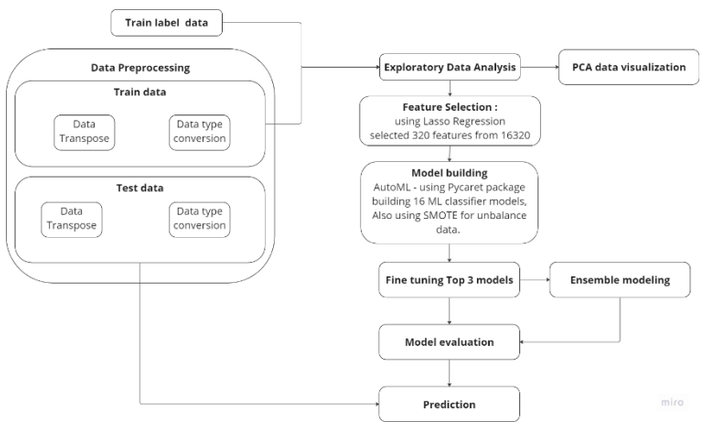
Predicting Cancer Types from Gene Expression Data: A Machine Learning Approach

Introduction:

This study presents a machine learning-based approach to predict the primary disease type of samples, focusing on breast cancer, lung squamous cell carcinoma, and lung adenocarcinoma. Gene expression data from "train_data.tsv" and "train_label.tsv" were employed for model training and evaluation. Performance metrics, including F1 score, Accuracy, AUC, Recall, Precision, Kappa, and MCC, were utilized to gauge model performance. Leveraging AutoML and ensemble techniques, an effective predictive model was developed. The analysis encompassed exploratory data analysis, dimensionality reduction, and feature selection techniques to enhance the prediction process. The results underscore the effectiveness of the model in accurately classifying cancer types, emphasizing the significance of robust pre-processing and thoughtful model selection.

Material and Methods:



Data Pre-processing:

The raw gene expression data from "train_data.tsv" and corresponding sample labels from "train_label.tsv" were merged to create a comprehensive dataset. Data integrity was ensured by transposing the data matrix, converting count values to float64, and encoding categorical labels numerically.

Exploratory Data Analysis:

Exploratory data analysis was conducted to gain insights into the distribution of classes and the relationships between gene expression and disease types. Visualization techniques were employed to uncover patterns, correlations, and target data or labels distribution.

Feature Selection:

To mitigate the challenges posed by the high dimensionality of gene expression data, Lasso Regression with 5-fold cross-validation was employed for feature selection. This technique helped identify a subset of genes that contribute significantly to classification accuracy [1].

Model Selection and Training:

An automated machine learning (AutoML) approach using Pycaret [2] was adopted to evaluate the performance of diverse classification algorithms. The SMOTE [3] technique was harnessed to tackle data imbalance. In our study, the distribution of cancer types exhibited varying proportions, with one class being more prevalent than others. This class imbalance can skew accuracy metrics, leading to misleadingly high results by favouring the majority class. The F1 score, by incorporating both precision and recall, provides a comprehensive insight into the model's ability to correctly classify all classes, considering both false positives and false negatives. Hence, he top-performing algorithms were chosen based on the F1 score to construct an ensemble model. The selected models were trained using the pre-processed and feature-selected dataset.

Model Evaluation:

Model evaluation involved assessing predictive performance using a range of metrics, including F1 score, Accuracy, AUC, Recall, Precision, Kappa, and MCC. The choice of metrics provided a comprehensive view of the model's strengths and limitations in classifying cancer types accurately.

Results and Discussion:

The developed machine learning model exhibited competitive predictive accuracy, with robust performance metrics including F1 score, Accuracy, AUC, Recall, Precision, Kappa, and MCC. This underscores the critical role of meticulous pre-processing, effective feature selection, and thoughtful model selection in achieving precise predictions. Presented below are the evaluation matrices for the top three models and the ensemble model:

Table 1: Evaluation matrices for top-3 and ensemble models

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Logistic Regression	0.9846	0.9969	0.9846	0.9851	0.9846	0.9729	0.9731
Linear Discriminant Analysis	0.9824	0.9973	0.9824	0.9832	0.9825	0.969	0.9693
Ridge Classifier	0.9824	0	0.9824	0.9827	0.9824	0.969	0.9691
Ensemble classifier	0.9846	0	0.9846	0.9851	0.9846	0.9729	0.9731

In consideration of these metrics, it is evident that the Logistic Regression model stands out, demonstrating the highest accuracy, recall, precision, F1-score, Kappa, and Matthews Correlation Coefficient (MCC). However, it is important to note that these metrics alone may not exclusively dictate the selection of a model for deployment or future applications. Consequently, a granular assessment of the models' predictive capabilities for individual categories was conducted. The subsequent matrices elucidate how each model performs in predicting specific categories:

Table 2: Performance matrices for top-3 and ensemble models

	Logistic Regression			**Linear Discriminant Analysis**			**Ridge Classifier**			**Ensemble classifier**		
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
breast invasive carcinoma	1	1	1	1	1	1	1	1	1	1	1	1
lung adenocarcinoma	0.94	0.99	0.96	0.93	0.99	0.96	0.94	0.98	0.96	0.94	0.99	0.96
lung squamous cell carcinoma	0.99	0.93	0.96	0.99	0.93	0.96	0.98	0.93	0.96	0.99	0.93	0.96

From the matrix comparison, it is apparent that both the Logistic Regression and Ensemble models yield commendable results, with only marginal disparities compared to the other two models. Thus, a judicious balance between performance and simplicity renders the Logistic Regression model a favourable choice. Alternatively, for those seeking a marginally more intricate model without compromising performance, the Ensemble Classifier presents a viable option. It is noteworthy that predictions from both models aligned significantly, with a concurrence rate of approximately 99.33%. The final prediction was executed using the ensemble model, and the outcomes were subsequently stored for reference.

Table 3: Predictions of Logistic regressor and ensemble model

	Ensemble classifier	**Logistic Regression**
breast invasive carcinoma	199	200
lung adenocarcinoma	211	207
lung squamous cell carcinoma	190	193

Conclusion and Future Perspective:

In conclusion, this study showcased the successful development of a machine learning model for predicting primary cancer types using gene expression data using AutoML and ensemble technique. The analysis revealed that a combination of effective pre-processing, feature selection, and model selection leads to improved classification accuracy. Future research avenues include exploring advanced dimensionality reduction techniques, integrating more complex ensemble or deep learning methods, utilizing explainable-AI and further investigating the biological implications of identified data clusters.

References:

1. V, Belitser E. Feature selection using lasso. VU Amsterdam Res. Paper Business Anal. 2017; [Google Scholar]
2. Moez Ali PyCaret: An open source, low-code machine learning library in Python, April 2020. [Link]
3. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–57

Code Availability:

1. Data_preprocessing_and_analysis.ipynb – Pre-processing, EDA and feature selection
2. autoML.ipynb – Model building, analysis, and evaluation.

Note: All graphs and matrices can be found in the notebooks.