Comprehensive Feature Selection and Biological Relevance Analysis in Multi-Disease Gene Expression Data

1st — Mashiyat Mahjabin Prapty
Department of CSE, BUET
1805117@ugrad.cse.buet.ac.bd

Supervisor — Mohammad Saifur Rahman
Department of CSE, BUET
mrahman@cse.buet.ac.bd

***Abstract*—Understanding the biological relevance of specific genes in disease progression is crucial for advancing diagnostics and treatment. Microarray technologies have created a bridge between computer science and biology. Due to the high dimensionality of these datasets, feature selection with various computational tools and algorithms has been the subject of study for several decades. In this study, we developed a robust two-step feature selection method to identify the most biologically relevant features associated with various diseases. Initially, we employed the Extreme Gradient Boosting(XGB) Classifier's in-built ranking mechanism to select the top 10% of features from each dataset. This step was followed by Recursive Feature Elimination with 10-fold cross-validation using a Logistic Regression wrapper (LR-RFECV) to refine the selection to the most optimal feature set.**

Our methodology was evaluated across an extensive range of binary and multiclass datasets, making this study one of the most comprehensive in the field. The comparative analysis of the proposed method against several state-of-the-art (SOTA) methods across various datasets demonstrates its robustness and superior performance. We focused on establishing the biological significance of the identified features. Utilizing explainable AI techniques, specifically SHAP, we interpreted the models and conducted enrichment analyses using different tools. These analyses highlight the utility of our pipeline in not only improving classification accuracy but also in providing meaningful biological insights.

***Index Terms*—Microarray, Cancer, Enrichment Analysis, SHAP, Feature Selection**

I. Introduction

Cancer research heavily relies on biomarkers like genes and proteins for diagnosis and prognosis, with microarray technology generating vast gene expression datasets [1]. However, the high dimensionality of these datasets poses challenges, addressed through gene selection methods that improve classification accuracy by reducing dimensionality [2]. AI and machine learning techniques, particularly feature selection methods such as filter, wrapper, embedded, ensemble, and hybrid approaches, are crucial in this process [3]. Recent advancements include graph-based methods [4] and explainable AI techniques like SHAP, which enhance model interpretability [5]. In our study, we used a two-step feature selection method combining XGBClassifier and LR-RFECV to identify biologically relevant features, supported by SHAP for interpretability and enrichment analysis for biological insight.

II. Methods

We worked with 11 binary class datasets and 10 multi-class datasets from different data repositories.

Our proposed pipeline has four major steps, namely:

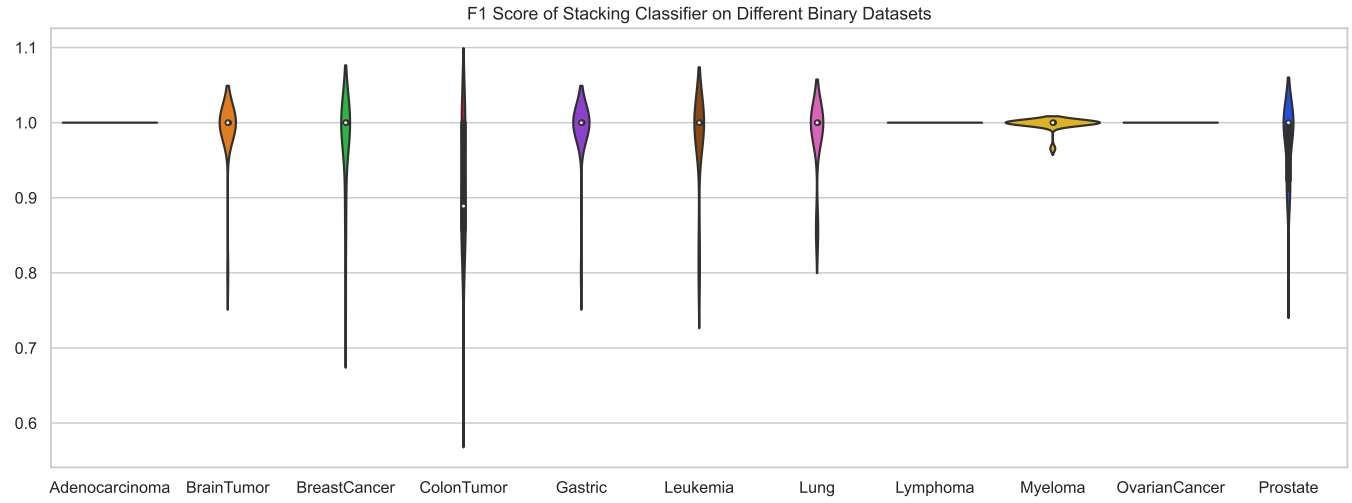
1) **Pre-processing**: Missing value imputation with mean; erroneous value detection and correction; standardization.
2) **Feature Selection with XGBoost Classifier**: Selecting the top 10% of features by fitting the data into an Extreme Gradient Boosting classifier based on the model's feature importance rank.
3) **10-fold RFECV with Logistic Regression**: RFECV is done with 10-fold cross-validation with a Logistic Regression classifier as the wrapper; LR is chosen for its simplicity and suitability with microarray datasets.
4) **Evaluation of the Obtained Gene Set**: Three types of evaluation are done-
   - Machine Learning Algorithms, namely Logistic Regression, Support Vector Machines, Random Forest, Voting (LR, SVM, RF), and Stacking (LR, SVM, RF; meta classifier - LR).
   - Use of Explainable AI (SHAP) to explain the importance of the features in ML models and understand the relevant biological significance with the associated disease.
   - Enrichment Analysis of some datasets to understand the biological relevance of the selected features with the associated disease using tools such as DAVID, Metascape, EnrichR, Toppgene, and g:Profiler.

III. Main Results

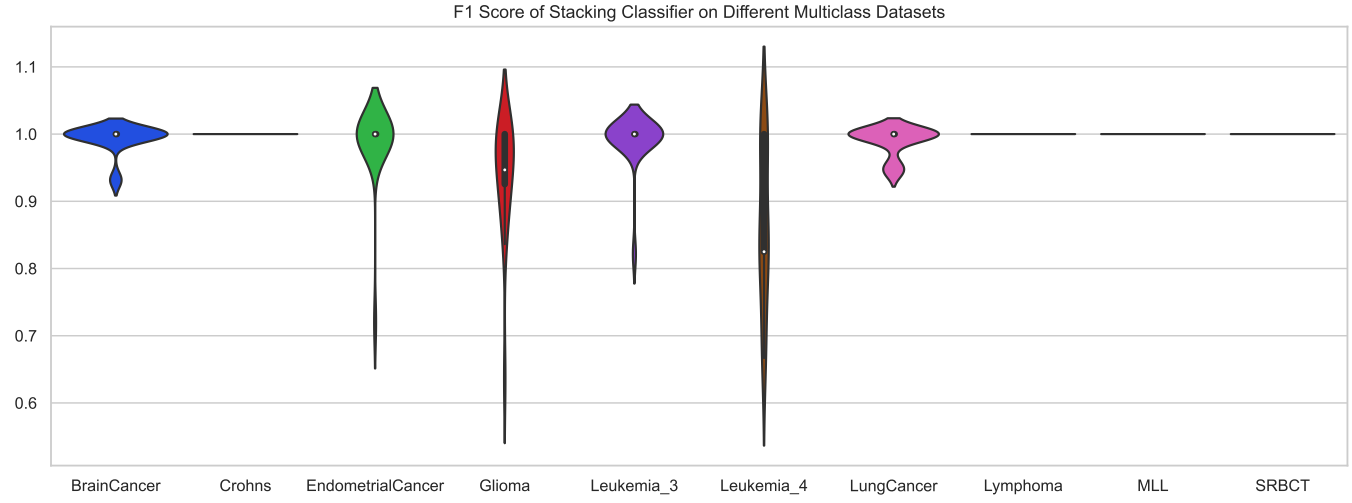
A. Results of Machine Learning Algorithm

1) Self Evaluation: We have added the violin plots of F1 score of all the datasets in Figure 1 in two subplots, Figure 1a showing the results of binary datasets and Figure 1b of multi-class datasets.

From Figure 1, we can see that the datasets Adenocarcinoma, Lymphoma(both binary and multi-class), Ovarian Cancer, Crohn disease, MLL, and SRBCT achieve a perfect score. A few datasets, such as Breast Cancer and Colon Tumor among the binary datasets, and Glioma and Leukemia4 among



(a) Binary Datasets



(b) Multi-class Datasets

Fig. 1: F1 Score of All Datasets Obtained from the Stacking Classifier

the multi-class datasets, do not achieve a stable score but still has an average score more than 85%. The rest of the datasets have a score of more than 90%.

2) Comparison with Previous Methods: We have conducted a comparison of performance with recent literature and confidently assert that our proposed method achieves state-of-the-art results. Furthermore, where additional opportunities for improvement exist, our method surpasses existing benchmarks.

For instance, in the Adenocarcinoma dataset, the proposed method using SVM and Stacking achieved a perfect accuracy of 1.00 with 17.85 features, outperforming methods like ILRC + SVM [10] (0.95 accuracy). In the Breast Cancer dataset, the proposed method (SVM) achieved an accuracy of 0.93 with 82.85 features, whereas the WCLFJHEF (SVC) [9] method achieved a higher accuracy of 0.99 with 25 features. Similarly, for the Leukemia dataset, the proposed method achieved an accuracy of 0.98 with 56.65 features, while the AltWOA method [11] achieved a perfect accuracy of 1.00 with 30 features.

B. SHAP-Based Interpretability of Model Predictions

We conducted SHAP analysis of three binary datasets (Adenocarcinoma, Leukemia, and Ovarian Cancer) and three multiclass datasets (Endometrial Cancer, Leukemia3, and MLL). For all the datasets, we performed an 80-20 stratified split, training the model on 80% of the data and conducting the analysis on the remaining 20%.

We used the KernelExplainer for the binary datasets, given that we employed a Logistic Regression model. KernelExplainer is versatile and works with any type of model. For the multiclass datasets, we used the TreeExplainer to take advantage of its specialized features for multi-class datasets. We

have used an XGBoost model for the multiclass datasets. Some key findings related to the Leukemia dataset are discussed below:

Leukemia: The summary plot of the Logistic Regression model for the Leukemia dataset is given in Figure 2.

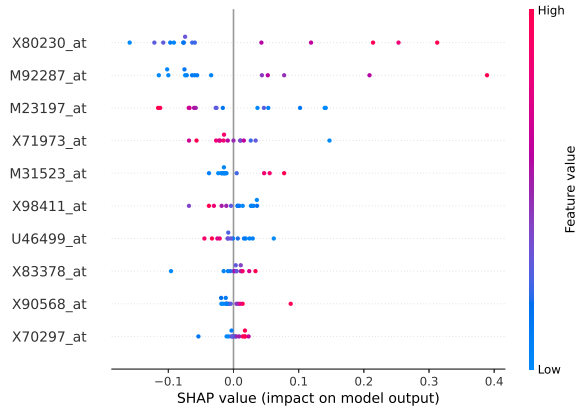


Fig. 2: SHAP summary plot for Leukemia dataset

The first prominent gene is X80230_at which has the gene symbol CDK9 (Cyclin Dependant Kinase 9). A study published in Blood [6] discusses the formation of novel CDK9 in AML and their role in leukemogenesis.

The other two prominent genes are M92287_at which is CCND3, the Cyclin D3 gene and M23197_at which is CD33 gene. Their correlation to Acute Leukemias has been topic of study in several papers such as [7] and [8].

C. Biological Relevance through Enrichment Analysis

Specifically, we selected two binary datasets (Adenocarcinoma and Prostate) with positive and negative classes to showcase our analysis. Notably, we refrained from conducting enrichment analysis on multiclass datasets due to the lack of coherent insights it provides. For binary datasets encompassing various cancer subtypes, we performed enrichment analysis. However, the pathways identified primarily underscored the relevance to the broader cancer category rather than individual subtypes.

We discuss some key findings from the Prostate dataset.

Prostate Cancer: The bar plot in Figure 3 illustrates the selected enrichment clusters for the prostate cancer dataset.

The genes that appeared most frequently in the disease-related enrichment terms are HPN, SPINK1, TGM2, TP63, AGR2, AMACR, and NPY. The relevance of these genes to prostate cancer has been discussed in many studies including and not limited to [12] [13] [14].

IV. Conclusion

This study introduces a two-step feature selection process that improves model performance and generalizability by identifying significant biomarkers from gene expression datasets, with SHAP values providing crucial feature importance insights. Additionally, the pipeline can be integrated into a

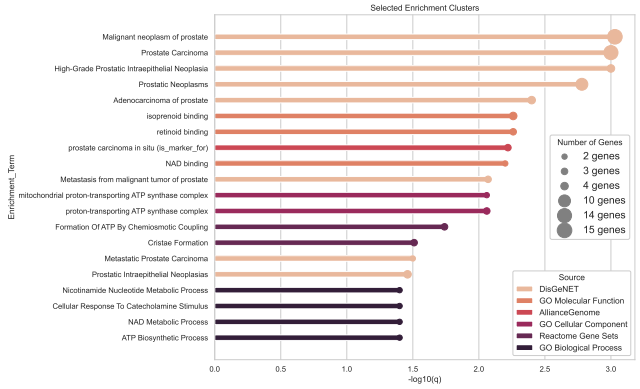


Fig. 3: Enrichment Clusters of Prostate Cancer Dataset

web-based platform, democratizing advanced bioinformatics analysis and fostering interdisciplinary collaboration.

References

[1] Moshood A. Hambali, Tinuke O. Oladele, Kayode S. Adewole, "Microarray cancer feature selection: Review, challenges and research directions", International Journal of Cognitive Computing in Engineering, Volume 1, 2020, Pages 78-97

[2] C. Tang et al., "Feature Selective Projection with Low-Rank Embedding and Dual Laplacian Regularization," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 9, pp. 1747-1760, 1 Sept. 2020

[3] Hakan Ezgi Kiziloz, "Classifier ensemble methods in feature selection", Neurocomputing, Volume 419, 2021, Pages 97-107

[4] Saeid Azadifar, Mehrdad Rostami, Kamal Berahmand, Parham Moradi, Mourad Oussalah, "Graph-based relevancy-redundancy gene selection method for cancer diagnosis", Computers in Biology and Medicine, Volume 147, 2022

[5] Scott M. Lundberg and Su-In Lee. 2017. "A unified approach to interpreting model predictions", In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc.

[6] Beauchamp EM, Abedin SM, Radecki SG, Fischietti M, Arslan AD, Blyth GT, Yang A, Lantz C, Nelson A, Goo YA, Akpan I, Eklund EA, Frankfurt O, Fish EN, Thomas PM, Altman JK, Platanias LC. Identification and targeting of novel CDK9 complexes in acute myeloid leukemia. Blood. 2019 Mar 14;133(11):1171-1185.

[7] Ketzer F, Abdelrasoul H, Vogel M, Marienfeld R, Müschen M, Jumaa H, Wirth T, Ushmorov A. CCND3 is indispensable for the maintenance of B-cell acute lymphoblastic leukemia. Oncogenesis. 2022 Jan 10;11(1):1.

[8] Liu J, Tong J, Yang H. Targeting CD33 for acute myeloid leukemia therapy. BMC Cancer. 2022 Jan 3;22(1):24.

[9] Sukriti Roy, Joginder Singh, Shubhra Sankar Ray, Weighted Combination of Łukasiewicz implication and Fuzzy Jaccard similarity in Hybrid Ensemble Framework (WCLFJHEF) for Gene Selection, Computers in Biology and Medicine, Volume 170, 2024.

[10] K. Yu, W. Xie, L. Wang, and W. Li, "Ilrc: a hybrid biomarker discovery algorithm based on improved l1 regularization and clustering in microarray data," BMC Bioinformatics, 2021.

[11] R. Kundu, S. Chattopadhyay, E. Cuevas, and R. Sarkar, "Altwoa: Altruistic whale optimization algorithm for feature selection on microarray datasets," Computers in Biology and Medicine, vol. 144, p. 105349, 2022.

[12] P. Kielb et el., "Novel histopathological biomarkers in prostate cancer: Implications and perspectives," Biomedicines, vol. 11, no. 6, 2023.

[13] J. A. Magee et el., "Expression Profiling Reveals Hepsin Overexpression in Prostate Cancer1," Cancer Research, vol. 61, pp. 5692–5696, 08 2001.

[14] S. R. Reynolds, Z. Zhang, L. A. Salas, and B. C. Christensen, "Tumor microenvironment deconvolution identifies cell-type-independent aberrant dna methylation and gene expression in prostate cancer," Clinical epigenetics, 2024.