Cancer Genomic Insights Platform

Akshay Sharma, Anmol Sidhu, Yogeshwar Shendye

Dalhousie University

Ak328835@dal.ca, an862548@dal.ca, yg755595@dal.ca

Abstract

Artificial intelligence and machine learning have proved to be very crucial tools in cancer studies for many various reasons, and one of them is extracting predictive indicators and information from these datasets which at times could have different modalities such as medical images, clinical records, genomic data, etc. While extracting patterns and information with the help of machine learning and artificial intelligence from images and clinical data has proven to be very effective, researchers have faced numerous challenges while applying those to genomic datasets [1]. Therefore, an explainable framework with the help of different visualizations aids not only the researchers in coming up with predictive frameworks to extract inference from the genomic data but also the clinical practitioners in deciding the level of transparency they want from the developed predictive models, thus making the whole process transparent, trustworthy and user-friendly. The proposed idea also paves the way for a collaborative environment between clinical researchers and practitioners.

1 Introduction:

The complexities of cancer biology demand advanced computational approaches for effective analysis and inferences. Modern machine learning and artificial intelligence which includes deep learning algorithms such as Deep Neural networks or transformers architecture have been very successful in extracting invaluable information from healthcare data which includes medical imaging, healthcare records data, etc. However, even after advancements showcased by teams such as AlphaFold [7] in protein folding, research work in domains such as Cancer Genomics, Immunogenic, and Single-cell genomics remains very challenging for researchers. With our proposed visual analytics system, we aim to address the problem while keeping explainable AI at the forefront which helps to instill trust in these systems. The dataset used for developing the framework has been compiled by the Indian Council of Medical Research and contains 802 samples with 20k features and 5 target cancer types.

The analysis carried out with the developed framework aims to solve the basic problems faced when making a predictive model for genomic data. It allows both the end-user and the model developer to identify which features are important for predicting a certain type of cancer. It also shows if these features have discriminative information for differentiating different types of cancer. These functionalities have been developed while keeping in mind the clinical relevance, for instance, different parts of sequences could be inspected with the developed system to check for their predictive and differentiating power for different cancers. A very prominent example of this is researchers using spike protein sub-sections instead of using full genome SARS-CoV-2 sequences [2]. The next part of our solution is aligned more with the needs of model developers, as it gives them the option to inspect the performance of different predictive algorithms, while also giving the option to check the impact of different data preprocessing/feature reduction techniques on the model performance. The model developers are also provided with the option of checking the effect of changing hyper-parameter values on the model performance, to get an intuitive understanding of the optimal hyper-parameter search process. Therefore, the framework provides a unique approach to address the needs of both the model developers and the users at the same time by providing a comprehensive and intuitive understanding of the full process of predictive model development.

2 Methods

2.1 Feature Importance & Dimensionality Reduction.

The framework is being developed using a dataset that contains 800 samples, each with 20,000 features. Given the high-dimensional nature of this data and the number of samples, it’s crucial to apply techniques for dimensionality reduction and feature importance. This is particularly important when the framework primarily uses traditional machine learning algorithms like Logistic Regression and other ensemble methods [3]. Thus applying dimensionality reduction and feature importance techniques to the dataset would help in not only addressing the issue of high dimensional data which often leads to overfitting but also reducing the computational costs of developing and deploying predictive models. We have utilized the following techniques in our developed framework.

- Mutual Information: Mutual Information is a type of feature selection based on its importance. It measures the dependency between the variables and is equal to zero if and only if the given variables are independent of each other. Higher values mean higher dependency within the variables. The advantage of using Mutual Information is that it could uncover any kind of dependency whether it is linear or non-linear between the variables.

- Principal Component Analysis: PCA is a widely used technique that transforms the data into a new coordinate system in which the greatest variance by any projection of the data comes to lie on the first component the second greatest variance on the second component, and so on. This helps in retaining the most important features that explain the most important features in the dataset. The technique improves the interpretability of the data while minimizing the information loss as it gives us the option of removing the lower-order components which often represent the noise and outliers in the dataset.

- Linear Discriminant Analysis: LDA is also one of the popular techniques that helps to find linear combinations of features that characterize or separate two or more classes. For our problem statement, the LDA could help to find the best set of features that best separate the different types of cancer based on the genomic information [4].

- t-Distributed Stochastic Neighbour Embeddings: t-SNE is a nonlinear dimensionality reduction technique while the above techniques are linear. The technique is considered highly suitable for visualizing high dimensional data types and could also be utilized to view if there are any natural clusters present within the genomic features for categorizing different cancer types.

- Uniform Manifold Approximation and Projection: UMAP is also a type of non-linear dimensionality reduction technique similar to t-SNE, but previous research has shown that UMAP retains more global structure of the features and is computationally less expensive than t-SNE [5].

2.2 Machine Learning Models

The predictive modeling for cancer-type prediction using the ICMR dataset presented us with a unique problem: building a generalized predictive model using a very high dimensional feature set and less number of instances available. More instances of the data would have allowed us to explore deep learning-based approaches but considering this along with the computational complexities of those models, we stick to conventional and ensemble methods available for multi-class classification tasks. We have used a range of machine learning methods, including Logistic Regression, XGBClassifier, RandomForest Classifier, Bagging Classifier, ADA Boost Classifier, and Decision Tree Classifier. The following points highlight the key reasons why these models are suitable for multiclass cancer classification.

- Logistic Regression: Despite being a very simple machine learning model, the logistic regression model can be very effective for genomic classification problems [6]. The model for our framework is trained using the one vs all scheme, where a separate model is trained for each class to predict whether a sample belongs to a class or not.

- XGBClassifier: XGBoost Classifier is a very powerful machine learning model based on the gradient boosting framework. Since it is an ensemble model, the technique is robust to overfitting making it a very suitable model candidate for problem statements.

- RandomForest Classifier: It is also a type of ensemble learning method, that constructs multiple decision trees and outputs the class that is the mode of the class output by individual trees. The classifier uses bagging, which makes the model quite robust to overfitting and is also very suitable for high dimensional data.

- Bagging Classifier: The Bagging classifier is based on the Bootstrap Aggregating algorithm, which decreases the variance of the prediction by generating additional data from the training data using a combination of repetitions to produce multi-sets of the original data making the model robust to overfitting.

- AdaBoost Classifier: It is also a type of boosting algorithm that constructs a classifier as a linear combination of simple classifiers. The model uses the idea of combining a set of weak classifiers into a strong one, which makes it a good candidate for a benchmarking model.

- Decision Tree Classifier: The decision Tree model is a very popular model for its easy understanding and interpretability, and can handle both numerical and categorical data. The good interpretability aspect of the model also makes it very suitable for our problem.

2.3 Visualization Modules

For the analysis of high-dimensional genomic data, the visualization modules play a very critical role. It not only helps in understanding the structure and characteristics within the data but also helps in interpreting the machine learning model development processes, thus making the visualization modules helpful for both the end user and the model developers. The following points discuss and highlight the key visualization modules used in our framework.

- Histogram for visualization of cancer classes & gene frequencies: A Histogram provides a visual representation of data distribution which is inferred from the shape of the histogram, making it an ideal choice for visualizing if any class imbalance is present between various cancer classes or checking the frequency of important genes across the dataset. A clear understanding of cancer class distribution and checking the anomalies or patterns of important genes would affect the model development process, thus making the histogram a simple yet very effective visualization module of our framework.

- Scatter plots for dimensionality reduction: A scatterplot is an effective visualization for understanding the relationship between two variables. Scatterplots are widely used for identifying clusters and groups in the data, and could also be used for observing trends and outliers in the data. For our framework, we use the scatterplot to get a visual representation of the high-dimensional genomic data in a reduced dimensional space. Each of the techniques being used has a different way of preserving certain characteristics of the original dataset, which could be uncovered with the help of scatterplots. This makes scatterplot very versatile and a very critical visualization module of our framework to visualize the high dimensional genomic data.

- Bar Plots for Model Evaluation Metrics: A Barplot is an easily interpretable visualization, which is generally used for comparing multiple categories at once. In our framework, we are using a barplot to compare the model performance with the help of the F1-score and a multiple bar plot for comparing the precision recall for each of the models. The bar plot makes it very easy and at the same time, very fast to identify which cancer types the models perform well.

- Heatmap for Confusion Matrix: A heatmap is used to represent the density of data points, with darker colors representing the higher values and vice versa. It allows an easy and quick comparison between the values for different categories. The heatmap further explains the model performance by providing the values of true positives, false positives, true negatives, and false negatives, which were used to calculate the f-1 scores.

- Parallel Plots for Hyperparameter Optimization: Parallel plots provide a very clear understanding when visualizing multi-dimensional data or when we are trying to understand complex relationships across different dimensions, for an effective understanding of the hyperparameter tuning process for each of the machine learning models, each vertical line represents a hyperparameter, and each horizontal line across these vertical lines represents one observation. This helps to determine if, for a particular model, a certain combination of hyperparameters tends to give better results.

2.4 Evaluation Metrics

In our project framework, we are utilizing data preprocessing techniques such as dimensionality reduction using PCA, UMAP, etc, and predictive modeling using machine learning models and hyper-parameter optimization. Since many elements in the proposed framework are unsupervised methods, the standalone evaluation of individual elements is out of the scope of this project. However, we evaluate the performance of these elements with the help of an intrinsic evaluation performed using the outputs of the predictive models. Thus, to evaluate the overall performance of the framework, we are utilizing the following metrics:

- Confusion Matrix: The confusion matrix table would provide information on the performance of a machine learning model in terms of correct and incorrect classification for each cancer class.

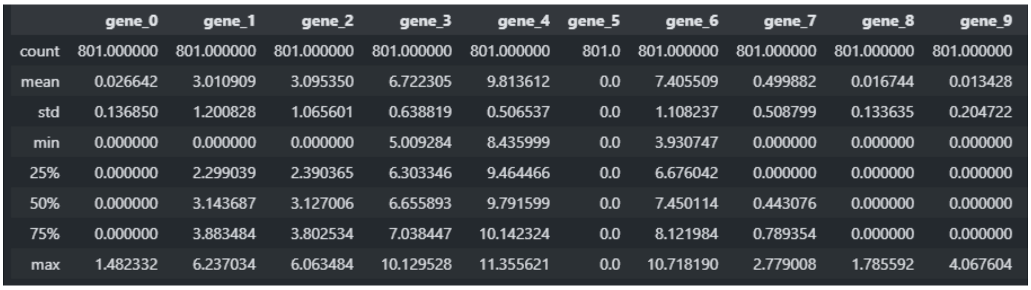
- F1 Scores: F1 Scores provide a macroscopic view of the model performance because they calculate the harmonic mean of precision and recall, and which can also signal an uneven class distribution.

- Precision & Recall: Precision is the ratio of true positives to the total predicted positives, while recall is the ratio of true positives to all observations. High precision relates to the low false positive ratio, while the high recall represents a low false negative rate.

Each of the above metrics prioritizes a different objective, for instance in cancer classification, the predictive model is required to have high precision to reduce the chances of false positives, which could lead to unnecessary medical procedures for the patient. Other metrics such as high recall would help to capture true positives as well as false negatives, which would be very useful for timely medical intervention. Therefore, to summarize, no single metric can give a complete picture of model performance, and each metric provides a different strength that could be utilized in conjunction to get a comprehensive understanding of model performance.

3 Experiment and Analysis

3.1 Exploratory Data Analysis



	gene_0	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	gene_9
count	801.000000	801.000000	801.000000	801.000000	801.000000	801.0	801.000000	801.000000	801.000000	801.000000
mean	0.026642	3.010909	3.095350	6.722305	9.813612	0.0	7.405509	0.499882	0.016744	0.013428
std	0.136850	1.200828	1.065601	0.638819	0.506537	0.0	1.108237	0.508799	0.133635	0.204722
min	0.000000	0.000000	0.000000	5.009284	8.435999	0.0	3.930747	0.000000	0.000000	0.000000
25%	0.000000	2.299039	2.390365	6.303346	9.464466	0.0	6.676042	0.000000	0.000000	0.000000
50%	0.000000	3.143687	3.127006	6.655893	9.791599	0.0	7.450114	0.443076	0.000000	0.000000
75%	0.000000	3.883484	3.802534	7.038447	10.142324	0.0	8.121984	0.789354	0.000000	0.000000
max	1.482332	6.237034	6.063484	10.129528	11.355621	0.0	10.718190	2.779008	1.785592	4.067604

Figure 1: Summary Statistics of the Dataset using the first 10 genes.

The dataset used for framework development has 20,532 gene features and 801 samples. We observe in Figure 1 that the mean and standard deviation across the gene features vary significantly, which suggests that the different gene features have different levels of expression and variability, and some genes might be highly expressed in certain samples. We also observe from Figure 1 that some gene features have a minimum value of zero, indicating that these genes might not be present in some samples. One very significant finding from summary statistics is that gene_5 has zero values throughout suggesting the presence of nonsignificant features in the feature set. It can also be inferred that for many gene features, there is a significant difference between the 75 percentile and maximum suggesting the presence of outliers and skewed distributions.

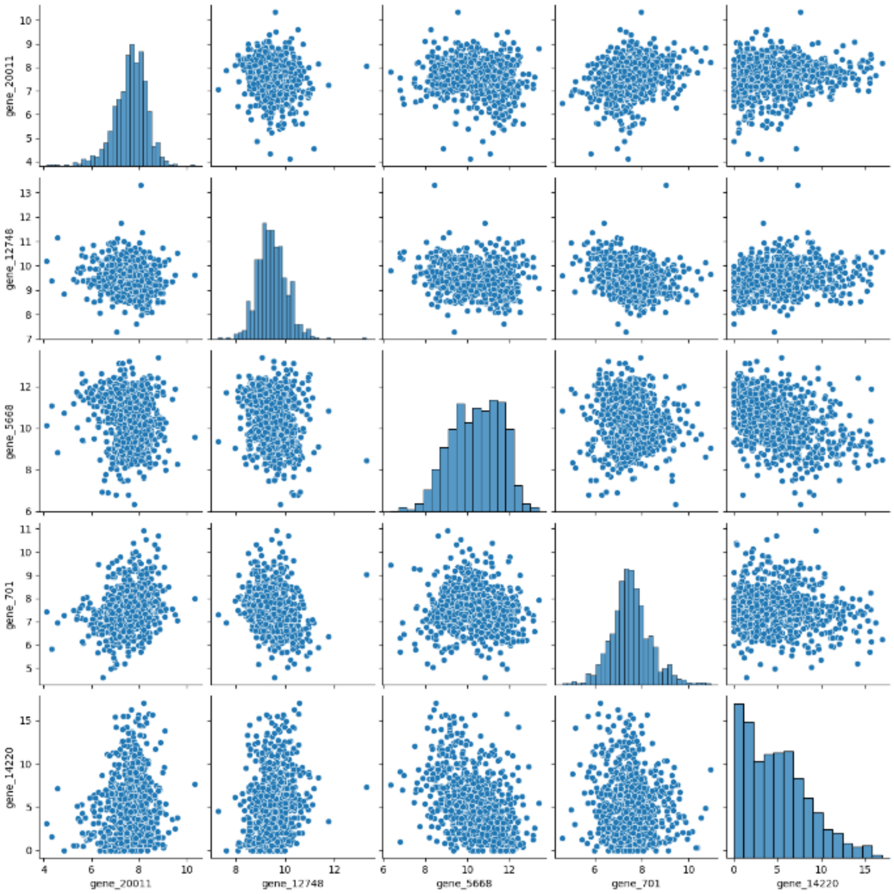


Figure 2: Pair Plot for a random sample of 5 gene features.

We observe in Figure 2 that 4 out of 5 random features selected have normal or near-normal distribution, which indicates that most gene expressions might be following a similar distribution pattern. It can also be inferred that some genes have a positive correlation between them, suggesting a biological interference among the genes. The distributional features play an important role in deciding the preprocessing steps required for implementing certain machine learning models.

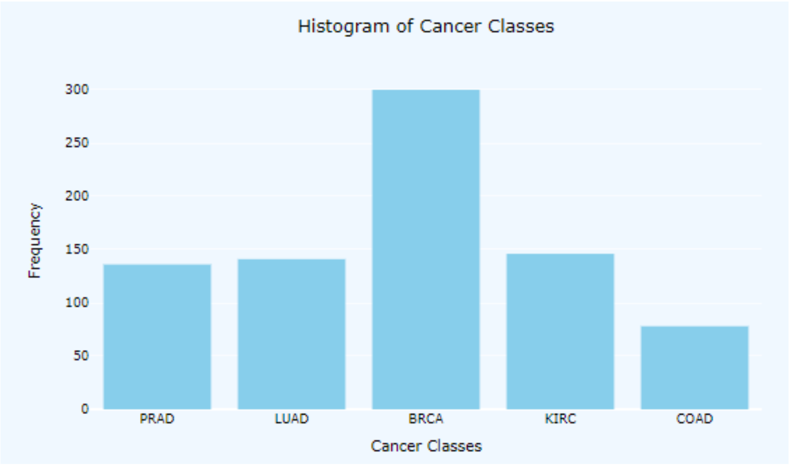
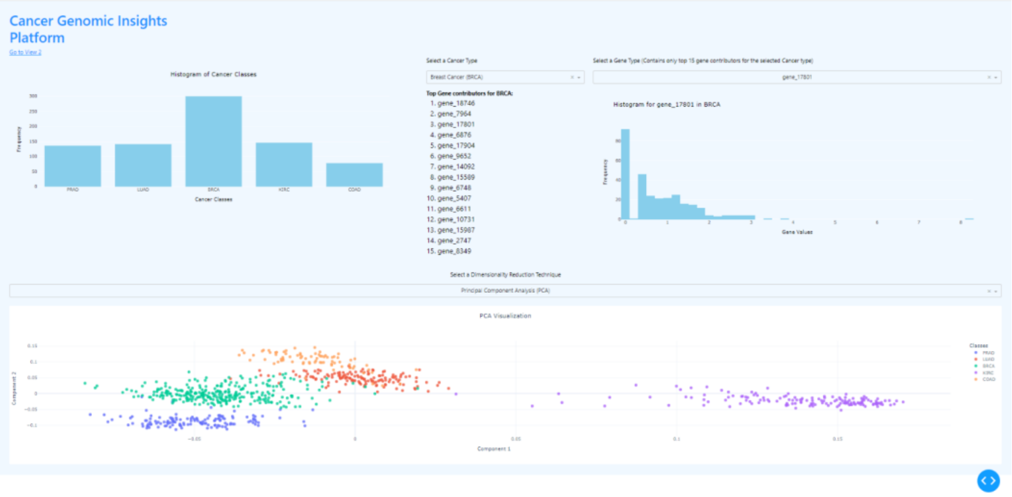


Figure 3: Distribution of Cancer Classes.

From Figure 3, It can be inferred that the dataset is imbalanced with a higher prevalence for the BRCA cancer class, and COAD is the least common class with almost 50 occurrences. This imbalance in the dataset affects the selection of machine learning algorithms and evaluation metrics. For instance, accuracy might not be a correct metric for this type of dataset, instead, precision and recall would be more appropriate.

To conclude the section, the exploratory data analysis step helped us to get an understanding of the dataset, which also influenced our selection of the machine learning model and the evaluation metrics for those predictive models.

3.1 Cancer Genomic Insights Dashboard – I



Cancer Genomic Insights Platform

Go to View 2

Select a Cancer Type: Breast Cancer (BRCA)

Select a Gene Type (Contains only top 15 gene contributors for the selected Cancer type): gene_17801

Top Gene contributors for BRCA:

1. gene_18746
2. gene_7964
3. gene_17801
4. gene_6876
5. gene_17904
6. gene_9652
7. gene_14092
8. gene_15589
9. gene_6748
10. gene_5407
11. gene_6611
12. gene_10731
13. gene_15987
14. gene_2747
15. gene_8349

Select a Dimensionality Reduction Technique: Principal Component Analysis (PCA)

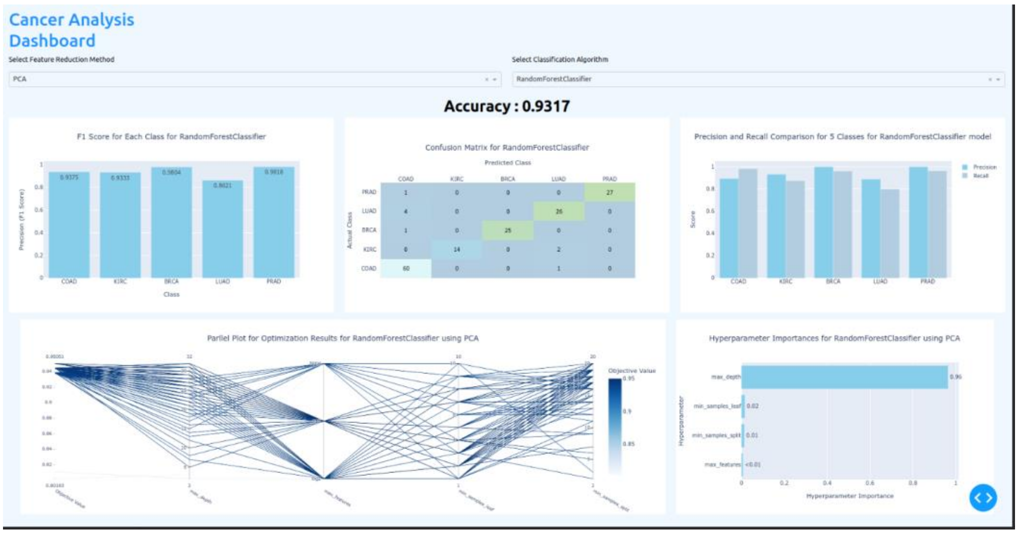
Figure 4: Cancer Genomic Insights Dashboard – I.

Following the Exploratory Data Analysis, we implemented the data preprocessing steps which included a mutual information test to extract key features for a given cancer class, and a set of dimensionality reduction techniques mentioned in Section 2.1. We have kept the class distribution in the dashboard, as it informs the clinical practitioners and model developers about the prevalent cancer class in the data. In the top-right corner are the results of the mutual information test, which identifies the most significant features that contribute to the variance and pattern in the dataset. In the current plot, we can see the top 10 genes for the BRCA cancer class, which might be crucial for predicting the BRCA class and could be potential natural biomarkers for this cancer class. In addition to this, we also plot distributions of gene expression values, which explains how the selected gene from the important features is expressed for the cancer class. For instance, in the current selection, the gene 17801 is not expressed highly i.e. the gene has lower values across the BRCA cancer class, and if certain expression levels are associated with the cancer class, then that could be used as a biomarker for diagnosing the BRCA cancer class.

Next, we implemented data dimensionality reduction techniques mentioned in Section 2.1. In the second half of Figure 4, we can see from the current selection of the PCA dimensionality reduction technique that the PCA is very effective at separating different cancer classes, indicating distinct genetic expression values and characteristics unique to each cancer class. Thus the analysis suggests that PCA is very effective at reducing the dimensionality of the feature space while preserving the differences between cancer types.

The results from Dashboard 1 guide the steps in the second dashboard. For example, the effectiveness of the PCA dimensionality reduction would be validated with the help of different machine-learning models and evaluation metrics as mentioned in sections 2.2 & 2.4. Results for other selections in Dashboard 1 are added in the appendix section of the report.

3.2 Cancer Genomic Insights Dashboard – II



Cancer Analysis Dashboard

Select Feature Reduction Method: PCA

Select Classification Algorithm: RandomForestClassifier

Accuracy : 0.9317

Confusion Matrix for RandomForestClassifier

Actual Class \ Predicted Class	COAD	KIRC	BRCA	LUAD	PRAD
PRAD	1	0	0	0	27
LUAD	4	0	0	26	0
BRCA	1	0	25	0	0
KIRC	0	14	0	2	0
COAD	60	0	0	1	0

F1 Score for Each Class for RandomForestClassifier
Precision (F1 Score)
1
0.8
0.6
0.4
0.2
0
0.9375
0.9333
0.9804
0.8621
0.9818
COAD
KIRC
BRCA
LUAD
PRAD
Class
Precision and Recall Comparison for 5 Classes for RandomForestClassifier model
Score
Precision
Recall
Parllel Plot for Optimization Results for RandomForestClassifier using PCA
Objective Value
0.95
0.9
0.85
Hyperparameter Importances for RandomForestClassifier using PCA
Hyperparameter
max_depth
0.96
min_samples_leaf
0.02
min_samples_split
0.01
max_features
<0.01
0
0.2
0.4
0.6
0.8
1
Hyperparameter Importance

Figure 5: Cancer Genomic Insights Dashboard – II.

For the 2nd dashboard, we implemented the machine learning model, and evaluation metrics mentioned in sections 2.2 & 2.4. We used the RandomForest Classifier, Bagging Classifier, AdaBoost classifier, and DecisionTree Classifier along with the data reduction techniques discussed and implemented in section 3.1. The model performance is evaluated using a set of evaluation metrics mentioned in section 2.4. We also implemented a hyperparameter tuning process with the help of the Optuna framework for each combination of machine learning model and dimensionality reduction techniques.

It can be seen in Figure 5 that the RandomForest model with PCA has a high F1 score for each cancer class indicating good performance for both the model and the dimensionality reduction technique. It can also be seen on the confusion matrix plot in Figure that the model is efficient in correctly classifying various cancer types, which further validates the findings from the F1 score results. We also note the differences between precision and recall for each cancer class in Figure 5, but neither of those is significantly lower than the other. Depending on the requirement, the user could choose to optimize either precision or recall.

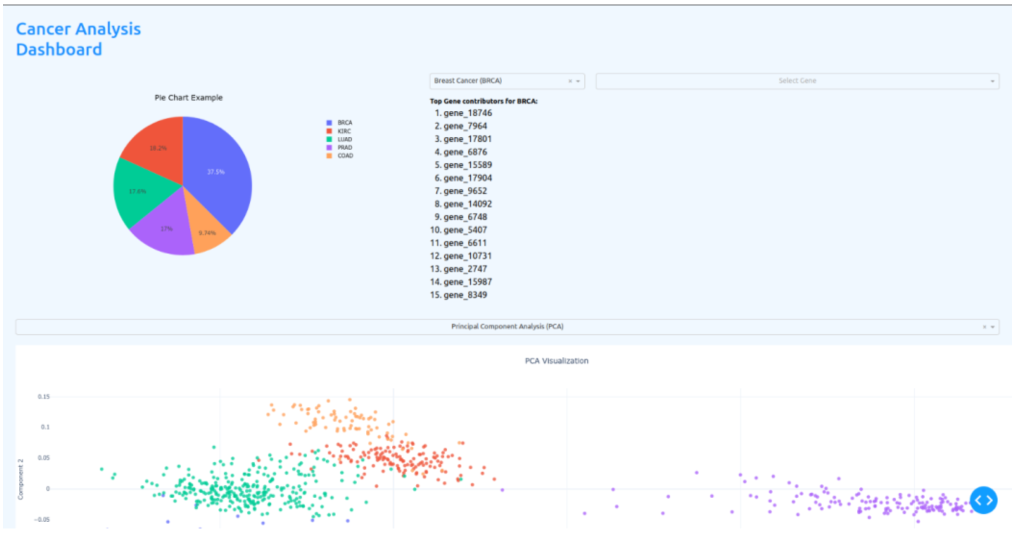
Next, we implement hyper-parameter optimization using Optuna Framework for each of the models, and dimensionality reduction techniques. It can be seen in Figure 5 that for the max_depth hyperparameter, a certain range of values positively affects the model performance while for other remaining hyperparameters, the instances are very well spread out indicating their less influence on the model performance. The Hyperparameter Importance Plot (Bar Plot) in Figure 5, shows the most significant hyperparameter for a given model, for instance, the max_depth has the longest bar in the plot, which indicates a high influence of the max_depth parameter on the model performance, which further validates the findings from the parallel plot graph.

To conclude, Dashboard II provides very crucial insights for understanding the model behavior and performance. This would help the model developers to identify the area of improvement further guiding the model development and deployment processes. It also underscores the importance of integrating multiple evaluation metrics with the model pipeline and visualizing the hyperparameter optimization process. This allows both the end user and model developer the option of choosing the metric as per their requirements, and also to spend more resources on fine-tuning a particular hyper-parameter if it shows a causal relationship with the model performance.

3.3 Testing and Evaluation

For Testing and Evaluation, we set up the following benchmarks:

- User Friendliness.
- Clarity of Visualization.
- How effective are the dashboards for decision-making?



Cancer Analysis Dashboard

Top Gene contributors for BRCA:

1. gene_18746
2. gene_7964
3. gene_17801
4. gene_6876
5. gene_15589
6. gene_17904
7. gene_9652
8. gene_14092
9. gene_6748
10. gene_5407
11. gene_6611
12. gene_10731
13. gene_2747
14. gene_15987
15. gene_8349

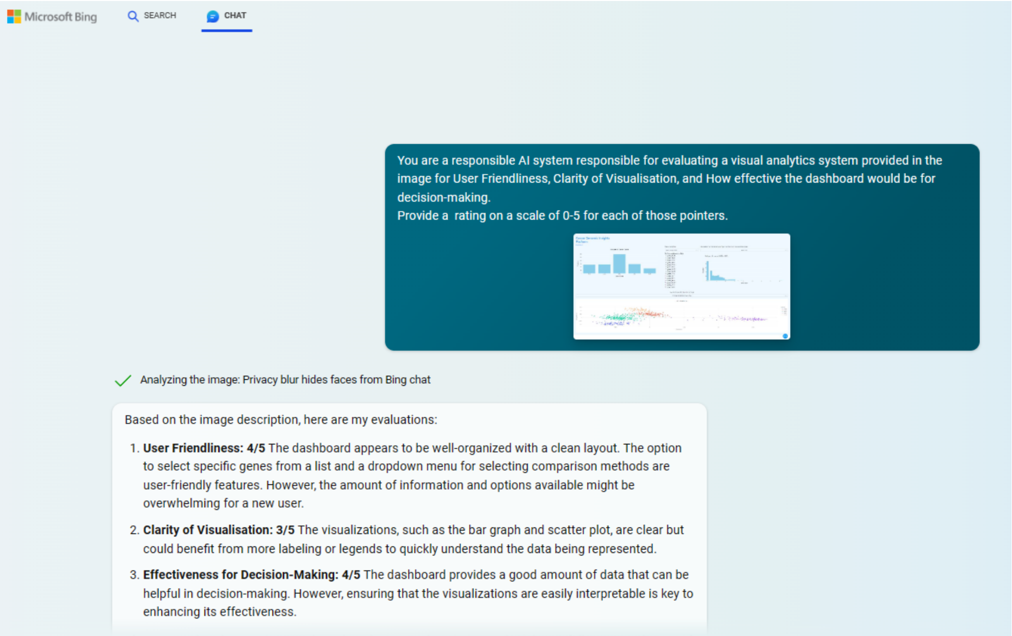
Figure 6: Dashboard I with alternate plots

We experimented with many variety of plots before finalizing the plot prioritizing the above-mentioned benchmarks for instance in Figure 6 we implemented a pie plot instead of a histogram plot. Based on the user feedback collected in class based on the above-mentioned benchmarks, we finalized the visualization plots mentioned in section 2.3. We also performed replication testing by generating the plots on different computer setups and performed manual integration testing to make sure that all of the project components were working as intended.

In addition to the above-mentioned tests, we also performed agent-based evaluation using a large language model. The motivation for these tests comes from the premise that human evaluation can be biased and subjective therefore testing the framework with both manual and automated testing would provide a more comprehensive picture of the framework's performance.

We used the GPT-4 model with the following prompt:

"You are a responsible AI system responsible for evaluating a visual analytics system provided in the image below for User Friendliness, Clarity of Visualisation, and How effective the dashboard would be for decision-making.
Provide a rating on a scale of 0-5 for each of those pointers.
"



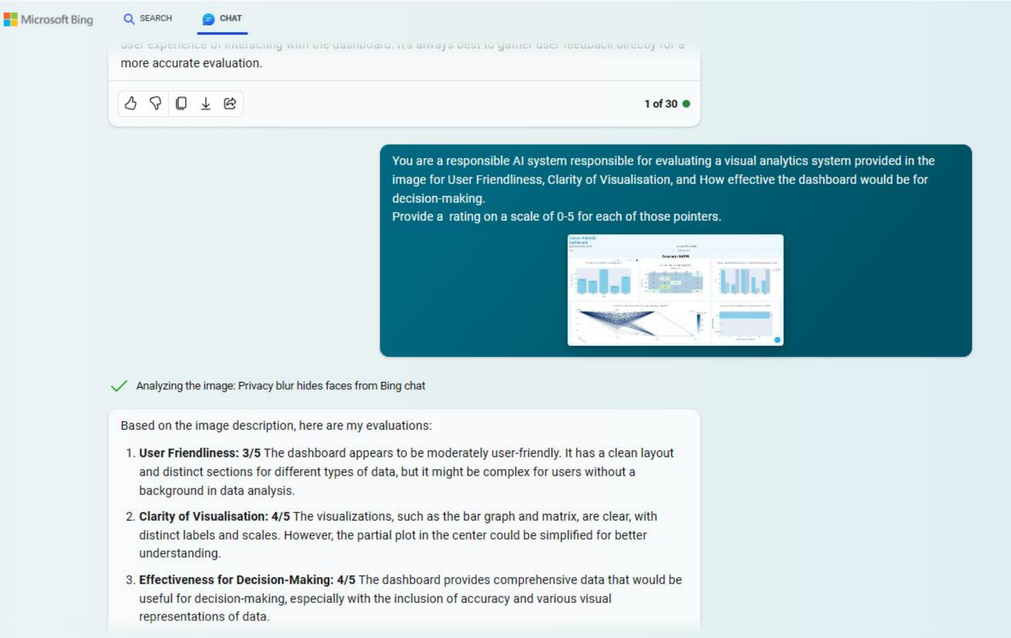
You are a responsible AI system responsible for evaluating a visual analytics system provided in the image for User Friendliness, Clarity of Visualisation, and How effective the dashboard would be for decision-making.
Provide a rating on a scale of 0-5 for each of those pointers.

✓ Analyzing the image: Privacy blur hides faces from Bing chat

Based on the image description, here are my evaluations:

1. **User Friendliness: 4/5** The dashboard appears to be well-organized with a clean layout. The option to select specific genes from a list and a dropdown menu for selecting comparison methods are user-friendly features. However, the amount of information and options available might be overwhelming for a new user.
2. **Clarity of Visualisation: 3/5** The visualizations, such as the bar graph and scatter plot, are clear but could benefit from more labeling or legends to quickly understand the data being represented.
3. **Effectiveness for Decision-Making: 4/5** The dashboard provides a good amount of data that can be helpful in decision-making. However, ensuring that the visualizations are easily interpretable is key to enhancing its effectiveness.

Figure 7: Agent-Based Evaluation of Dashboard I using GPT 4.



user experience of interacting with the dashboard. It's always best to gather user feedback directly for a more accurate evaluation.

You are a responsible AI system responsible for evaluating a visual analytics system provided in the image for User Friendliness, Clarity of Visualisation, and How effective the dashboard would be for decision-making.
Provide a rating on a scale of 0-5 for each of those pointers.

Analyzing the image: Privacy blur hides faces from Bing chat

Based on the image description, here are my evaluations:

1. **User Friendliness: 3/5** The dashboard appears to be moderately user-friendly. It has a clean layout and distinct sections for different types of data, but it might be complex for users without a background in data analysis.
2. **Clarity of Visualisation: 4/5** The visualizations, such as the bar graph and matrix, are clear, with distinct labels and scales. However, the partial plot in the center could be simplified for better understanding.
3. **Effectiveness for Decision-Making: 4/5** The dashboard provides comprehensive data that would be useful for decision-making, especially with the inclusion of accuracy and various visual representations of data.

Figure 8: Agent-Based Evaluation of Dashboard II using GPT 4.

It can be seen in Figures 7 & 8 that agent-based evaluation has given satisfactory scores on all of the benchmarks mentioned above. Therefore considering the manual testing with human feedback and automated testing with GPT-4, we deem the developed framework to be very efficient at conveying information to the end user.

6 Discussion

The project underscores the importance of explainability and transparency in artificial intelligence and machine learning-based frameworks. This is particularly crucial in complex fields such as genomic studies. The project demonstrates how effective visualizations, such as histograms of cancer classes and PCA scatter plots, can aid in communicating insights from genomic data to both clinical practitioners and predictive model developers.

While the project addresses several challenges in genomic data analysis, there are still challenges and limitations with the predictive modeling and explainability aspect of genomic data. The complexities of genomic data pose challenges for both the end user and the model developer to get a clear understanding of patterns and information inside them.

Dimensionality-reduction algorithms, such as PCA, UMAP, T-SNE, and LDA, have been applied to aid in understanding these complexities. However, these methods might not capture all the important information present in the high-dimensional genomic data. Therefore, the application of more sophisticated frameworks that include deep learning-based methods might provide a clearer and more informative understanding of this data.

Deep learning methods, such as LSTM or transformer-based architectures [9], have the potential to capture complex patterns in the genomic data and improve the classification accuracy of different cancer types.

The project could also benefit from real-time data updation or integration of the framework with other databases, which might also aid in setting up performance monitoring for the framework.

The framework evaluation is performed using a user study and automated evaluation using a large language model. The user study was conducted with a diverse group of users to assess the usability and effectiveness of the framework. The large language model was used to evaluate the framework’s ability to generate accurate and relevant insights from the genomic data.

However, the evaluation process could be improved by conducting user studies with domain experts and clinical practitioners. Their expertise could provide valuable insights into the practical utility and usability of the framework, leading to more targeted improvements.

In conclusion, this project represents a first step towards making genomic data more accessible and understandable to a wide range of users. Future work could focus on addressing the identified challenges and further enhancing the capabilities of the framework.

7 Future Work

Building upon the current work and limitations of the current implementation of the framework there are several areas in which the project could be improved. Some of those ideas are:

- Deep Learning Methods for Classification: Deep learning methods have shown remarkable success in various fields such as genomics [9], machine vision, etc. Future work could be based on implementing more sophisticated predictive models based on these architectures, which might capture more complex patterns in the data, and improve the classification accuracy of different cancer types.

- Integration of Large Language Models: LLMs such as GPT-4 could be integrated with the framework, which would assist the user in navigating and interpreting the results from the dashboard. The framework would greatly benefit from natural language explanations of complex genomic data insights, making the information accessible to both clinical practitioners and model developers. This would further improve the explainability and transparency of the framework.

- Dashboard Improvements & Real-time Data Integration: The current dashboard provides clear and effective visualizations. However, these could be further improved with the help of more sophisticated visualizations such as Network Diagrams [10] which would improve the explainability of the results as well as the aesthetics of the dashboard. Future work could also focus on enhancing the interactive features of the dashboard, which would allow the user to explore data with more flexibility. Further, integrating a real-time database with the framework would allow the user to analyze the most recent data, which could improve the accuracy and relevance of the analysis.

- Evaluation with Domain Experts: While the current evaluation strategy involves user studies and automated evaluation using a large language model, future work could involve conducting user studies with the target users of the framework: domain experts and clinical practitioners. The feedback from the target users would improve the practical utility and usability of the framework, leading to more targeted improvements.

References:

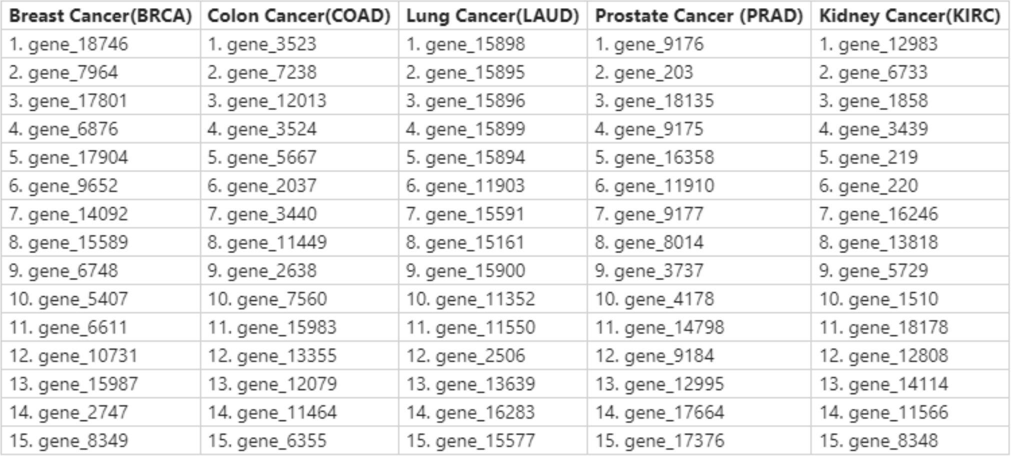
1. Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2021). Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-021-00434-9
2. Kumavath, R., Barh, D., Andrade, B. S., Imchen, M., Aburjaile, F. F., Ch, A., Rodrigues, D. L. N., Tiwari, S., Alzahrani, K. J., Góes-Neto, A., Weener, M. E., Ghosh, P., & Azevedo, V. (2021). The Spike of SARS-CoV-2: Uniqueness and Applications. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.663912
3. Debie, E., & Shafi, K. (2017). Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. *Pattern Analysis and Applications*, *22*(2), 519–536. https://doi.org/10.1007/s10044-017-0649-0
4. Debie, E., & Shafi, K. (2017). Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. *Pattern Analysis and Applications*, *22*(2), 519–536. https://doi.org/10.1007/s10044-017-0649-0
5. Becht, E., Charles-Antoine Dutertre, Kwok, I., Ng, L., Florent Ginhoux, & Newell, E. (2018). Evaluation of UMAP as an alternative to t-SNE for single-cell data. *BioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/298430
6. Bazzoli, C., & Lambert-Lacroix, S. (2018). Classification based on extensions of LS-PLS using logistic regression: application to clinical and multiple genomic data. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2311-2
7. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., & Back, T. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

8. Abass, Y. A., & Adeshina, S. A. (2021). Deep Learning Methodologies for Genomic Data Prediction: Review. *Journal of Artificial Intelligence for Medical Sciences*, *2*(1-2), 1. https://doi.org/10.2991/jaims.d.210512.001

9. Choi, S. R., & Lee, M. (2023). Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. *Biology*, *12*(7), 1033. https://doi.org/10.3390/biology12071033

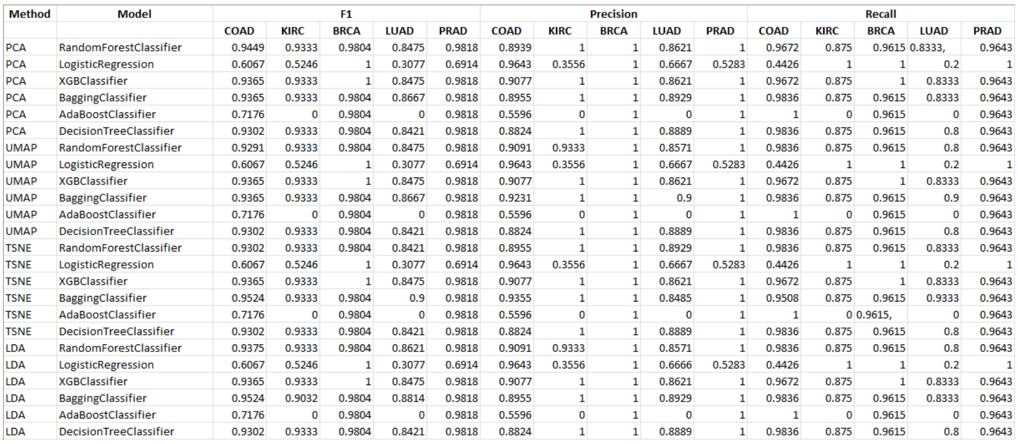
10. Koutrouli, M., Karatzas, E., Katerina Papanikolopoulou, & Pavlopoulos, G. A. (2020). *NORMA-The network makeup artist: a web tool for network annotation visualization*. https://doi.org/10.1101/2020.03.05.978585

Appendix:



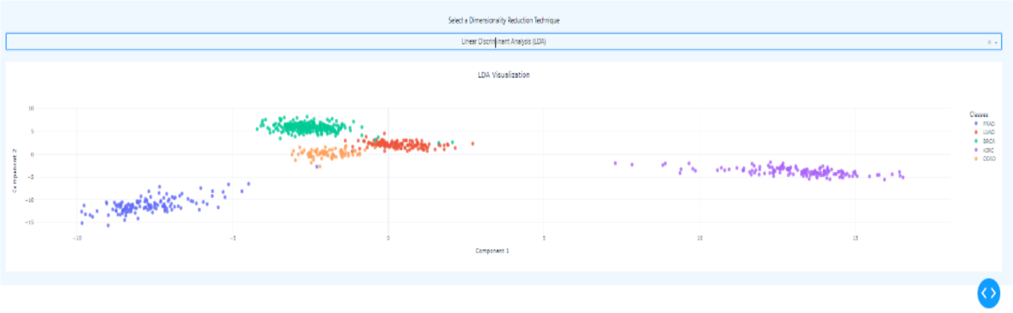
Breast Cancer(BRCA)	Colon Cancer(COAD)	Lung Cancer(LAUD)	Prostate Cancer (PRAD)	Kidney Cancer(KIRC)
1. gene_18746	1. gene_3523	1. gene_15898	1. gene_9176	1. gene_12983
2. gene_7964	2. gene_7238	2. gene_15895	2. gene_203	2. gene_6733
3. gene_17801	3. gene_12013	3. gene_15896	3. gene_18135	3. gene_1858
4. gene_6876	4. gene_3524	4. gene_15899	4. gene_9175	4. gene_3439
5. gene_17904	5. gene_5667	5. gene_15894	5. gene_16358	5. gene_219
6. gene_9652	6. gene_2037	6. gene_11903	6. gene_11910	6. gene_220
7. gene_14092	7. gene_3440	7. gene_15591	7. gene_9177	7. gene_16246
8. gene_15589	8. gene_11449	8. gene_15161	8. gene_8014	8. gene_13818
9. gene_6748	9. gene_2638	9. gene_15900	9. gene_3737	9. gene_5729
10. gene_5407	10. gene_7560	10. gene_11352	10. gene_4178	10. gene_1510
11. gene_6611	11. gene_15983	11. gene_11550	11. gene_14798	11. gene_18178
12. gene_10731	12. gene_13355	12. gene_2506	12. gene_9184	12. gene_12808
13. gene_15987	13. gene_12079	13. gene_13639	13. gene_12995	13. gene_14114
14. gene_2747	14. gene_11464	14. gene_16283	14. gene_17664	14. gene_11566
15. gene_8349	15. gene_6355	15. gene_15577	15. gene_17376	15. gene_8348

Table 1: Top 10 important genes inferred from Mutual Information test for each cancer class



Method	Model	F1					Precision					Recall				
		COAD	KIRC	BRCA	LUAD	PRAD	COAD	KIRC	BRCA	LUAD	PRAD	COAD	KIRC	BRCA	LUAD	PRAD
PCA	RandomForestClassifier	0.9449	0.9333	0.9804	0.8475	0.9818	0.8939	1	1	0.8621	1	0.9672	0.875	0.9615	0.8333,	0.9643
PCA	LogisticRegression	0.6067	0.5246	1	0.3077	0.6914	0.9643	0.3556	1	0.6667	0.5283	0.4426	1	1	0.2	1
PCA	XGBClassifier	0.9365	0.9333	1	0.8475	0.9818	0.9077	1	1	0.8621	1	0.9672	0.875	1	0.8333	0.9643
PCA	BaggingClassifier	0.9365	0.9333	0.9804	0.8667	0.9818	0.8955	1	1	0.8929	1	0.9836	0.875	0.9615	0.8333	0.9643
PCA	AdaBoostClassifier	0.7176	0	0.9804	0	0.9818	0.5596	0	1	0	1	1	0	0.9615	0	0.9643
PCA	DecisionTreeClassifier	0.9302	0.9333	0.9804	0.8421	0.9818	0.8824	1	1	0.8889	1	0.9836	0.875	0.9615	0.8	0.9643
UMAP	RandomForestClassifier	0.9291	0.9333	0.9804	0.8475	0.9818	0.9091	0.9333	1	0.8571	1	0.9836	0.875	0.9615	0.8	0.9643
UMAP	LogisticRegression	0.6067	0.5246	1	0.3077	0.6914	0.9643	0.3556	1	0.6667	0.5283	0.4426	1	1	0.2	1
UMAP	XGBClassifier	0.9365	0.9333	1	0.8475	0.9818	0.9077	1	1	0.8621	1	0.9672	0.875	1	0.8333	0.9643
UMAP	BaggingClassifier	0.9365	0.9333	0.9804	0.8667	0.9818	0.9231	1	1	0.9	1	0.9836	0.875	0.9615	0.9	0.9643
UMAP	AdaBoostClassifier	0.7176	0	0.9804	0	0.9818	0.5596	0	1	0	1	1	0	0.9615	0	0.9643
UMAP	DecisionTreeClassifier	0.9302	0.9333	0.9804	0.8421	0.9818	0.8824	1	1	0.8889	1	0.9836	0.875	0.9615	0.8	0.9643
TSNE	RandomForestClassifier	0.9302	0.9333	0.9804	0.8421	0.9818	0.8955	1	1	0.8929	1	0.9836	0.875	0.9615	0.8333	0.9643
TSNE	LogisticRegression	0.6067	0.5246	1	0.3077	0.6914	0.9643	0.3556	1	0.6667	0.5283	0.4426	1	1	0.2	1
TSNE	XGBClassifier	0.9365	0.9333	1	0.8475	0.9818	0.9077	1	1	0.8621	1	0.9672	0.875	1	0.8333	0.9643
TSNE	BaggingClassifier	0.9524	0.9333	0.9804	0.9	0.9818	0.9355	1	1	0.8485	1	0.9508	0.875	0.9615	0.9333	0.9643
TSNE	AdaBoostClassifier	0.7176	0	0.9804	0	0.9818	0.5596	0	1	0	1	1	0	0.9615,	0	0.9643
TSNE	DecisionTreeClassifier	0.9302	0.9333	0.9804	0.8421	0.9818	0.8824	1	1	0.8889	1	0.9836	0.875	0.9615	0.8	0.9643
LDA	RandomForestClassifier	0.9375	0.9333	0.9804	0.8621	0.9818	0.9091	0.9333	1	0.8571	1	0.9836	0.875	0.9615	0.8	0.9643
LDA	LogisticRegression	0.6067	0.5246	1	0.3077	0.6914	0.9643	0.3556	1	0.6666	0.5283	0.4426	1	1	0.2	1
LDA	XGBClassifier	0.9365	0.9333	1	0.8475	0.9818	0.9077	1	1	0.8621	1	0.9672	0.875	1	0.8333	0.9643
LDA	BaggingClassifier	0.9524	0.9032	0.9804	0.8814	0.9818	0.8955	1	1	0.8929	1	0.9836	0.875	0.9615	0.8333	0.9643
LDA	AdaBoostClassifier	0.7176	0	0.9804	0	0.9818	0.5596	0	1	0	1	1	0	0.9615	0	0.9643
LDA	DecisionTreeClassifier	0.9302	0.9333	0.9804	0.8421	0.9818	0.8824	1	1	0.8889	1	0.9836	0.875	0.9615	0.8	0.9643

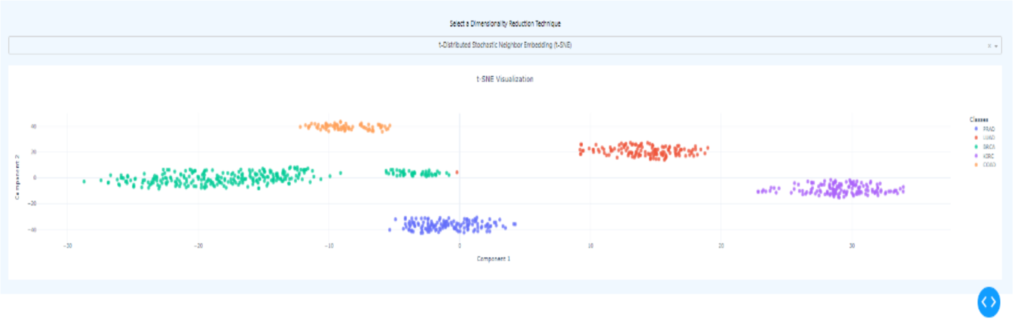
Table 2: Predictive Model Performance Comparison.



Select a Dimensionality Reduction Technique

Linear Discriminant Analysis (LDA)

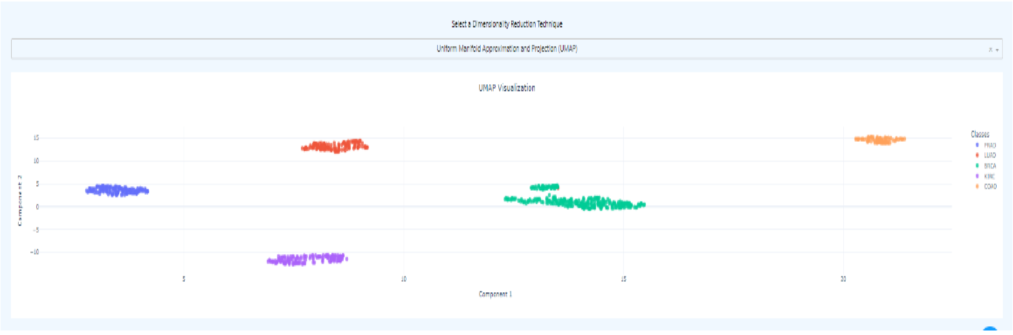
Figure 1: Scatter Plot for LDA Dimensionality Reduction.



Select a Dimensionality Reduction Technique

t-Distributed Stochastic Neighbor Embedding (t-SNE)

Figure 2: Scatter Plot for t-SNE Dimensionality Reduction.



Select a Dimensionality Reduction Technique

Uniform Manifold Approximation and Projection (UMAP)

Figure 3: Scatter Plot for UMAP Dimensionality Reduction.