TCGA: Cancer Genomics Project

FALL - 2023

In collaboration with

Cedars Sinai

Dr. Jianchao Han
Advisor

Dr. Mohsen Beheshti
Committee Member

Dr. Ryan Urbanowicz
Committee Member

By:

Sai Keertana Padmanabham
212231853

CSUDH

AI Campus

Goal

The goal of the "TCGA Cancer Genomics Project" is to utilize machine learning to precisely classify cancer types based on genetic data, contributing to personalized medicine and treatment strategies.

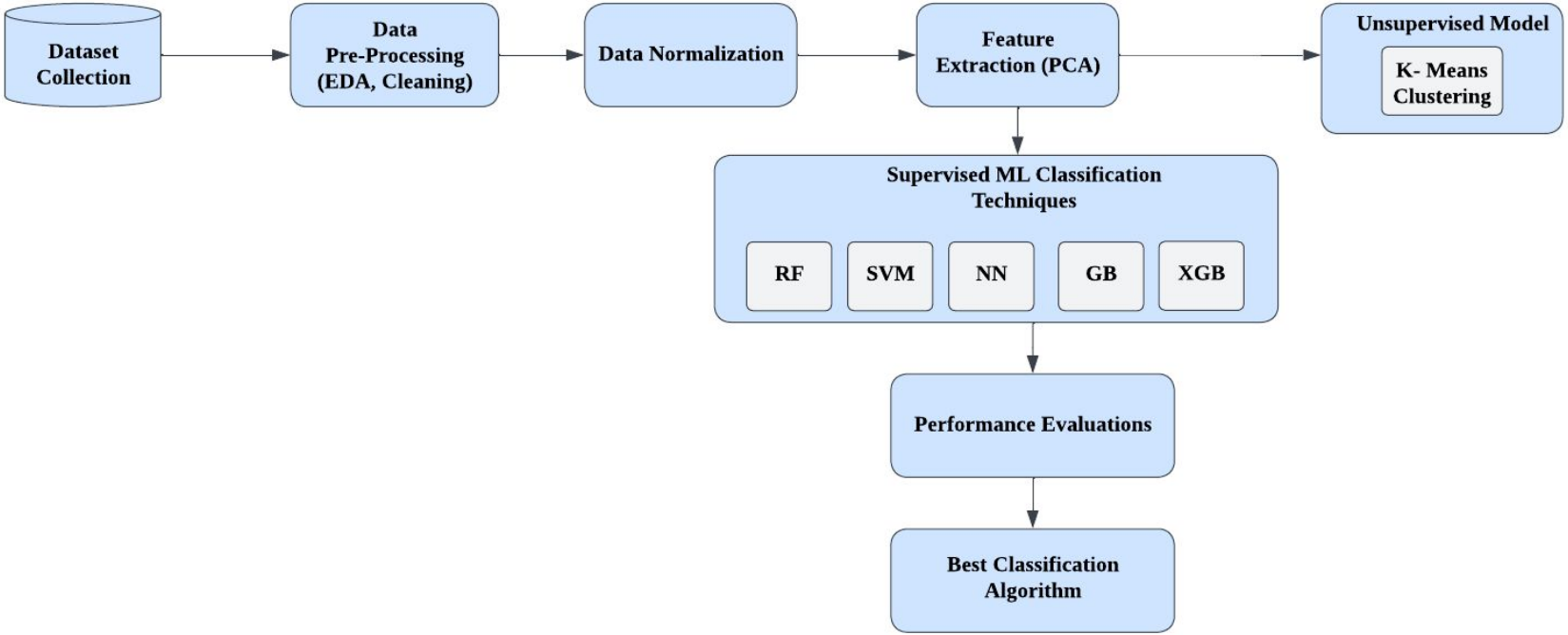
Problem Statement

This project aims to accurately classify cancer types using advanced machine learning algorithms, enhancing our understanding of tumor genetics for improved diagnosis and personalized treatment.

Related Work

The studies and advancements in cancer genomics and machine learning, focusing on genetic expression analysis for cancer classification and treatment personalization.

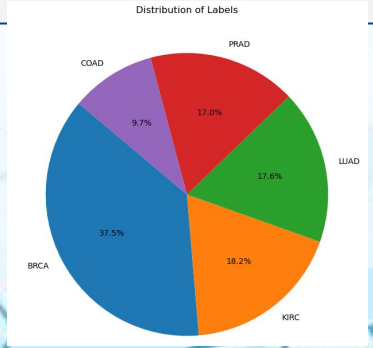
System Design



Methodology Overview

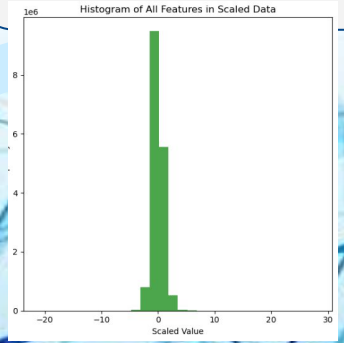
Data Preprocessing

Performing EDA analysis for the dataset provided, examined the descriptive statistics and removed the Inconsistencies.



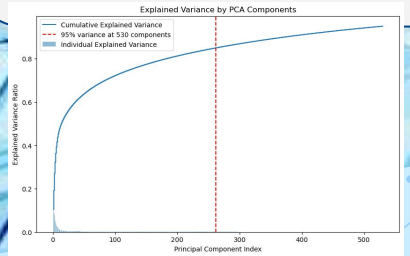
Data Normalization

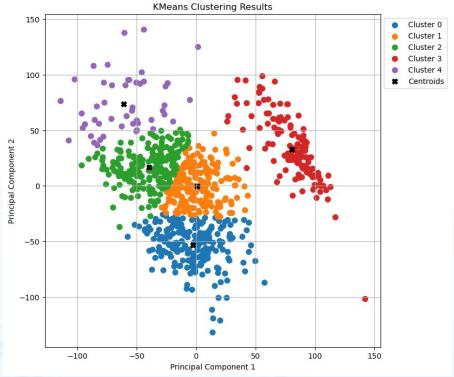
Ensuring the data is scaled to zero mean and unit variance to enhance the accuracy and efficiency of the machine learnin models.



Feature Extraction

The PCA was applied to our dataset to reduce its dimensionality, ensuring that 95% of the variance is retained for effective analysis and modeling.





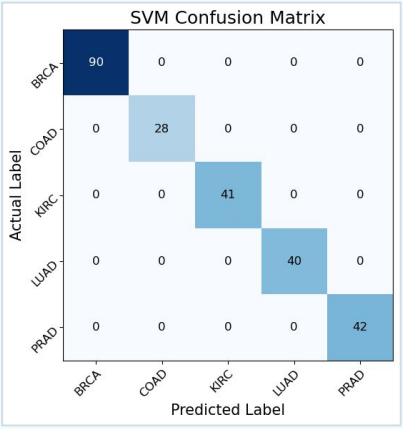
Unsupervised Learning

K-means clustering is used to uncover hidden patterns and structures in the data. By grouping similar data points into clusters, the algorithm reveals underlying patterns that might not be immediately apparent.

Performance

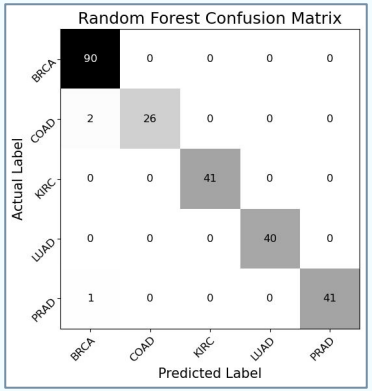
Accuracy

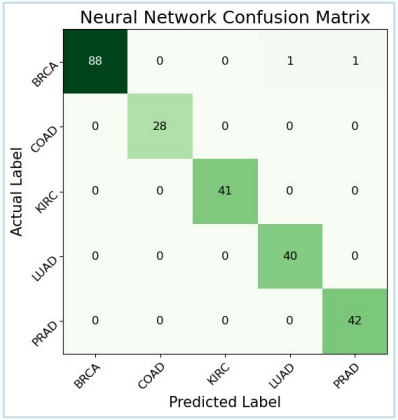
Random Forest: 98.7%
SVM: 100%
Neural Network: 99.1%
Gradient Boost: 99.1%
XGB: 98.1%

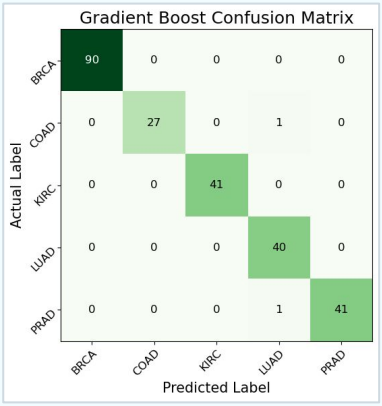


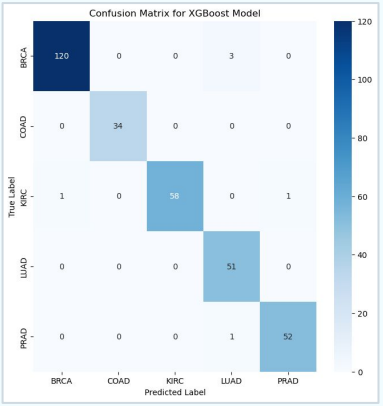
Supervised Learning

(1) Random Forest
(2) Support Vector Machine
(3) Neural Networks
(4) Gradient Boost
(5) XGBoost



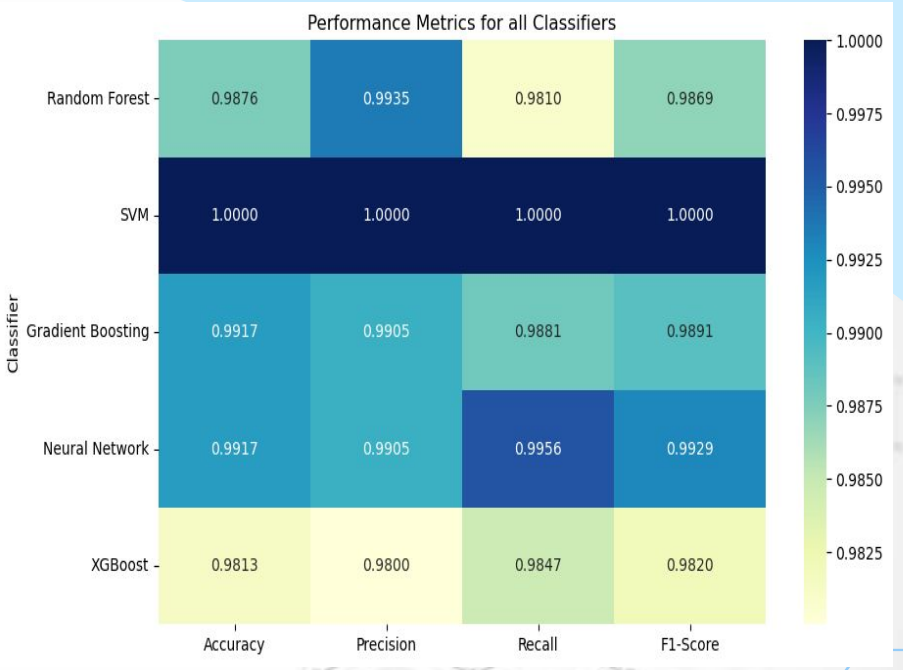






Results Analysis

Our comparative performance metrics showcase the robustness of the classifiers used in the TCGA Cancer Genomics Project. The SVM classifier outperformed others with perfect scores across all metrics. Random Forest and XGBoost showed impressive results, with both achieving above 98% across all the metrics, demonstrating their effectiveness in handling complex genomic data. Gradient Boosting and Neural Networks also performed admirably, each displaying over 99% accuracy, which underscores the potential of these models in predictive accuracy and reliability in cancer genomics classification



Future Work

Mainly focus on expanding my dataset to include more diverse genetic profiles, enhancing the model's predictive power across a broader range of cancer types. also plan to integrate deep learning approaches to uncover more complex patterns in the data that traditional algorithms might miss. Lastly, we will seek to collaborate with biologists and oncologists to translate these computational findings into actionable medical interventions.