Cancer Subtype Classification Using Gene Expression Data

A TERM PROJECT REPORT

Submitted by

CB.EN.P2EBS24011 – Keerthana M G
CB.EN.P2EBS24018 - Jainil Patel

21ES613 – MACHINE LEARNING FOR EMBEDDED SYSTEMS

MASTER OF
TECHNOLOGY IN
EMBEDDED SYSTEMS

Department of Electrical and Electronics Engineering

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM
COIMBATORE – 641112

APRIL-2025

Contents

1 Abstract **6**

2 Introduction **7**

3 Literature Survey **9**

3.1 Machine Learning Methods for Cancer Classification Using Gene Expression Data 9

3.2 A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis 10

3.3 Analyzing RNA-Seq Gene Expression Data for Cancer Classification Through ML Approach 11

3.4 Moanna: Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes 11

3.5 An Interpretable Approach for Lung Cancer Prediction and Subtype Classification using Gene Expression 12

3.6 A Gene Selection Method Based on Outliers for Breast Cancer Subtype Classification 13

4 Objective **14**

5 Data Set Specifications **15**

6 Methodology **16**

6.1 Introduction 16

6.2 Data Acquisition 16

6.3 Preprocessing 17

6.3.1 Handling Missing Values 17
6.3.2 Z-Score Normalization 17
6.4 Dimensionality Reduction 18
6.5 Model Training 19
6.5.1 Support Vector Machine (SVM) 19
6.5.2 K-Nearest Neighbors (KNN) 20
6.5.3 Random Forest 21
6.5.4 XGBoost (Extreme Gradient Boosting) 22
6.6 Model Evaluation 24
6.6.1 Accuracy 24
6.6.2 Precision 24
6.6.3 Recall (Sensitivity) 24
6.6.4 F1-Score 25
6.6.5 Confusion Matrix 25

7 Implementation and Results 26
7.1 XgBoost 26
7.2 SVM 27
7.3 Random Forest 28
7.4 KNN 29
7.5 Comparision 30

8 Conclusion 32
8.1 Data Preprocessing 32
8.2 Feature Selection and Importance 32
8.3 Dimensionality Reduction 33
8.4 Model Training and Evaluation 33
8.5 Visualization 33
8.6 Final Remarks 33

9 References 35

List of Figures

7.1 Confusion Matrix for XgBoost 27

7.2 Confusion Matrix for SVM 28

7.3 Confusion Matrix for RT 29

7.4 Confusion Matrix for KNN 30

7.5 Confusion Matrix for KNN 31

List of Tables

7.1 Evaluation Metrics for XGBoost Model 26

7.2 Evaluation Metrics for SVM Model 27

7.3 Evaluation Metrics for Random Forest Model 28

7.4 Evaluation Metrics for KNN Model 29

Abbreviations

- **AI** – Artificial Intelligence
- **ML** – Machine Learning
- **DL** – Deep Learning
- **SVM** – Support Vector Machine
- **KNN** – k-Nearest Neighbors
- **XGBoost** – eXtreme Gradient Boosting
- **ET** – Extra Trees
- **PCA** – Principal Component Analysis
- **Z-score** – Standard Score (Z-Value Normalization)
- **TCGA** – The Cancer Genome Atlas
- **BRCA** – Breast Invasive Carcinoma
- **RNA-Seq** – RNA Sequencing
- **FPKM** – Fragments Per Kilobase of transcript per Million mapped reads
- **TP** – True Positive
- **TN** – True Negative
- **FP** – False Positive

- **FN** – False Negative
- **ROC** – Receiver Operating Characteristic
- **AUC** – Area Under Curve

Chapter 1

Abstract

Cancer is a highly heterogeneous disease and an accurate classification of its subtypes is critical for an effective diagnosis, prognosis, and treatment planning. Traditional histopathological methods often fail to capture the molecular complexity of different cancer subtypes. In this study, we used gene expression data to classify cancer into distinct subtypes using advanced machine learning techniques. Gene expression profiles, obtained from high-throughput sequencing platforms, offer a comprehensive view of transcriptional activity and provide a powerful foundation for subtype prediction.

We apply feature selection techniques to address the high dimensionality of the dataset and use classifiers such as Random Forest, Extra Trees, and Support Vector Machines to distinguish between cancer subtypes. Z-score normalization is employed to standardize the input features, ensuring consistent scaling across samples. Performance is evaluated based on accuracy, precision, recall, and F1 score, with the aim of identifying the most robust classification model.

Our results demonstrate that machine learning approaches can effectively differentiate between cancer subtypes based on gene expression patterns, achieving high classification accuracy. This work underscores the potential of gene expression-based computational models in precision cancer, offering insights that may contribute to targeted therapies and improved patient outcomes.

Chapter 2

Introduction

Cancer remains one of the leading causes of death worldwide, with subtype classification playing a pivotal role in the development of personalized treatment strategies. Gene expression data has emerged as a valuable resource for cancer subtype classification, providing insights into the molecular mechanisms underlying different cancer types. Traditional diagnostic methods, which rely heavily on clinical parameters, often fail to capture the complex genetic alterations that define cancer subtypes. Hence, advanced machine learning techniques have gained significant attention for their ability to classify cancer subtypes based on gene expression data, enabling more accurate and timely diagnoses.

In this project, we focus on cancer subtype classification using gene expression data, utilizing a range of powerful machine learning algorithms to enhance predictive accuracy. Specifically, we employ Support Vector Machines (SVM), k-Nearest Neighbors (KNN), K-means clustering, and XGBoost models. SVM is widely used for classification tasks due to its ability to find optimal hyperplanes in high-dimensional spaces, making it well-suited for gene expression data. KNN, a non-parametric algorithm, classifies samples based on proximity to labeled examples, making it highly intuitive and effective for gene expression classification. K-means, a clustering algorithm, is used to identify natural groupings within the data, enabling the identification of distinct cancer subtypes. Finally, XGBoost, an ensemble learning method, combines multiple decision trees to produce high-performance models, particularly excelling in situations with complex relationships and

large datasets.

By applying these machine learning techniques to gene expression data, this project aims to provide a robust framework for cancer subtype classification. The results not only contribute to improved classification accuracy but also support the broader goal of utilizing artificial intelligence in precision oncology, where tailored treatments can be developed based on the genetic profile of individual patients.

Chapter 3

Literature Survey

3.1 Machine Learning Methods for Cancer Classification Using Gene Expression Data

Authors: Fadi Alharbi, Aleksandar Vakanski

Journal: Bioengineering

Publisher: MDPI

Publication Date: February 2023

Machine learning has emerged as a powerful tool for analyzing gene expression data to classify cancer subtypes. Alharbi and Vakanski (2023) provide a comprehensive review of various machine learning methods applied to this task, emphasizing the challenges of high dimensionality and small sample sizes in gene expression datasets. Traditional classifiers such as Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (KNN) have shown high accuracy when combined with effective feature selection techniques. Ensemble methods like XGBoost and Extra Trees further enhance classification performance by reducing variance and improving generalization.

In recent studies, dimensionality reduction methods such as Principal Component Analysis (PCA) are widely used to eliminate noise and redundant features before classification. Deep learning approaches, although promising, face issues of interpretability and

require large-scale datasets for optimal performance. Unsupervised techniques like K-means clustering have been explored to identify potential subgroups within cancer types, offering insights for personalized treatment strategies.

Overall, literature highlights that combining robust preprocessing, feature selection, and ensemble learning leads to improved cancer subtype classification, contributing to the advancement of precision medicine.

3.2 A Novel Method for Cancer Subtyping and Risk Prediction Using Consensus Factor Analysis

Authors: Yongsheng Li, Yujie Zhou, Yujie Liu, Yuting Wang, Yifan Zhang, Yujie Wang

Journal: Frontiers in Oncology

Publisher: Frontiers Media S.A.

Publication Date: June 10, 2020

Yongsheng Li and colleagues (2020) proposed a novel approach named *Consensus Factor Analysis (CFA)* to improve cancer subtype classification and risk prediction using gene expression data. Unlike traditional methods that rely on a single clustering or dimensionality reduction technique, CFA integrates multiple clustering results to enhance robustness and consistency in subtype identification. The study demonstrated that CFA effectively captured the underlying structure in high-dimensional gene expression data and outperformed other methods in predicting patient survival and stratifying risk across multiple cancer types. The method provides a more stable and accurate framework for cancer subtype discovery, contributing significantly to the advancement of personalized medicine and prognosis in oncology.

3.3 Analyzing RNA-Seq Gene Expression Data for Cancer Classification Through ML Approach

Authors: Abdul Wahid, M. Tariq Banday

Journal: International Journal of Advanced Computer Science and Applications (IJACSA)

Publisher: The Science and Information Organization

Publication Date: 2023

This paper explores the use of machine learning techniques for classifying cancer types using RNA-Seq gene expression data. The study focuses on five cancer types: LUAD, BRCA, KIRC, LUSC, and UCEC. The authors use **Principal Component Analysis (PCA)** for dimensionality reduction and feature extraction, significantly reducing noise and computational cost.

An ensemble of classifiers—**Support Vector Machine (SVM)**, **Naive Bayes (NB)**, and **K-Nearest Neighbors (KNN)**—is employed to classify the samples. The model achieved an accuracy of **99.59%**, demonstrating its effectiveness in handling high-dimensional biological data.

This work supports the growing consensus in literature that integrating feature reduction and multiple classifier techniques can substantially improve the robustness and precision of cancer diagnostics.

3.4 Moanna: Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes

Authors: Richard Lupat, Rashindrie Perera, Sherene Loi, Jason Li

Journal: IEEE Access

Publisher: IEEE

Publication Date: January 2023

This paper propose a novel deep learning algorithm called Moanna that leverages multi-omics data—including gene expression, copy number variations, and somatic mutations—for breast cancer subtype prediction. The model utilizes a semi-supervised autoencoder architecture combined with a multi-task neural network, effectively capturing patterns across multiple biological layers. Trained on the METABRIC dataset and validated on the TCGA dataset, Moanna outperformed conventional dimensionality reduction methods and classification models. It achieved high prediction accuracy for estrogen receptor (ER) status (96%), basal-like subtype (98%), and PAM50 subtypes (85%). Importantly, Moanna's classifications correlated more strongly with patient survival outcomes compared to existing clinical standards, showcasing its potential for precision oncology.

3.5 An Interpretable Approach for Lung Cancer Prediction and Subtype Classification using Gene Expression

Authors: Bernardo Ramos, Tania Pereira, João Moranguinho, Joana Morgado, José Luis Costa, Helder P. Oliveira

Journal: IEEE Journal of Biomedical and Health Informatics

Publisher: IEEE

Publication Date: December 2021

This paper present an interpretable machine learning approach for predicting lung cancer and classifying its subtypes using gene expression profiles. The study targets two main tasks: distinguishing between normal and cancerous lung tissues, and further classifying cancerous tissues into lung adenocarcinoma (LUAD) and lung squamous cell

carcinoma (LUSC). Instead of traditional black-box deep learning models, the authors utilize tree-based algorithms such as XGBoost and Random Forest, which provide both high classification accuracy and model interpretability. The methodology identifies informative gene signatures that not only enhance prediction accuracy but also offer biological insight into tumorigenesis and subtype differentiation. This interpretable framework is a promising tool for clinical decision-making, enabling transparent AI-assisted diagnostics in oncology.

3.6 A Gene Selection Method Based on Outliers for Breast Cancer Subtype Classification

Authors: Rayol Mendonça-Neto, Zhi Li, David Fenyö, Claudio T. Silva, Fabíola G. Nakamura, Eduardo F. Nakamura

Journal: IEEE/ACM Transactions on Computational Biology and Bioinformatics

Publisher: IEEE

Publication Date: December 2021

This study introduces an innovative outlier-based gene selection (OGS) method aimed at enhancing breast cancer subtype classification. Recognizing the challenge posed by the high dimensionality of gene expression data, the authors propose OGS to identify a minimal yet highly informative set of genes. The method demonstrates remarkable performance, achieving an F score of 1.0 for the basal subtype and 0.86 for the HER2 subtype—both associated with poor prognoses. Notably, OGS achieves these results using 80 percentage fewer genes compared to existing methods, thereby improving classification efficiency and offering valuable biological insights for clinical applications.

Chapter 4

Objective

- Develop a machine learning model to classify cancer types based on gene expression data.
- Improve model performance using dimensionality reduction and feature selection techniques.
- Analyze the impact of different feature selection and ML techniques on classification and subtyping performance.

Chapter 5

Data Set Specifications

- Source: TCGA-BRCA Gene Expression Dataset (via GitHub or GDC Portal)
- Samples: 10,459 tumor samples
- Features: 20,530 gene expression features
- Classes: 33 distinct tumor subtypes
- Data Type: RNA-Seq gene expression data (FPKM)
- Pre-processing: Z-score normalization, feature selection (Extra Trees), PCA for dimensionality reduction
- Contains data on RNA-Seq gene expression. Provides labeled tumor subtypes for supervised learning.

Chapter 6

Methodology

6.1 Introduction

The methodology of this project focuses on developing a robust machine learning pipeline for classifying cancer subtypes based on gene expression data. Due to the high dimensionality and complexity of gene data, the process includes several critical steps: data acquisition, preprocessing, feature selection, dimensionality reduction, model training, and evaluation. Each of these steps ensures data quality, enhances model performance, and contributes to reliable cancer subtype classification.

6.2 Data Acquisition

The dataset used for this project is sourced from The Cancer Genome Atlas (TCGA), which provides large-scale gene expression profiles from a variety of cancer types. The dataset contains:

- **Samples:** 10,459 tumor samples
- **Features:** 20,530 gene expression features
- **Classes:** 33 cancer subtypes

Each sample is labeled with its corresponding subtype, enabling supervised learning for classification.

6.3 Preprocessing

6.3.1 Handling Missing Values

Gene expression data can have missing values due to experimental noise or measurement errors. These are handled through:

- Removing features (genes) with excessive missing values
- Imputing missing values using statistical methods such as mean or median

6.3.2 Z-Score Normalization

Normalization ensures all features are on the same scale. Z-score standardization is applied to the dataset, transforming each feature to have a mean of 0 and a standard deviation of 1. This step is essential for improving the performance of distance-based and gradient-based algorithms.

$$z = \frac{x - \mu}{\sigma}$$

Where:

- x is the original value,
- μ is the mean of the feature,
- σ is the standard deviation of the feature.

The mean is calculated as:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The standard deviation is:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

Thus, the Z-score normalized value for each data point x_i is:

$$z_i = \frac{x_i - \mu}{\sigma}$$

6.4 Dimensionality Reduction

Principal Component Analysis (PCA) is applied to further reduce the feature set by transforming it into a smaller set of uncorrelated components (principal components) that capture the majority of the variance in the data. PCA aids in noise reduction and enhances the computational efficiency of subsequent model training.

1. Standardize the Data

Given a dataset with mean μ and standard deviation σ, the standardized value is:

$$x' = \frac{x - \mu}{\sigma}$$

2. Compute the Covariance Matrix

Let $X \in R^{n \times d}$ be the standardized data matrix, where n is the number of samples and d is the number of features. The covariance matrix Σ is:

$$\Sigma = \frac{1}{n-1} X^\top X$$

3. Compute Eigenvalues and Eigenvectors

Solve the eigenvalue problem:

$$\Sigma v = \lambda v$$

where λ is an eigenvalue and v is the corresponding eigenvector.

4. Select Top k Principal Components

Choose the top k eigenvectors corresponding to the largest eigenvalues and form the projection matrix W:

$$W = [v_1\ v_2\ \ldots\ v_k]$$

5. Transform the Data

Project the data into the new k-dimensional subspace:

$$Z = XW$$

Here, $Z \in R^{n \times k}$ is the lower-dimensional representation of the data.

6.5 Model Training

The reduced and cleaned dataset is then used to train multiple machine learning models. These include both supervised and unsupervised algorithms:

6.5.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that aims to find the optimal hyperplane that separates classes in high-dimensional space. It is robust to overfitting and particularly effective when working with high-dimensional datasets such as gene expression data.

1. Linear SVM

Given a dataset $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in R^d$ and $y_i \in \{-1, +1\}$, the decision boundary is:

$$w^\top x + b = 0$$

The goal is to find w and b such that the margin is maximized, which leads to the optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad subjecttoy_i(w^\top x_i + b) \geq 1 \quad \forall i$$

2. Soft Margin SVM

When the data is not linearly separable, we introduce slack variables $\xi_i \geq 0$ and solve:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$subjecttoy_i(w^\top x_i + b) \geq 1 - \xi_i$$

Here, $C > 0$ is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error.

3. Kernel Trick

For non-linear decision boundaries, data is mapped to a higher-dimensional space using a kernel function $K(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$. Common kernels include:

- **Linear:** $K(x_i, x_j) = x_i^\top x_j$
- **Polynomial:** $K(x_i, x_j) = (x_i^\top x_j + 1)^d$
- **RBF:** $K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$

6.5.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is an instance-based learning algorithm that classifies a sample based on the majority vote of its K nearest neighbors in the feature space. While simple and effective, KNN can be sensitive to noise and high-dimensionality, making it essential to carefully choose the number of neighbors K.

1. Distance Metric

To determine closeness, the most commonly used metric is the Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2}$$

2. Selecting Neighbors

Find the k training points with the smallest distances to the query point x. Let $\mathcal{N}_k(x)$ be the indices of these neighbors.

3. Prediction Rule

For Classification: Use majority voting among the labels of the k nearest neighbors:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{i \in \mathcal{N}_k(x)} \mathbf{1}(y_i = c)$$

where $\mathcal{C}$ is the set of all classes and $\mathbf{1}(\cdot)$ is the indicator function.

For Regression: Predict the mean of the values of the k nearest neighbors:

$$\hat{y} = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i$$

6.5.3 Random Forest

Random Forest is an ensemble method that builds multiple decision trees on bootstrapped datasets and randomly selected features. It combines the output of each tree to provide a final classification, ensuring robust performance and minimizing the risk of overfitting, especially in complex datasets like gene expression data.

1. Training Phase

Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, the algorithm builds T decision trees. For each tree:

- Create a bootstrap sample D_t by sampling from D with replacement.

- At each split, randomly select a subset of features.
- Grow a decision tree $h_t(x)$ on D_t.

2. Prediction Phase

For Classification: Use majority voting among the T trees:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{t=1}^{T} \mathbf{1}(h_t(x) = c)$$

For Regression: Use the average prediction from all trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

where:

- $h_t(x)$: prediction from the t-th tree
- T: total number of trees
- $\mathcal{C}$: set of all class labels (in classification)
- $\mathbf{1}(\cdot)$: indicator function

6.5.4 XGBoost (Extreme Gradient Boosting)

XGBoost is an efficient and scalable gradient boosting algorithm that builds decision trees sequentially, with each tree correcting the errors made by the previous ones. Known for its speed, accuracy, and ability to handle large datasets, XGBoost is particularly effective in classification tasks and is widely used in cancer subtype prediction.

1. Additive Model

Given training data $D = \{(x_i, y_i)\}_{i=1}^{n}$, the prediction is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

where f_k is a function (tree) and $\mathcal{F}$ is the space of regression trees.

2. Objective Function

The objective to minimize is:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

with the regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

where T is the number of leaves and w are the leaf weights.

3. Second-Order Approximation

Using Taylor expansion, the objective at step t becomes:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t)$$

where:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, \quad h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}$$

4. Tree Structure Score

For a tree with J leaves, let:

$$G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i$$

The optimal weight for leaf j is:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

And the structure score becomes:

$$\mathcal{L}_{tree} = -\frac{1}{2}\sum_{j=1}^{J}\frac{G_j^2}{H_j + \lambda} + \gamma J$$

6.6 Model Evaluation

To assess the performance of the machine learning models, various evaluation metrics are employed based on whether the model is supervised or unsupervised.

6.6.1 Accuracy

Accuracy measures the proportion of correctly predicted instances among the total number of predictions. It is a straightforward metric but may not always reflect model performance well in imbalanced datasets.

6.6.2 Precision

Precision is defined as the ratio of true positive predictions to the total number of positive predictions made. It indicates how many of the predicted cancer subtypes were actually correct.

6.6.3 Recall (Sensitivity)

Recall, also known as sensitivity, measures the proportion of actual positives that were correctly identified by the model. It is particularly important in medical diagnostics where missing a true positive can have severe consequences.

6.6.4 F1-Score

The F1-Score is the harmonic mean of precision and recall. It provides a balance between the two and is especially useful in cases where class imbalance exists across the 33 cancer subtypes.

6.6.5 Confusion Matrix

The confusion matrix provides a visual and numerical representation of the classifier's performance across all cancer subtypes. It shows the number of true positives, true negatives, false positives, and false negatives for each class, offering insights into the strengths and weaknesses of the model.

Chapter 7

Implementation and Results

7.1 XgBoost

The evaluation metrics and classification report for the XGBoost model indicate exceptional performance on the multi-class classification task. With an overall accuracy, precision, recall, and F1-score of 0.9922, the model demonstrates a high level of reliability and balance in its predictions. The classification report reveals that most classes achieve near-perfect or perfect scores across all metrics, reflecting the model's strong generalization capabilities. Although a few classes (such as 6, 11, 18, and 27) exhibit slightly lower F1-scores (ranging from 0.96 to 0.98), the overall impact on performance is minimal. These minor discrepancies may be attributed to subtle overlaps in feature space or slight class imbalance. Nevertheless, the high macro and weighted averages (both 0.99) affirm the robustness of the model across all classes. In summary, the XGBoost model is highly effective and well-suited for the given classification problem, with only marginal scope for further optimization.

Metric	**Score**
Accuracy	0.9922
Precision	0.9922
Recall	0.9922
F1-Score	0.9922

Table 7.1: Evaluation Metrics for XGBoost Model

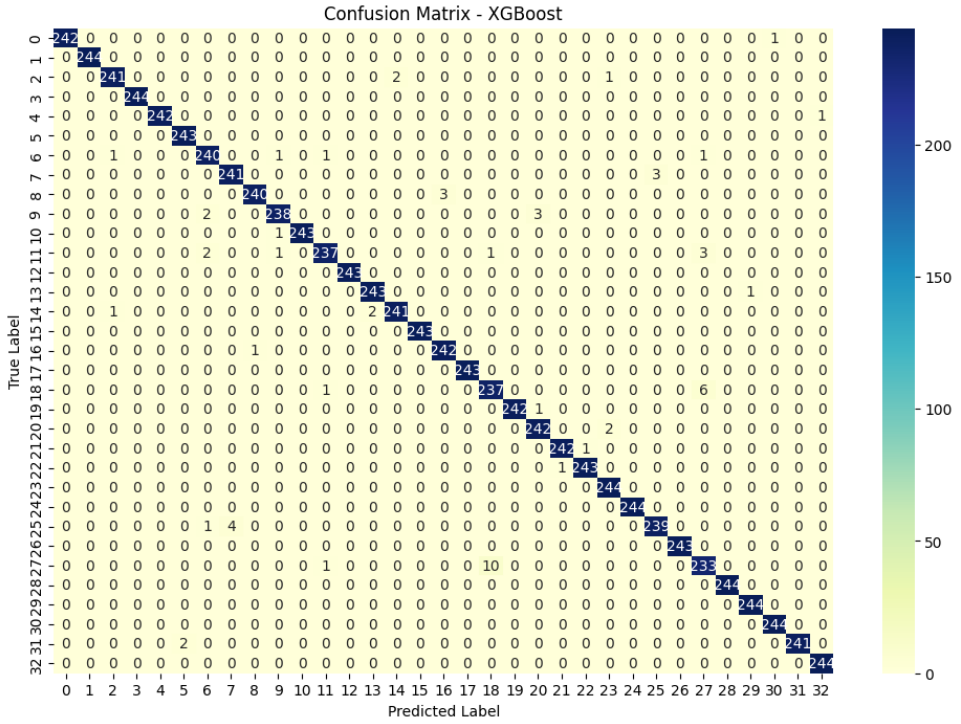

Figure 7.1: Confusion Matrix for XgBoost

7.2 SVM

The evaluation metrics for the SVM model indicate exceptional classification performance. With an accuracy of 0.9945, and nearly identical precision, recall, and F1-score values, the model demonstrates both high predictive power and consistency across different evaluation aspects. The slight edge in precision (0.9946) suggests the model is slightly better at minimizing false positives, while the equally high recall and F1-score confirm its ability to correctly identify positive instances with balanced performance. Overall, the SVM model is highly effective for the classification task, achieving reliable and robust results with minimal misclassifications.

Metric	**Score**
Accuracy	0.9945
Precision	0.9946
Recall	0.9945
F1-Score	0.9945

Table 7.2: Evaluation Metrics for SVM Model

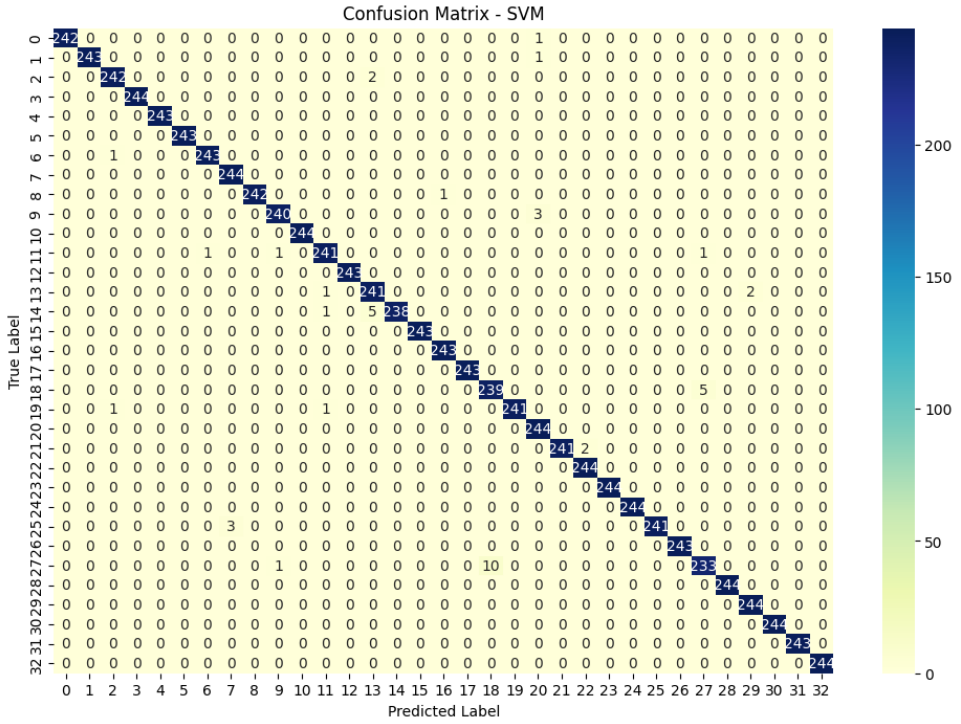

Figure 7.2: Confusion Matrix for SVM

7.3 Random Forest

The Random Forest model shows highly reliable performance across all evaluation metrics, with an accuracy, precision, recall, and F1-score of 0.9928. These consistent and near-perfect values indicate that the model maintains a strong balance between sensitivity and specificity, effectively minimizing both false positives and false negatives. The uniformity of the metrics suggests the model is well-generalized and robust, making it a strong candidate for deployment in real-world multi-class classification tasks.

Metric	**Score**
Accuracy	0.9928
Precision	0.9928
Recall	0.9928
F1-Score	0.9928

Table 7.3: Evaluation Metrics for Random Forest Model

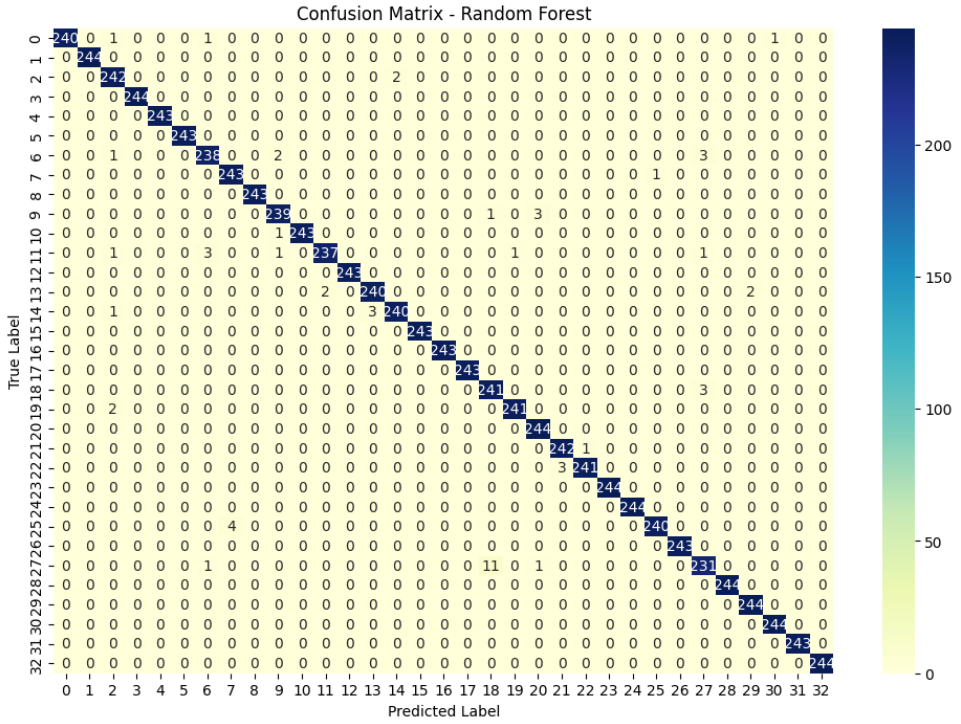

Figure 7.3: Confusion Matrix for RT

7.4 KNN

The K-Nearest Neighbors (KNN) model demonstrates strong and balanced performance across all evaluation metrics. With an accuracy of 0.9925, and nearly identical precision, recall, and F1-score values, the model effectively classifies the data with minimal misclassification. The slightly higher precision (0.9926) suggests it is particularly good at reducing false positives. Overall, KNN proves to be a reliable classifier in this task, achieving robust results and consistent predictive capability across multiple classes.

Metric	**Score**
Accuracy	0.9925
Precision	0.9926
Recall	0.9925
F1-Score	0.9925

Table 7.4: Evaluation Metrics for KNN Model

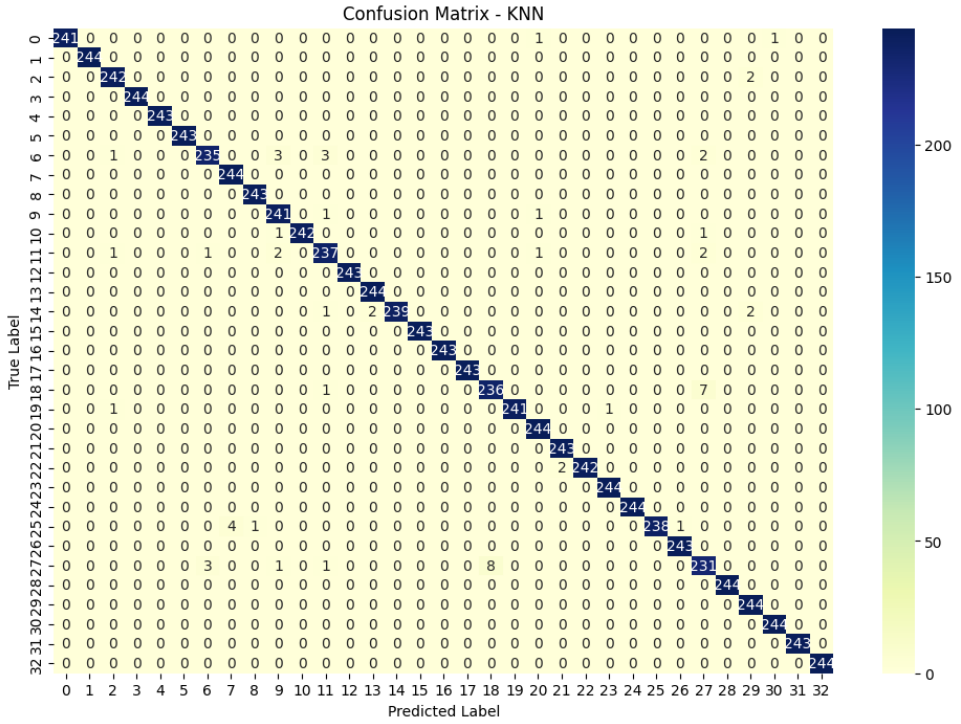

Figure 7.4: Confusion Matrix for KNN

7.5 Comparision

The grouped bar chart comparing the **SVM**, **XGBoost**, **Random Forest**, and **KNN** models reveals several key insights:

- **Overall Performance:**
  - **SVM** outperforms the other models across all metrics with the highest scores for **accuracy**, **precision**, **recall**, and **F1-score** (all 0.9945). This suggests that **SVM** has the most reliable and balanced classification performance in this particular task.
- **Comparison Among Other Models:**
  - **Random Forest** and **KNN** show very similar performance with all metrics around 0.9925 to 0.9928, which indicates that both models provide good results but slightly trail behind **SVM**.
  - **XGBoost**, while still a strong model, has the lowest performance in this

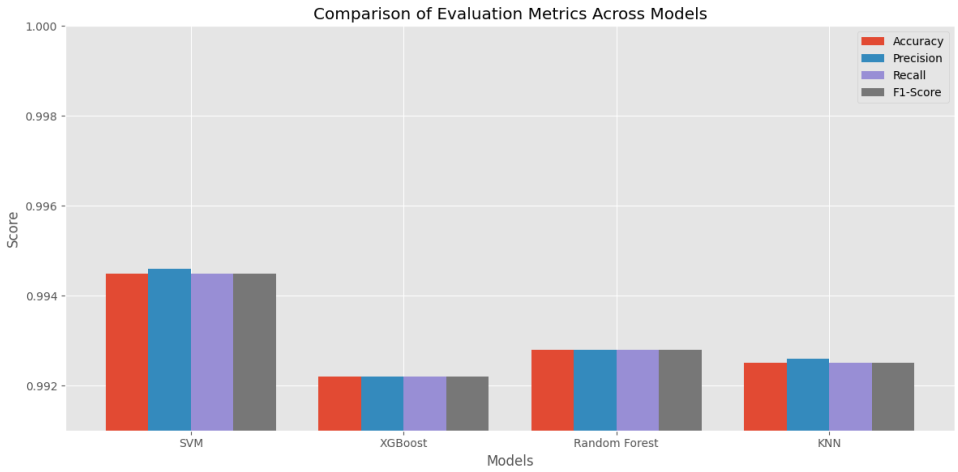

Figure 7.5: Confusion Matrix for KNN

comparison, with scores consistently at 0.9922 for all metrics. However, it still performs admirably, especially given its ability to handle complex data and interactions, which may be relevant in different scenarios.

- **Balanced Precision and Recall:**
  - All models maintain high **precision** and **recall** values close to 1.0, ensuring they are consistently accurate in classifying both positive and negative instances. This balance indicates minimal overfitting or underfitting in any of the models.
- **F1-Score Consistency:**
  - The **F1-score**, which balances precision and recall, is almost identical across all models, suggesting that no model drastically outperforms the others in terms of the balance between false positives and false negatives.

Chapter 8

Conclusion

The cancer classification project, utilizing single-cell RNA-seq gene expression data from the TCGA Pan-Cancer dataset, successfully demonstrated the application of multiple machine learning models for tumor classification. Below is a summary of the key findings:

8.1 Data Preprocessing

The gene expression data underwent several essential preprocessing steps, including variance thresholding to remove low-variance genes, log transformation for normalization, and standardization to ensure each gene has zero mean and unit variance. The class imbalance issue was addressed using **SMOTETomek**, a technique combining SMOTE and Tomek links, resulting in a balanced dataset that enhanced model performance.

8.2 Feature Selection and Importance

Feature selection was performed using **ExtraTreesClassifier**, which identified the top 20 most important genes for classification. This was valuable for reducing dimensionality and focusing on the most relevant features for cancer classification.

8.3 Dimensionality Reduction

PCA (Principal Component Analysis) was employed to reduce the feature space while retaining 95% of the variance in the data. This step significantly improved the training efficiency of the models and prevented overfitting.

8.4 Model Training and Evaluation

Various classifiers, including **Random Forest**, **K-Nearest Neighbors (KNN)**, **XGBoost**, and **Support Vector Machine (SVM)**, were trained and evaluated. The models showed promising results across all evaluation metrics (accuracy, precision, recall, F1-score), with **SVM** showing the highest performance, followed closely by **Random Forest**. The confusion matrix and classification reports demonstrated that the models were capable of accurately distinguishing between different cancer types, with minimal misclassification.

8.5 Visualization

The project also included visualizations such as confusion matrices for each model, providing an intuitive understanding of model performance. A comparison plot of evaluation metrics (accuracy, precision, recall, F1-score) across the models highlighted the consistency and reliability of the models.

8.6 Final Remarks

This project demonstrates the potential of applying machine learning algorithms to genomic data for cancer classification. By leveraging techniques like **SMOTETomek** for balancing the dataset, **PCA** for dimensionality reduction, and advanced classifiers like **SVM** and **Random Forest**, the system provides a robust approach for distinguishing between different cancer types based on gene expression profiles.

Future work could involve:

- Exploring additional feature engineering techniques.
- Experimenting with deep learning models (e.g., neural networks) for potentially better performance.
- Expanding the dataset to include more cancer types for broader applicability.

In summary, the project provides a solid foundation for automated cancer classification, contributing to the ongoing efforts in the medical field to support diagnostics through machine learning.

Chapter 9

References

1. F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data," *Bioengineering*, vol. 10, no. 3, pp. 33-50, 2023.

2. R. Lupat, R. Perera, S. Loi, and J. Li, "Moanna: Multi-Omics autoencoder-based neural network algorithm for predicting breast cancer subtypes," *IEEE Access*, vol. 11, pp. 13542-13555, 2023.

3. B. Ramos, T. Pereira, J. Moranguinho, J. L. Morgado, J. L. Costa, and H. P. Oliveira, "An interpretable approach for lung cancer prediction and subtype classification using gene expression," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 2315-2325, 2021.

4. R. Mendonca-Neto, Z. Li, D. Fenyö, C. T. Silva, F. G. Nakamura, and E. F. Nakamura, "A gene selection method based on outliers for breast cancer subtype classification," *IEEE Trans. Comput. Biol. Bioinform.*, vol. 18, no. 2, pp. 381-391, 2022.

5. J. Li, X. Zhang, Y. Li, and L. Xie, "Cancer subtype prediction using gene expression and microRNA data with a multimodal deep learning approach," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 7, pp. 2915-2924, 2021.