Lab Report

for

CSE 4754: Bioinformatics Lab

on

Tumor vs Normal Sample Classification Using Machine Learning Models

Group - 04

Tasnimul Hasnat 19004113

Nur Nasrum 190041106

Khaja Abdus Sami 190041136

This report outlines the methodology and findings of a study aimed at classifying tumor and normal samples from four types of The Cancer Genome Atlas (TCGA) projects: UCEC, KICH, LIHC, and ESCA. The objective was to classify between normal samples and four cancer classes using various machine learning models.

Data Collection

The 4 datasets contain genetic expression for 4 different cancer types in cells, sourced from TCGA projects. **UCEC** or **Uterine Corpus Endometrial Carcinoma** focuses on uterine corpus endometrial carcinoma, a type of cancer that originates in the lining of the uterus. It provides comprehensive genomic information, including DNA sequencing data, gene expression profiles, and clinical data. **KICH** or **Kidney Chromophobe** pertains to kidney chromophobe, a subtype of renal cell carcinoma. This dataset offers a wealth of genomic and clinical data for researchers studying kidney cancer. **LIHC** or **Liver Hepatocellular Carcinoma** focuses on hepatocellular carcinoma, the most common type of primary liver cancer. This dataset contains genomic information, including DNA sequencing data and gene expression profiles, along with clinical details from individuals diagnosed with liver hepatocellular carcinoma. **ESCA** or **Esophageal Carcinoma** is dedicated to esophageal carcinoma, a type of cancer that affects the esophagus. This dataset encompasses genomic and clinical data, offering insights into the molecular landscape of esophageal cancer.

Data Separation

With the help of the R programming language, normal samples and tumorous samples were separated for all four cancer types. The tumorous samples were then identified as their own classes and all the other normal samples were combined together in one single class. The data and their corresponding labels for each sample were then saved in two separate csv files called **data.csv** and **labels.csv.** With this we conclude our preparation of the dataset on which we will try different classification techniques.

Classification Models

Now we move on to Google collab to try out different machine learning and deep learning models to classify the cancer types. Overall we tried six different methods of classification on the dataset. They are,

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. K-Nearest Neighbors (KNN)
5. Gradient Boosting
6. Support Vector Machine (SVM)

Model Evaluation Metrics

The above models are evaluated based on the following metrics:

- **Accuracy:** The ratio of correctly predicted observations to the total observations.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positive observations.
- **Recall (Sensitivity):** The ratio of correctly predicted positive observations to all observations in the actual class.
- **F1 Score:** The harmonic mean of **Precision** and **Recall**.
- **Weighted and Macro Average Scores:** These provide an overall measure of the model's performance, taking into account the balance of classes in the dataset.

Comparative Analysis

The performance of each model was compared using the aforementioned metrics. The comparison chart presents a comprehensive view of how each model fares against the others in terms of accuracy, precision, recall, F1 score, and weighted/macro averages.

From all the six different classification techniques, Gradient Boosting gave the best result of them all with an accuracy of **98.46%** higher than all other models.

Classification performance

	Accuracy	Macro Precision	Weighted Precision	Macro Recall	Weighted Recall	Macro F1-Score	Weighted F1-Score
Logistic Regression	98.08%	0.9582	0.9820	0.9812	0.9808	0.9690	0.9809
Decision Tree	95.38%	0.9271	0.9541	0.9420	0.9538	0.9341	0.9537
Random Forest	98.08%	0.9595	0.9814	0.9759	0.9808	0.9672	0.9808
KNN	95.77%	0.9274	0.9643	0.9659	0.9577	0.9432	0.9588
Gradient Boosting	98.46%	0.961	0.9851	0.9786	0.9846	0.9729	0.9846
SVM	98.08%	0.9582	0.9820	0.9812	0.9808	0.9690	0.9809

Conclusion

This study demonstrates the effectiveness of various machine learning models in classifying tumor and normal samples from TCGA datasets. The comparative analysis provides insights into which models are more suited for this kind of classification task, based on accuracy, precision, recall, and other relevant metrics of which we found Gradient Boosting gave the best result for this particular classification task.

—XXX—