process for specific iterations. The regularization method is the prevalent embedded technique, which penalizes features by assigning a coefficient threshold. Several regularization algorithms include LASSO, Elastic Net, and Select From Model.

The current study employed four distinct feature selection methods, namely SelectFromModel [60], Lasso [83, 85], SelectKBest [86], and ElasticNet [87], to ascertain the most pertinent genes for the categorization of 33 diverse cancer types. The techniques were applied to the dataset that underwent standard scaler normalization to determine the top 100, 500, and 1000 genes. Subsequently, the chosen genes were employed in a classification framework to categorize the 33 distinct cancer types present in the dataset.

SelectFromModel: The SelectFromModel technique is a feature selection approach that relies on a specified estimator to identify the most significant features [60, 88]. The estimator undergoes training on the dataset, following which the feature importance scores are computed. Subsequently, a predetermined threshold is established to elect the most salient features that surpass the said threshold. This approach is advantageous due to its ability to autonomously identify significant features without necessitating extensive prior expertise. The function was provided with an estimator and a predetermined limit on the number of features to be considered for selection.

LASSO: The LASSO technique is a regression analysis approach that incorporates variable selection and regularization in order to enhance the predictive accuracy and interpretability of the resulting statistical model [89]. Its primary utility lies in feature selection. An $alpha$ value of 0.5 was employed, and a max_iter of 1000 was utilized to select the leading 100, 500, and 1000 genes.

SelectKBest: In contrast, SelectKBest represents a feature selection approach based on filtering [90]. The process involves the selection of k features with the highest score, as determined by a designated scoring function. The scoring function may encompass a range of statistical tests or $feature\_ranking$ techniques that assess the significance of the features. The $F-test$ scoring function,

specifically f_{classif}, was employed for classification purposes. The number of features varied from 100, 500, and 1000.

Elastic Net: The Elastic Net algorithm [91] is designed to optimize the trade-off between precision and weight magnitude in a linear model by utilizing both $L1$ and $L2$ regularization techniques [92, 93]. This approach has gained significant popularity in the field of bioinformatics for feature selection due to its tendency to generate parsimonious models with minimal non-zero weights. The $L1$ ratio was established at 0.4, and the top genes were selected based on the number of features.

All feature selection methods that were taken into consideration have been implemented within the machine learning package known as sci-kit-learn.

4.3 Classifiers

The process of selecting an appropriate classification model is a critical aspect of machine learning, given its potential to significantly influence the accuracy of the outcomes. Diverse classification algorithms exhibit distinct advantages and limitations that may impact their efficacy when applied to a particular dataset. This study outlines a pipeline in which seven distinct classification methodologies were employed, namely Random Forest, XGBoost, Logistic Regression, SVM, MLP, 1-D CNN, and TabNet, with the aim of determining the optimal models for our dataset. The selection of these models was based on their widespread usage, prior research, and their capacity to effectively manage data with a high number of dimensions [44, 41, 94, 95, 33, 96].

The performance of each model was assessed through the implementation of a stratified k-fold cross-validation technique, where k was set to 5. The utilization of the stratified k-fold technique is recommended for the proposed pipeline due to its ability to preserve the consistent distribution of each class across all folds. This approach minimizes the potential for bias and enhances the dependability of the outcomes. The data were partitioned into training and testing sets in an 80:20 ratio to assess the models' ability to generalize to new data.

Two normalization techniques, namely min-max and standard scaler, were employed in the analysis. Additionally, the baseline data was retained without any normalization applied. The purpose of this study was to conduct a performance comparison between models trained on normalized data and those trained on baseline data. The utilized feature selection algorithms encompassed SelectFromModel, Lasso, SelectKBest, and ElasticNet. Every algorithm possesses a distinct methodology for selecting the most significant features from a given dataset. Experiments were carried out without the implementation of any feature selection algorithm in order to establish a baseline performance for the classification models. Through the utilization of various feature selection algorithms and classification models, we have identified the optimal combinations for cancer classification. The aforementioned methodology not only enhances the precision of the categorization process but also diminishes the complexity of the issue, rendering it practically manageable to process vast sets of data. Figure 4.3 represents the classification procedure for different classifier models.

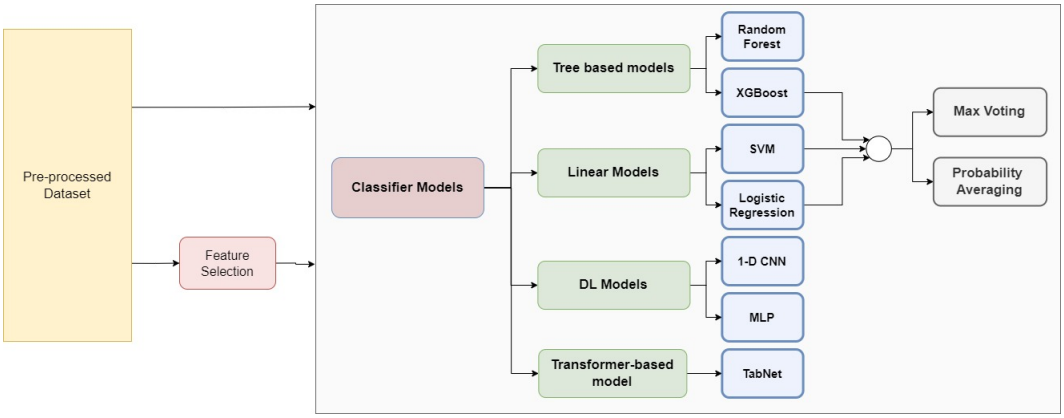


Figure 4.3: Classification Architecture

In order to maintain consistency in the comparison, the parameters used for model fitting were held constant throughout all experiments. The aforementioned analysis aided in identifying the feature selection algorithm that potentially enhanced the classification performance, as well as the classification model that exhibited superior performance on the given dataset.

Two commonly used tree-based models, namely XGBoost and Random Forest, were utilized in our experimentation. XGBoost and Random Forest are preferred

choices for gene classification in tabular data due to their high accuracy, feature importance analysis, robustness, and flexibility. These algorithms have proven to be effective in handling the complexities of gene expression data and providing reliable predictions for cancer-type classification[97, 98, 99, 100].

XGBoost: XGBoost is a gradient-boosting algorithm known for its superior performance in tabular data classification tasks. It relies on decision trees and employs a boosting methodology to enhance its efficacy. Boosting is an iterative technique that involves training multiple weak models and subsequently aggregating their predictions to generate a more resilient model. The XGBoost methodology utilizes optimized gradient boosting [101] by means of parallel processing, tree pruning, missing value handling, and regularization techniques to mitigate the risks of bias and overfitting (Figure 4.4). Moreover, XGBoost exhibits excellent performance on structured data and has demonstrated superior classification capabilities compared to numerous other methodologies. The XG-

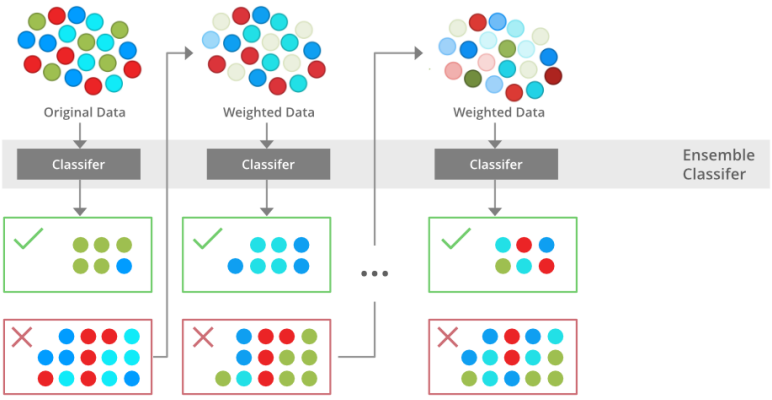


Figure 4.4: XGBoost

Boost hyperparameters were configured in the following manner for our experiments: $objective =$ "$multi : softprob$", $max\_depth = 4$, $learning\_rate = 0.1$, $n\_estimators = 1000$, and $early\_stopping = 10$. The concept of early stopping involves the cessation of the training process prior to its natural conclusion in the event that the performance on a validation set fails to exhibit improvement beyond a predetermined threshold over a specified number of iterations. The imple-

mentation of techniques aimed at mitigating overfitting and minimizing training time can be beneficial.

Random Forest: The Random Forest algorithm operates by generating numerous decision trees during the training phase and amalgamating their forecasts to yield a conclusive output. The construction of each decision tree involves the random selection of a subset of features and a subset of samples from the training data. The algorithm proceeds to partition the data into increasingly smaller subsets by utilizing the chosen features in a recursive manner, until either the maximum depth is attained or a predetermined stopping criterion is satisfied. The ultimate outcome of the Random Forest algorithm is established through the consolidation of prognostications from all decision trees. In the context of multiclass classification, the algorithm is designed to assign the class label that receives the greatest number of votes from the decision trees Figure 4.5. The hyperparameters for the Random Forest algorithm were configured to have a value of $n\_estimators = 20$. The aforementioned value was employed to achieve optimal accuracy while avoiding overfitting the model.

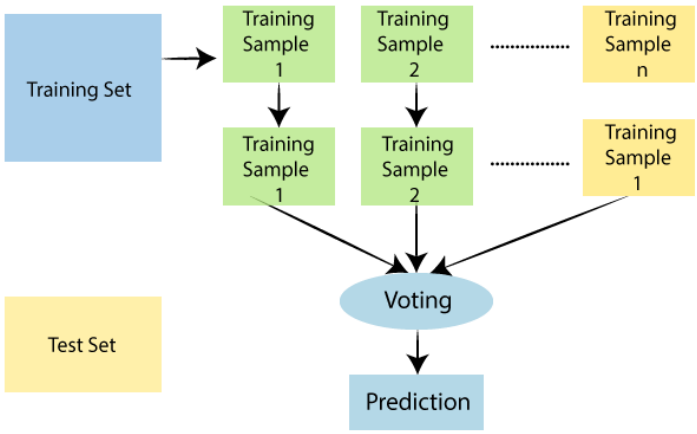


Figure 4.5: Random Forest

Logistic Regression and Support Vector Machines (SVM) are commonly employed algorithms in gene classification within the scope of traditional Machine Learning methodologies. The interpretability, feature selection capabilities, ro-

bustness to outliers, and scalability of decision trees make them a suitable choice for gene classification tasks [41, 102, 103, 104, 51, 105].

Logistic Regression: In the context of multiclass classification, logistic regression is utilized to estimate the probabilities of each class. This is achieved by fitting numerous binary logistic regression models. The algorithm acquires knowledge of the weights or coefficients linked to the features. Logistic regression calculates the probability of each class for new instances by utilizing the acquired weights and input features. It then predicts the class with the highest probability to make accurate predictions. The logistic regression model postulates a linear association between the predictor variables and the natural logarithm of the probability of the response variable (Figure 4.6). In our logistic regression model, we have set the *max_iter* parameter to 1000, which indicates the utmost number of weight updates the algorithm will perform during training. The aforementioned parameter has an impact on the convergence of the algorithm and establishes the number of iterations executed to minimize the loss function.

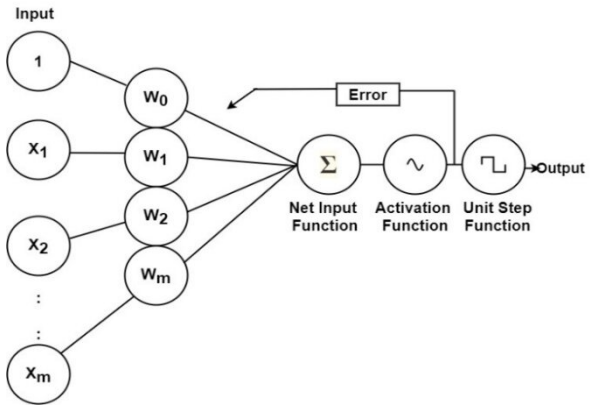


Figure 4.6: Logistic Regression

The regularization strength in logistic regression is governed by the *C* parameter. Regularization is a method employed to avoid overfitting in machine learning models by incorporating a penalty term into the loss function. A diminutive *C* value augments the regularization potency, thereby mitigating the influence of superfluous or cacophonous attributes and fostering more parsimonious models.

The value of C has been assigned as 100.
The logistic regression's penalty parameter plays a crucial role in determining the regularization type that is implemented. $l2$ regularization [106], also known as $L2$ or ridge regularization [107], has been selected for our model. The implementation of $L2$ regularization promotes the reduction of weights to a more moderate and equitable scale while avoiding the constraint of complete elimination. The utilization of this technique aids in the management of model complexity and the reduction of the impact of outliers or values that deviate significantly from the norm.

Support Vector Machines (SVM): The Support Vector Machines (SVM) algorithm is a robust machine learning technique that is extensively employed for multiclass classification in the analysis of gene cancer [51, 105, 108]. The method efficiently segregates diverse categories of gene expression data by identifying an optimal hyperplane that exhibits a maximal margin, defined as the spatial gap between the hyperplane and the nearest data points belonging to each category. Within the domain of gene cancer multiclass classification, the Support Vector Machine (SVM) algorithm is of significant importance in identifying the support vectors, which are the data points that are in the closest proximity to the decision boundary. It utilizes the gene expression levels linked with various cancer types to map the gene expression data to a feature space with high dimensions and then proceeds to identify the hyperplane that optimally separates the classes. Figure 4.7 shows the architecture of SVM In our particular analysis, we utilized the SVM implementation referred to as **svm.svc**, with specific parameter configurations. The selection of a '*linear*' kernel is based on the assumption that the gene expression data can be separated linearly in the feature space, which facilitates the implementation of a linear decision boundary. The value of the regularization parameter 'C' was established as 1, with the intention of achieving an equilibrium between the objectives of maximizing the margin and minimizing the training error. This approach was adopted to prevent the occurrence of overfitting or underfitting.

In order to achieve accurate convergence in the optimization process, the toler-

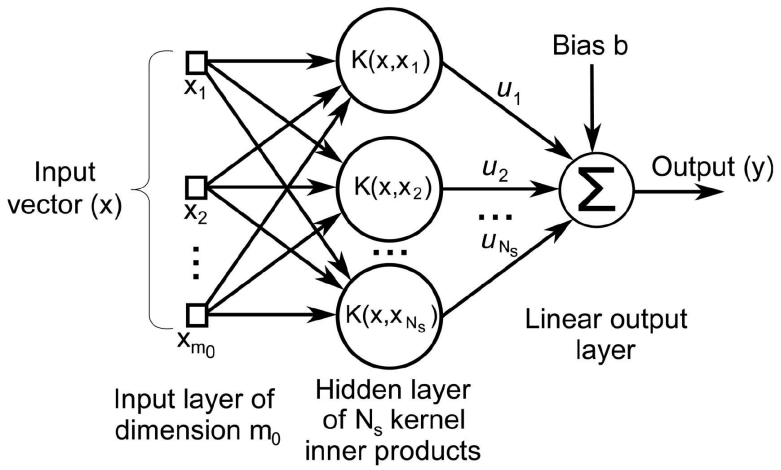


Figure 4.7: Support Vector Machines (SVM)

ance parameter '*tol*' was established at a value of $1e-5$. The aforementioned value serves as a determinant for the point at which the optimization procedure concludes, upon the occurrence of a decrease in the objective function that is less than or equal to the designated tolerance.

Moreover, the parameter '*decision_function_shape*' was configured as '*ovo*' (one-vs-one) [109], which employs a technique of generating binary classifiers for each pair of classes, and the ultimate prediction is determined by a voting mechanism. The aforementioned methodology effectively addresses the task of multiclass classification, while duly considering the unique cancer subtypes that are evident in the gene expression dataset. The objective of our study was to achieve efficient separation of various cancer classes using gene expression data through the configuration of SVM with specific parameter values. This was accomplished by leveraging the assumption of linear separability, appropriate regularization, precise optimization convergence, and efficient handling of multiclass classification.

The utilization of deep learning has emerged as an effective strategy for the classification of cancer owing to its ability to acquire intricate patterns and extract significant features from data with high dimensionality [38, 37, 110, 111, 42]. Deep learning models have the ability to automatically extract pertinent features

from vast genomic data sets, such as gene expression profiles or DNA sequences, without the need for manual feature engineering. The aforementioned capability enables a more extensive examination of the data, which has the potential to reveal concealed associations and biomarkers that could aid in precise cancer categorization. For the classification of 33 types of cancer using deep learning, two commonly used models are MLP (Multilayer Perceptron) and 1D-CNN (1-Dimensional Convolutional Neural Network).

Multilayer Perceptron (MLP): The MLP is a type of neural network that follows a feedforward architecture [112], comprising distinct layers of input, hidden, and output nodes. This neural network architecture is commonly regarded as a superficial variant of deep learning, characterized by a limited number of concealed layers (Figure 4.8). In the context of analyzing gene expression data, individual gene expression profiles are considered as input, and the output layer generates class probabilities for the given sample [113, 114, 115].

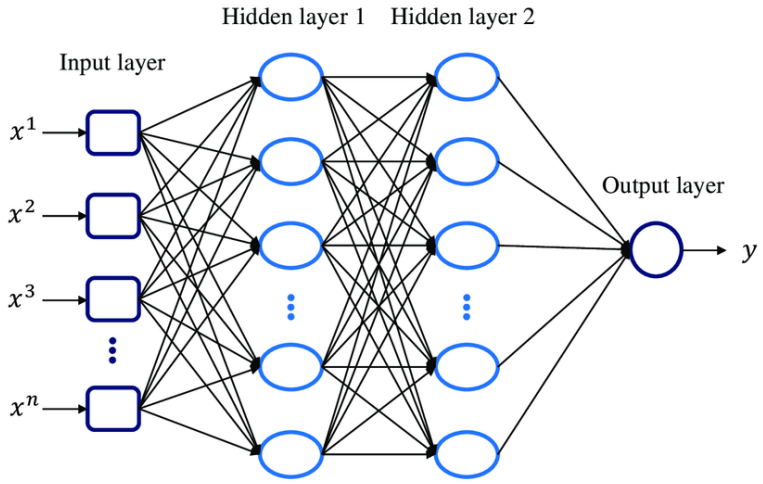


Figure 4.8: Multilayer Perceptron (MLP)

The MLPClassifier has been configured with multiple parameter values. A regularization strength of 0.001 is indicative of a modest $alpha$ value that serves to strike a balance between effectively capturing patterns in the data and mitigating the risk of overfitting. The selection of a $learning\_rate$ initialization value of 0.001 is intended to promote a consistent and incremental approach to model

convergence throughout the training process. The Rectified Linear Unit (*ReLU*) [116] activation function is recognized for its ability to efficiently capture non-linear associations within the data.

The Multilayer Perceptron (MLP) is composed of three hidden layers, each comprising 100 neurons. This architecture enables the model to acquire intricate representations by means of multiple layers of abstraction. The model's weight optimization process has been implemented using the '*Adam*' [117] solver, which is a highly effective stochastic gradient-based optimizer. A maximum of 200 iterations is set ('*max_iter*') to prevent superfluous training time and mitigate the risk of overfitting. The rationale behind selecting these parameters is to achieve an equilibrium between the intricacy of the model and its ability to perform well on unseen data.

1D-CNN (1-Dimensional Convolutional Neural Network): Several convolutional neural networks (CNN) models have been suggested for the purpose of predicting cancer types [42, 118]. The convolutional neural network architecture (Figure 4.9) utilized in this study takes gene expression data in the form of a vector and applies one-dimensional kernels. The 1D-CNN architecture that was implemented comprises multiple layers aimed at effectively modeling gene expression data for the purpose of cancer classification. The gene expression data is received by the input layer in the form of a sequence, with a maximum length of '*maxlen*'. In order to standardize the input, a *BatchNormalization* [119] layer is implemented, succeeded by a *Dropout* [120] layer with a rate of 0.3 to enforce regularization of the model and mitigate overfitting.

The normalized input is then sent through a *Weight Normalized* [121] *Dense layer* with 4096 units, which assists in capturing complex data relationships. The resulting output of the aforementioned layer is transformed into a tensor with dimensions of $(256, 16)$. Following this, an additional layer of *Batch Normalization* and *Dropout* is implemented for the purpose of regularization. In this study, a *Weight Normalized* 1*D Convolutional* layer is employed with 64 filters, a *kernel* size of 3, and Rectified Linear Unit (*ReLU*) activation. Subsequently, the feature

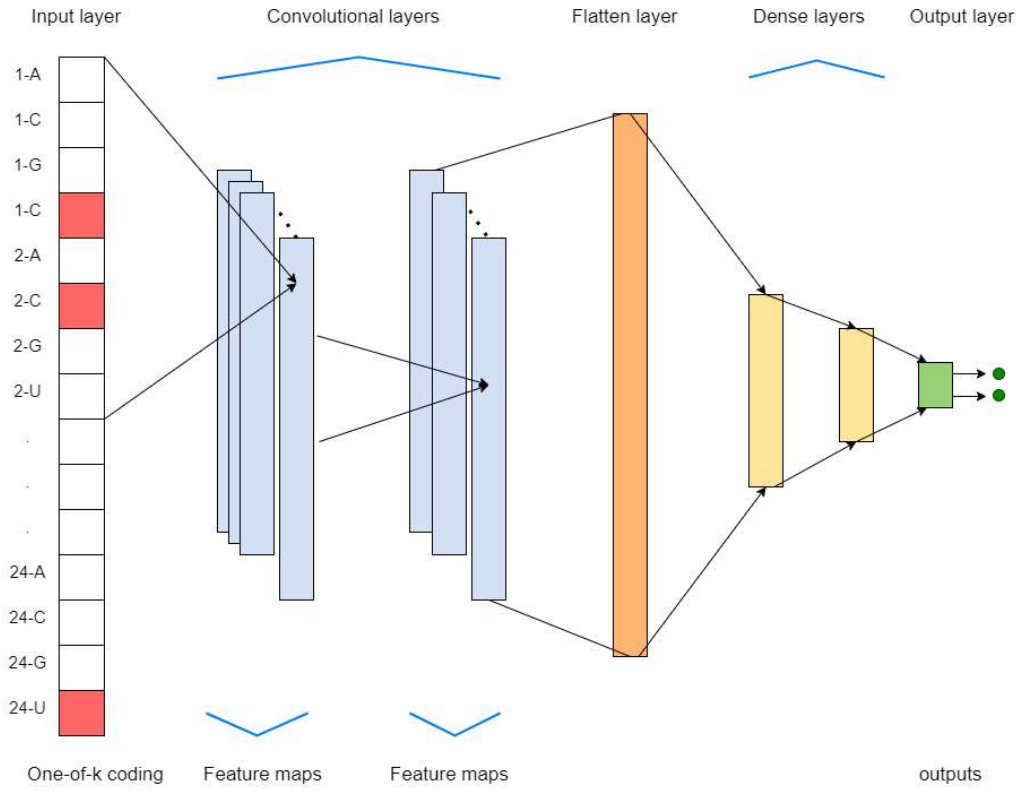


Figure 4.9: 1D-CNN

maps undergo *Average Pooling* for the purpose of downsampling. The incorporation of a residual connection serves to capture supplementary information from the preceding layer. An additional grouping of techniques includes Batch Normalization, Dropout, and Weight Normalized methods. The application of convolutional layers is succeeded by an element-wise multiplication operation between the output of said layers and the feature maps obtained from the preceding stage. Subsequently, the feature maps undergo a process of Max Pooling to achieve additional downsampling, following which the resulting tensor is flattened. The flattened tensor undergoes *Batch Normalization* and *ReLU* activation. Subsequently, the compressed tensor undergoes processing via a sequence of three fully connected layers, each comprising 512, 256, and 128 units, respectively. The activation function employed in each layer is Rectified Linear Unit (*ReLU*). The ultimate stratum comprises a set of '*num_classes*' entities that employ $softmax$ activation to generate the anticipated probabilities for every category. The compiled model employs the *sparse* categorical cross-entropy

loss function, utilizes the *Adam* optimizer with a $learning\_rate$ of 0.001, and assesses performance using accuracy as the evaluation metric.

The 1D-CNN architecture is designed to effectively capture relevant patterns and features in gene expression data for accurate cancer classification. This is achieved through the incorporation of normalization, dropout, weight normalization, convolutional, pooling, and fully connected layers. Although deeper convolutional neural network (CNN) models tend to exhibit higher accuracy in computer vision tasks when dealing with limited sample sizes in cancer-type prediction, it is advisable to use shallower models to mitigate the risk of overfitting and minimize the resources required for training [119, 120].

The efficacy of the transformer architecture in cancer classification [122, 123, 124] is attributed to its capacity to apprehend distant dependencies and acquire contextual associations from gene expression data. The self-attention mechanism is employed by transformers to represent the interrelationships among various positions in the input sequence. Moreover, transformers utilize numerous attention heads and layers, facilitating the acquisition of hierarchical representations and the incorporation of both local and global dependencies. Based on the aforementioned reasons, we have made the decision to incorporate the Attentive Interpretable Tabular Learning neural network (TabNet) as the classification model for our study.

Attentive Interpretable Tabular Learning neural network (TabNet): The effectiveness of the TabNet architecture in cancer classification can be attributed to its distinctive amalgamation of attention mechanisms and tabular data handling [125, 126, 127]. The TabNet model integrates a revised version of the Transformer attention mechanism, enabling the model to discern and focus on significant features and acquire intricate associations within the data (Figure 4.11). Furthermore, TabNet employs a feature selection mechanism in its training process that dynamically chooses a subset of features to concentrate on. The incorporation of interpretability into the model not only enhances its performance in terms of generalization but also mitigates the risk of overfitting.

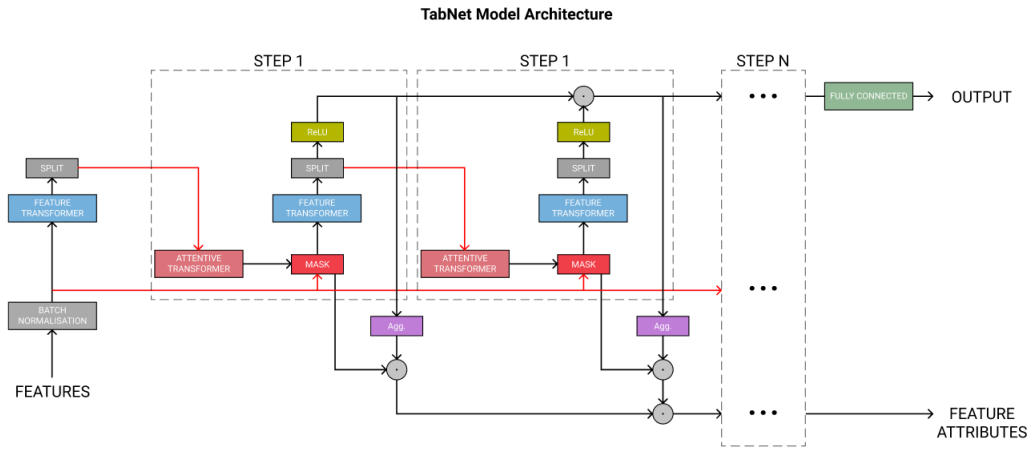


Figure 4.10: TabNet

The *Adam* optimizer with default parameters was utilized to train the model, and a *StepLR* scheduler was implemented to modulate the learning rate throughout the training process. The scheduling algorithm was set up with a step size of 10 and a *decay rate* (*gamma*) of 0.9. Through the implementation of a *dynamic learning rate*, the model demonstrated efficient convergence towards an optimal solution. In order to regulate the training procedure, a threshold of 150 *epochs* was established and a strategy of early stopping was executed with patience of 60 *epochs*. In order to enhance memory utilization, a *batch size* of 512 and a *virtual batch size* of 512 were employed. The weighting of the samples was established as 1, signifying that equivalent significance was attributed to all occurrences in the training dataset.

4.4 Ensemble Approach

The utilization of ensemble approaches in combination with classification models can yield significant advantages. The amalgamation of various discrete models into a unified ensemble model can yield a substantial enhancement in the overall efficacy and precision of the classification undertaking [128, 129]. Furthermore, the utilization of ensemble approaches has the potential to enhance stability and dependability through the mitigation of the influence of outliers or noisy data points. We have implemented two types of ensemble approaches -

Max-Voting and Probability-Averaging on three of our top-performed classification models- Logistic regression, SVM, and XGBoost.

Max voting employs multiple independent classifiers or models for training and utilizes majority voting for making predictions. Each model generates a prediction and the class label with the highest number of votes is chosen as the ultimate prediction. The ensemble model utilizes the notion that various models possess unique advantages and disadvantages. By amalgamating their forecasts, the ensemble model gains from their combined knowledge.

Probability averaging entails training multiple classifiers and averaging their predicted probabilities for each class. The approach of averaging the predicted probabilities of each model is utilized to obtain a more precise estimation of class probabilities, rather than relying on the majority vote. This methodology incorporates model confidence and assigns greater weight to dependable predictions, leading to improved accuracy and calibrated probability estimation.

We have chosen an odd number of classification models to apply ensemble methodologies for mainly two reasons. First, when making predictions, an odd number of models allows for the prospect of a majority vote. When each model in the ensemble predicts a class label, an odd number of models ensures that there is always a majority class when voting. This can aid in decision-making and reduce the likelihood of ties, which can occur when an even number of models are considered. Second, ensembling an odd number of models can help in handling outliers or noisy predictions. In the case of outlier predictions from a single model, having an odd number of models permits the ensemble to effectively disregard these outliers and rely on the predictions of the majority of models.

4.5 Feature Attribution

Feature attributions indicate how much each feature in the model contributed to the predictions for each given instance [70]. When there is a request for predictions, we get predicted values as appropriate for our model. When we request explanations, we get the predictions along with feature attribution information

[130]. With the help of feature attribution, we will be identifying the cancer-specific genes and also the patient-specific gene set for each cancer. We will be using SHAP [131] for feature attribution in our work.

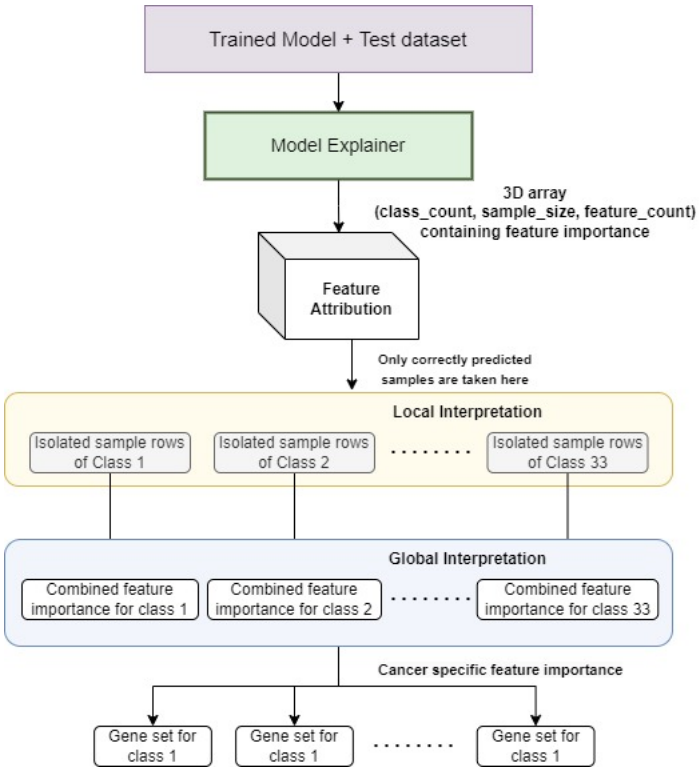


Figure 4.11: Explainability analysis

The SHAP technique is a mathematical approach utilized to elucidate the predictions generated by machine learning models. This approach is grounded in the principles of game theory. It can elucidate the prognostications of any machine-learning model through the computation of the individual contribution of each feature to the prediction. Some advantages of SHAP [132, 133] are:

- The utilization of Shapely values ensures that the forecast is equitably allocated among diverse characteristics.
- The global interpretation is obtained by calculating the Shapely values for an entire dataset and subsequently aggregating them.
- This technique establishes a connection with other interpretability methods, such as LIME.

- SHAP has a lightning-fast Tree-based model explainer.

4.5.1 Calculating the Feature Attribution Score

When calculating feature attributions using SHAP, different types of SHAP explainers need to be used depending on the category of the trained model. Here are some commonly used SHAP explainers for different types of models:

- TreeExplainer: This explainer is suitable for tree-based models such as decision trees. We used this for Random Forest and XGBoost.
- LinearExplainer: This explainer is specific to linear models and provides SHAP values based on the coefficients of the linear model. We used this one for Logistic Regression and SVM.
- DeepExplainer: This explainer is designed for deep learning models. We used it on 1D CNN and MLP.

To calculate feature attributions using SHAP, we typically need to pass test data to the explainer along with the trained model. The explainer then uses the model and the test data to generate SHAP values, which represent the contribution of each feature to the output prediction for each instance in the test data. SHAP gives a 3D array with dimensions ($classes$, $sample\_size$, $feature\_size$). So from this array, we have to extract the relevant rows for our work.

4.5.2 Identifying the Correctly Predicted Samples

When identifying global and local specific gene sets for each cancer, we only focused on the correctly predicted samples made by the model. This approach aims to extract the genes that positively contribute to accurate predictions, as the genes from incorrectly predicted samples could introduce noise or incorrect signals to the important gene set.

By considering only the correctly predicted samples, you can prioritize the genes that consistently and positively influence the model's ability to make accurate

predictions for each cancer type [134]. This helps to reduce the impact of potential false positives or misleading gene associations that may arise from incorrectly predicted samples.

Focusing on correctly predicted samples enables the identification of gene sets that are more likely to be biologically relevant and specific to each cancer type [135]. These genes can potentially offer insights into the underlying molecular mechanisms and pathways associated with the specific cancer types, aiding in further understanding and targeted research in cancer biology.

4.5.3 Extracting the Feature Attribution for a Relevant Sample List of Each Cancer

SHAP array contains feature attribution for each cancer for each of the samples. But the samples do not belong to every class. So we have taken the sample list from each cancer of the main dataset and according to that, we have separated the rows from the SHAP array for each cancer. Then from this list, we discarded the wrong predicted sample lists. On these sets, we have done further calculations which lead to global and patient-specific gene sets.

This approach allows for a more focused and accurate exploration of the gene expression patterns associated with each cancer type. By considering only the correctly predicted samples and focusing on relevant feature attributions, we can obtain more meaningful and biologically significant gene sets.

4.6 Cancer-specific Gene Set

We have collected the gene sets for the correctly predicted samples of each cancer and these scores combined give an idea of the global importance of these genes. To determine the global importance of genes, we calculated the median of each column (gene) across the collected gene sets specific to particular cancer [136]. By calculating the median, we obtained a representative value that reflects the central tendency of the gene's importance across the correctly predicted samples.

After obtaining the median values for each gene, we sorted them in descending order for each cancer type. This sorting allows us to rank the genes based on their importance scores, with higher scores indicating greater importance in the context of a specific cancer type. So if we now select the top 500 cells from a cancer row, these will be the top 500 important features of that specific cancer.

4.7 Patient-specific Gene Set

We applied feature attribution for each sample in the dataset. Then we will get the specific genes for each patient for all 33 cancers. For a specific cancer, the results of all patient-specific gene sets will be analyzed which can provide a set of genes that will be greatly helpful for gene therapy and can be a stepping stone for precision medicine.

To determine the patient-specific genes, we sort the contribution values in descending order for each cancer type from the lists of scores obtained after performing earlier tasks. This sorting allows us to rank the genes based on their importance scores for a specific patient.

4.8 Statistical Validation

Statistical validation is essential in data analysis to establish the significance of the results and eliminate the possibility of random chance. Validation is crucial in studying cancer-specific gene sets to confirm their biological significance.

Differential Gene Expression (DGE) analysis is a frequently employed approach for detecting genes that display varied expression levels across distinct sample groups [137, 138]. DGE analysis enabled the identification of genes exhibiting significant expression level variations across various cancer types. This analysis facilitated the identification of genes that may have a significant role in contributing to individual types of cancer. DESeq2, a commonly used software tool, was employed for statistical genomic analysis and quantification of differential gene expression [133, 99].

4.8.1 DESeq2

DESeq2 is a feasible tool for analyzing differential gene expression in RNA-seq data. DESeq2 employs negative binomial generalized linear models to detect differential expressions. Using empirical Bayes approaches, it calculates priors for log fold change and dispersion as well as posterior estimates for these values [72]. DESeq2 involves several steps for differential expression analysis, as illustrated in Figure 4.12. DESeq2 employs *normalization* factors (size factors) to model raw counts and address discrepancies in library depth. The method will estimate gene-wise dispersion and subsequently reduce these estimates to enhance the accuracy of dispersion estimates for count simulation. Finally, it employs the Wald test or the Likelihood Ratio Test to fit the negative binomial model and conduct hypothesis testing.

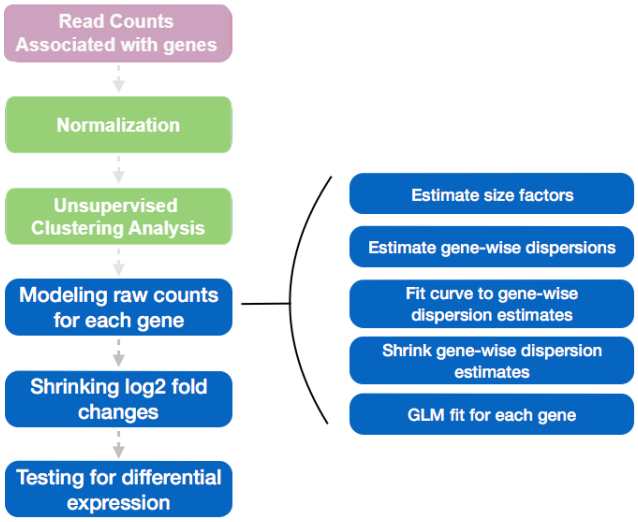


Figure 4.12: DESeq2 working principle

In this analysis, we utilized raw counts of gene expression values of both tumor samples and healthy tissue samples [139, 140]. The low counts were removed, keeping rows that have at least 10 reads. The factor level was set to "healthy tissue". A Differential Gene Expression (DGE) analysis was conducted on individual samples of each cancer type to identify cancer-specific genes that have a significant impact on each type of cancer. These genes represent the overall behavior of the population with respect to each cancer type and can be viewed as

Index	Cancer type	Tumor sample count	Healthy tissue sample count	Total differentially expressed gene
1	BLCA	364	19	802
2	BRCA	984	112	707
3	CESC	261	4	1337
4	CHOL	33	11	1904
5	COAD	287	43	878
6	ESCA	185	13	818
7	HNSC	462	44	889
8	KICH	62	27	1449
9	KIRC	477	74	821
10	KIRP	238	31	768
11	LIHC	297	50	737
12	LUAD	505	61	1026
13	LUSC	491	53	1769
14	PRAD	428	50	266
15	READ	89	12	883
16	STAD	382	35	637
17	THCA	443	55	445
18	UCEC	143	25	1205

Table 4.1: Cancer wise total number of differentially expressed genes using DESeq2

global features. We used a threshold of $|log2Fold-change| > 3$ and an adjusted $p-value < 0.001$ to identify differentially expressed genes. Table 4.1 presents a comprehensive summary of the sample sizes for both tumor and healthy tissues, along with the aggregate count of genes that exhibit differential expression across 18 distinct cancer types.

Chapter 5

Experimental Results

5.1 Experimental Setup

In this section, we present the experimental setup and describe the various factors involved in the training and testing phases of our cancer-specific classification pipeline.

5.1.1 Environment

In our work, we utilized both R and Python for different stages of the project. RStudio was employed to apply DESeq2, a popular R package, for obtaining global important features specific to individual cancers. On the other hand, the Python environment was utilized for training the classification model and conducting the remaining experiments. The specific details of the Python environment used are as follows:

- Processor: Intel® Core™ $i9-12900K$ (12 cores, 24 threads)
- RAM: 128 GB
- GPU: 2 $\times$ NVIDIA GeForce RTX 3090
- GPU Memory: 2 $\times$ 24 GB

5.1.2 Dataset Split

The performance of each model was assessed through the implementation of a stratified $k-fold$ cross-validation technique, where k was set to 5. The utilization of the stratified $k-fold$ technique is recommended for the proposed pipeline due to its ability to preserve the consistent distribution of each class across all folds. This approach minimizes the potential for bias and enhances the dependability of the outcomes. In accordance with the Pareto principle [141],

which is also referred to as the 80/20 rule or the principle of factor sparsity, the data in the sample were divided into two sets: 80% for training and 20% for testing, as documented in reference [142]. A validation set comprising 20% of the data from the training set is utilized in relation to the entire dataset.

5.2 Evaluation Metrics

5.2.1 Accuracy

Accuracy is a machine learning evaluation metric that measures the ratio of correct predictions made by a model to the total number of predictions. The accuracy is determined by the formula $N/M * 100\%$, where N represents the number of correct predictions and M represents the total number of samples. Equation 3 represents the detailed equation. of accuracy score.

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{3}$$

The accuracy score is determined by evaluating the model's performance on a test set comprising previously unseen data samples. This guarantees a more dependable evaluation of the model's capacity to generalize and provide precise predictions on unobserved data. Accuracy measurement assesses the model's performance in accurately classifying data. A higher accuracy score reflects the model's effectiveness in making accurate classifications by indicating a higher percentage of correct predictions.

5.2.2 Precision

Precision is defined as the ratio of true positives to the total number of positive predictions made across all classes (Equation 4). In a multi-class problem, precision is employed to assess the proportion of correctly classified instances for each class relative to all instances that were classified as belonging to that particular class. In the context of imbalanced classification problems featuring multiple classes, precision is computed by dividing the sum of true positives for

all classes by the sum of true positives and false positives for all classes. The outcome corresponds to a numerical value ranging from 0.0, indicating the absence of precision, to 1.0, representing complete or flawless precision.

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4}$$

5.2.3 Recall

Precision is defined as the ratio of true positives to the total number of positive predictions made across all classes (Equation 5). In a multi-class problem, precision is employed to assess the proportion of correctly classified instances for each class relative to all instances that were classified as belonging to that particular class. In the context of imbalanced classification problems featuring multiple classes, recall is computed by dividing the sum of true positives for all classes by the sum of true positives and false negatives for all classes.

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5}$$

Here, recall refers to situations where a gene for a certain cancer type, such as BLCA, is wrongly predicted or categorized as CHOL rather than the actual cancer type BLCA. In other words, recall assesses the proportion of predicted and classified cancer type T genes among the actual cancer type T genes. The outcome corresponds to a numerical value ranging from 0.0, indicating the absence of precision, to 1.0, representing complete or flawless precision.

5.2.4 F1 score

The F1 score is a widely employed metric for the assessment of cancer classification models. The harmonic mean of precision and recall is a metric that offers a balanced evaluation of a model's performance with respect to positive and negative predictions. Precision measures the capacity of the model to accurately identify positive cases within the predicted positives. Recall quantifies the capacity of the model to accurately detect positive instances in the presence of

both true positives and false negatives.

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

The F1 score is a metric that integrates precision and recall into a single measure, assigning equal weight to both performance indicators. The formula for calculating a specific metric involves multiplying precision and recall, doubling the product, and then dividing the result by the sum of precision and recall (Equation 6). The F1 score is a metric that varies between 0 and 1, where a value of 1 represents ideal precision and recall. The F1 score holds significance in the classification of cancer owing to its ability to account for both false positives and false negatives, which are critical factors in medical diagnosis. It provides a balanced assessment of the model's overall accuracy and is commonly used when there is an imbalance between positive and negative cases.

5.3 Performance Analysis

In our study, we conducted experiments on our dataset with seven different classification methods. These algorithms are examples of numerous machine learning method branches that are frequently used for tabular classification tasks. The objective was to investigate and evaluate how various model types performed in our particular scenario.

Two of the algorithms we employed, XGBoost [63] and Random Forest [45], utilized tree-based models. Tree-based models are well-suited for tabular data analysis because of their propensity to capture intricate linkages and interactions between attributes. Additionally, we used the Logistic Regression [61] and Support Vector Machine (SVM) [62] classical machine learning techniques. These algorithms have a strong foundation in statistical modeling and classification problems and are widely employed. They deliver outcomes that are easy to understand and are proficient at working with huge datasets.

We used two separate methods—Multi-Layer Perceptron (MLP) [64] and 1-D Convolutional Neural Network [65] (1-D CNN)—to combine deep learning tech-

niques. Deep learning models are excellent at automatically deriving complex representations and patterns from incoming data. While 1-D CNN uses convolutional layers to identify regional patterns in sequential data, MLP is a flexible architecture with tightly connected layers. In addition, we used the TabNet [67] transformer-based paradigm. Transformers have drawn a lot of interest in problems involving natural language processing, but they also have potential in tabular data analysis. To handle tabular data, TabNet adapts the transformer design, and it has proven to be quite effective in a number of classification tasks.

We tried to cover the main machine learning methodologies widely used for tabular classification by incorporating these various algorithms. This gave us a chance to compare how well they performed and determine which approaches worked best for the particular dataset and classification issue we were dealing with.

5.3.1 Performance on the Baseline Dataset, $n_{features}$=19238

We implemented the aforementioned classifier models using the baseline data consisting of 19,238 gene features. This gene feature dataset served as the input for our classification models, allowing us to leverage genetic information to predict and classify the target classes. The results given by the classification algorithm at the initial stage are shown in Table 5.1.

In the analysis of the gene expression tabular dataset using machine learning models, it was observed that Logistic Regression achieved the highest accuracy of 96.43%. SVM, XGBoost, and MLP also performed well, with accuracies over 95%. Based on the high accuracies achieved by these models, the top three models, namely Logistic Regression, SVM, and XGBoost were selected for ensembling. Two ensembling techniques were applied: Probability Averaging and Max Voting. When using Probability Averaging, which combines the predicted probabilities from the individual models, the ensemble accuracy improved to 96.45%. This indicates that combining the predictions from multiple models can enhance the overall performance. Similarly, with Max Voting, which selects the majority class prediction among the individual models, the ensemble accuracy

Dataset	Normalization	Feature Selection	Number of features	Classifier Models	Accuracy	Precision	Recall	F1-Score	Ensemble Accuracy (Probability Averaging)	Ensemble Accuracy (Max Voting)
TCGA Pan-Can Dataset [56]	—	—	19238	Logistic Regression	96.43%	0.9501	0.9420	0.9391	96.45%	96.33%
				SVM	96.30%	0.9418	0.9322	0.9387		
				XGBoost	95.52%	0.9331	0.9189	0.9203		
				MLP	95.34%	0.8597	0.8669	0.8673		
				Tabnet	94.56%	0.9183	0.9218	0.9129		
				1D-CNN	94.50%	0.8320	0.8660	0.844		
				Random Forest	92.17%	0.9082	0.9115	0.9033		
	Standard Scalar	SelectFrom Model	500	Logistic Regression	96.39%	0.9484	0.9388	0.9382	**96.60%**	**96.54%**
				SVM	95.87%	0.9352	0.9373	0.9344		
				XGBoost	95.10%	0.9331	0.9392	0.9315		
				Random Forest	93.00%	0.8994	0.8736	0.87526		

Table 5.1: Summary of classification model performance

was slightly lower at 96.33%. Nonetheless, it still outperformed the standalone models. Probability averaging outperforms max voting in ensemble methods due to its ability to capture more nuanced information from individual models. In probability averaging, the predicted probabilities of each class from multiple models are averaged, resulting in a more precise estimation of the class probabilities. This approach takes into account the confidence levels of each model and provides a more reliable prediction. On the other hand, max voting simply selects the class with the highest number of votes among the ensemble of models. While it can be effective when the models are highly accurate and have similar performance, it may not take into account the relative confidence or certainty of each model's prediction. Overall, the ensembling of the top three models resulted in slightly improved accuracy compared to the highest accuracy achieved by a standalone model (Logistic Regression). This highlights the effectiveness of ensembling techniques in harnessing the predictive power of multiple models.

It is important to note that accuracy may not provide a complete assessment of the model's performance in the case of datasets that are imbalanced, where one class is much more abundant than the other. Accuracy is influenced by the class distribution and tends to favor the majority class, which can result in misleading outcomes. The F1 score is a more informative metric for assessing model performance in such situations. The F1 score evaluates the model's performance on imbalanced datasets by taking into account both false positives and false negatives. The F1 score is a useful metric in imbalanced datasets where the minority class is of interest. The evaluation metric takes into account both recall, which is the ability to accurately identify positive instances, and precision, which is the ability to minimize false positives. The F1 score offers a balanced assessment of a model's performance on imbalanced datasets by weighing the trade-off between two metrics. The table shows that the F1 score for Probability averaging is 93.99% and for Max Voting is 93.86%. The high F1 scores indicate that the models can classify each of the classes accurately.

5.3.2 Performance on the Normalized Dataset

We evaluated classifier models using two data preparation techniques: the min-max scaler and the standard scalar. Each preprocessing strategy was tested along with each classifier model to see how they affected performance improvement. We wanted to investigate potential gains in model performance by using the min-max scaler, which scales features to a defined range (usually 0 to 1), and the standard scalar, which adjusts features to have zero mean and unit variance.

When compared to the min-max scaler, the use of the data preprocessing technique, StandardScaler, with the classifier models resulted in a considerable improvement in overall accuracy. The models' accuracy ranged from 96.89% to 91.65%, with Logistic Regression scoring the highest at 96.89%, followed by SVM and XGBoost. Several things have contributed to this progress. To begin, StandardScaler's Gaussian Distribution Assumption assumes that the data has a Gaussian distribution. When the data roughly follows this distribution, using StandardScaler produces superior results. Furthermore, certain algorithms, such as logistic regression and neural networks with weight regularization, are more compatible with StandardScaler since they rely on the mean and variance of the features. Another advantage of StandardScaler is that it keeps the features' interpretability by retaining their original distribution while scaling them. This can be critical in situations when feature interpretability is critical. MinMaxScaler, on the other hand, scales the data to a defined range, which may change the interpretation of the features because they are mapped to a new scale.

Based on the prior evaluation findings, the top three models for ensembling were chosen: Logistic Regression, SVM, and XGBoost. To combine the predictions of these models, two ensembling approaches, Probability Averaging, and Max Voting, were used.

The ensemble accuracy decreased to 96.31% when Probability Averaging was used. This technique computes the average probability for each class by combining the expected probabilities from each individual model. The ensemble accuracy indicates that the ensemble model outperformed the individual models

in terms of accuracy. Max Voting, which chooses the majority class prediction among the different models, produced a slightly lower ensemble accuracy of 96.23%. The prediction of each model is considered in this technique, and the class with the most votes is chosen as the ensemble prediction.

However, it is important to emphasize that in this circumstance, the ensemble of models did not produce sufficient results. Despite minor accuracy disparities between the two ensembling procedures, neither approach considerably improved the overall accuracy of the individual models.

5.3.3 Performance on the Dataset with Normalization and Feature Selection

In the analysis of the gene expression tabular dataset using ML models, the data was first standardized using Standard Scalar normalization. Four feature selection techniques were applied: Lasso, SelectFromModel, Select-K-Best, and ElasticNet. Subsequently, four models (Logistic Regression, SVM, XGBoost, and Random Forest) were trained on the selected features, with 100, 500, and 1000 features chosen for each technique. It is important to note that the feature selection methods used are not compatible with 1D CNN and MLP models.

This resulted in a total combination of 48 models. Among all the models, the SelectFromModel technique with 500 features yielded the best performance, followed by Lasso, Select-K-Best, and ElasticNet performing the worst. For Logistic Regression, the accuracy achieved was 96.39%, which was the highest among the four models considered (Table 5.1). The confusion matrix is shown in Figure 5.1. SelectFromModel outperformed the other techniques because it has the ability to automatically select the most relevant features based on their importance in the classification task. It uses a machine learning model, such as a decision tree or random forest, to determine the feature importance and selects the top features accordingly. This approach allows SelectFromModel to capture the most discriminative information from the input data, leading to improved classification accuracy.

Lasso, on the other hand, performs well but not as effectively as SelectFromModel. Lasso employs $L1$ regularization, which encourages sparse feature se-

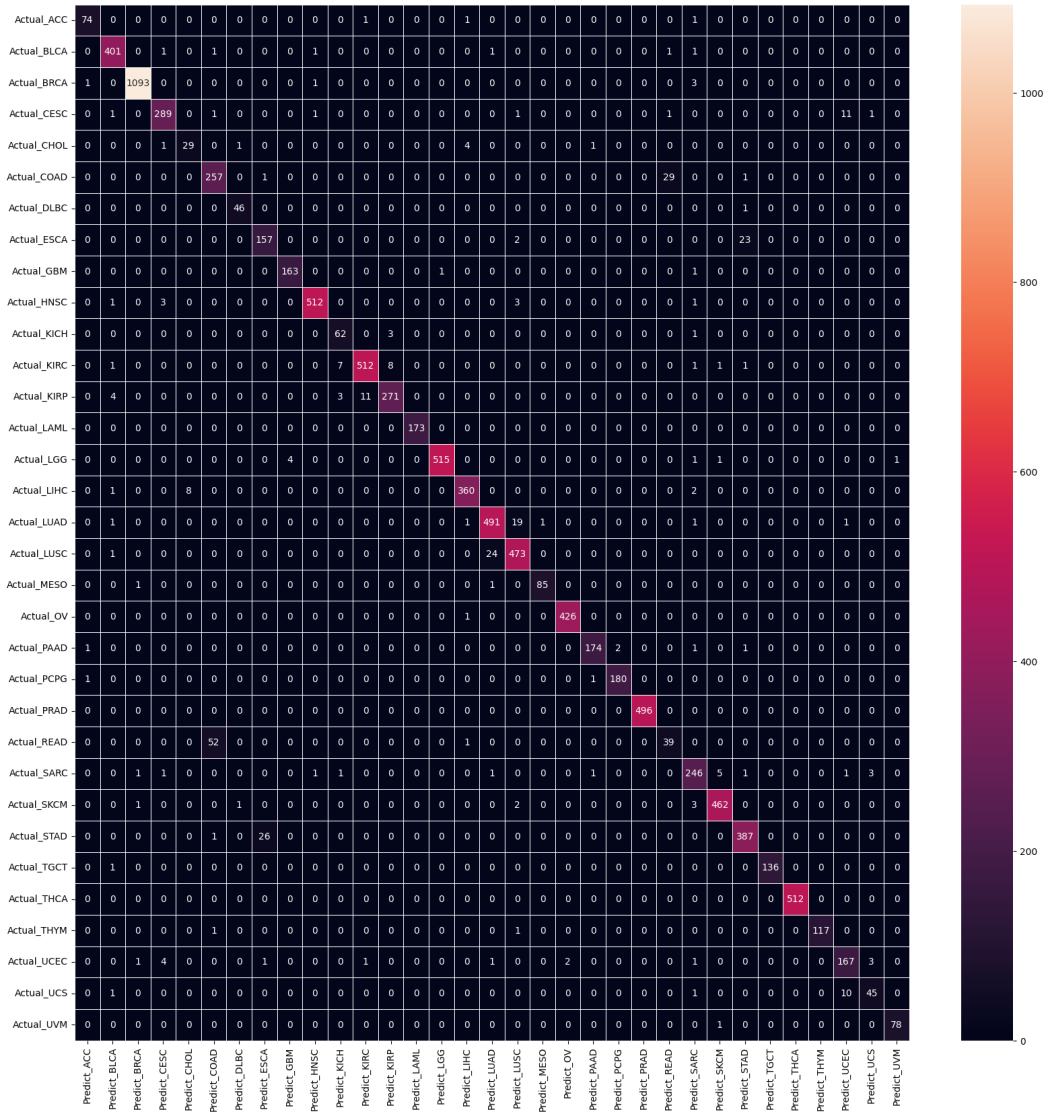


Figure 5.1: Confusion matrix for Logistic Regression SelectFromModel

lection by shrinking the coefficients of less important features to zero. While Lasso can effectively reduce the number of features, it may still retain some less informative features, leading to slightly lower performance compared to SelectFromModel. Select-K-Best is a simple feature selection technique that selects the K-best features based on a univariate statistical measure, such as chi-square or mutual information. While Select-K-Best can be useful in reducing the dimensionality of the data, it may not capture the complex interactions between features and their relationship with the target variable as effectively as SelectFromModel or Lasso. ElasticNet, combining $L1$ and $L2$ regularization [143], aims to balance

between Lasso and Ridge regression. However, in this particular classification task, it did not perform as well as the other techniques. This could be due to the nature of the data and the specific characteristics of the cancer classification problem, where the $L1$ regularization of Lasso or the feature importance-based selection of SelectFromModel may be more suitable.

Ensembling techniques were further applied to enhance the performance. Probability averaging, which combines the predicted probabilities from the individual models, yielded an accuracy of 96.60%. Max Voting, which selects the majority class prediction, achieved an accuracy of 96.54%. These ensemble results surpassed the performance of all standalone models as well as the ensembling results obtained using the raw data and only normalized data.

In addition to accuracy, the $F1-scores$ were also evaluated. The $F1-score$ for probability averaging was 94.14, and for max voting, it was 93.75. These scores provide an assessment of the model's precision and recall, considering both false positives and false negatives.

The observed improvement in ensembling results demonstrates the effectiveness of combining the predictions from multiple models, particularly when using the SelectFromModel technique with 500 features. This suggests that this combination of feature selection and ensembling can provide valuable insights into gene expression patterns and enhance the overall predictive performance.

5.3.4 Comparison with State-of-the-Art Approaches

Our present study involved a comparative analysis aimed at assessing the efficacy of the proposed architecture for multiclass cancer classification. We conducted a comparative analysis with various state-of-the-art [41, 37, 38] techniques that have been established on the TCGA pan-cancer dataset and executed a classification task on 33 distinct cancer categories. Table 5.2 displays the accuracy of performance for each of the architectures, including the architecture proposed by us. The findings unambiguously indicate that the architecture we have proposed exhibits superior performance compared to all of the pre-existing methodologies.

Number of Cancer Types	References	Number of Genes	Avg. Accuracy
33	Hsu, Yi-Hsin et al. [41] - Linear SVM	9900	0.9498
	Lyu et al. [37] - Deep Learning based algorithm	10381	0.9559
	de Guia, Joseph M. et al. [38]- DeepGx	20531	0.9565
	Proposed architecture	500	**0.966**

Table 5.2: Comparison with other state-of-the-art architectures

One crucial advantage of the proposed architecture is its ability to achieve superior performance while utilizing a significantly reduced number of features. Our architecture necessitates only 500 features, which is significantly lower in comparison to alternative approaches. The decrease in the number of features significantly aids in lessening the computational resources necessary for classification assignments. Overall, the proposed architecture exhibits superior performance accuracy and efficient utilization of computational resources, rendering it a promising solution for multi-class cancer classification.

5.3.5 Feature Attribution Validation with Deseq2 ($n_{features=19238}$)

In the analysis of the gene expression tabular dataset using ML models, the SHAP (Shapley Additive exPlanations) technique was applied to the trained models. This allowed the identification of cancer-specific gene sets for each model. To validate these gene sets, a comparison was made with the gene sets provided by the statistical analysis tool Deseq2 (Figure 5.2).

Deseq2 is known to provide important gene sets specific to each cancer type. For each model, the intersected genes were obtained by comparing the gene lists generated by Deseq2 with the top 500 features from the lists obtained from the model for each cancer type. The counts of intersected genes are summarized in the table. The significant number of common genes observed between the gene sets obtained from the ML models and the gene sets provided by Deseq2

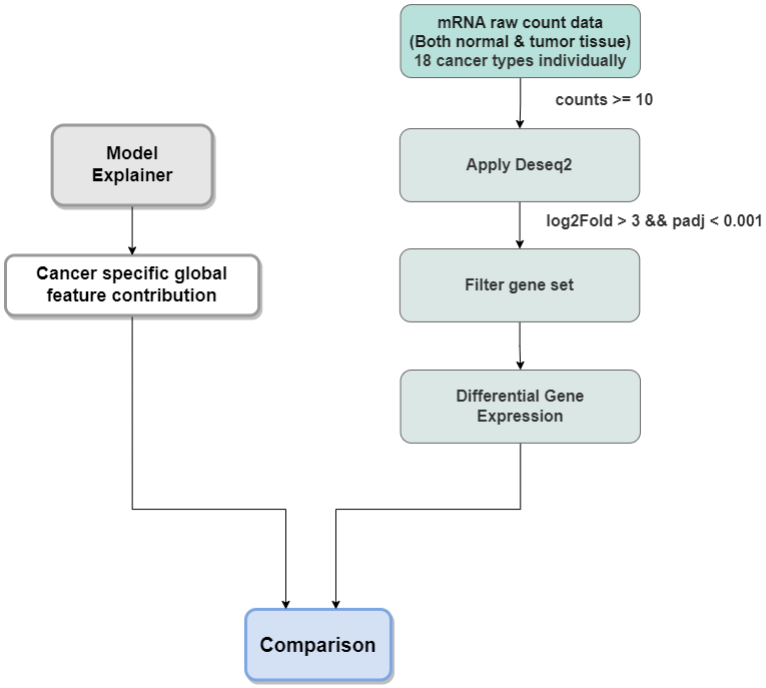


Figure 5.2: Feature attribution validation with Deseq2

indicates that these genes are likely to be important for the corresponding cancer types.

This finding suggests that the ML models were successful in capturing and identifying genes that have a strong association with specific cancer types. By using SHAP and comparing the results with Deseq2, confidence is gained in the importance and relevance of the identified gene sets.

These common genes can provide valuable insights into the underlying biology of the specific cancer types and potentially serve as potential biomarkers or therapeutic targets. Further analysis and exploration of these gene sets can contribute to a better understanding of the molecular mechanisms and pathways involved in cancer development and progression. It is important to note that the validation of the gene sets using Deseq2 adds credibility to the findings, but further experimental validation or integration with other genomic datasets may be required to confirm their functional relevance in the context of cancer.

5.3.6 Feature Attribution Validation with Deseq2 ($n_{features=500}$)

In addition to applying SHAP to the trained models, SHAP was also applied specifically to the SelectFromModel (SFM) model with 500 features. From the feature contribution values obtained through SHAP, class-specific gene sets were derived for each of the models, using the same methodology as before. These gene sets were then compared with the gene sets provided by Deseq2 to validate their relevance (Table 5.3). The table presents the counts of common genes between the intersected gene sets obtained from the SFM model and the gene sets provided by Deseq2 for each cancer type.

The intersected gene set represents the genes that are deemed most important for each specific cancer type. The fact that these gene sets align with the gene sets derived from Deseq2 confirms the significance and relevance of these genes as validated by an established statistical analysis tool.

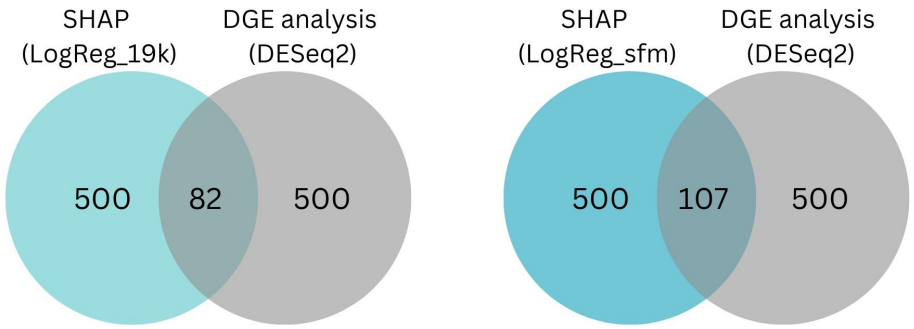
By leveraging the SHAP values and comparing the results with Deseq2, we can establish a stronger foundation for the importance of the identified gene sets. These common genes, supported by the consensus between the SFM model and Deseq2, hold potential implications for understanding the molecular mechanisms and pathways specific to each cancer type.

5.3.7 Comparison of the Common Genes

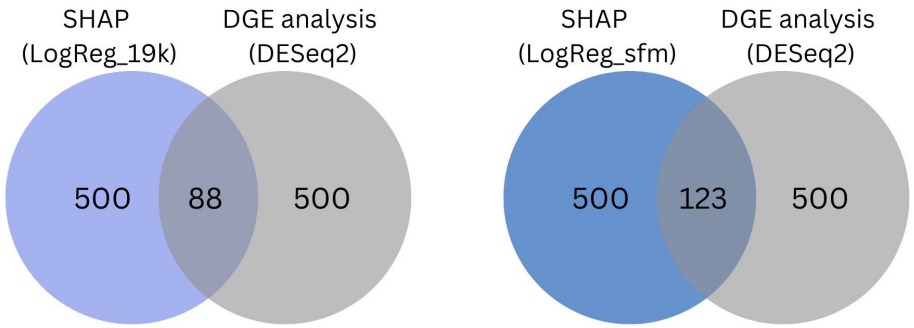
The present investigation utilized a two-stage methodology to ascertain prevalent genes linked to distinct types of cancer. Initially, the top 500 genes were chosen for each type of cancer by utilizing the Shap values derived from the three most effective classifiers. These classifiers employed a feature set comprising 19,238 features. The selection of these classifiers was based on their efficacy in precisely categorizing the various types of cancer. Following this, we implemented an identical methodology to identify the leading 500 genes utilizing the Shap values derived from the top three classifiers. However, in this instance, we employed a truncated feature set consisting of 500 features. Furthermore, the DESeq2 approach was employed to identify the 500 most distinctive genes for every type of

cancer. Through the intersection of gene sets derived from both methodologies, we have successfully identified a set of genes that were consistently detected by both the proposed architecture and DESeq2.

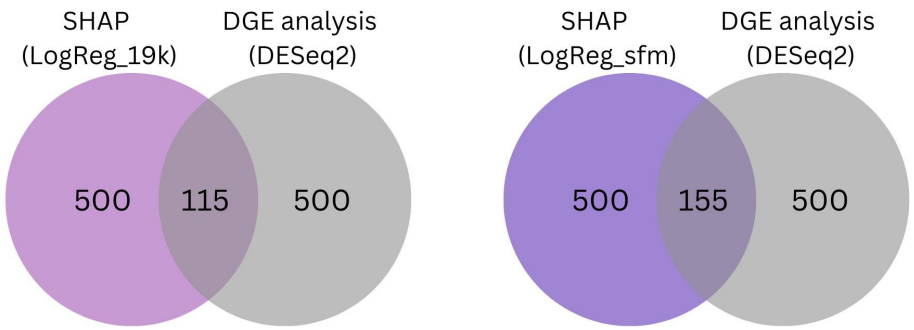
Figure 5.3 displays a comparison between the prevalent genes detected through the integration of Logistic regression with selectFromModel for 500 features (LogReg_sfm) and DESeq2 in our proposed architecture, and those identified in the baseline dataset with 19,238 features (LogReg_19k). The diagram is centered on four distinct types of cancer, namely BLCA, COAD, LUAD, and UCEC.



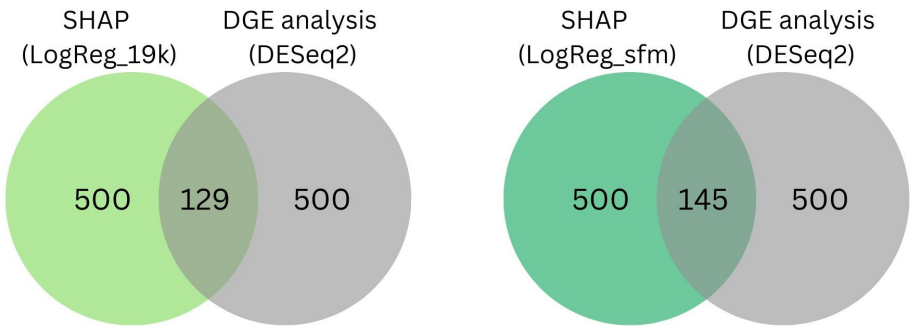
(a) BLCA cancer



(b) COAD cancer



(c) LUAD cancer



(d) UCEC cancer

Figure 5.3: Continued: Common gene set from both statistical DESeq2 analysis and models trained on two different types of features

The data presented in the figure indicates that the employment of the proposed architecture results in a greater number of shared genes when compared to the number of shared features detected within the baseline dataset. The aforementioned observation suggests that the amalgamation of Logistic regression and selectFromModel techniques amplifies the detection of common genetic characteristics among the chosen cancer categories. Table 5.3 shows the detailed common gene set for each cancer type.

The study implicated that the count of common features across different models is higher for the case of 500 features compared to 19,238 features. This find-

Cancer Name	Top 500 common feature contribution from models trained on 19238 features			Top 500 common feature contribution from models trained using SelectFromModel feature selection approach		
	LogReg	XGB	SVM	LogReg_sfm	XGB_sfm	SVM_sfm
BLCA	82	20	69	107	101	89
BRCA	68	14	67	85	71	63
CHOL	187	55	143	112	233	122
COAD	88	27	71	123	129	128
ESCA	79	14	61	110	92	90
HNSC	103	29	88	131	106	100
KICH	220	46	190	139	170	139
KIRC	115	23	109	97	90	81
KIRP	60	15	62	91	83	72
LIHC	35	15	30	114	87	60
LUAD	115	31	112	155	142	158
LUSC	252	57	199	239	240	168
PRAD	45	7	28	56	42	52
READ	101	19	91	113	110	91
STAD	69	21	83	107	92	100
THCA	83	6	55	40	64	59
UCEC	129	30	86	145	174	132

Table 5.3: Top 500 globally common feature contribution using DESeq2 and machine learning approaches

ing validates the effectiveness of feature selection when resource constraints are present. It indicates that by utilizing a smaller subset of top features obtained from the SelectFromModel (SFM) models, comparable results can be achieved compared to using the entire set of 19,238 features. The higher count of common features implies that the selected 500 features capture the most relevant and informative aspects of the gene expression data. These features exhibit similar characteristics and associations with the gene sets provided by Deseq2, further confirming their importance in cancer analysis. This observation has practical implications, especially when resources such as computational power or data storage are limited. By performing feature selection and focusing on the top features, it becomes possible to streamline the analysis process while maintaining comparable performance to using the full set of features.

The results suggest that the selected subset of 500 features, derived from the SFM models, captures the crucial aspects of the gene expression patterns and is just as effective in identifying important genes associated with specific cancer types as using the larger set of 19,238 features. This finding supports the notion that feature selection techniques can offer a practical solution for resource-constrained scenarios, where it is desirable to reduce the dimensionality of the data while maintaining or even improving the predictive power and biological relevance of the analysis.

It is important to note that the effectiveness of feature selection may vary depending on the specific dataset and analysis context. Further validation and experimentation, as well as consideration of domain knowledge, are essential to ensure the reliability and generalizability of the selected feature subsets in different scenarios.

Chapter 6

Conclusion

mRNA gene expression can be a valuable tool for cancer classification, as the expression levels of certain genes are often altered in cancer cells and can be used to distinguish different types of cancer [61, 108]. Feature attribution techniques are a useful tool for making machine learning models more explainable, particularly in the context of cancer classification. These techniques can help to identify which features (such as specific genes) are most important for making a prediction and how they contribute to the final prediction made by the model.

In this paper, we have implemented an Explainable AI-based panCancer classification approach using gene expression analysis which will help to detect the type of cancer prevailing in individuals accurately within a very short time. We have implemented 7 classifier algorithms to classify 33 different kinds of cancers, among which two are tree-based models (XGBoost and Random Forest), two are traditional machine learning algorithms (Logistic Regression and SVM), and two deep learning-based algorithms (MLP and 1-D CNN) and a transformer-based model (TabNet). In our gene expression analysis study using machine learning models, we employed four feature selection techniques: Lasso, SelectFromModel, Select-K-Best, and ElasticNet. Among these techniques, SelectFromModel with 500 features achieved the best performance, followed by Lasso, Select-K-Best, and ElasticNet. We applied ensemble methods of probability averaging and max voting, with probability averaging achieving the highest accuracy of 96.60%. The F1 scores also showed the effectiveness of combining predictions from multiple models. Validating the selected features using SHAP values and comparing them with gene sets from DESeq2 analysis confirmed their significance and relevance. The common genes identified between SelectFromModel and DESeq2 provided insights into cancer-specific molecular mechanisms and pathways. Feature selection proved effective in capturing relevant aspects of gene expression data while reducing dimensionality, highlighting its importance

in maintaining predictive power and biological relevance.

Future work can include concentrating on SSGSEA implementation in precision medicine. SSGSEA enables the identification of patient-specific gene sets, revealing important insights into the molecular mechanisms underlying the diseases of individual patients. By employing SSGSEA to gene expression data, unique gene sets that are specifically active or suppressed in each patient's tumor can be identified, thereby facilitating the comprehension of disease progression. Integrating patient-specific gene set information with pathway analysis provides a comprehensive view of dysregulated biological processes, thereby facilitating the development of targeted therapeutic strategies and individualized treatments. Implementing SSGSEA in precision medicine has the potential to revolutionize cancer treatment by customizing approaches based on the tumor characteristics of each individual patient, thereby enhancing patient outcomes.

References

[1] Hyuna Sung et al. “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[2] Kristen Trost Mantlo. “Understanding Young Adult Survivors of Childhood Cancers’ Participation in Late Effects Screening: A Mixed Methods Approach”. PhD thesis. Old Dominion University, 2019.

[3] David A Hanauer et al. “Bioinformatics approaches in the study of cancer”. In: *Current molecular medicine* 7.1 (2007), pp. 133–141.

[4] Yi-Ping Phoebe Chen and Feng Chen. “Identifying targets for drug discovery using bioinformatics”. In: *Expert opinion on therapeutic targets* 12.4 (2008), pp. 383–389.

[5] Jun Wang et al. “Regulatory roles of long noncoding RNAs implicated in cancer hallmarks”. In: *International journal of cancer* 146.4 (2020), pp. 906–916.

[6] Aisha Patel. “Benign vs malignant tumors”. In: *JAMA oncology* 6.9 (2020), pp. 1488–1488.

[7] David V Schapira et al. “Intensive care, survival, and expense of treating critically III cancer patients”. In: *Jama* 269.6 (1993), pp. 783–786.

[8] Julie Eggert. “Genetics and genomics in oncology nursing: what does every nurse need to know?” In: *Nursing Clinics* 52.1 (2017), pp. 1–25.

[9] Ying Lu and Jiawei Han. “Cancer classification using gene expression data”. In: *Information Systems* 28.4 (2003), pp. 243–268.

[10] Kirk J Mantione et al. “Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq”. In: *Medical science monitor basic research* 20 (2014), p. 138.

[11] Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47–e47.

[12] Christine H Chung, Philip S Bernard, and Charles M Perou. “Molecular portraits and the family tree of cancer”. In: *Nature genetics* 32.4 (2002), pp. 533–540.

[13] Douglas Hanahan and Robert A Weinberg. “Hallmarks of cancer: the next generation”. In: *cell* 144.5 (2011), pp. 646–674.

[14] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *nature* 542.7639 (2017), pp. 115–118.

[15] E Farshi. “Peptide-Based mRNA Vaccines”. In: *J Gastro Hepato* 9.16 (2023), pp. 1–6.

[16] Davide Ruggero and Pier Paolo Pandolfi. “Does the ribosome translate cancer?” In: *Nature Reviews Cancer* 3.3 (2003), pp. 179–192.

[17] Wengong Si et al. “The role and mechanisms of action of microRNAs in cancer drug resistance”. In: *Clinical epigenetics* 11.1 (2019), pp. 1–24.

[18] Edward L Tatum. “Molecular biology, nucleic acids, and the future of medicine”. In: *Perspectives in biology and medicine* 10.1 (1966), pp. 19–32.

[19] Lela Buckingham. “Fundamentals of Nucleic Acid Biochemistry: An Overview”. In: ().

[20] Francis Crick. “Central dogma of molecular biology”. In: *Nature* 227.5258 (1970), pp. 561–563.

[21] Robert G Roeder. “Transcriptional regulation and the role of diverse coactivators in animal cells”. In: *FEBS letters* 579.4 (2005), pp. 909–915.

[22] Melissa J Moore. “From birth to death: the complex lives of eukaryotic mRNAs”. In: *Science* 309.5740 (2005), pp. 1514–1518.

[23] Richard W Carthew and Erik J Sontheimer. “Origins and mechanisms of miRNAs and siRNAs”. In: *Cell* 136.4 (2009), pp. 642–655.

[24] David P Bartel. “MicroRNAs: genomics, biogenesis, mechanism, and function”. In: *cell* 116.2 (2004), pp. 281–297.

[25] John L Rinn and Howard Y Chang. “Genome regulation by long noncoding RNAs”. In: *Annual review of biochemistry* 81 (2012), pp. 145–166.

[26] Timothy H Bestor. “The DNA methyltransferases of mammals”. In: *Human molecular genetics* 9.16 (2000), pp. 2395–2402.

[27] M Perou Charles et al. “Molecular portraits of human breast tumours”. In: *Nature* 406.6797 (2000), pp. 747–752.

[28] Gyöngyi Munkácsy, Libero Santarpia, and Balázs Győrffy. “Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Tumor Microenvironment Features”. In: *Biomedicines* 10.2 (2022), p. 248.

[29] George A Calin and Carlo M Croce. “MicroRNA signatures in human cancers”. In: *Nature reviews cancer* 6.11 (2006), pp. 857–866.

[30] Barbara Pardini et al. “Noncoding RNAs in extracellular fluids as cancer biomarkers: the new frontier of liquid biopsies”. In: *Cancers* 11.8 (2019), p. 1170.

[31] Hui Li et al. “A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells”. In: *Science* 321.5894 (2008), pp. 1357–1361.

[32] Kenzui Taniue and Nobuyoshi Akimitsu. “Fusion genes and RNAs in cancer development”. In: *Non-coding RNA* 7.1 (2021), p. 10.

[33] Konstantina Kourou et al. “Machine learning applications in cancer prognosis and prediction”. In: *Computational and structural biotechnology journal* 13 (2015), pp. 8–17.

[34] Meriem Amrane et al. “Breast cancer classification using machine learning”. In: *2018 electric electronics, computer science, biomedical engineerings’ meeting (EBBT)*. IEEE. 2018, pp. 1–4.

[35] Sara Tarek, Reda Abd Elwahab, and Mahmoud Shoman. "Gene expression based cancer classification". In: *Egyptian Informatics Journal* 18.3 (2017), pp. 151–159. ISSN: 1110-8665. DOI: https://doi.org/10.1016/j.eij.2016.12.001. URL: https://www.sciencedirect.com/science/article/pii/S1110866516300809.

[36] Maxim D Podolsky et al. "Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels". In: *Asian Pacific journal of cancer prevention* 17.2 (2016), pp. 835–838.

[37] Boyu Lyu and Anamul Haque. "Deep learning based tumor type classification using gene expression data". In: *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 2018, pp. 89–96.

[38] Joseph M de Guia, Madhavi Devaraj, and Carson K Leung. "DeepGx: deep learning using gene expression for cancer classification". In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2019, pp. 913–920.

[39] Yuanyuan Li et al. "A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data". In: *BMC genomics* 18.1 (2017), pp. 1–13.

[40] Pradipta Maji and Chandra Das. "Relevant and significant supervised gene clusters for microarray cancer classification". In: *IEEE Transactions on nanobioscience* 11.2 (2012), pp. 161–168.

[41] Yi-Hsin Hsu and Dong Si. "Cancer type prediction and classification based on rna-sequencing data". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 5374–5377.

[42] Milad Mostavi et al. "Convolutional neural network models for cancer type prediction based on gene expression". In: *BMC medical genomics* 13 (2020), pp. 1–13.

[43] Jean-François Laplante and Moulay A Akhloufi. “Predicting cancer types from miRNA stem-loops using deep learning”. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 5312–5315.

[44] Kazi Ferdous Mahin et al. “PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning”. In: *Genomics* 114.2 (2022), p. 110264.

[45] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[46] Yoav Freund, Robert Schapire, and Naoki Abe. “A short introduction to boosting”. In: *Journal-Japanese Society For Artificial Intelligence* 14.771-780 (1999), p. 1612.

[47] Anestis Antoniadis, Sophie Lambert-Lacroix, and Frédérique Leblanc. “Effective dimension reduction methods for tumor classification using gene expression data”. In: *Bioinformatics* 19.5 (2003), pp. 563–570.

[48] D Pavithra and B Lakshmanan. “Feature selection and classification in gene expression cancer data”. In: *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)*. IEEE. 2017, pp. 1–6.

[49] Yongjun Piao and Keun Ho Ryu. “Detection of differentially expressed genes using feature selection approach from RNA-seq”. In: *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2017, pp. 304–308.

[50] Hanaa Salem, Gamal Attiya, and Nawal El-Fishawy. “Classification of human cancer diseases by gene expression profiles”. In: *Applied Soft Computing* 50 (2017), pp. 124–134.

[51] Isabelle Guyon et al. “Gene selection for cancer classification using support vector machines”. In: *Machine learning* 46.1 (2002), pp. 389–422.

[52] Alejandro Lopez-Rincon et al. “Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection”. In: *BMC bioinformatics* 20.1 (2019), pp. 1–17.

[53] Pilar Garcıa-Dıaz et al. “Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data”. In: *Genomics* 112.2 (2020), pp. 1916–1925.

[54] Yu-Heng Lai et al. “Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning”. In: *Scientific reports* 10.1 (2020), p. 4679.

[55] Dejun Zhang et al. “Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer”. In: *Ieee Access* 6 (2018), pp. 28936–28944.

[56] JN Weinstein. “TCGAR Network, EA Collisson et al.,“The cancer genome atlas pan-cancer analysis project,”” in: *Nature Genetics* 45.10 (2013), pp. 1113–1120.

[57] Yingdong Zhao et al. “TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository”. In: *Journal of translational medicine* 19.1 (2021), pp. 1–15.

[58] Cole Trapnell et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. In: *Nature biotechnology* 28.5 (2010), pp. 511–515.

[59] Bo Li and Colin N Dewey. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. In: *BMC bioinformatics* 12 (2011), pp. 1–16.

[60] Mia Huljanah et al. “Feature selection using random forest classifier for predicting prostate cancer”. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 546. 5. IOP Publishing. 2019, p. 052031.

[61] David G Kleinbaum et al. *Logistic regression*. Springer, 2002.

[62] Lipo Wang. *Support vector machines: theory and applications*. Vol. 177. Springer Science & Business Media, 2005.

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

[64] Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal. “Introduction to multi-layer feed-forward neural networks”. In: *Chemometrics and intelligent laboratory systems* 39.1 (1997), pp. 43–62.

[65] Serkan Kiranyaz et al. “1D convolutional neural networks and applications: A survey”. In: *Mechanical systems and signal processing* 151 (2021), p. 107398.

[66] Ravisutha Sakrepatna Srinivasamurthy. “Understanding 1D Convolutional Neural Networks Using Multiclass Time-Varying Signalss”. PhD thesis. Clemson University, 2018.

[67] Sercan Ö Arik and Tomas Pfister. “Tabnet: Attentive interpretable tabular learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. 2021, pp. 6679–6687.

[68] Hans Hersbach. “Decomposition of the continuous ranked probability score for ensemble prediction systems”. In: *Weather and Forecasting* 15.5 (2000), pp. 559–570.

[69] Robi Polikar. “Ensemble based systems in decision making”. In: *IEEE Circuits and systems magazine* 6.3 (2006), pp. 21–45.

[70] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).

[71] Hao Luo et al. “DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools”. In: *Nucleic acids research* 49.D1 (2021), pp. D677–D686.

[72] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), pp. 1–21.

[73] Sanjaya K Panda, Subhrajit Nag, and Prasanta K Jana. “A smoothing based task scheduling algorithm for heterogeneous multi-cloud environment”. In: *2014 International Conference on Parallel, Distributed and Grid Computing*. IEEE. 2014, pp. 62–67.

[74] SGOPAL Patro and Kishore Kumar Sahu. “Normalization: A preprocessing stage”. In: *arXiv preprint arXiv:1503.06462* (2015).

[75] C Saranya and G Manikandan. “A study on normalization techniques for privacy preserving data mining”. In: *International Journal of Engineering and Technology (IJET)* 5.3 (2013), pp. 2701–2704.

[76] Md Manjurul Ahsan et al. “Effect of data scaling methods on machine learning algorithms and model performance”. In: *Technologies* 9.3 (2021), p. 52.

[77] Ekaba Bisong and Ekaba Bisong. “Introduction to Scikit-learn”. In: *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (2019), pp. 215–229.

[78] VN Ganapathi Raju et al. “Study the influence of normalization/transformation process on the accuracy of supervised classification”. In: *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE. 2020, pp. 729–735.

[79] Zaneta Swiderska-Chadaj et al. “Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer”. In: *Scientific Reports* 10.1 (2020), pp. 1–14.

[80] Junfang Wu and Chao Li. “Feature selection based on features unit”. In: *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE. 2017, pp. 330–333.

[81] Huiqing Liu, Jinyan Li, and Limsoon Wong. “A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns”. In: *Genome informatics* 13 (2002), pp. 51–60.

[82] Zena M Hira and Duncan F Gillies. “A review of feature selection and feature extraction methods applied on microarray data”. In: *Advances in bioinformatics* 2015 (2015).

[83] Valeria Fonti and Eduard Belitser. “Feature selection using lasso”. In: *VU Amsterdam research paper in business analytics* 30 (2017), pp. 1–25.

[84] Jason Brownlee. “An introduction to feature selection”. In: *Machine learning process* 6 (2014).

[85] R Muthukrishnan and R Rohini. “LASSO: A feature selection technique in predictive modeling for machine learning”. In: *2016 IEEE international conference on advances in computer applications (ICACA)*. IEEE. 2016, pp. 18–20.

[86] Anamika Chauhan et al. “Detection of lung cancer using machine learning techniques based on routine blood indices”. In: *2020 IEEE international conference for innovation in technology (INOCON)*. IEEE. 2020, pp. 1–6.

[87] Hui Zou and Trevor Hastie. “Regression shrinkage and selection via the elastic net, with applications to microarrays”. In: *JR Stat Soc Ser B* 67 (2003), pp. 301–20.

[88] Asma Agaal and Mansour Essgaer. “Influence of Feature Selection Methods on Breast Cancer Early Prediction Phase using Classification and Regression Tree”. In: *2022 International Conference on Engineering & MIS (ICEMIS)*. IEEE. 2022, pp. 1–6.

[89] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[90] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[91] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *Journal of statistical software* 33.1 (2010), p. 1.

[92] Artem Sokolov et al. “Pathway-based genomics prediction using generalized elastic net”. In: *PLoS computational biology* 12.3 (2016), e1004790.

[93] Amrita Basu et al. “RWEN: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines”. In: *Bioinformatics* 34.19 (2018), pp. 3332–3339.

[94] Mahmood Khalsan et al. “A survey of machine learning approaches applied to gene expression analysis for cancer prediction”. In: *IEEE Access* 10 (2022), pp. 27522–27534.

[95] Jose Liñares-Blanco, Alejandro Pazos, and Carlos Fernandez-Lozano. “Machine learning analysis of TCGA cancer data”. In: *PeerJ Computer Science* 7 (2021), e584.

[96] Ahsan Bin Tufail et al. “Deep learning in cancer diagnosis and prognosis prediction: a minireview on challenges, recent trends, and future directions”. In: *Computational and Mathematical Methods in Medicine* 2021 (2021).

[97] Vabiyana Safira Desdhanty and Zuherman Rustam. “Liver cancer classification using random forest and extreme gradient boosting (xgboost) with genetic algorithm as feature selection”. In: *2021 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE. 2021, pp. 716–719.

[98] Bong-Hyun Kim, Kijin Yu, and Peter CW Lee. “Cancer classification of single-cell gene expression data by neural network”. In: *Bioinformatics* 36.5 (2020), pp. 1360–1366.

[99] Sk Md Mosaddek Hossain et al. “Pan-cancer classification by regularized multi-task learning”. In: *Scientific reports* 11.1 (2021), p. 24252.

[100] Yulin Zhang et al. “A novel XGBoost method to identify cancer tissue-of-origin based on copy number variations”. In: *Frontiers in genetics* 11 (2020), p. 585029.

[101] Jerome H Friedman. “Stochastic gradient boosting”. In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378.

[102] Xiaobo Zhou, Kuang-Yu Liu, and Stephen TC Wong. “Cancer classification and prediction using logistic regression with Bayesian gene selection”. In: *Journal of Biomedical Informatics* 37.4 (2004), pp. 249–259.

[103] Zakariya Yahya Algamal and Muhammad Hisyam Lee. “Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification”. In: *Expert Systems with Applications* 42.23 (2015), pp. 9326–9332.

[104] Zakariya Yahya Algamal and Muhammad Hisyam Lee. “Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification”. In: *Computers in biology and medicine* 67 (2015), pp. 136–145.

[105] Lingyun Gao et al. “Hybrid method based on information gain and support vector machine for gene selection in cancer classification”. In: *Genomics, proteomics & bioinformatics* 15.6 (2017), pp. 389–395.

[106] Ahmed Arafa et al. “Regularized logistic regression model for cancer classification”. In: *2021 38th National Radio Science Conference (NRSC)*. Vol. 1. IEEE. 2021, pp. 251–261.

[107] Trevor Hastie. “Ridge regularization: An essential concept in data science”. In: *Technometrics* 62.4 (2020), pp. 426–433.

[108] Enrique Alba et al. “Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms”. In: *2007 IEEE congress on evolutionary computation*. IEEE. 2007, pp. 284–290.

[109] Mikel Galar et al. “An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes”. In: *Pattern Recognition* 44.8 (2011), pp. 1761–1776.

[110] Luıs A Vale Silva and Karl Rohr. “Pan-cancer prognosis prediction using multimodal deep learning”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 568–571.

[111] Niousha Bagheri Khoulenjani et al. “Cancer miRNA biomarkers classification using a new representation algorithm and evolutionary deep learning”. In: *Soft Computing* 25 (2021), pp. 3113–3129.

[112] Thomas Serre, Aude Oliva, and Tomaso Poggio. “A feedforward architecture accounts for rapid categorization”. In: *Proceedings of the national academy of sciences* 104.15 (2007), pp. 6424–6429.

[113] U Ravindran and C Gunavathi. “A survey on gene expression data analysis using deep learning methods for cancer diagnosis”. In: *Progress in Biophysics and Molecular Biology* 177 (2023), pp. 1–13.

[114] Pablo Guillen and Jerry Ebalunode. “Cancer classification based on microarray gene expression data using deep learning”. In: *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE. 2016, pp. 1403–1405.

[115] Feng Gao et al. “DeepCC: a novel deep learning-based framework for cancer molecular subtype classification”. In: *Oncogenesis* 8.9 (2019), p. 44.

[116] Bing Xu et al. “Empirical evaluation of rectified activations in convolutional network”. In: *arXiv preprint arXiv:1505.00853* (2015).

[117] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[118] Mohanad Mohammed et al. “A stacking ensemble deep learning approach to cancer type classification based on TCGA data”. In: *Scientific reports* 11.1 (2021), pp. 1–22.

[119] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[120] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[121] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[122] Madhuri Gokhale, Sraban Kumar Mohanty, and Aparajita Ojha. "GeneViT: Gene vision transformer with improved DeepInsight for cancer classification". In: *Computers in Biology and Medicine* 155 (2023), p. 106643.

[123] Anwar Khan and Boreom Lee. "Gene transformer: Transformers for the gene expression-based classification of lung cancer subtypes". In: *arXiv preprint arXiv:2108.11833* (2021).

[124] Ting-He Zhang et al. "Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions". In: *Cancers* 14.19 (2022), p. 4763.

[125] Faysal Bin Rahman, Farhan Anjum, and Musaddiq Hasan Fatin Khan. "Detection of Lung Adenocarcinoma Cancer based on RNA-seq gene expression data using LIMMA and TabNet". PhD thesis. Department of Computer Science and Engineering (CSE), Islamic University of ..., 2022.

[126] R Tyler McLaughlin et al. "Fast, accurate, and racially unbiased pan-cancer tumor-only variant calling with tabular machine learning". In: *NPJ Precision Oncology* 7.1 (2023), p. 4.

[127] Ahmad Nasimian et al. "A deep tabular data learning model predicting cisplatin sensitivity identifies BCL2L1 dependency in cancer". In: *Computational and Structural Biotechnology Journal* (2023).

[128] Yawen Xiao et al. "A deep learning-based multi-model ensemble method for cancer prediction". In: *Computer methods and programs in biomedicine* 153 (2018), pp. 1–9.

[129] Aik Choon Tan and David Gilbert. "Ensemble machine learning on gene expression data for cancer classification". In: (2003).

[130] Eloise Withnell et al. "XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data". In: *Briefings in Bioinformatics* 22.6 (2021), bbab315.

[131] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[132] Wilson E Marcılio and Danilo M Eler. “From explanations to feature selection: assessing SHAP values as feature selection mechanism”. In: *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*. Ieee. 2020, pp. 340–347.

[133] Masrur Sobhan and Ananda Mohan Mondal. “Explainable Machine Learning to Identify Patient-specific Biomarkers for Lung Cancer”. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2022, pp. 3152–3159.

[134] Katsuya Futagami et al. “Pairwise acquisition prediction with SHAP value interpretation”. In: *The Journal of Finance and Data Science* 7 (2021), pp. 22–44.

[135] Melvyn Yap et al. “Verifying explainability of a deep learning tissue classifier trained on RNA-seq data”. In: *Scientific reports* 11.1 (2021), p. 2641.

[136] Michael Chromik. “Making SHAP Rap: Bridging local and global insights through interaction and narratives”. In: *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*. Springer. 2021, pp. 641–651.

[137] A Stupnikov et al. “Robustness of differential gene expression analysis of RNA-seq”. In: *Computational and structural biotechnology journal* 19 (2021), pp. 3470–3481.

[138] Adam McDermaid et al. “Interpretation of differential gene expression results of RNA-seq data: review and integration”. In: *Briefings in bioinformatics* 20.6 (2019), pp. 2044–2054.

[139] Qingguo Wang et al. “Enabling cross-study analysis of RNA-Sequencing data”. In: *BioRxiv* (2017), p. 110734.

[140] Qingguo Wang et al. “Unifying cancer and normal RNA sequencing data from different sources”. In: *Scientific data* 5.1 (2018), pp. 1–8.

[141] Nick Bunkley. “Joseph Juran, Pioneer in Quality Control, Dies”. In: *The New York Times* 103 (2008).

[142] V Roshan Joseph. “Optimal ratio for data splitting”. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15.4 (2022), pp. 531–538.

[143] Denny Wu and Ji Xu. “On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10112–10123.

Chapter A

Supplementary Data

A.1 Top 50 Genes for 17 Cancers

Table A.1 displays the 50 most prominent gene sets across 17 distinct types of cancer. The gene sets were generated through the identification of shared genes resulting from the implementation of logistic regression and DESeq2 methodologies. The selection of logistic regression with 500 features was based on its superior performance in terms of accuracy compared to the other six classifiers evaluated in this investigation. The inclusion of cancer types in the analysis was determined by the dataset provided by Q. Wang et al [139, 140] to ensure a comprehensive representation of diverse cancer types.

Cancer Name	Top 50 genes
BLCA	HMGCLL1, SCARA5, MAGEC2, BMP5, GRP, SPRR2F, CSAG1, NBPF4, PM20D1, POU4F1, TBX5, TBX20, KISS1R, POTEE, PCP4, HAND1, DMRTA2, TLX3, VCX3B, SCN7A, KRTDAP, HPSE2, MAGEA4, TUBA3E, KLK2, DES, MMRN1, OR2B6, SPRR2E, EN2, BARX1, MAGEA10, PRAME, MAGEA6, XAGE1A, PAGE2, ESX1, ATP4A, MAGEA3, CTAG1B, RHAG, SERTM1, ASB5, ACTC1, HSFY1, CTAG1A, MAGEA2B, HOXC9, TCF21, CILP
BRCA	SCARA5, MAGEC2, MRGPRX2, CSAG1, NBPF4, FOXG1, MYH1, GALNT15, KISS1R, URAD, COL6A6, CPB1, TLX3, A2ML1, HBG1, TMEM215, ALDH1L1, MAGEA4, R3HDML, DES, MMRN1, OR2B6, MAGEA10, GNGT1, PRAME, SLC30A8, MAGEA6, GATA4, PAGE2, CPA1, HMX3, HBG2, CSN1S1, MAGEA3, CTAG1B, NKX3-2, SERTM1, LHX2, BRINP3, CLDN19, MAGEA2B, SULT1C3, MAGEA12, NBPF6, CCDC60, HPSE2, POU4F1, SMYD1, PAX1, DPT

Cancer Name	Top 50 genes
CHOL	GRP, TTC36, PITX2, PGLYRP1, KISS1R, MROH2B, HOXC13, COL6A6, XKR6, FRMD7, FCN2, NETO1, MT-ND6, ALDH1L1, MYO3A, REG1B, HOXC8, PTPN5, SPRR3, FSTL4, OR2B6, EN2, PRAME, TMEM178B, TGM3, C4BPA, ASIP, SIX2, AGMO, UGT2A2, NPTX2, IL13RA2, UGT2A1, TTR, SLC28A1, C1orf100, HOXD3, NKX3-2, EPHX3, HOXC9, PGC, HOXD9, ADAMTS16, SLC5A12, NDST3, HOXD4, PON3, GNG5P2, CALML3, WNT3A
COAD	CLCA1, HMGCLL1, AQP2, AKAP4, SCARA5, IFITM5, CAMKV, STMN2, CASR, DHRS7C, MRGPRX2, GOLGA8F, CSAG1, FOXG1, TBX20, SOX14, HOXC13, HAND1, SP8, DMRTA2, TLX3, SCN7A, A2ML1, HPSE2, FGF3, MAGEA4, NEFM, HBE1, DES, SPRR3, IRX5, SPRR2E, EN2, GNGT1, PTF1A, SLC30A8, MAGEA6, GATA4, SIX2, KRT24, UPK1A, IZUMO2, MAGEA3, RSPO2, TBX5, CDH19, RHAG, SERPINA9, TCP11, CLDN8
ESCA	CCKAR, G6PC2, SCARA5, PAX3, HOXD11, MAGEC2, STMN2, RSPO2, SPRR2F, CSAG1, LEFTY1, PM20D1, POU4F1, HOXD13, KISS1R, POTEE, HOXA11, CT45A6, SP8, NOS2, NMUR2, SCN7A, HOXD10, MAGEA4, REG1B, CD200R1L, HOXC8, KLK2, DES, OR2B6, EN2, MAGEA10, GNGT1, IL17A, FABP7, PRAME, MAGEA6, C4BPA, AGMO, IL13RA2, LAIR2, UGT2A1, ATP4A, MUC5B, ALX1, MAGEA3, CTAG1B, PNLDC1, HPSE2, LMOD2
HNSC	HMGCLL1, CSN3, HOXD11, MAGEC2, KRT40, DHRS7C, GOLGA8F, CSAG1, NRAP, HOXD13, KRT4, POTEE, HOXA11, CT45A6, LRRC14B, SP8, TLX3, SCN7A, KRT85, FCN2, NETO1, HOXD10, STATH, GADL1, SCGB3A2, ZNF541, MAGEA4, CAV3, DRD5, KRT36, HBE1, HOXC8, SPRR3, OR2B6, MAGEA10, GNGT1, FABP7, PRAME, NKX2-4, QRFPR, MAGEA6, TGM3, GATA4, CRNN, XAGE1A, PAGE2, ART3, UGT2A1, UPK1A, SLC28A1
KICH	MYO1H, AQP2, CCKAR, RAG2, FAM183A, CAMKV, PRL, KRT40, CASR, RHOXF2B, CSAG1, PITX2, PM20D1, LEFTY1, SPATA22, HELT, NKX6-1, LRRC14B, CPB1, NMUR2, VCX3B, KRT85, CDH19, AGTR2, HPSE2, GADL1, TMEM215, CAV3, NEFM, PRND, HBE1, KLK2, UNCX, DES, LHFPL4, OR14I1, PSG2, IRX5, GNGT1, NEUROD1, TAS1R1, NPHS2, NPAP1, ADCYAP1, FAM217A, UGT2A2, KRT24, IL13RA2, UPK1B, UGT1A7

Cancer Name	Top 50 genes
KIRC	AQP2, AMELY, RAG2, IFITM5, STMN2, KRT40, CASR, CSAG1, POU4F1, TBX5, HELT, HOXD13, KISS1R, CRX, FRMD7, NMUR2, SCN7A, GADL1, SCGB3A2, MYO3A, REG1B, KRT36, KLRF2, TUBA3E, UNCX, MAGEA10, FABP7, NPHS2, SLC30A8, FAM217A, GATA4, SAG, UGT2A2, NPTX2, KRT24, LAIR2, CPA1, UGT2A1, ALX1, MAGEA3, SERPINA9, CLDN8, LHX2, BRINP3, CLDN19, PM20D1, MC2R, DMP1, NRK, CA1
KIRP	ZIC4, AQP2, CCKAR, AKAP4, RAG2, BMP5, IFITM5, CAMKV, ADAM7, TTC36, CASR, PM20D1, HELT, HOXC13, FRMD7, NMUR2, SCN7A, KRT85, CDH19, AGTR2, HPSE2, GADL1, REG1B, UNCX, REN, EN2, NEUROD1, MFRP, PRAME, NPHS2, NKX2-4, FAM217A, GATA4, C4BPA, ASIP, ZNF705D, UGT2A2, IL13RA2, CPA1, HMX3, MUC5B, TTR, DEFB103B, CLDN8, SLC6A15, ELSPBP1, CLDN19, HSFY1, CNTN6, TCF21
LIHC	ZIC4, MAGEC2, GRP, RBMY1J, RHOXF2B, CSAG1, PITX2, HNRNPCL1, KISS1R, PCP4, HOXA11, HOXC13, COL6A6, SP8, VCX3B, CACNG1, THBS4, FCN2, HOXD10, MAGEA4, DRGX, REG1B, HOXC8, KLK2, LHFPL4, FSTL4, EN2, MAGEA10, GNGT1, PRAME, NPFFR2, MAGEA6, TGM3, SIX2, XAGE1A, DEFA3, PXDNL, PAGE2, MAP7D2, CPA1, LGALS14, ALX1, HOXD3, MAGEA3, CTAG1B, NKX3-2, HOXD8, SLC6A15, BLOC1S5-TXNDC5, BRINP3
LUAD	ZIC4, HOXD11, MAGEC2, IFITM5, THEG, SPRR2F, CSAG1, NBPF4, FOXG1, PITX2, POU4F1, SOX14, GUCA1A, HELT, HOXD13, KISS1R, POTEE, PCP4, HOXA11, HOXC13, ISL1, RAX, ITLN2, LRRC14B, SP8, DMRTA2, HTR3C, IRX4, OTX2, A2ML1, NETO1, RXFP1, HBG1, PCDH8, MAGEA4, CAV3, PTPN5, DES, LHFPL4, SPRR3, OR2B6, EN2, BARX1, MAGEA10, GNGT1, NEUROD1, PRAME, NPFFR2, TFAP2D, NKX2-4
LUSC	ZIC4, PLCZ1, CSN3, SCARA5, LEFTY2, EVX2, PAX3, HOXD11, CAMKV, CEACAM8, RSPO2, THEG, UPK1B, SPRR2F, GOLGA8F, CSAG1, NBPF4, OR7C1, SERPINB13, FOXG1, MYH1, POU4F1, PITX2, IAPP, TBX20, SOX14, GUCA1A, HOXD13, PGLYRP1, KISS1R, PNLDC1, POTEE, HOXA11, NKX6-1, HOXC13, COL6A6, ISL1, GAGE12E, CT45A6, ITLN2, LRRC14B, RAX, SP8, DMRTA2, TLX3, HTR3C, VCX3B, CT45A2, MAGEC2, SCN7A

Cancer Name	Top 50 genes
PRAD	SIM1, HOXC4, SPRR3, AQP2, OR2B6, WNT9B, PRSS1, GSTM1, PATE1, GKN1, PRAME, NPFFR2, XAGE1B, HOXB5, PAX1, PON3, LCN1, B3GNT6, GC, FOXG1, MUC6, POU4F1, XAGE1A, KRT24, IAPP, TMEM114, HOXC6, CST2, KISS1R, TRIM49, DAZ1, NKX6-1, HOXC13, IL36RN, ATP6V1G3, KRT13, MAGEA3, CCDC83, NKX2-5, HOXB6, HOXC12, C10orf99, DAZ4, ELSPBP1, LCN15, MAGEA6, NETO1, MAGEC2, OTP, NKX2-3
READ	HMGCLL1, AKAP4, SCARA5, EVX2, CAMKV, RSPO2, CASR, DHRS7C, CSAG1, PITX2, TBX20, SOX14, MROH2B, HAND1, SP8, CPB1, SCN7A, CDH19, AGTR2, HPSE2, MAGEA4, REG1B, NEFM, PRND, R3HDML, DES, MMRN1, SPRR3, IRX5, EN2, GNGT1, IL17A, PRAME, PTF1A, PGPEP1L, MAGEA6, SIX2, KRT24, MAGEA3, RHAG, CLDN8, ASB5, BRINP3, MAGEA2B, KRTAP13-2, CA1, SNTG2, PGC, DMRTA2, MAGEA12
STAD	CLCA1, AQP2, HOXA9, AKAP4, SCARA5, HOXD11, MAGEC2, TTC36, DHRS7C, CSAG1, SERPINB13, LEFTY1, HOXD13, KRT4, POTEE, NKX6-1, HOXC13, ITLN2, SP8, CPB1, MC2R, VCX3B, IRX4, A2ML1, KRTDAP, HPSE2, FGF3, MAGEA4, DRGX, NEFM, HOXC8, SPRR3, OR2B6, EN2, MAGEA10, GNGT1, PRAME, PTF1A, MAGEA6, TGM3, CRNN, C4BPA, FLG, PAGE2, UGT1A7, ATP4A, UPK1A, SLC28A1, HOXA11, KRT13
THCA	IBSP, CBLN1, REN, FSTL4, SPRR3, CLDN10, DSC3, UGT3A2, PRSS1, SCARA5, PTGER1, CEACAM8, STMN2, CD207, DPT, PSG8, B3GNT6, SFTPA1, SPRR1B, NPTX2, BPIFB1, OBP2B, PSG4, KISS1R, HOXA11, IL36RN, CDKN2A, MS4A15, SFTPB, NETO1, CCL21, SLC6A15, CGA, TMEM215, DMP1, COL6A5, NAPSA, CST2, SFTPA2, MMRN1
UCEC	CBLN1, ZIC4, SCARA5, LEFTY2, BMP5, MAGEC2, CAMKV, ADAM29, CSRP3, SERPINB13, FOXG1, LEFTY1, POU4F1, PITX2, TBX20, GALNT15, HOXD13, KISS1R, KRT4, POTEE, HOXC13, RAX, CT45A6, SP8, FRMD7, DMRTA2, TLX3, SCN7A, OTX2, HPSE2, FGF3, MAGEA4, NEFM, PRND, OSR1, DES, MMRN1, SPRR3, FSTL4, OR2B6, SPRR2E, BARX1, MAGEA10, GNGT1, FABP7, PRAME, EXD1, ADCYAP1, SLC30A8, MAGEA6

Table A.1: Cancer-specific top 50 gene set

A.2 Empirical Results

Table A.2 displays a subset of our meta-analysis conducted on the TCGA gene expression dataset using our chosen classifiers (Logistic Regression, SVM, XGBoost, MLP, Random Forest, and 1D-CNN), four feature selection techniques (Lasso, SelectFromModel, Select-K-Best, and Elasticnet), and two normalization techniques (Standard Scaler and MinMax Scaler) respectively. Among the classifiers, Logistic Regression achieved the highest accuracy of 96.43%, while SVM, XGBoost, and MLP also performed well with accuracies over 95%. By employing data preprocessing techniques with the classifier models, we found that standard scaling had a significant positive impact on overall accuracy compared to the min-max scaler. The accuracy of the models ranged from 96.89% to 91.65%, with Logistic Regression reaching the highest accuracy of 96.89%, followed by SVM and XGBoost. Additionally, the classifiers were trained with each of the feature selection techniques selecting 100, 500, and 1000 features each time. From all possible combinations, SelectFromModel with 500 features, led to improved performance compared to other techniques like Lasso, Select-K-Best, and ElasticNet. SelectFromModel outperforms other feature selection strategies because it automatically selects the most relevant characteristics based on their importance in the classification task and assesses feature relevance while selecting the top features based on the respective machine learning models. This allows the approach to extract discriminative information from the input data, which improves classification accuracy. These findings underscore the importance of data normalization and feature selection in enhancing the performance of machine learning models for gene expression analysis.

Normalization	Feature Selection Method	Number of Features	Classifier	Accuracy	F1-Score
None	None	19238	Logistic Regression	0.9643	0.9643
		19238	SVM	0.9630	0.9630
		19238	XGBoost	0.9552	0.9552
		19238	MLP	0.9534	0.9534
		19238	Random Forest	0.9217	0.9217
		19238	1D-CNN	0.9450	0.8440
Standard Scaler	None	19238	Logistic Regression	0.9689	0.9391
		19238	SVM	0.9603	0.9387
		19238	XGBoost	0.9552	0.9233
		19238	MLP	0.9479	0.9203
		19238	Random Forest	0.9165	0.8673
		19238	1D-CNN	0.9460	0.8420
MinMax Scaler	None	19238	Logistic Regression	0.9617	0.9617
		19238	SVM	0.9620	0.9620
		19238	XGBoost	0.9547	0.9235
		19238	Random Forest	0.9223	0.9223
		19238	MLP	0.8446	0.8446
		19238	1D-CNN	0.9496	0.8450
Standard Scaler	Lasso	100	Logistic Regression	0.8990	0.8990
		100	Random Forest	0.9071	0.9072
		100	MLP	0.9204	0.9204
		100	SVM	0.9137	0.9138
		100	XGBoost	0.9244	0.9244
		500	Logistic Regression	0.9506	0.9506
		500	Random Forest	0.9192	0.9192
		500	MLP	0.9370	0.9370
		500	SVM	0.9498	0.9498
		500	XGBoost	0.9474	0.9474
		1000	Logistic Regression	0.9547	0.9547
		1000	Random Forest	0.9207	0.9207
		1000	MLP	0.9472	0.9473
		1000	SVM	0.9538	0.9538
		1000	XGBoost	0.9487	0.9487

Normalization	Feature Selection Method	Number of Features	Classifier	Accuracy	F1-Score
Standard Scaler	SelectFromModel	100	Logistic Regression	0.9201	0.9201
		100	SVM	0.9371	0.9371
		100	XGBoost	0.9263	0.9263
		100	Random Forest	0.8970	0.8970
		100	MLP	-	-
		500	Logistic Regression	0.9631	0.9631
		500	SVM	0.9587	0.9587
		500	XGBoost	0.9510	0.9510
		500	Random Forest	0.9300	0.9300
		500	MLP	-	-
		1000	Logistic Regression	0.9636	0.9636
		1000	SVM	0.9638	0.9638
		1000	XGBoost	0.9539	0.9539
		1000	Random Forest	0.9286	0.9286
		1000	MLP	-	-
Standard Scaler	Select-K-Best	100	Logistic Regression	0.8812	0.8812
		100	Random Forest	0.8729	0.8729
		100	SVM	0.9027	0.9027
		100	XGBoost	0.9009	0.9009
		100	1D-CNN	0.7930	0.7930
		100	MLP	0.9021	0.9027
		500	Logistic Regression	0.9441	0.9441
		500	Random Forest	0.9107	0.9107
		500	SVM	0.9424	0.9424
		500	XGBoost	0.9371	0.9371
		500	1D-CNN	0.8852	0.8852
		500	MLP	0.9399	0.9399
		1000	Logistic Regression	0.9551	0.9551
		1000	Random Forest	0.9139	0.9139
		1000	SVM	0.9514	0.9514
		1000	XGBoost	0.9416	0.9416
		1000	MLP	0.9469	0.9469

Normalization	Feature Selection Method	Number of Features	Classifier	Accuracy	F1-Score
Standard Scaler	ElasticNet	100	Logistic Regression	0.8843	0.8843
		100	Random Forest	0.8803	0.8803
		100	SVM	0.9070	0.9070
		100	XGBoost	0.9164	0.9164
		100	MLP	0.9160	0.9160
		500	Logistic Regression	0.9428	0.9428
		500	Random Forest	0.8997	0.8997
		500	SVM	0.9429	0.9429
		500	XGBoost	0.9358	0.9358
		500	MLP	0.9371	0.9371
		1000	Logistic Regression	0.9428	0.9428
		1000	Random Forest	0.8997	0.8997
		1000	SVM	0.9429	0.9429
		1000	XGBoost	0.9358	0.9358
		1000	MLP	0.9371	0.9371

Table A.2: Empirical Results