CZ-01

Biomedical Deep Learning - A staged approach using trustworthy deep learning for multi-omics data classification

Abstract

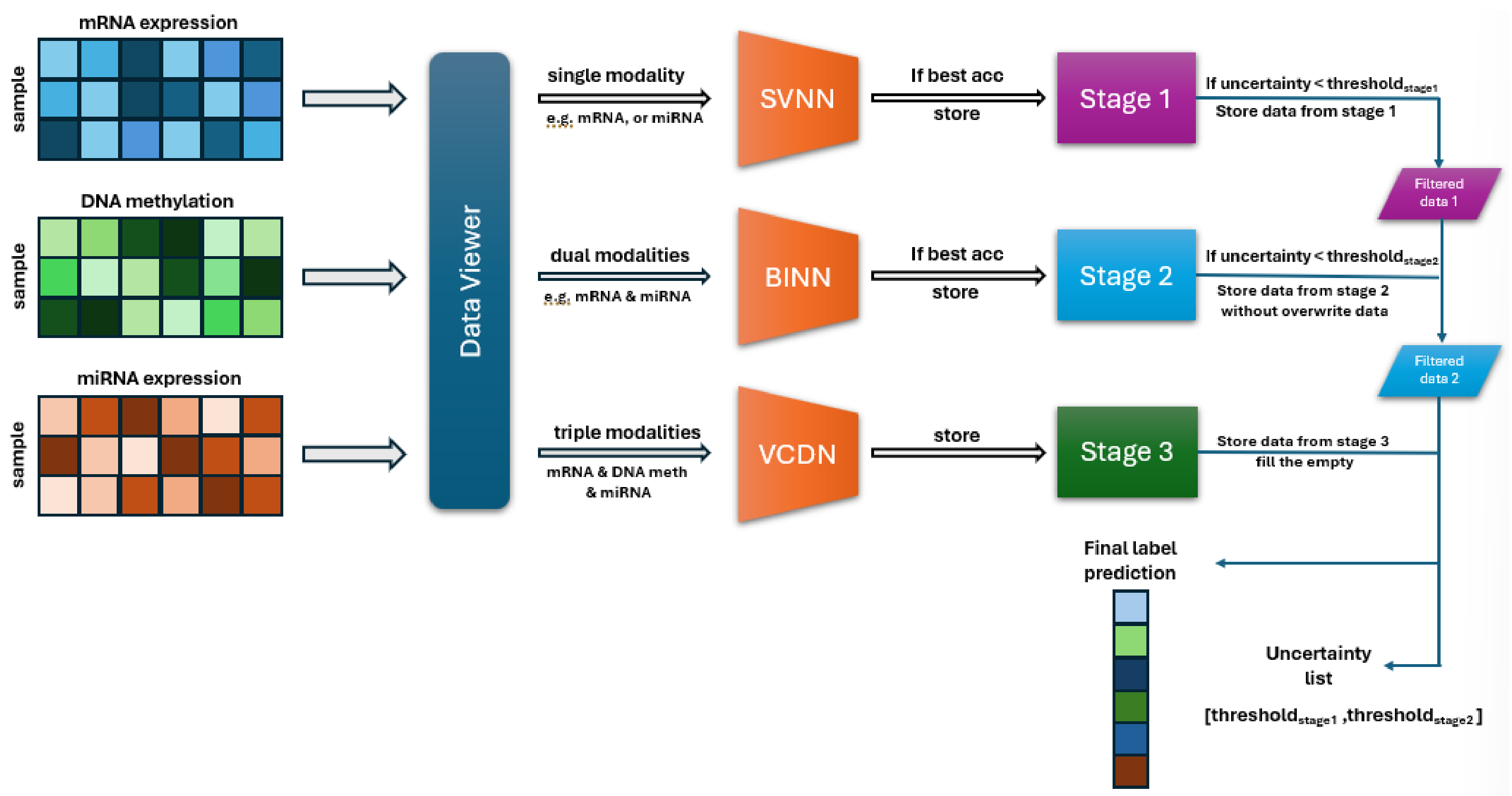
Genetic data like mRNA, miRNA, and DNA methylation provide valuable insights into disease mechanisms and improve diagnostic accuracy. Combining these data types enables a multi-dimensional approach to biomarker discovery, which can lead to earlier, more precise diagnoses. However, integrating multiple modalities raises clinical costs. Unlike past methods, our model selectively uses partial modalities when feasible. Utilizing subjective logic and trustworthy deep learning in a staged approach, we predict disease risk. Our research developed effective modality combinations for single and bi-view models, optimized a multi-perception layer for single-view classification, and created methods to quantify and manage uncertainty in incomplete multi-omics integration.

Introduction

Omics research (genomics, transcriptomics, proteomics, etc.) has transformed understanding of complex biological systems, supporting breakthroughs in biomarker discovery and personalized medicine. Single-omics approaches often miss complex biological interactions, limiting their ability to fully capture disease mechanisms. Multi-omics integration combines diverse data, enhancing insights by providing a holistic view; for example, combining genomics and proteomics uncovers regulatory mechanisms and improves predictions in diseases like cancer. Despite its advantages, multi-omics faces challenges, including data heterogeneity, computational demands, and high costs. This study introduces a staged approach for selectively integrating mRNA, DNA methylation, and microRNA data based on predictive uncertainty, aiming to enhance accuracy while reducing costs. By dynamically adjusting thresholds, the model adapts to real-time performance, avoiding unnecessary data use and enabling a more efficient multi-omics analysis tailored to specific clinical needs.

Materials and Methods

In this section, we introduce SATD for AD diagnosis, which is designed as a binary classification task. An overview of SGUQ is shown in Figure 1.



Results

Base Omics Selection

- The best base omics for BRCA, ROSMAP, LGG are mRNA, for KIPAN is DNA methylation.

Dataset	Features	Classifier dims	Accuracy	F1 weighted	F1 macro	Uncertainty
BRCA	mRNA	1024-512-256	0.8397	0.8434	0.7926	0.4381
BRCA	methy	1024-512-256	0.7443	0.7256	0.6233	0.4917
BRCA	miRNA	512-512-256	0.7175	0.6874	0.5434	0.5617
BRCA	mRNA, methy	1024-512-256	0.8168	0.8161	0.7723	0.1910
BRCA	mRNA, miRNA	128-128-128	0.7977	0.7699	0.6258	0.4148
BRCA	methy-miRNA	512-256-128	0.7443	0.7059	0.5634	0.4364
KIPAN	mRNA	512-512-256	0.9645	0.9612	0.9157	0.0675
KIPAN	mehy	1024-512-256	1.0	1.0	1.0	0.0966
KIPAN	miRNA	1024-512-256	0.9746	0.9742	0.9597	0.0461
KIPAN	mRNA, methy	512-512-256	1.0	1.0	1.0	0.0086
KIPAN	mRNA, miRNA	1024-512-256	0.9848	0.9843	0.9669	0.0181
KIPAN	methy-miRNA	256-128-32	1.0	1.0	1.0	0.0639

Dataset	Features	Classifier dims	Accuracy	F1	AUC	Uncertainty
ROSMAP	mRNA	256-128-32	0.8381	0.8411	0.8382	0.5608
ROSMAP	mehy	512-512-256	0.7524	0.7347	0.7549	0.3513
ROSMAP	miRNA	128-64-32	0.7524	0.7592	0.7522	0.5606
ROSMAP	mRNA, methy	1024-512-256	0.8571	0.8598	0.8573	0.1440
ROSMAP	mRNA, miRNA	256-256-128	0.8667	0.8679	0.8671	0.3765
ROSMAP	methy-miRNA	256-128-32	0.7714	0.7736	0.7718	0.2502
LGG	mRNA	256-256-128	0.8289	0.8375	0.8281	0.4317
LGG	mehy	256-128-64	0.8026	0.80	0.8035	0.6951
LGG	miRNA	64-64-32	0.8223	0.8280	0.8221	0.5870
LGG	mRNA, methy	256-256-128	0.8289	0.8333	0.8288	0.2316
LGG	mRNA, miRNA	128-128-128	0.8487	0.8535	0.8484	0.3521
LGG	methy-miRNA	64-64-32	0.7960	0.7947	0.7968	0.5952

Highest Accuracy Multi-omics Selection

- The best performance for multi-omics combination based on each dataset (BRCA, ROSMAP, LGG, KIPAN).

Dataset	Features	Classifier dims	Accuracy	F1 weighted	F1 macro	Uncertainty
KIPAN	mehy	1024-512-256	1.0	1.0	1.0	0.0966
KIPAN	mRNA, methy	512-512-256	1.0	1.0	1.0	0.0086
KIPAN	mRNA, methy, miRNA	512-256-128	0.9848	0.9847	0.9821	0.0
BRCA	mRNA	1024-512-256	0.8397	0.8434	0.7926	0.4381
BRCA	mRNA, methy	1024-512-256	0.8168	0.8161	0.7723	0.1910
BRCA	mRNA, methy, miRNA	64-64-32	0.8855	0.8878	0.8579	0.0

Dataset	Features	Classifier dims	Accuracy	F1	AUC	Uncertainty
LGG	mRNA	256-256-128	0.8289	0.8375	0.8281	0.4317
LGG	mRNA, miRNA	128-128-128	0.8487	0.8535	0.8484	0.3521
LGG	mRNA, methy, miRNA	256-128-32	0.8289	0.8289	0.8295	0.0
ROSMAP	mRNA	256-128-32	0.8381	0.8411	0.8382	0.5608
ROSMAP	mRNA, miRNA	256-256-128	0.8667	0.8679	0.8671	0.3765
ROSMAP	mRNA, methy, miRNA	512-256-128	0.8476	0.8545	0.8469	0.0

Accuracy & Threshold Determination

- The best uncertainty thresholds(threshold 1 and threshold 2) based on the selected result performance.

Dataset	Best accuracy	F1 weighted	F1 macro	Best threshold 1	Best threshold 2
BRCA	0.8855	0.8434	0.7926	0.0980	0.0020
KIPAN	1.0	1.0	1.0	0.0395	0.0718

Dataset	Best accuracy	F1	AUC	Best threshold 1	Best threshold 2
ROSMAP	0.8857	0.8411	0.8382	0.2673	0.4309
LGG	0.8487	0.8375	0.8281	0.1470	0.5043

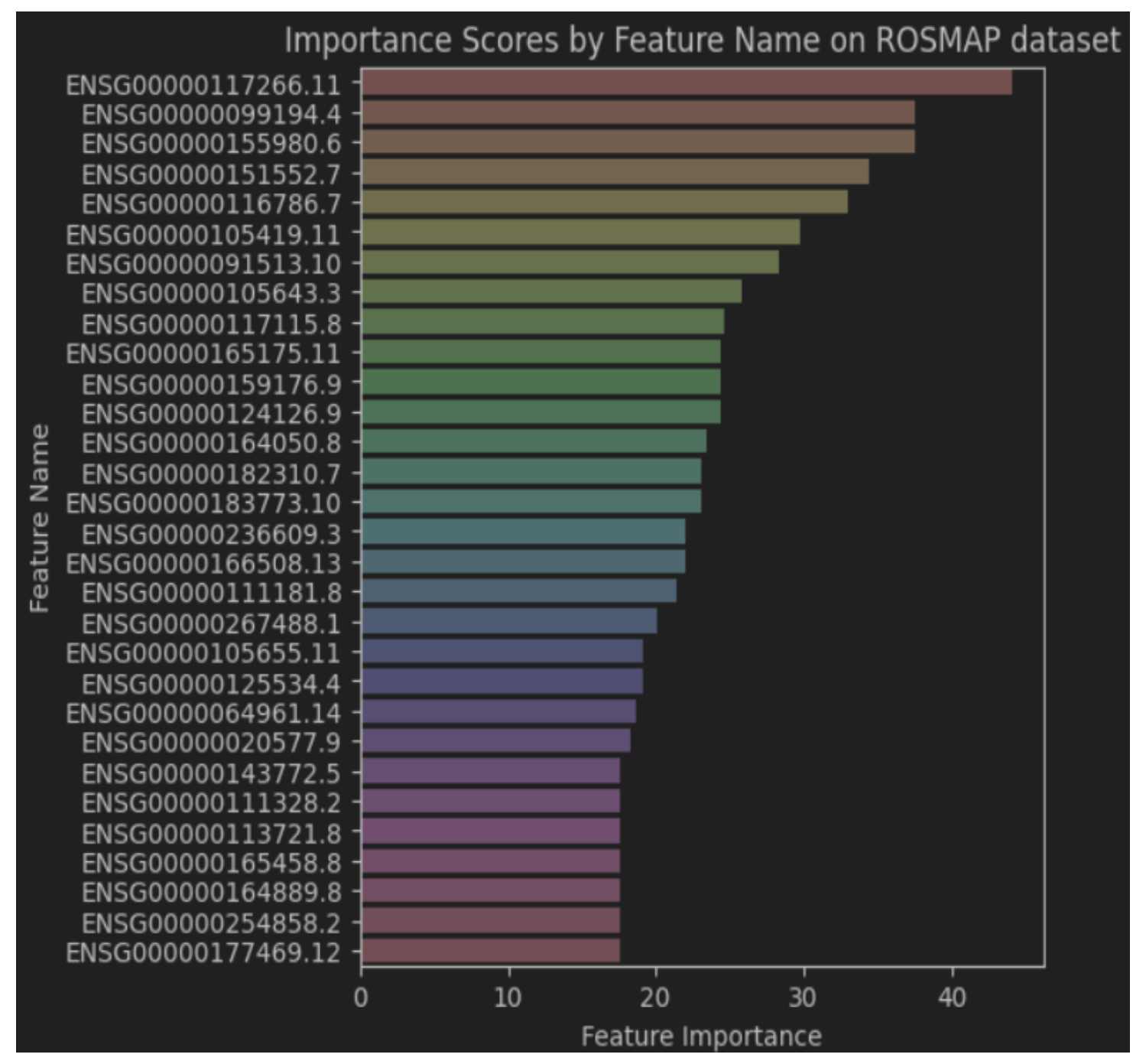
Modality Usage Percentage

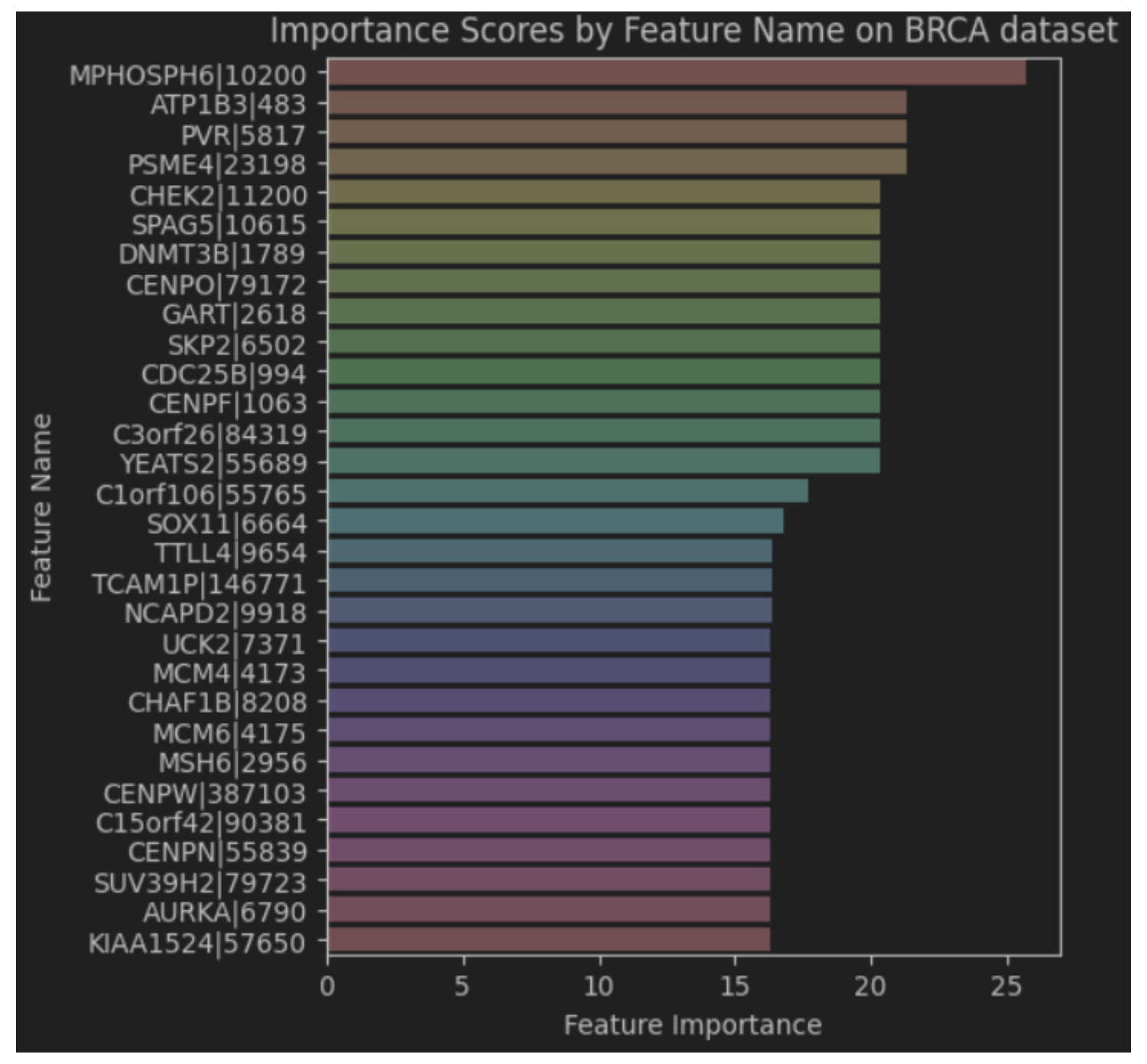
- The percentage of each modality usage for each dataset (BRCA, ROSMAP, LGG, KIPAN).

Dataset	Stage 1	Stage 2	Stage 3
ROSMAP	mRNA | 0.95%	mRNA | miRNA | 65.71%	mRNA | methy | miRNA | 33.33%
BRCA	mRNA | 0.38%	mRNA | methy | 0.38%	mRNA | methy | miRNA | 99.24%
LGG	mRNA | 0.66%	mRNA | miRNA | 99.34%	mRNA | methy | miRNA | 0%
KIPAN	methy | 0.51%	mRNA | methy | 98.48%	mRNA | methy | miRNA | 1.02%

Feature Performance Analysis

- The feature performance example based on ROSMAP and BRCA dataset.





Conclusions

In this study, we introduced a novel staged deep learning approach that achieves low-cost, high-performance classification across various disease datasets through a phased integration of multi-omics data. Based on our experimental results, we observed that most datasets achieved high-performance predictions even with limited omics data, while still ensuring enhanced predictive accuracy for integrated multi-omics data. Specifically, in stage 2, the ROSMAP dataset used 65.71% of the data with an accuracy of 0.8857. The LGG dataset used 99.34% of the data in stage 2, achieving an accuracy of 0.8487. The KIPAN dataset used 98.48% of the data in stage 2 with a perfect accuracy of 1.0. However, the BRCA dataset did not meet the experimental criteria, with 99.24% data usage in stage 3 and an accuracy of 0.8855. Collectively, these findings highlight the potential of our staged deep learning framework for diverse disease diagnostics and demonstrate broader applications of multi-view data in clinical decision-making. Our work paves the way for more efficient and cost-effective strategies in disease detection and management.

Contact Information

Chen Zhao: czhao4@kennesaw.edu
Tianze Liu: tliu11@students.kennesaw.edu
Yongbo An: yan2@students.kennesaw.edu

References

Wang, T., Shao, W., Huang, Z. et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun 12, 3445 (2021). https://doi.org/10.1038/s41467-021-23774-w

P. P. Liang, Z. Liu, R. Salakhutdinov, and L.-P. Morency, "Multibench: Multiscale benchmarks for multimodal representation learning," arXiv preprint arXiv:2102.02051v1, 2021. [Online]. Available: https://arxiv.org/abs/2102.02051v1

A. Vaswani et al., "Attention is all you need," arXiv preprint arXiv:1806.01768, 2017. [Online]. Available: https://arxiv.org/abs/1806.01768

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author(s): Yongbo An, Tianze Liu
Advisors(s): Chen Zhao