Project Report

Gene expression cancer ICMR

Name: Prajyot Kore
Roll No: 22N0044
Mail: 22N0044@iitb.ac.in

Source:

UCI Link: gene expression cancer RNA-Seq - UCI Machine Learning Repository

Kaggle Link: https://www.kaggle.com/datasets/shibumohapatra/icmr-data/c

1.1 DESCRIPTION

ICMR wants to analyze different types of cancers, such as breast cancer, renal cancer, colon cancer, lung cancer, and prostate cancer becoming a cause of worry in recent years. They would like to identify the probable cause of these cancers in terms of genes responsible for each cancer type. This would lead us to early identification of each type of cancer reducing the fatality rate.

1.2 Dataset Details:

The input dataset contains 802 samples for the corresponding 802 people who have been detected with different types of cancer. Each sample contains expression values of more than 20K genes. Samples have one of the types of tumours: BRCA, KIRC, COAD, LUAD, and PRAD.

The dataset you described appears to be related to cancer classification and gene expression analysis. Here's an extended description of the tumors associated with the dataset:

1. BRCA (Breast Cancer): Breast cancer is one of the most common cancers among women. It originates in the breast tissue and can occur in both men and women, although it is more prevalent in women. The dataset likely includes samples from individuals with breast cancer, and the analysis aims to identify the genes associated with this type of cancer.

2. KIRC (Renal Cancer): Renal cell carcinoma, or kidney cancer, occurs in the lining of small tubes in the kidney. It is one of the common forms of kidney cancer. The dataset may contain samples from individuals with renal cancer, and the analysis aims to uncover the genetic factors associated with this type of cancer.

3. COAD (Colon Cancer): Colorectal cancer includes colon cancer (affecting the large intestine) and rectal cancer (affecting the rectum). Colon cancer is a major cause of cancer-related deaths. The dataset could include samples from individuals with colon cancer, with a focus on identifying the genes responsible for this type of cancer.

4. LUAD (Lung Cancer - Adenocarcinoma): Lung adenocarcinoma is a subtype of non-small cell lung cancer. It originates in the cells lining the airways and is one of the most common types of lung cancer. The dataset may contain samples from individuals with lung adenocarcinoma, and the analysis aims to uncover the genetic factors specific to this form of lung cancer.

5. PRAD (Prostate Cancer): Prostate cancer occurs in the prostate, a small gland in men that produces seminal fluid. It is one of the most common cancers in men. The dataset could include samples from individuals with prostate cancer, focusing on identifying the genes associated with this type of cancer.

2. Objective:

The primary objective of this analysis is to unravel the genetic foundation underlying various types of cancers, including breast cancer (BRCA), renal cancer (KIRC), colon cancer (COAD), lung adenocarcinoma (LUAD), and prostate cancer (PRAD). This exploration is conducted through the examination of gene expression data gathered from 802 individuals. The ultimate goal is to pinpoint specific genes or genetic signatures that can serve as indicators for each distinct cancer type. The insights derived from this analysis have the potential to revolutionize cancer diagnosis and treatment, leading to earlier identification and intervention, thereby mitigating the fatality rates and enhancing patient outcomes.

Dimensionality Reduction:

1. PCA, LDA, and t-SNE: The initial phase of our analysis centers on the dimensionality reduction of the dataset, which initially comprises an extensive 20,531 gene expression attributes. Recognizing that not all of these genes are imperative for the analysis of cancer types, we will employ dimensionality reduction techniques, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and t-distributed Stochastic Neighbor Embedding (t-SNE). These methods will help us distill the most influential attributes from the dataset, thus simplifying subsequent classification modeling.

- Input: The complete dataset with 20,531 genes.

- Output: The result of each dimensionality reduction method, which will be a reduced set of genes to be used in the subsequent modeling phase.

Building Classification Model(s) with Feature Selection:

1. Multiclass SVM, Random Forest, and Deep Neural Network: The second phase of our analysis is dedicated to constructing robust classification models for the precise identification of each cancer type. Simultaneously, we will carry out feature selection to identify the genes that play a pivotal role in classifying the distinct cancer types. The following sub-tasks are to be accomplished:

- Build Classification Models: We will create classification models, including Multiclass Support Vector Machines (SVM), Random Forest, and Deep Neural Networks. These models will be instrumental in classifying the input data into one of the five cancer types: BRCA, KIRC, COAD, LUAD, or PRAD.

- Feature Selection: Feature selection will be carried out to identify the genes that have the most discriminative power for classifying each cancer type. This will help streamline the model and reduce computational complexity.

The overarching objective is to empower the medical community with precise diagnostic tools and therapeutic insights by unearthing the genetic underpinnings of different cancer types, leading to improved patient outcomes and potentially reducing cancer-related fatalities.

Previous Work Done:

1. gp33-icmr | Kaggle
2. healthcare | Kaggle

3. Analysis:

The data we have, has 20,534 columns with 800 columns, as mentioned in the description. And distribution of classes is as given below:

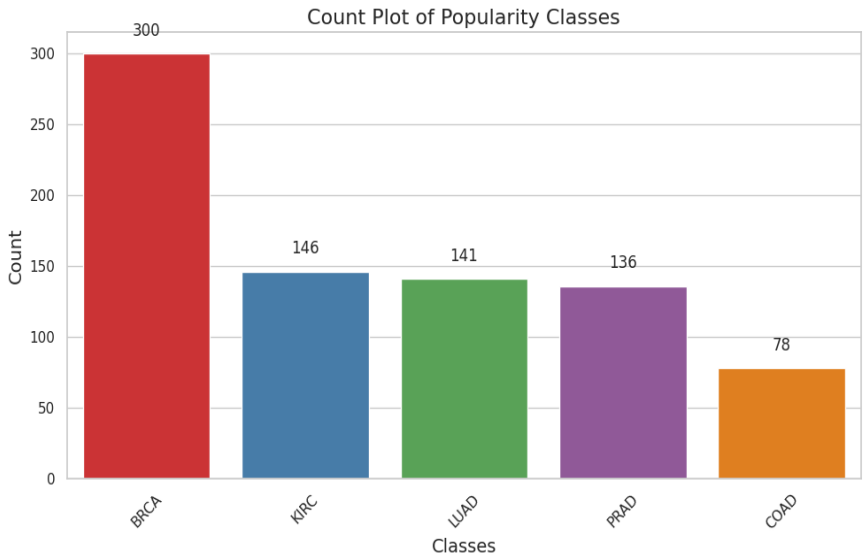


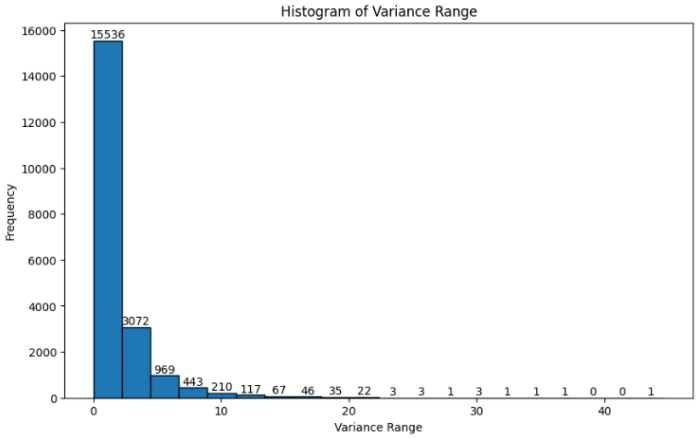
Fig. 1 Frequency of Classes

frequency plot, we can see that instances for class BRCA is more than that of other followed by KIRC, LUAD and PRAD while, COAD having least number of instances.

3.1 Pre-Processing:

Before going further, we did data preprocessing, which involves removing **Null**, **duplicate** Values: fortunately, there were no null values and duplicates.

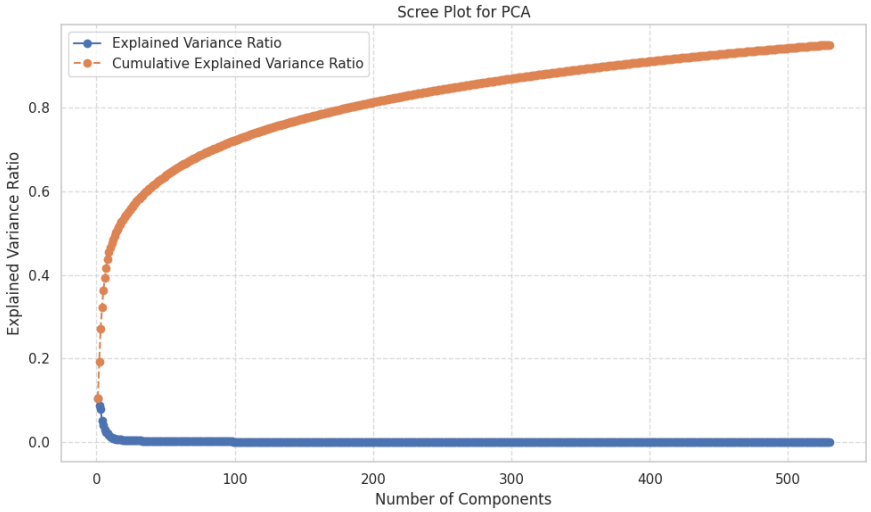
Variance Threshold:



We can see that class BRCA has highest count and COAD has lowest count and KIRC, LUAD and PRAD almost has same count, but this imbalance can lead to bias into models.From graph its easy to see that, most of columns has low variance. Features with high variance have data points that are spread out over a wide range. These features are considered significant because they contain valuable information and can help discriminate between different classes or categories. Similarly, Features with low variance have data points that are concentrated or have little variation. Such features may not carry much discriminative power and can be considered less significant.

3.2 Dimension Reduction Using PCA:

Having high dimensionality or feature space can lead to model to perform poorly known as “Curse-of-Dimensionality”, to deal with this, we will reduce the feature space by performing the dimension reduction techniques, like **Principal Component Analysis (PCA)**, and we will consider 95% variance explainability, as shown in figure,



Fig, 3 Scree Plot with Cummulative Explained variance Plot

From the figure, it's evident that nearly 550 Principal Components (PCs) out of the original 20,534 columns capture 95% of the variability, resulting in a significant reduction in dimension by over 97%.

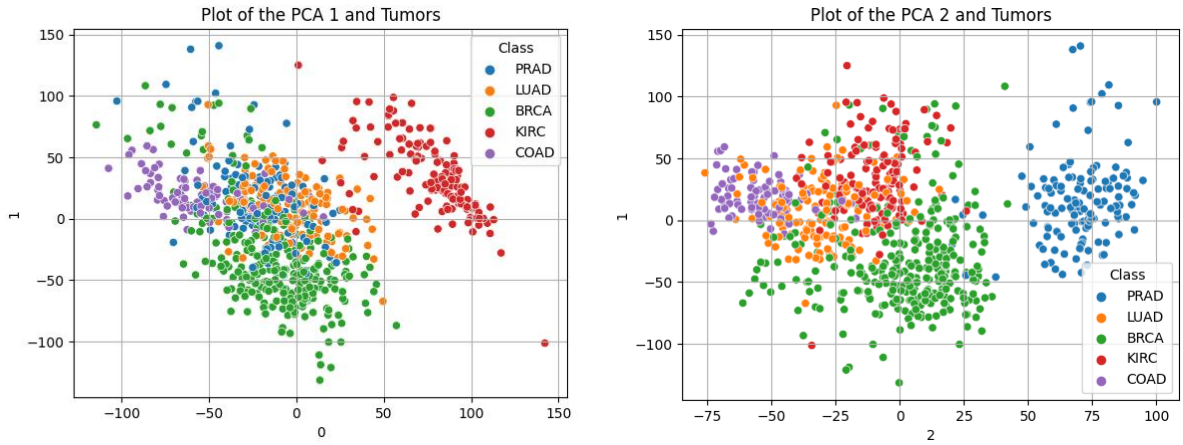


Fig. 4 Plot of Principal Components and Different Tumours for first 3 PCs

Remark:

- From the first scatter plot (PC1 vs. PC2):
- Class KIRC is notably more separable than other classes, indicating effective feature differentiation.
- PC1 and PC2 hold valuable information for distinguishing KIRC from other cancer types.
- From the second scatter plot (PC2 vs. PC3):
- Class PRAD exhibits enhanced separability, emphasizing the significance of PC2 and PC3.
- PC2 and PC3 capture distinctive features for characterizing PRAD, making it more distinguishable.

3 Model Fitting:

We first fit the models on PCA data then, will go for more complex methods like ensemble methods and Neural Network then we will compare the results of these different models

4 Evaluation Metrics:

$$\text{Precision} = \underset{\text{Micro-average}}{} \frac{TP_A + TP_B + \dots TP_N}{TP_A + FP_A + TP_B + FP_B + \dots TP_N + FP_N}$$

$$\text{Recall} = \underset{\text{Micro-average}}{} \frac{TP_A + TP_B + \dots TP_N}{TP_A + FN_A + TP_B + FN_B + \dots TP_N + FN_N}$$

- **Class Imbalance Awareness:**

The presence of class imbalance in the dataset highlights that there aren't enough representatives of minority classes. This imbalance can lead to skewed results and incorrect predictions.

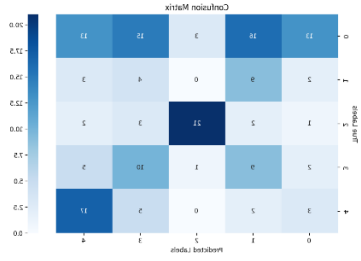
- **Deceptive Metrics:**
  Metrics such as overall accuracy and ROC-AUC can be deceptive in this context. They don't provide a clear picture of how well the model performs for minority classes due to the data bias towards majority classes.
- **Need for Better Metrics:**
  To address the problem of class imbalance and obtain more meaningful insights, it's essential to consider better metrics.
- **Accuracy of Each Class:**
  Evaluating the accuracy for each individual class allows us to understand how well the model performs for specific cancer types. This is particularly crucial when dealing with imbalanced data.
- **Micro Average F1 Score:**
  Micro-averaging is a valuable approach to calculate the F1 score. It aggregates true positives, false positives, and false negatives across all classes, giving equal weight to each instance, regardless of class labels and class sizes.
- **Macro vs. Micro Averaging:**
  Macro-averaging calculates performance metrics for each class and then takes the mean. It assigns equal weight to all classes, irrespective of their prevalence.
  Micro-averaging, on the other hand, considers the overall counts of true positives, false positives, and false negatives, giving equal weight to each instance, regardless of class labels.
- **More Promising Predictions:**
  Micro Average F1 Score, being sensitive to class imbalance, can lead to more reliable predictions and inferences. It provides a balanced assessment of model performance.

Pros and Cons of Micro Average Score:

- **Macro-averaging** calculates each class's performance metric (e.g., precision, recall) and then takes the arithmetic mean across all classes. So, the macro-average gives equal **weight to each class**, regardless of the number of instances.
- **Micro-averaging**, on the other hand, aggregates the counts of true positives, false positives, and false negatives across all classes and then calculates the performance metric based on the total counts. So, the micro-average gives equal **weight to each instance**, regardless of the class label and the number of cases in the class.
- So, rather than constraining to the one particular metrics we will use two **Accuracy of Each Class** and **Micro Average F1 Score**.

Now, we will fit different models and compare the results:

5.1 Discriminant Analysis:

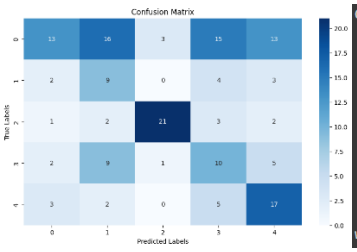


Classification Report:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	60
1	0.95	1.00	0.97	18
2	1.00	0.97	0.98	29
3	1.00	0.93	0.96	27
4	1.00	1.00	1.00	27
accuracy			0.98	161
macro avg	0.98	0.98	0.98	161
weighted avg	0.98	0.98	0.98	161

Class-wise Accuracies:
Class 0: 1.00
Class 1: 1.00
Class 2: 0.97
Class 3: 0.93
Class 4: 1.00

Fig.6 Classification Report and Accuracy of each class for LDA



Classification Report:

	precision	recall	f1-score	support
0	0.62	0.22	0.32	60
1	0.24	0.50	0.32	18
2	0.84	0.72	0.78	29
3	0.27	0.37	0.31	27
4	0.42	0.63	0.51	27
accuracy			0.43	161
macro avg	0.48	0.49	0.45	161
weighted avg	0.53	0.43	0.43	161

Class-wise Accuracies:
Class 0: 0.22
Class 1: 0.50
Class 2: 0.72
Class 3: 0.37
Class 4: 0.63

Fig.7 Classification Report and Accuracy of each class for QDA

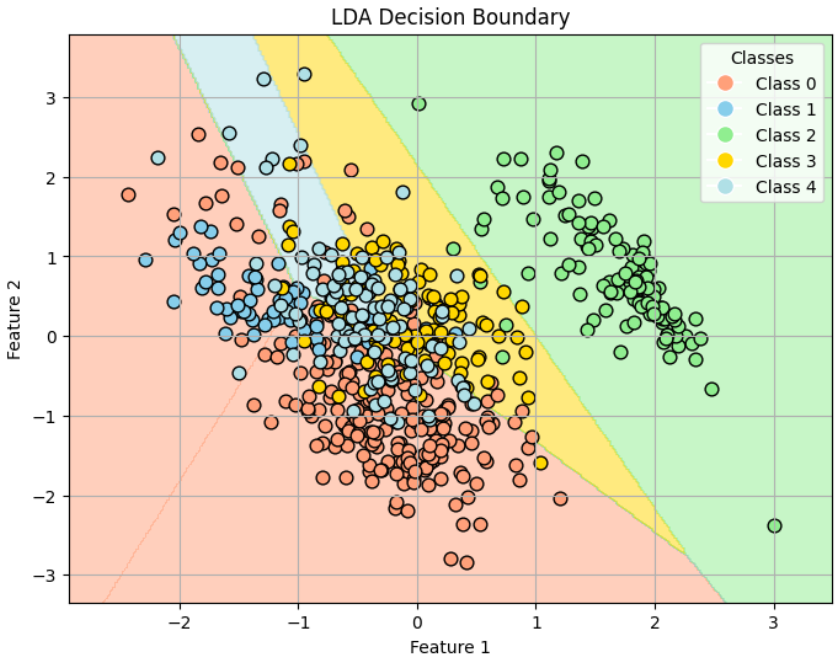


Fig. 8 Decision Boundary for PCA 1 and PCA 2 for LDA

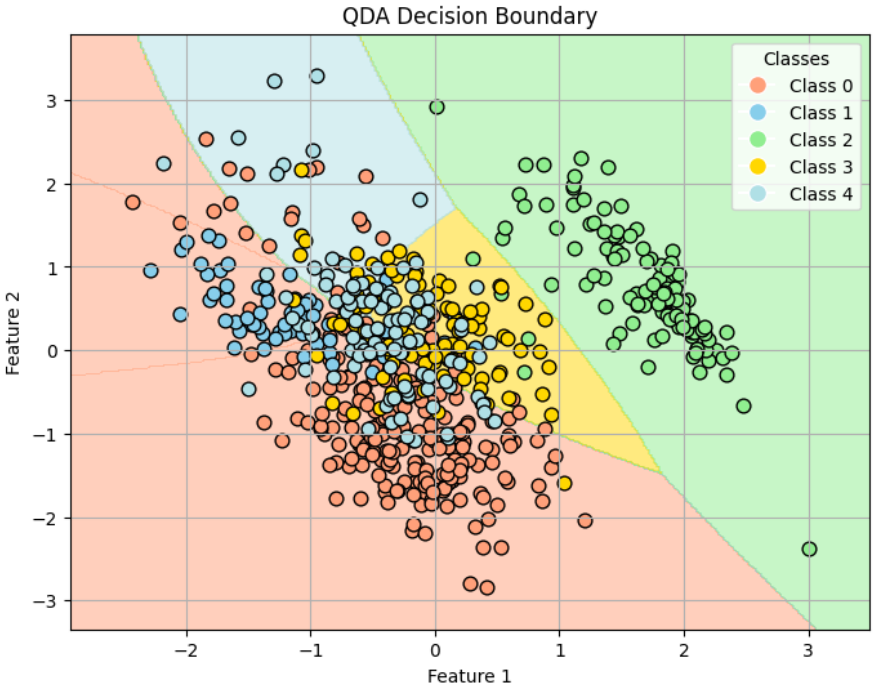


Fig. 9 Decision Boundary for PCA 1 and PCA 2 for QDA

Remark:

- LDA Demonstrated Higher Accuracy: consistently achieved better accuracy in classifying the cancer types, indicating its proficiency in capturing underlying data patterns.
- Simpler Decision Boundaries: LDA's linear boundaries were more interpretable and effectively separated cancer types, contributing to its superior performance.
- Reduced Risk of Overfitting: LDA's linear nature reduced model complexity, enhancing generalization and mitigating overfitting risks.
- Assumptions Matched the Data: LDA's assumption of a common covariance matrix for classes aligned well with our dataset, favoring its performance over QDA.
- Less Susceptible to Multicollinearity: LDA's robustness to multicollinearity issues provided stable and reliable results.
- Advantage in High-Dimensional Data: LDA excelled in high-dimensional scenarios, making it a suitable choice for datasets with a large number of genes.
- Consistent Results: Across multiple evaluations, LDA consistently outperformed QDA in terms of accuracy and precision, demonstrating its reliability.
- Interpretability: LDA's linear decision boundaries offered insights into significant genes and features, enhancing interpretability.

Next, we will explain the Top Performer:

5.2 Tree using Decision Tree with max depth 3:

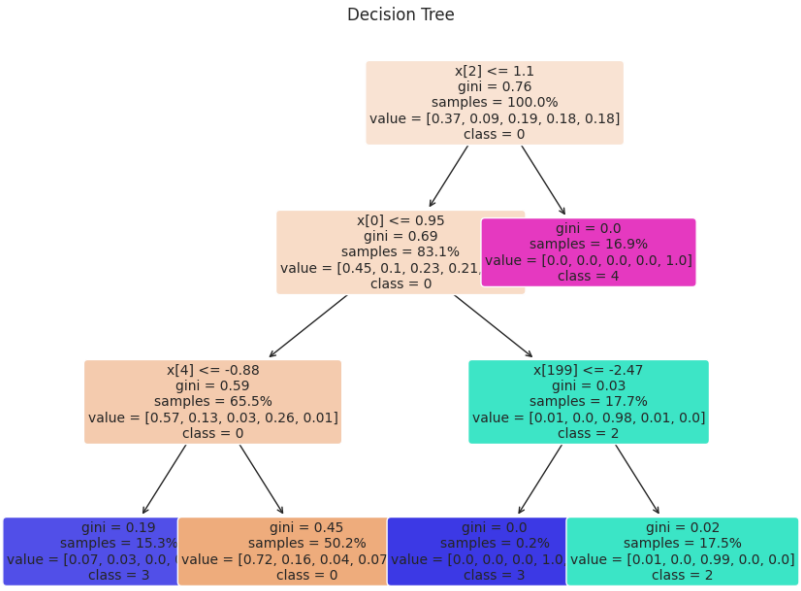


Fig. 7 Tree Plot from Decision Tree of Max depth 3

Decision Tree are famous for interpretability and simplicity.

From the graph top nodes, closer to the root, represent the most influential features for classification, while the leaf nodes represent the final decisions, indicating that 3rh PC being important one, followed by 1st and 4 and 199th PC, and it does make sense as 1st PC is responsible for higher variability explanation than others.

5.3 Confusion Matrix and Classification report for xgboost and Random Forest:

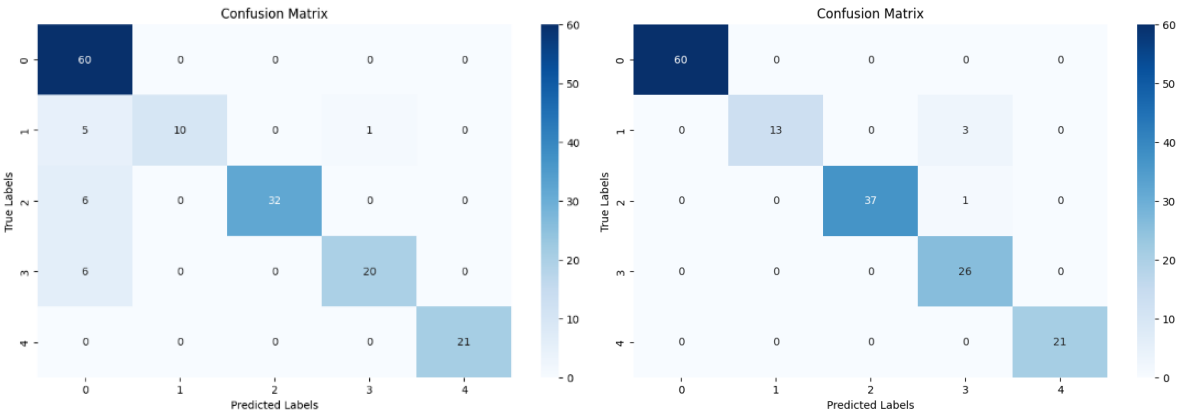


Fig. 6 Confusion Matrices for Random Forest(left) and Xgboost(right)

Classification Report:

	precision	recall	f1-score	support
0	0.78	1.00	0.88	60
1	1.00	0.62	0.77	16
2	1.00	0.84	0.91	38
3	0.95	0.77	0.85	26
4	1.00	1.00	1.00	21
accuracy			0.89	161
macro avg	0.95	0.85	0.88	161
weighted avg	0.91	0.89	0.89	161

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	60
1	1.00	0.81	0.90	16
2	1.00	0.97	0.99	38
3	0.87	1.00	0.93	26
4	1.00	1.00	1.00	21
accuracy			0.98	161
macro avg	0.97	0.96	0.96	161
weighted avg	0.98	0.98	0.98	161

Fig. 7 Classification Reports for Random Forest(left) and Xgboost(right)

From the Confusion Matrix and Classification Report we can see that Random Forest is not performing well for cancer COAD and KIRC whereas Xgboost is performing well on each of classes. It is as expected as class COAD followed by KIRC has less instances compared to other.

5.4 Classification Report and Confusion Matrix for Bagging Classifiers:

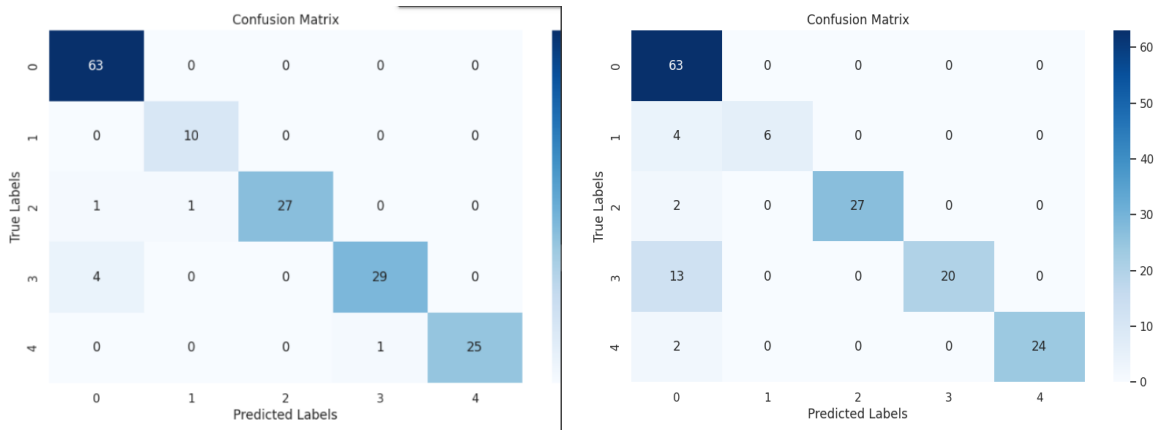


Fig. 9, Confusion Matrix for Decision Tree (left) and Random Forest (Right)

Remark:

- From Confusion Matrix we can see that for class 3 which is COAD, we can see some of the instances of this being falsely predicted as BRC.
- probably due to the class Imbalance, as COAD has less instances and BRCA has more instances that any other classes

Classification Report:

	precision	recall	f1-score	support
BRCA	0.93	1.00	0.96	63
COAD	0.91	1.00	0.95	10
KIRC	1.00	0.93	0.96	29
LUAD	0.97	0.88	0.92	33
PRAD	1.00	0.96	0.98	26
accuracy			0.96	161
macro avg	0.96	0.95	0.96	161
weighted avg	0.96	0.96	0.96	161

Classification Report:

	precision	recall	f1-score	support
BRCA	0.75	1.00	0.86	63
COAD	1.00	0.60	0.75	10
KIRC	1.00	0.93	0.96	29
LUAD	1.00	0.61	0.75	33
PRAD	1.00	0.92	0.96	26
accuracy			0.87	161
macro avg	0.95	0.81	0.86	161
weighted avg	0.90	0.87	0.87	161

Fig. 10, **Classification Report for Decision Tree (left) and Random Forest (Right)**

5.5 AdaBoost:

For adaboost we used tree based models with different depth 1,2 and 3 respectively, and accuracy of each class is as given below:

- Depth-1 Decision Stumps: The first AdaBoost model employed depth-1 decision stumps. These are extremely shallow decision trees with a single split, which means they make simple, one-level decisions.
- Depth-2 Decision Trees: The second AdaBoost model utilized depth-2 decision trees. These trees can make more complex decisions with up to two splits.
- Depth-3 Decision Trees: The third AdaBoost model featured depth-3 decision trees. These trees are capable of making more intricate decisions with up to three splits, offering greater complexity.

5.6.1 Neural Network:

We tried out ANN with some dropout layers and Neural Network tends to overfit the data, model summary is as given below:

- The initial layer includes 64 neurons with the ReLU activation function. ReLU, or Rectified Linear Unit, is a common choice for deep learning models, known for its ability to handle complex patterns in the data.
- Dropout layers were incorporated after each hidden layer to prevent overfitting. A dropout rate of 0.5 was applied, which means that half of the neurons were randomly deactivated during training, promoting robustness.
- The final layer contains a number of neurons equal to the classes in the dataset, employing the sigmoid activation function, suitable for multi-class classification problems.

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_10 (Dense)            (None, 64)                33984

 dropout_8 (Dropout)         (None, 64)                0

 dense_11 (Dense)            (None, 64)                4160

 dropout_9 (Dropout)         (None, 64)                0

 dense_12 (Dense)            (None, 64)                4160

 dropout_10 (Dropout)        (None, 64)                0

 dense_13 (Dense)            (None, 64)                4160

 dropout_11 (Dropout)        (None, 64)                0

 dense_14 (Dense)            (None, 5)                 325

=================================================================
Total params: 46789 (182.77 KB)
Trainable params: 46789 (182.77 KB)
Non-trainable params: 0 (0.00 Byte)
```

Fig. 11 Model Summary for NN

Model Compilation and Training:

- The model was compiled using the RMSprop optimizer and the categorical cross-entropy loss function, which is appropriate for multi-class classification tasks.
- Training was performed with a batch size of 10 and over 40 epochs, with the validation dataset used to monitor model performance during training.

5.6.2 Model Performance Summary:

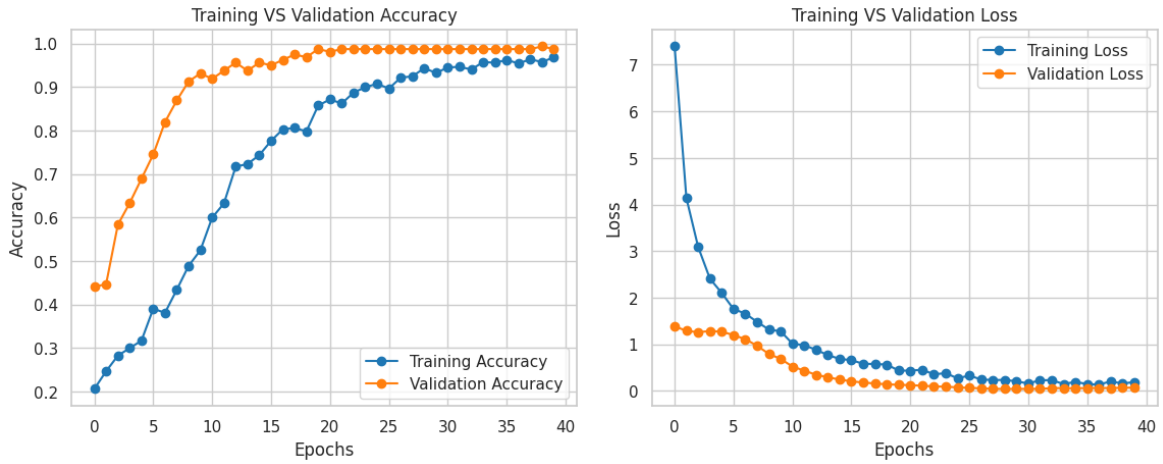


Fig. 12 Training and Validation Accuracy and Loss using Neural Network

Because of small test data, instead of dividing the data into three sets like training, validation and testing set, I divided data into training and Validation data and corresponding loss and accuracy is as given in the above plots, and it can seen that from plots that as epochs increasing accuracy getting better while loss getting smaller, until epoch 20, after this epochs validation accuracy and loss becomes stagnant and there is not significant reduction in validation loss and elevation in validation accuracy.

Performance of model:

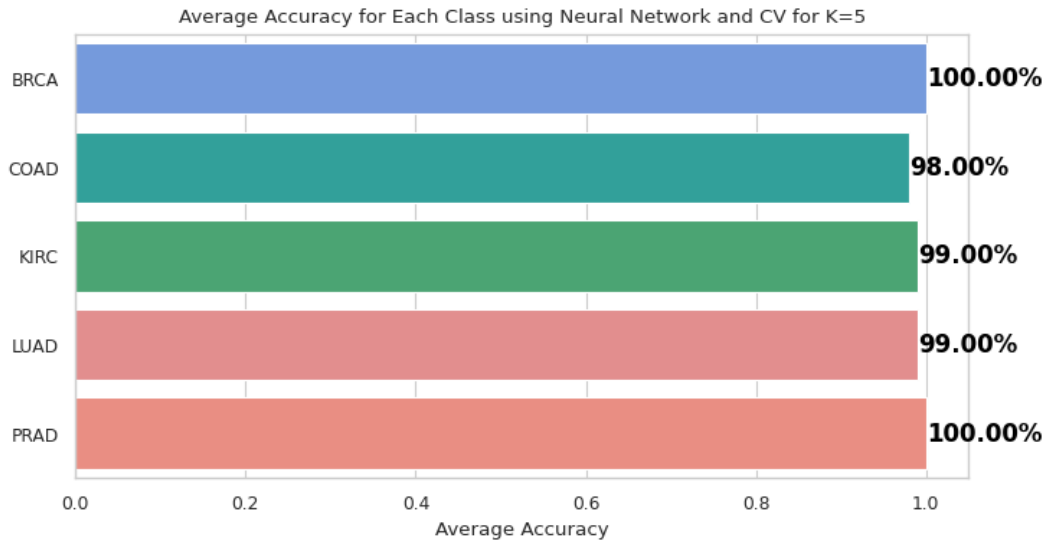
Fig 13. Confusion Matrix, Classification Report and Class wise Accuracy using NN

Remarks:

- Neural Network's Strong Performance: The neural network model demonstrates robust and reliable performance with just two instances misclassified, highlighting its effectiveness in classifying cancer types.
- Expected Misclassifications: Minimal misclassifications are common in real-world scenarios, and encountering only two misclassified instances among 802 samples is well within the expected range.
- Model Reliability and Generalization: The neural network's low misclassification rate underlines its reliability and generalization capabilities, making it a dependable tool for cancer type identification, instilling confidence among healthcare professionals and patients.

5.6.3 K-Fold Cross Validation for K = 5:

- Model Architecture: The neural network architecture used in this study is a Sequential model. It starts with an Input layer, specifying the input shape as a tuple, which adapts to the dataset's feature space.
- Hidden Layers with Dropout: Two dense hidden layers follow the input layer. The first dense layer consists of 128 neurons with a Rectified Linear Unit (ReLU) activation function, which helps in capturing complex patterns in the data. Between these dense layers, dropout layers with a dropout rate of 0.3 are introduced to reduce overfitting. These layers randomly deactivate a portion of neurons during training, encouraging the model to generalize better.
- Output Layer for Multiclass Classification: The final dense layer consists of the number of neurons equal to the number of classes (num_classes). It uses the sigmoid activation function, suitable for multiclass classification tasks. This output layer provides the class probabilities for each cancer type.



Fig, 14 Accuracy of Each Class of K- Cross Validation with k= 5 using Shallow Neural Network

It is quite impressive that k- Fold Cross Validation gave us better result even though Neural Network used for cross validation is shallower than originally employed Neural Network and these results are more reliable.

6. Comparison of Model Performance:

6.1 Models:

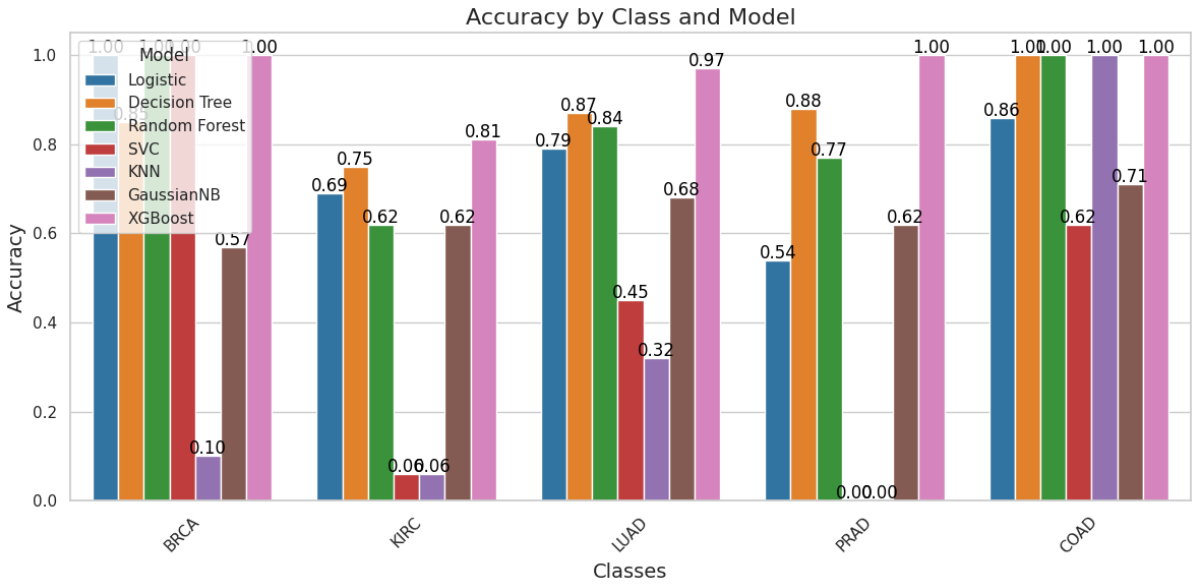
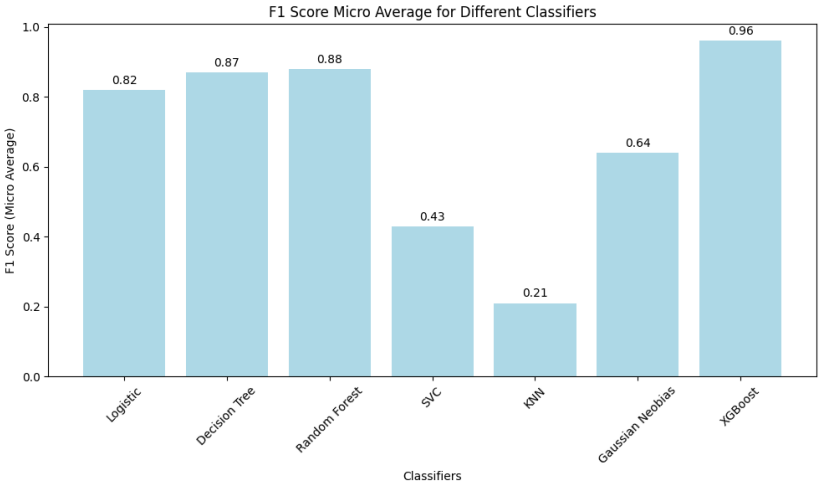


Fig. 4 Accuracy of Each class for different models ranging from classical models

Remark:

- COAD Performance: Across all models, the classification performance for colon cancer (COAD) is notably strong, highlighting well-captured distinguishing features.
- PRAD Classification Challenge: Models like KNN and SVC (linear) exhibit weaker performance in prostate cancer (PRAD) classification. This suggests the need for fine-tuned hyperparameters and different kernel methods, especially for the Support Vector Machine (SVM) classifier.
- Differential Performance for KIRC and LUAD: Kidney renal clear cell carcinoma (KIRC) and lung adenocarcinoma (LUAD) display varying degrees of performance, with KIRC showing relatively better performance, indicating more distinctive features.
- Balanced BRCA Performance: Breast cancer (BRCA) demonstrates balanced performance across models, without significant outliers, showcasing effective feature capture.
- Ensembled Models Shine: Ensembled models, such as XGBoost and Random Forest, consistently outperform individual classifiers, highlighting the strength of ensemble methods in handling diverse cancer types.



From fig 5 its easy to see that, F1 Score Micro Avg is highest for xgboost followed by Random Forest, Decision Tree, while for KNN and SVC its lowest implies model performed poorly.

Fig. 5 F1 Score Micro Average for Different Classifier

6.2 Results from Bagging Classifier with different Base Estimator/ weak Learner:

Here, we used 10 number of estimators.

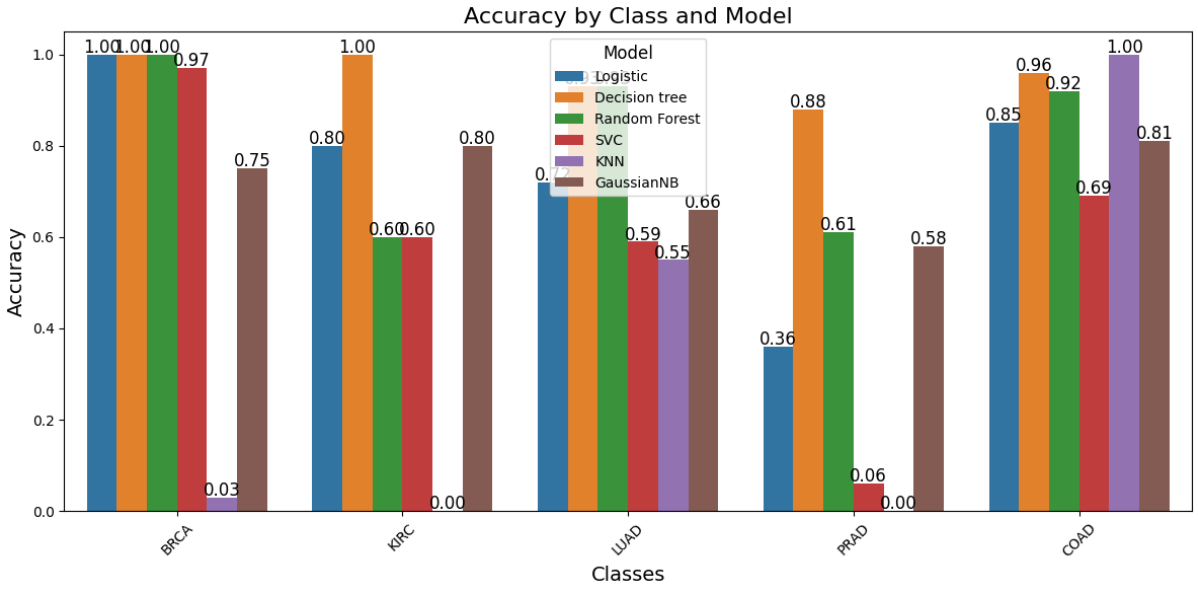


Fig. 8, Accuracy of each Class for different models

From plot it's clear that for class BRCA all model Bagging classifier is performing well except, KNN its obvious as we do have more observations for this class, also, its interesting that Decision tree as weak learner performing better than other.

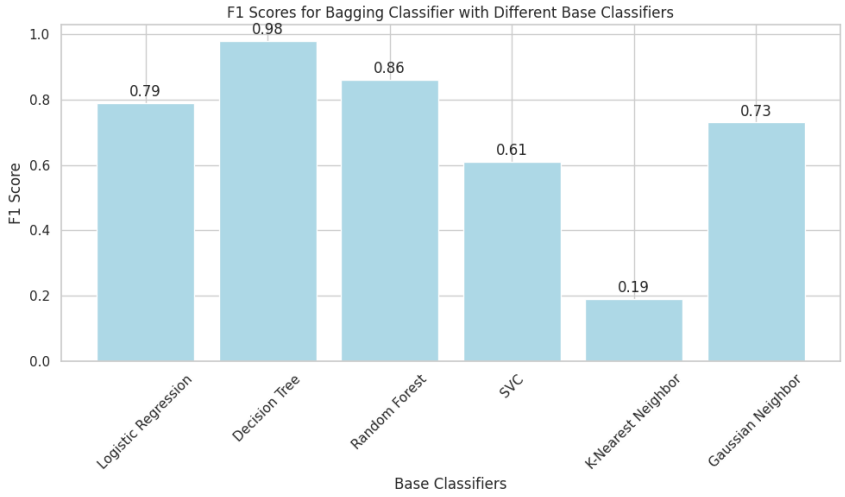


Fig. 8, Micro Score Average F1 of Bagging Classifier using different Weak Learner

From this plot it can see that Decision Tree has highest F1 score micro average, followed by Random Forest and logistic regression,

6.3 AdaBoosts:

For adaboost we used tree based models with different depth 1,2 and 3 respectively, and accuracy of each class is as given below:

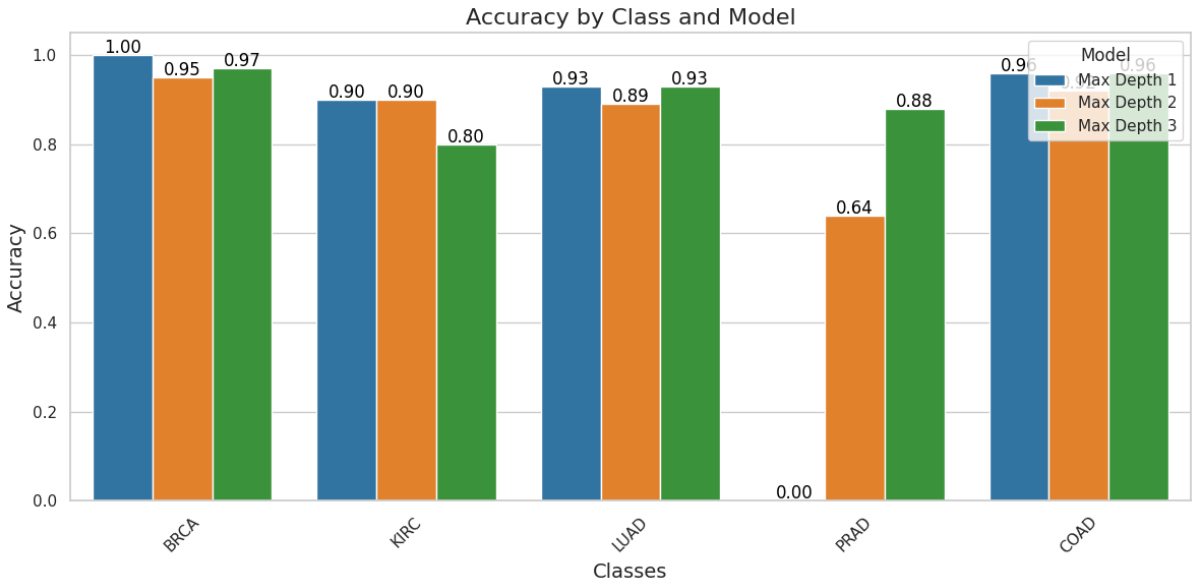
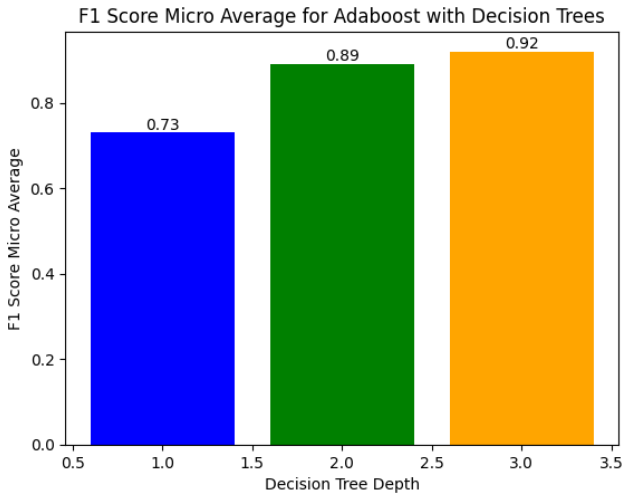


Fig. 12, Accuracy of each class for AdaBoost with weak learner being Decision Tree for max depth 1,2 and 3

From accuracy plot we see that weak learner decision tree with max depth 3 tend to perform better on each of class followed by decision tree with max depth 2.



From given graph its easy to that AdaBoost with Decision Tree with max depth 3 is performing better compared to decision tree with max depth 2 and decision tree with max depth 1.

Fig. 13, F1 Score Micro Average of each class for AdaBoost with weak learner being Decision Tree for max depth 1,2 and 3

Possible Reason why Decision Tree with Max depth 3 as weak learner working better:

- Model Complexity: Decision trees with different maximum depths represent varying degrees of model complexity. A decision tree with depth 1 is a simple model and can underfit the data by not capturing complex patterns. A decision tree with depth 3, on the other hand, can capture more complex patterns in the data.
- Adaptive Boosting: AdaBoost works by giving more weight to misclassified samples in each iteration. It adapts to the "hard" examples in the dataset. If the decision tree with max depth 3 is able to correct more misclassifications from the previous weak learners (which might have been too simplistic), it can improve the overall ensemble's performance.
- Combining Weak Classifiers: AdaBoost combines multiple weak classifiers to create a strong ensemble. If the decision tree with max depth 3 is a better weak learner compared to depth 2 and 1, it can lead to a more accurate ensemble.
- Data Complexity: The dataset itself plays a crucial role. If the dataset contains complex patterns that require a more complex model to capture, a decision tree with depth 3 can be more effective.

7. Conclusion:

- The analysis aimed to uncover the genetic underpinnings of various cancer types, including breast cancer, renal cancer, colon cancer, lung adenocarcinoma, and prostate cancer. It utilized gene expression data from 802 individuals, and the primary goal was to identify specific genes or genetic signatures as indicators for each distinct cancer type.

- The analysis began with dimensionality reduction techniques, including Principal Component Analysis (PCA). This techniques helped distill the most influential attributes from the initial dataset of 20,531 gene expression attributes.
- Classification models were constructed, including Multiclass Support Vector Machines (SVM), Random Forest, Deep Neural Networks, and ensemble methods like AdaBoost, Bagging, and XGBoost. Feature selection was performed to identify the most discriminative genes for classifying each cancer type.
- Evaluation metrics included class imbalance awareness, accuracy of each class, and Micro Average F1 Score. It was noted that deceptive metrics like overall accuracy and ROC-AUC could be misleading in the presence of class imbalance.
- Neural networks were employed for the classification task, and their performance was impressive, with minimal misclassifications.
- K-Fold Cross Validation (K=5) further improved the performance and reliability of the neural network model.
- The analysis compared various models, demonstrating that Bagging classifiers and ensemble methods like XGBoost outperformed individual classifiers. AdaBoost with Decision Trees as weak learners exhibited strong performance, especially when the decision trees had a depth of 3.

8. Scope for Improvement:

- Address Class Imbalance: While the performance was strong for most cancer types, it's important to address the class imbalance, particularly for COAD and KIRC. Techniques like oversampling, undersampling, or synthetic data generation can help mitigate this issue.
- Hyperparameter Tuning: Further fine-tuning of hyperparameters for classifiers with weaker performance, such as KNN and SVC, could potentially improve their accuracy for prostate cancer (PRAD).
- Interpretability: Consider investigating the most influential genes or features for each cancer type to gain biological insights. Tools like SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) can aid in feature interpretability.
- Data Augmentation: If feasible, increasing the dataset size or gathering additional samples can lead to more robust and generalizable models.
- Exploration of More Complex Models: While the shallow neural network exhibited strong performance, it might be worthwhile to explore more complex architectures, such as deep learning models with convolutional layers and recurrent layers for gene expression analysis.
- Ensembling: Combining multiple models through ensembling methods like stacking or blending can potentially further improve performance and model robustness.