AI Campus

CSC 590 Master's Project

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

Presented By:

Sai Keertana Padmanabham
212231853

CSUDH

AGENDA

01 Introduction

02 Objective

03 System Design

04 Methodologies

05 Results

06 Conclusion

INTRODUCTION

- The Genomics Project provides a deep dive into the field of gene expression, utilizing data that was first made accessible by the TCGA Pan Cancer research project and then collected by the UCI Machine Learning Repository.
- The data is a part of the RNA-Seq (HiSeq) PANCAN dataset, representing a random selection of gene expressions from patients with different Cancer types.

Cancer types in Focus are:

- BRCA
- KIRC
- COAD
- LUAD
- PRAD

- A key focus is feature selection to identify genes that accurately distinguish different cancer types, blending data analytics with cancer genomics for enhanced understanding.

OBJECTIVE

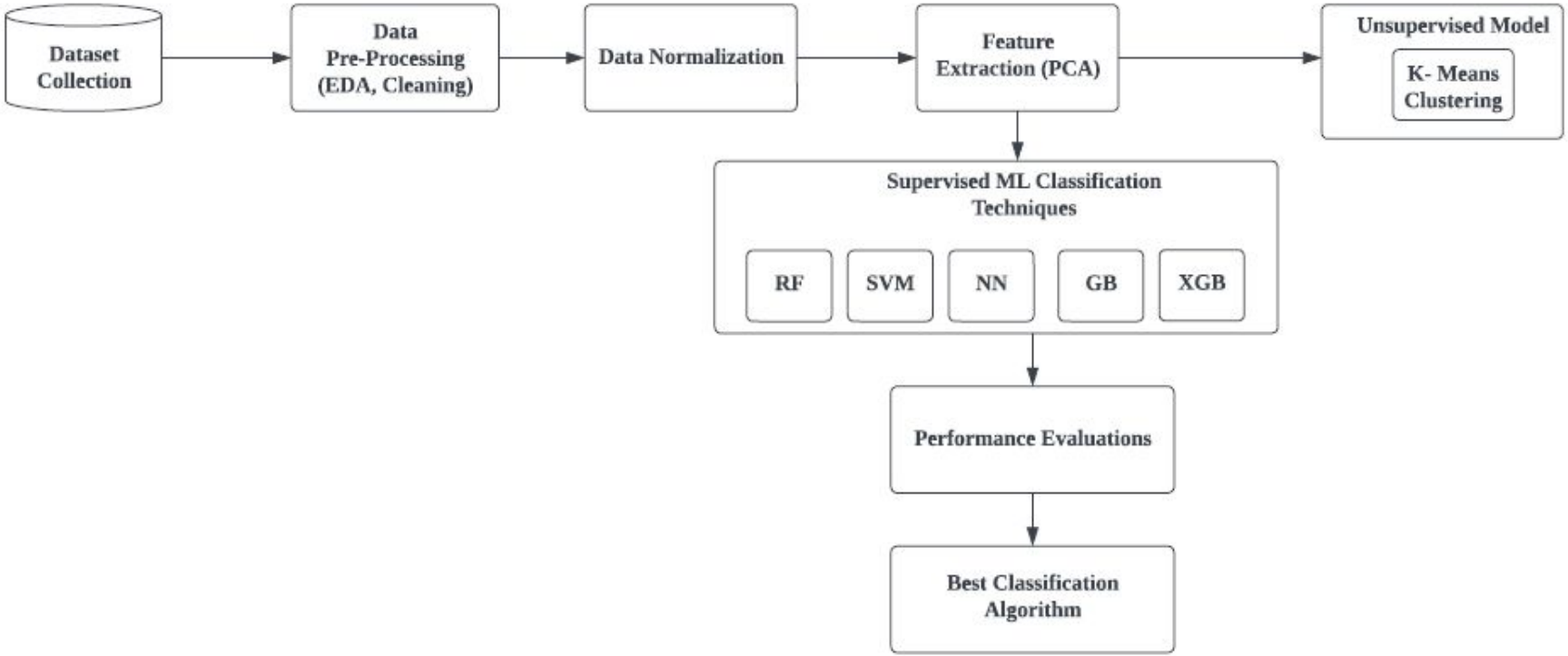
The primary objectives of the TCGA Research Network, which are to create a comprehensive understanding of genetic irregularities across different tumor generations, are in line with this idea.

- **Unsupervised Learning:** This technique is to analyze gene expression data, uncovering hidden patterns and structures, providing insights into the complex genetic interactions underlying various cancer types.

  K- Means Clustering
- **Supervised Learning:** This method used to accurately classify cancers based on gene expression, using known labels in training data to predict new, unseen data categories.

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

SYSTEM DESIGN



DATA COLLECTION

- The dataset is extracted from The Cancer Genome Atlas (TCGA), accessible through the Genomic Data Commons (GDC) portal.

- The dataset contains gene expression data from patients diagnosed with various cancer types, including BRCA, KIRC, COAD, LUAD, and PRAD. The dataset contains approximately **801** records and **20531** attributes, each representing a different gene and its expression level.

Labels Data

	Unnamed: 0	Class
0	sample_0	PRAD
1	sample_1	LUAD
2	sample_2	PRAD
3	sample_3	PRAD
4	sample_4	BRCA
...	...	...
796	sample_796	BRCA
797	sample_797	LUAD
798	sample_798	COAD
799	sample_799	PRAD
800	sample_800	PRAD

801 rows × 2 columns

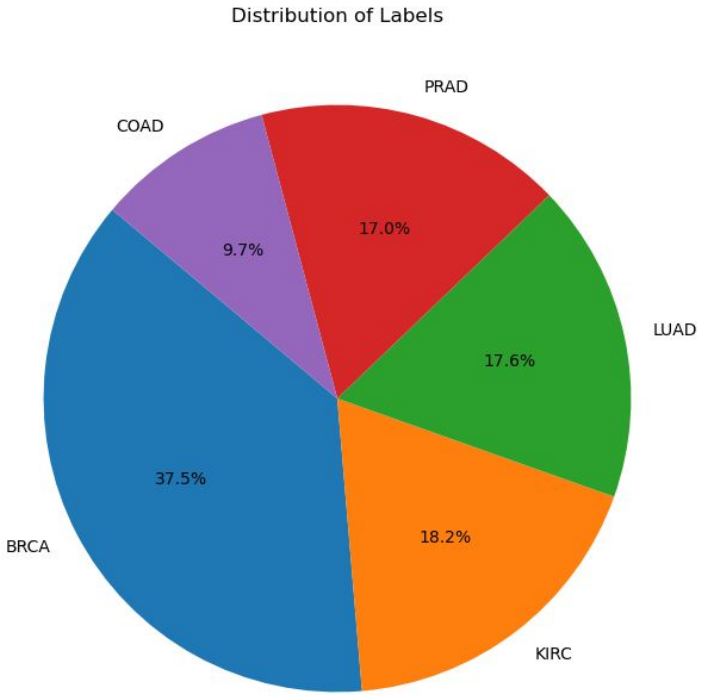
	Unnamed: 0	gene_0	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	...	gene_20521	gene_20522	gene_20523	gene_20524
0	sample_0	0.0	2.017209	3.265527	5.478487	10.431999	0.0	7.175175	0.591871	0.0	...	4.926711	8.210257	9.723516	7.220030
1	sample_1	0.0	0.592732	1.588421	7.586157	9.623011	0.0	6.816049	0.000000	0.0	...	4.593372	7.323865	9.740931	6.256586
2	sample_2	0.0	3.511759	4.327199	6.881787	9.870730	0.0	6.972130	0.452595	0.0	...	5.125213	8.127123	10.908640	5.401607
3	sample_3	0.0	3.663618	4.507649	6.659068	10.196184	0.0	7.843375	0.434882	0.0	...	6.076566	8.792959	10.141520	8.942805
4	sample_4	0.0	2.655741	2.821547	6.539454	9.738265	0.0	6.566967	0.360982	0.0	...	5.996032	8.891425	10.373790	7.181162
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
796	sample_796	0.0	1.865642	2.718197	7.350099	10.006003	0.0	6.764792	0.496922	0.0	...	6.088133	9.118313	10.004852	4.484415
797	sample_797	0.0	3.942955	4.453807	6.346597	10.056868	0.0	7.320331	0.000000	0.0	...	6.371876	9.623335	9.823921	6.555327
798	sample_798	0.0	3.249582	3.707492	8.185901	9.504082	0.0	7.536589	1.811101	0.0	...	5.719386	8.610704	10.485517	3.589763
799	sample_799	0.0	2.590339	2.787976	7.318624	9.987136	0.0	9.213464	0.000000	0.0	...	5.785237	8.605387	11.004677	4.745888
800	sample_800	0.0	2.325242	3.805932	6.530246	9.560367	0.0	7.957027	0.000000	0.0	...	6.403075	8.594354	10.243079	9.139459

801 rows × 20532 columns

Features Data

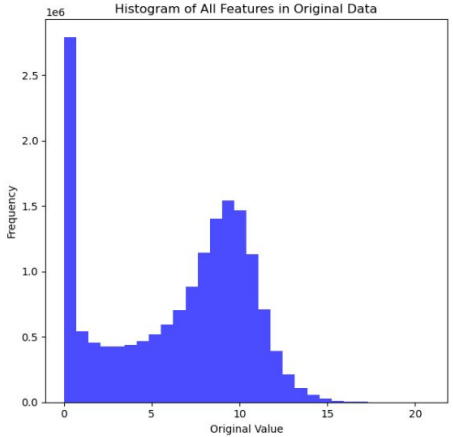
DATA PRE-PROCESSING

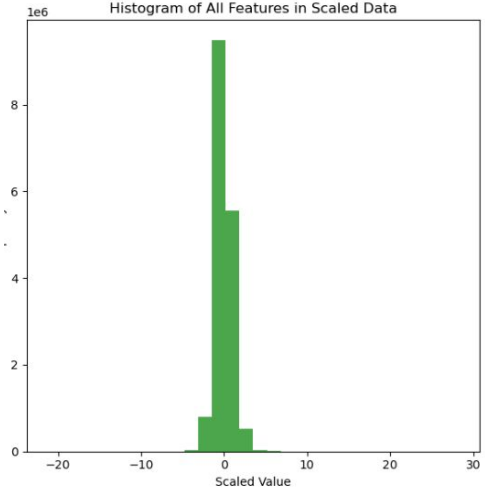
- The data and labels in the Cancer Genomics dataset are complete, with no missing entries, according to the Exploratory Data Analysis (EDA) that looked for any missing values.
- Checked descriptive statistics that give a thorough summary of the data, including the quartiles, count, mean, and standard deviation.
- Once the data is loaded, removed all the null or missing values present in the dataset, eliminating these values can adversely affect the accuracy of our machine learning models.



DATA NORMALIZATION

- Normalization helps in making the training process well-behaved, improving the accuracy and efficiency of the machine learning models.
- I have used the StandardScaler from the sklearn.preprocessing library.
- It typically involves scaling the data so that it fits within a specific range or has specific statistical properties, such as a zero mean and a standard deviation of one.



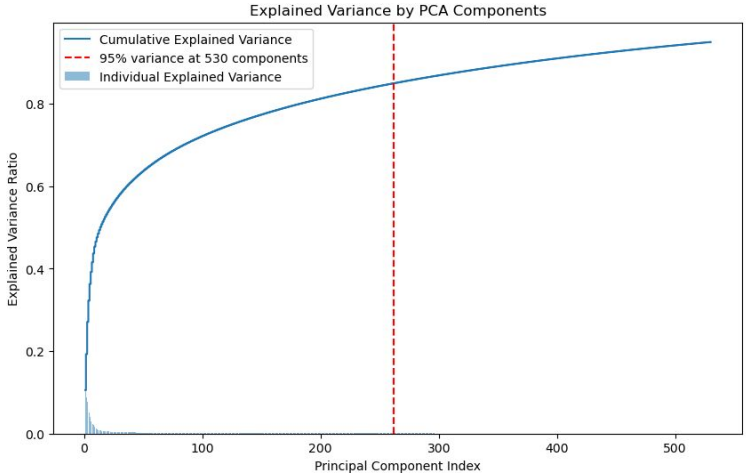


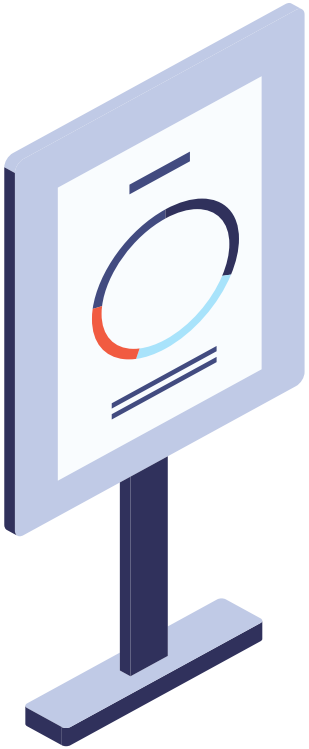
FEATURE EXTRACTION

- PCA is most commonly used for reducing the dimensionality of large data sets. By transforming the data into fewer dimensions, PCA helps in simplifying the data structure without losing significant information.

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

```
Original number of features: 20531
Reduced number of features: 530
```



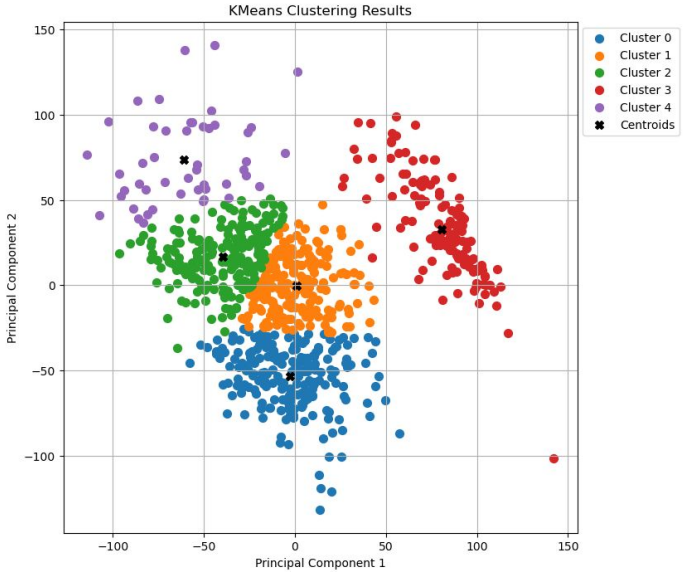


METHODOLOGIES

UNSUPERVISED LEARNING
K - MEANS CLUSTERING

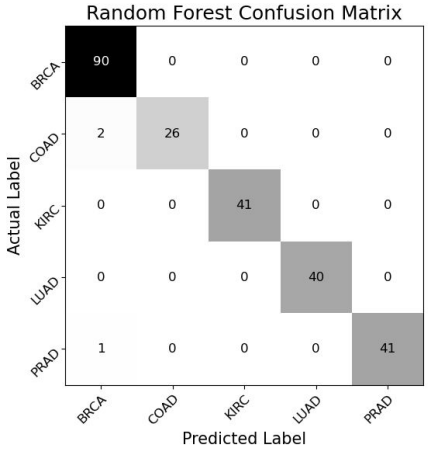
- K-means clustering is used to uncover hidden patterns and structures in the data. By grouping similar data points into clusters, the algorithm reveals underlying patterns that might not be immediately apparent.

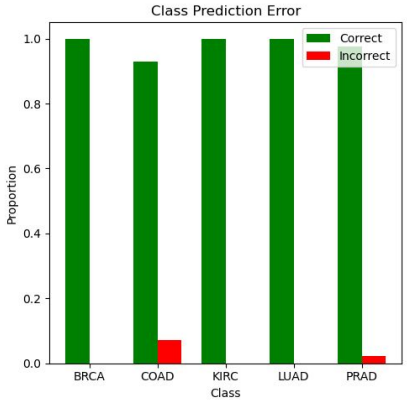
- For high dimensional dataset, clustering simplifies this data by segmenting it into distinct groups, making it easier to analyze and interpret.



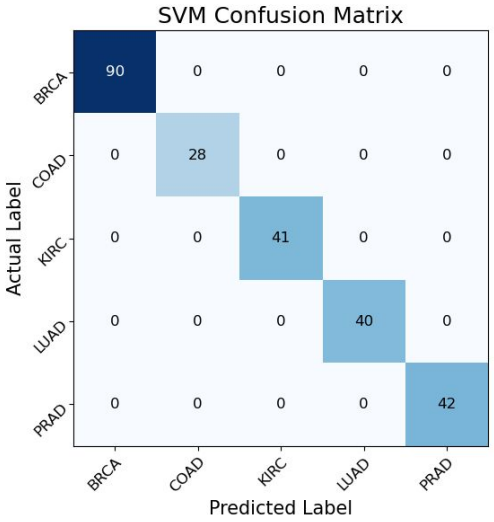
RANDOM FOREST

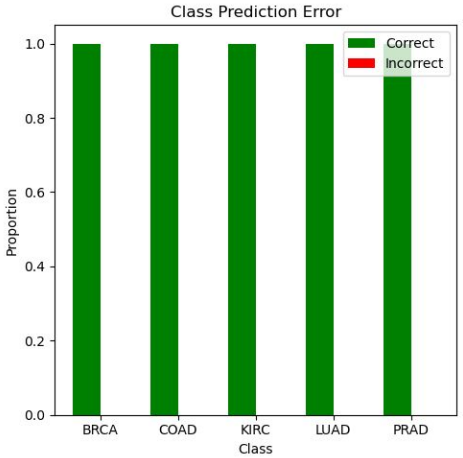
- Random Forest is well-suited for high-dimensional datasets, common in genomics. It can handle thousands of input variables without variable deletion, crucial for analyzing comprehensive gene expression data.
- The Random Forest classifier demonstrated high effectiveness, with an overall accuracy of **98.7%** in classifying different cancer types.





SUPPORT VECTOR MACHINE

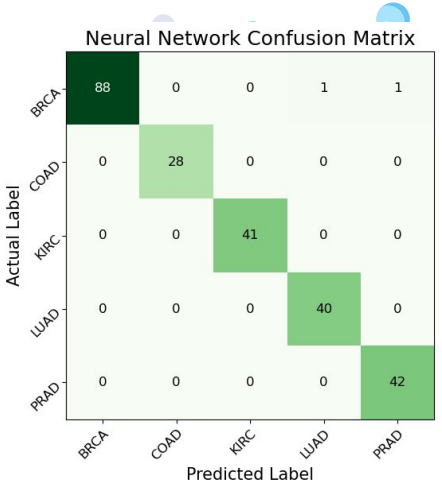


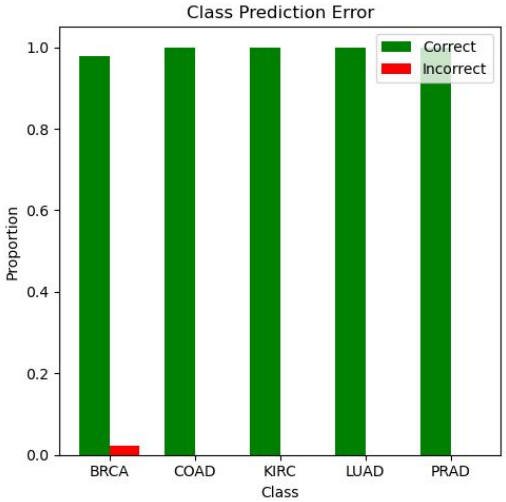


- SVM works by finding the hyperplane that best separates the classes with the maximum margin.
- The SVM classifier achieved a remarkable accuracy of **100%.**
- Along with accuracy, other metrics like precision, recall, and F1-score were used to evaluate the SVM classifier and got 100% scores.

NEURAL NETWORK

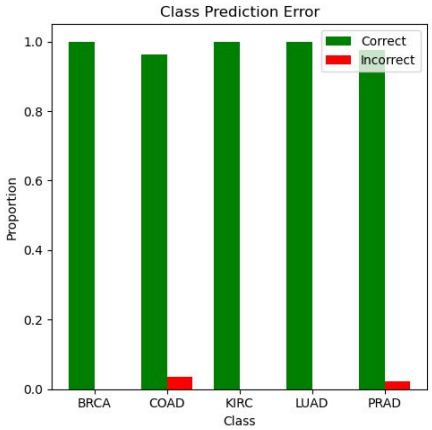
- Neural networks offer multiple advantages, one of which is their capacity to represent complex, nonlinear relationships found in data.
- Neural networks are a great tool for capturing details of the complex relationships between various genes and how they affect various types of cancer.
- The Neural Network model achieved an accuracy of **99.17%.**

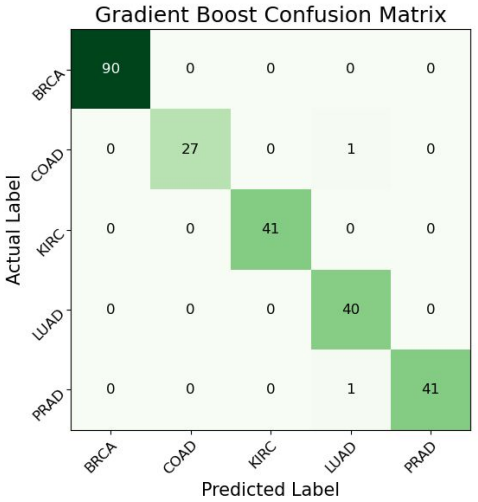




GRADIENT BOOST

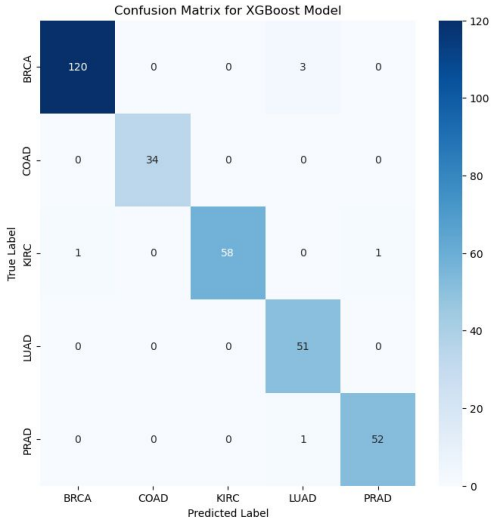
- Gradient Boost is renowned for its strong performance, especially in complex datasets.
- Gradient Boost builds the model in a stage-wise fashion, learning from the errors of the previous trees.
- The Gradient Boost classifier achieved a **99.17%** accuracy rate.

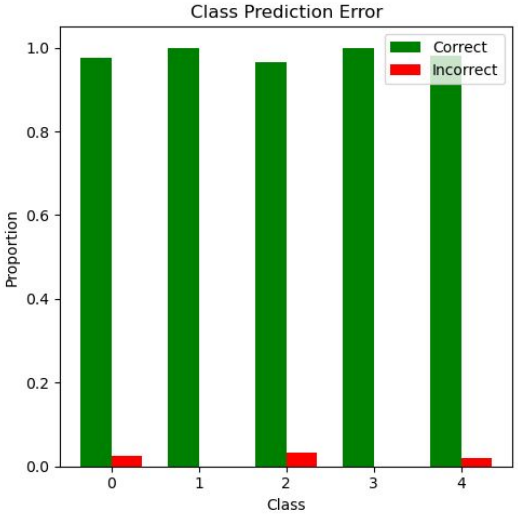




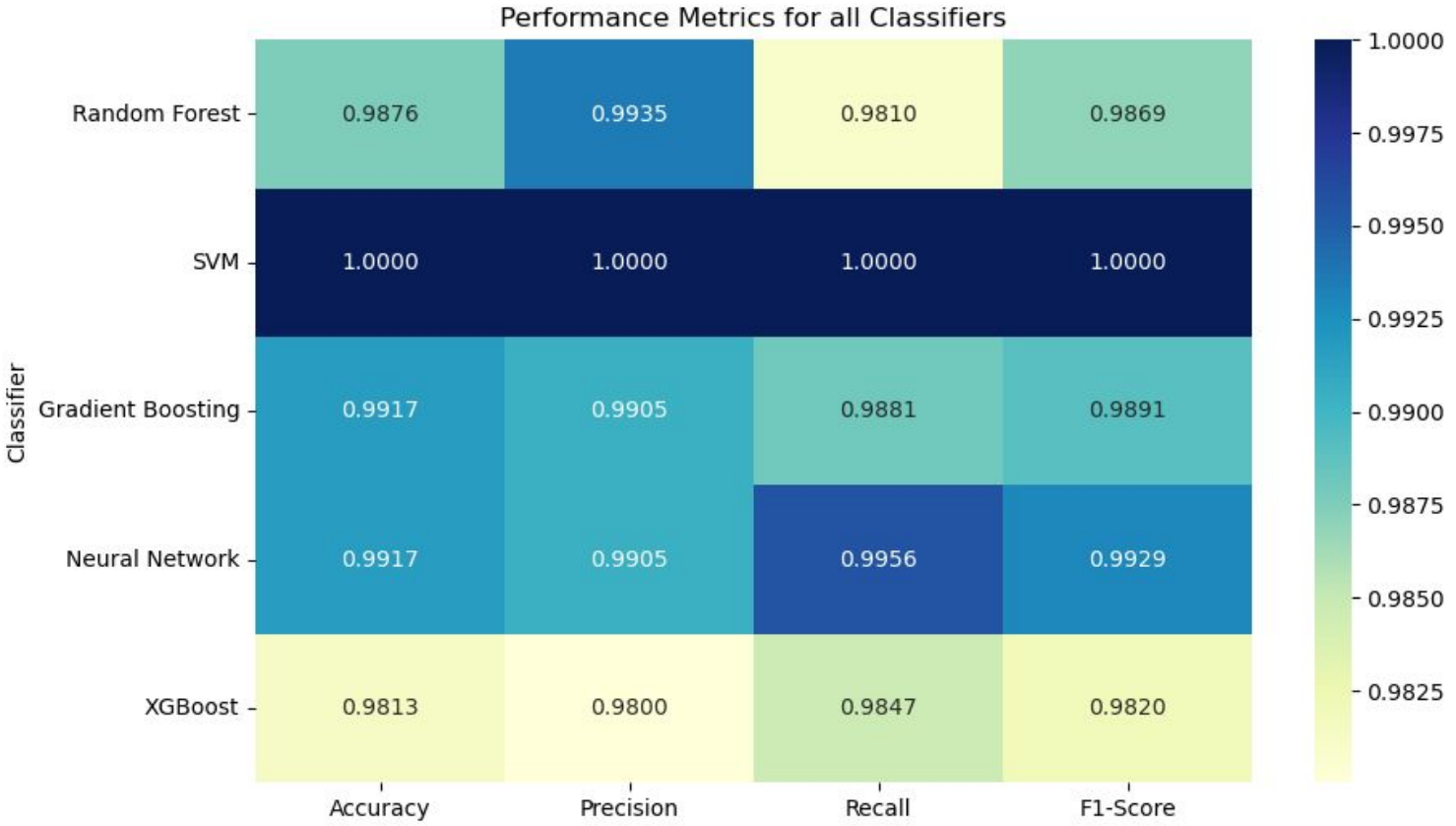
EXTREME GRADIENT BOOST

- XGBoost provides several hyperparameters, like learning rate and max depth, that can be finely tuned to optimize performance for the specific characteristics of the genomic data.
- XGBoost builds an ensemble of decision trees in a sequential manner.
- The XGBoost classifier achieved a high accuracy of **98.13%.**





COMPARISON


Performance Metrics for all Classifiers
Classifier
Random Forest
SVM
Gradient Boosting
Neural Network
XGBoost
0.9876
0.9935
0.9810
0.9869
1.0000
1.0000
1.0000
1.0000
0.9917
0.9905
0.9881
0.9891
0.9917
0.9905
0.9956
0.9929
0.9813
0.9800
0.9847
0.9820
Accuracy
Precision
Recall
F1-Score
1.0000
0.9975
0.9950
0.9925
0.9900
0.9875
0.9850
0.9825

CONCLUSION

- This research leveraged advanced machine learning algorithms, significantly contributing to the field of cancer genomics.
- Employed a variety of classifiers (Random Forest, SVM, Neural Networks, Gradient Boost, XGBoost) to navigate the complex genomic patterns associated with different cancer types.
- The classifiers demonstrated exceptional accuracy and reliability, particularly notable in the SVM models, indicating a significant step forward in precise cancer classification.

Thank You...