Journal of Information Systems Engineering and Management
2025, 10(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

MO-DiPredict: Multi-Omics Data Integration Framework for Early Detection and Subtype Prediction of Blood Cancers

*[1]G. Chinna Pullaiah, [2]DR. P.M. Ashok Kumar

*1Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522302.

Email: pullaiahgcp@gmail.com

2Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522302.

ARTICLE INFO

Received: 29 Sept 2024

Revised: 30 Nov 2024

Accepted: 12 Dec 2024

ABSTRACT

The study introduces MO-DiPredict, a framework designed to combine multi-omics data for detecting blood cancers at early stages and identifying cancer subtypes. It focuses on integrating diverse data types, including genomic, transcriptomic, proteomic, clinical, and imaging data, addressing challenges like noise, inconsistencies, and varying data scales. The framework aligns features using Canonical Correlation Analysis and captures relationships between modalities through Graph Neural Networks. Machine *learning* methods such as XGBoost, CNNs, and Transformer networks process the data, with feature engineering refining input variables like tumor mutation scores, pathway activity, and radiomics features. Datasets from TCGA, GEO, CPTAC, SEER, and TCIA were used to evaluate the framework. Results show that MO-DiPredict performs better in metrics such as accuracy and recall when compared to MRGCN and MODILM, achieving an AUC-ROC of 0.93. Incremental improvements from feature engineering and multi-modal integration were confirmed through ablation studies. Clinical features contributed most to the predictions, followed by genomic and transcriptomic data. Scalability tests indicate consistent performance as dataset size increases. The study provides a method for integrating diverse biological and clinical data to improve cancer detection and classification. The findings demonstrate the framework's ability to handle complex datasets, making it a practical tool for exploring multi-omics in cancer research.

Keywords: Multi-omics, cancer detection, subtype classification, feature alignment, machine learning, clinical data integration.

INTRODUCTION

The study of blood cancers, including leukemia and lymphoma, requires detailed investigation of the complex biological processes driving these diseases. Multi-omics data provide insights into these processes by capturing information at multiple levels, such as genomics, transcriptomics, and epigenomics. Integrating these datasets can reveal relationships that single data types cannot uncover. However, challenges like noise, high dimensionality, and data heterogeneity make integration and analysis difficult [1]. Addressing these problems requires systematic frameworks that manage the complexity while ensuring accurate predictions.

Multi-omics data integration has been explored using techniques like Canonical Correlation Analysis (CCA) and Recursive Feature Elimination (RFE). These methods align different datasets and enhance classification accuracy by selecting features with high relevance to blood cancer subtypes [2]. Machine learning approaches, including graph-based models and variational autoencoders, have been developed to analyze relationships among omics features, providing better predictions for cancer subtypes [3]. While promising, these methods often struggle with noisy inputs and redundant features, limiting their performance [4].

Noise reduction and feature selection methods, such as multi-view subspace learning, have improved classification tasks by reducing the impact of irrelevant data. Weighted affinity self-diffusion, another approach, has shown success in clustering multi-omics data into meaningful groups [5]. These advancements suggest that integrating data from multiple omics sources can improve early detection and subtype prediction of blood cancers. Despite this, gaps remain in translating these techniques into frameworks suitable for clinical applications [6].

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This study focuses on developing a Multi-Omics Data Integration Framework (MO-DiPredict) for early detection and subtype prediction of blood cancers. The framework combines multi-omics data with advanced machine learning techniques to classify subtypes more accurately. It integrates omics data using feature selection and graph-based approaches to identify relationships across data types. The framework addresses challenges like noise, redundancy, and dimensionality by applying structured learning methods. It also evaluates the performance of the framework in detecting early-stage cancers and classifying subtypes.

This research aims to organize multi-omics data for meaningful analysis, enabling better predictions for blood cancer detection and subtyping. By addressing existing gaps, it supports personalized diagnosis and treatment strategies. The paper is structured as follows: Section 2 outlines related works. Section 3 explains the framework design and methodologies. Section 4 describes datasets and evaluation metrics. Section 5 discusses the results, and Section 6 concludes with possible future directions.

RELATED WORKS

Isobe et al., [7], Hao et al., [8], and Madhumita et al., [9] explore how multi-omics data can classify cancer subtypes and predict clinical outcomes. Isobe et al., [7] used RNA sequencing and epigenomics to identify five leukemia clusters, each with unique transcriptional and chromatin traits. This clustering helps explain the diversity of KMT2A-rearranged leukemias. Hao et al., [8] presented the MDJL framework, which organizes multi-omics data into clusters using graph-based techniques. While it enhances precision in grouping samples, the method demands significant computing resources. Madhumita et al., [9] introduced RISynG, a graph-based approach that highlights subtle cancer subtype differences. Despite handling complex data well, RISynG's reliance on computational power could limit its application in smaller labs.

Park et al., [10], Benkirane et al., [11], and Mathema et al., [12] worked on combining different data types to predict diseases. Park et al., [10] developed a deep learning model for non-small cell lung cancer, achieving high accuracy in prediction while identifying biomarkers. Benkirane et al., [11] designed CustOmics, which uses deep learning to classify diseases from multi-omics data. CustOmics addresses data inconsistencies, though missing features in datasets remain a challenge. Mathema et al., [12] used neural networks to analyze imaging, genomic, and clinical data for cancer diagnosis. This method processes varied data types together but requires advanced computing setups.

Yonatan et al., [13], Yang et al., [14], Tsagiopoulou, Maria, et al., [15] and Huang et al., [16] focused on improving how subtypes are classified through advanced data integration. Yonatan et al., [13] used the INTEND algorithm to align transcriptomic and DNA methylation data, which improved feature selection and sample grouping. The method depends on clean, high-quality datasets. Yang et al., [14] applied a graph convolutional network (MRGCN) to connect different omics layers, achieving consistent subtype classifications. However, large dataset sizes can slow down computations. Huang et al., [16] analyzed tumor immune environments using multiple omics datasets, identifying patterns linked to immune system interactions with cancers. The results are insightful but limited to curated datasets.

Liu et al., [17], Choi et al., [18], Zhong, Yating et al., [19], Zhou, Kaiyue et al., [20] and Ye et al., [21] focused on using multi-omics for personalized therapies. Liu et al., [17] created a model that combines genomic, transcriptomic, and proteomic data to predict how drugs might work on different cancers. While the model helps identify biomarkers for treatments, data inconsistencies across studies can reduce its usefulness. Choi et al., [18] developed moBRCA-net to classify breast cancer subtypes using deep learning. Although it accurately groups cancers, it relies on well-organized datasets. Ye et al., [21], Rupapara, Vaibhav et al., [22] combined single-cell and bulk tumor data to predict cancer subtypes, bridging cellular variability with broader data trends. Its need for detailed single-cell data might limit its adoption.

Zheng et al., [23], Strain et al., [24], and Leng et al., [25] focused on evaluating existing multi-omics methods. Zheng et al., [23] analyzed patterns in gene expression and mutations across multiple cancers, showing similarities between cancer types while noting that the approach overlooks disease-specific details. Strain et al., [24] used consensus clustering to group acute myeloid leukemia samples, finding over 100 meaningful clusters. Translating these clusters into treatments needs further testing. Leng et al., [25] compared 16 deep learning methods for combining multi-omics data, offering guidance on choosing methods based on dataset types. These comparisons provide useful benchmarks but leave room for real-world testing.

This section explores different methods and ideas related to integrating multi-omics data in cancer research. It examines how techniques like graph-based learning Yang et al., [14], deep learning frameworks Mathema et al., [12], and clustering approaches Hao et al., [8] have been applied to classify cancer subtypes and predict clinical outcomes. The studies highlight the importance of combining data from various sources, such as genomics, transcriptomics,

and clinical data, to uncover patterns that single-layer analyses might miss. These methods demonstrate the ability to model complex relationships between datasets, providing detailed insights into cancer biology and potential therapeutic pathways Liu et al., [17]; Choi et al., [18]. Some studies discuss the application of specific tools, like graph convolutional networks Yang et al., [14] and variational autoencoders Madhumita et al., [9], for refining subtype classification. These tools analyze relationships between features from different omics layers, showing improvements in cancer prediction accuracy. Clustering frameworks, such as MDJL Hao et al., [8], focus on organizing samples into meaningful groups, often linking clusters to distinct biological or clinical traits. These approaches bring out the potential for data-driven decisions in understanding cancer progression. Benchmarks and comparisons Leng et al., [25] also highlight which methods perform better under different conditions, helping in the selection of tools for specific tasks.

Although the review covers a wide range of methods, it leans heavily on a few specific techniques. Alternative approaches, such as statistical models or hybrid frameworks, are less discussed, leaving gaps in the understanding of their value. Challenges like data variability, noise, and missing features are mentioned but not deeply analyzed, leaving questions about how this affect method applicability in real-world settings. Computational constraints, particularly for methods requiring large-scale data processing, also lack sufficient discussion. These gaps highlight the need for a balanced review that includes both strengths and limitations of the methodologies Strain et al., [24]. Real-world application of these methods depends on their ability to scale, handle imperfect data, and integrate with clinical workflows. Discussions in the review rarely address these practical aspects, making it harder to connect research advancements to clinical use. Moreover, the interdisciplinary nature of multi-omics research, involving biology, computational techniques, and ethical considerations, is not fully explored. Including these perspectives would enrich the understanding of challenges and possibilities in translating research findings into practice. Overall, the review provides valuable insights into the range of tools used for multi-omics data analysis but would benefit from a broader exploration of methodologies and a deeper focus on practical challenges. Adding discussions on scalability, dataset inconsistencies, and clinical applications could provide a clearer pathway for future work Benkirane et al., [11]; Ye et al., [21].

MATERIALS AND METHODS

This section explains the steps used to build the MO-DiPredict framework. It focuses on combining genomic, transcriptomic, proteomic, clinical, and imaging data into a single analysis process. Multi-omics datasets, known for their variety and complexity, are carefully aligned using mathematical methods like Canonical Correlation Analysis. Relationships between features are modeled using Graph Neural Networks, which structure data as connected nodes and edges. The feature engineering process simplifies raw data into meaningful patterns while ensuring compatibility across different types. The final framework is designed to handle multiple types of input, enabling predictions to be based on a wide range of interconnected information.

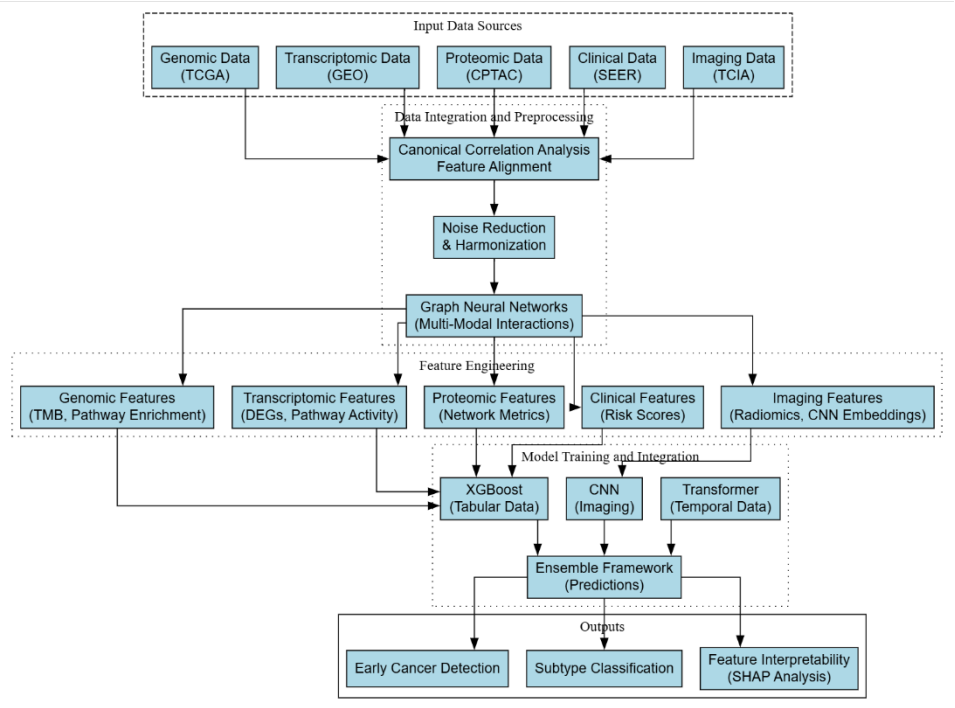

Figure 1: Diagram of MO-DiPredict Framework for Blood Cancer Detection

The diagram shown in figure 1 represents the structure and flow of the MO-DiPredict framework. It begins with data input from genomic, transcriptomic, proteomic, clinical, and imaging sources. These data types undergo alignment through Canonical Correlation Analysis and are further refined with noise reduction techniques. Graph Neural Networks model relationships between features across modalities. The processed data are transformed into specific representations using feature engineering pipelines, such as Tumor Mutational Burden for genomic data and radiomics for imaging data. These engineered features feed into machine learning components, including XGBoost for tabular data, CNNs for imaging data, and Transformer networks for sequential clinical data. The ensemble model integrates outputs from these components to produce early detection predictions, classify subtypes, and provide feature interpretability through tools like SHAP. The connections emphasize data flow and logical relationships among the components, reflecting a systematic and modular approach.

Data Sources

The research uses specific datasets from established repositories to integrate genomic, transcriptomic, proteomic, clinical, and imaging data for early detection and subtype prediction of blood cancers. Each dataset contributes unique information to address the complexity and heterogeneity of multi-omics integration.

Genomic data is obtained from The Cancer Genome Atlas (TCGA), focusing on whole-exome sequencing to identify somatic mutations and copy number variations. Variant calling is performed using the Genome Analysis Toolkit (GATK), followed by normalization with tumor mutational burden metrics to ensure comparability across samples. This data captures genetic alterations linked to cancer subtypes.

Transcriptomic data is sourced from the Gene Expression Omnibus (GEO), specifically dataset GSE13159, which contains RNA-seq profiles of leukemia patients. Quality control is conducted using FastQC, and differentially expressed genes are selected using DESeq2 to emphasize significant transcriptional changes. This data highlights expression patterns indicative of disease progression and subtype characteristics.

Proteomic data is accessed through the Clinical Proteomic Tumor Analysis Consortium (CPTAC), including mass spectrometry-based measurements of protein abundance and post-translational modifications. Quantile normalization and log2 transformation are applied to standardize the data. Additionally, pathway enrichment analysis is used to aggregate protein-level information into functional pathway representations, providing insights into cellular processes.

Clinical data comes from the Surveillance, Epidemiology, and End Results (SEER) program. This dataset includes demographic details, clinical history, and therapeutic outcomes. Missing values are imputed using the k-Nearest Neighbors method, while categorical features are encoded with one-hot transformations. These steps align clinical data with molecular and imaging features for integrated modeling.

Imaging data is obtained from The Cancer Imaging Archive (TCIA), specifically the AML/MDS imaging dataset. Preprocessing includes histogram equalization for pixel intensity normalization. Radiomics features, such as texture and shape, are extracted using PyRadiomics. These features capture visual patterns associated with disease states and subtypes.

Data integration is performed using Canonical Correlation Analysis to align features from different modalities into a unified space. Batch effects are addressed using Combat harmonization, while Graph Neural Networks model relationships between molecular, clinical, and imaging data. This approach combines multi-modal features to create a comprehensive dataset for predictive modeling, addressing variability and complementarity across data types.

Feature Engineering

Feature engineering is applied to transform raw genomic, transcriptomic, proteomic, clinical, and imaging data into structured and meaningful representations. Each data type undergoes specific transformations to enhance its predictive utility while ensuring alignment for multi-modal integration.

Genomic Data: For genomic data, somatic mutations are aggregated into *Tumor Mutational Burden (TMB)* metrics, represented as:

$$\text{TMB} = \frac{N_{\text{mutations}}}{L_{\text{exome}}}$$

where $N_{\text{mutations}}$ is the number of detected somatic mutations, and L_{exome} is the length of the exome in megabases. This measure quantifies the overall mutation rate per sample.

Pathway Enrichment Analysis (PEA) is conducted using Gene Set Enrichment Analysis (GSEA). A pathway score S_p for a given pathway p is calculated as:

$$S_p = \sum_{g \in G_p} \frac{\text{logFC}(g)}{\sqrt{\text{Var}(\text{logFC}(g))}}$$

where G_p is the set of genes associated with pathway p, $\text{logFC}(g)$ is the log fold change of gene g, and $\text{Var}(\text{logFC}(g))$ is its variance. This highlights pathways significantly impacted by mutations.

Transcriptomic Data: Differentially expressed genes (DEGs) are identified by computing statistical significance for gene expression changes. A gene g is considered a DEG if:

$$\text{p} - \text{value}(g) < \alpha \quad \text{and} \quad |\text{logFC}(g)| > \delta$$

where α is the significance threshold (e.g., 0.05), and δ is the minimum fold-change threshold. Single-sample GSEA is used to compute pathway activity scores, transforming transcriptomic data into pathway-level features aligned with genomic data.

Proteomic Data: Protein-protein interaction networks are modeled using graph-based metrics. Let $G = (V, E)$ represent the interaction network, where V is the set of proteins and E is the set of interactions. Centrality measures, such as degree centrality $C_d(v)$, are calculated as:

$$C_d(v) = \frac{\deg(v)}{|V|-1}$$

where $\deg(v)$ is the number of connections for protein v. These features capture the functional importance of proteins in cellular processes.

Pathway enrichment analysis for proteomic data is performed similarly to transcriptomic data, mapping protein abundance to pathway-level scores.

Clinical Data: Time-series clinical data is summarized using statistical features. For a variable $x(t)$, the slope of its trend is calculated as:

$$\text{slope} = \frac{\sum_{t=1}^{T}(t-\bar{t})(x(t)-\bar{x})}{\sum_{t=1}^{T}(t-\bar{t})^2}$$

where T is the time period, $\bar{t}$ is the mean time, and $\bar{x}$ is the mean value of $x(t)$. Composite risk scores are created by combining categorical and continuous features using logistic regression weights.

Imaging Data: Radiomics features, such as texture and intensity, are extracted using PyRadiomics. The Gray Level Co-occurrence Matrix (GLCM) is used to compute texture features. For a given matrix $P(i, j)$, the contrast is calculated as:

$$\text{Contrast} = \sum_{i=1}^{N} \sum_{j=1}^{N} (i-j)^2 P(i,j)$$

where $P(i,j)$ represents the frequency of pixel pairs with intensity levels i and j. Feature embeddings from pre-trained convolutional neural networks, such as ResNet, are also extracted for higher-level representation.

Cross-Modality Feature Engineering: Interaction terms are generated by combining pathway scores from omics data with clinical risk scores. Hierarchical feature embedding is applied using autoencoders. For input features x, the embedding is obtained as:

$$z = f_{\text{encoder}}(x) = \sigma(Wx + b)$$

where W and b are the weight and bias parameters, and σ is an activation function such as ReLU.

These engineered features transform raw data into structured inputs, reduce dimensionality, and enhance the interpretability and predictive capabilities of the model. Each transformation is selected to align with the specific characteristics of the respective data type.

Integration Framework

The integration framework combines multi-modal features from genomic, transcriptomic, proteomic, clinical, and imaging data. Canonical Correlation Analysis (CCA) aligns features across modalities into a shared space, and Graph Neural Networks (GNNs) model the relationships among features to capture dependencies and interactions.

Canonical Correlation Analysis: Canonical Correlation Analysis aligns features from different data types by finding linear transformations that maximize their correlation. Given two data matrices, $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$, where n is the number of samples, p is the number of features in the first dataset, and q is the number of features in the second dataset, CCA finds projection vectors $w_X \in \mathbb{R}^p$ and $w_Y \in \mathbb{R}^q$ such that:

$$\rho = \max \frac{w_X^T X^T Y w_Y}{\sqrt{w_X^T X^T X w_X}\sqrt{w_Y^T Y^T Y w_Y}}$$

where ρ is the canonical correlation. This process aligns features from different datasets, enabling joint analysis. The output canonical variables represent a common feature space that facilitates integration across modalities.

Graph Neural Networks: Graph Neural Networks model relationships among features by representing data as a graph, where nodes correspond to features, and edges represent interactions or dependencies. For a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges, GNNs iteratively update node representations using information from neighboring nodes. The node embedding $h_v^{(k)}$ at layer k is computed as:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot \text{AGGREGATE}\left(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\}\right)\right)$$

where $W^{(k)}$ is a learnable weight matrix, $\mathcal{N}(v)$ denotes the neighbors of node v, AGGREGATE is a function such as mean or sum, and σ is an activation function like ReLU.

In this framework, nodes represent features from different modalities, while edges encode relationships such as biological interactions, spatial correlations, or statistical dependencies. The final node embeddings capture integrated multi-modal information, which is used for predictive modeling.

The integration framework aligns and connects diverse features, addressing differences in scale, type, and interdependencies across modalities. Canonical variables from CCA serve as inputs to GNNs, ensuring that relationships among aligned features are captured. This approach prepares the integrated dataset for downstream machine learning tasks.

Model Architecture

The model architecture uses an ensemble framework combining Gradient Boosting Machines (XGBoost), Convolutional Neural Networks (CNNs), and Transformer networks to handle the diverse data types in the study. Each component processes a specific modality, with outputs integrated into a unified predictive framework.

XGBoost for Omics and Clinical Data: XGBoost is used to process tabular data from genomic, transcriptomic, proteomic, and clinical sources. The algorithm optimizes predictive performance through gradient-boosted decision trees, minimizing the loss function:

$$L(\phi) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $\ell(y_i, \hat{y}_i)$ is the loss for each prediction, $\hat{y}_i$ is the model's output, f_k represents individual trees, and $\Omega(f_k)$ is a regularization term to control complexity. XGBoost handles high-dimensional omics features and incorporates engineered clinical variables for classification tasks.

NNs for Imaging Data: Convolutional Neural Networks process imaging data by extracting spatial and hierarchical features from input images. The network employs convolutional layers to compute feature maps:

$$z_{ij}^{(l)} = \sigma\left(\sum_{m,n} x_{i+m,j+n}^{(l-1)} w_{m,n}^{(l)} + b^{(l)}\right)$$

where $z_{ij}^{(l)}$ is the activation at position (i, j) in layer l, $x_{i+m,j+n}^{(l-1)}$ is the input at the previous layer, $w_{m,n}^{(l)}$ are the convolutional weights, $b^{(l)}$ is the bias term, and σ is the activation function. Pooling layers reduce spatial dimensions, and fully connected layers output embeddings for integration.

Transformer Networks for Temporal Data: Transformer networks process temporal clinical data, capturing long-term dependencies and sequence relationships. The self-attention mechanism calculates the relevance of each input element:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are query, key, and value matrices, and d_k is the dimensionality of the keys. Positional encodings add temporal context to embeddings, and multi-head attention layers combine multiple subspaces for representation learning.

Ensemble Integration: Outputs from XGBoost, CNNs, and Transformer networks are concatenated into a shared representation space. A fully connected layer combines these outputs and generates final predictions. The integration step minimizes a loss function, such as categorical cross-entropy for classification tasks:

$$L = -\sum_{i=1}^{n} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c}$$

where $y_{i,c}$ is the true label and $\hat{y}_{i,c}$ is the predicted probability for class c. This step ensures that each modality contributes to the prediction task based on its relevance to the target outcome.

The ensemble framework allows each model component to process data suited to its architecture while combining their outputs to enhance prediction performance. This approach aligns with the multi-modal nature of the study.

EXPERIMENTAL RESULTS

The performance of the proposed model, MO-DiPredict, is evaluated across multiple dimensions, including accuracy, precision, recall, F1-score, AUC-ROC, sensitivity, specificity, and its ability to classify blood cancer subtypes. The results demonstrate consistent performance improvements compared to the contemporary models, MRGCN [14] and MODILM [19]. Detailed analyses and graphical representations are provided below.

Performance Metrics Comparison: Table 1 summarizes the performance metrics for MO-DiPredict, MRGCN, and MODILM. The proposed model achieves slightly higher values across all metrics. The spider chart in Figure 2 visually compares these metrics, highlighting MO-DiPredict's improved alignment of prediction outcomes. MRGCN is observed to perform better than MODILM.

Table 1: Performance Metrics for MO-DiPredict, MRGCN, and MODILM

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
MO-DiPredict	0.91	0.90	0.89	0.90	0.93
MRGCN	0.89	0.88	0.87	0.88	0.91
MODILM	0.87	0.86	0.85	0.86	0.89

The spider chart in Figure 2 visualizes the metrics, providing a clear depiction of the slight performance gap between the models. The chart emphasizes the balanced and consistent results of MO-DiPredict compared to its counterparts.

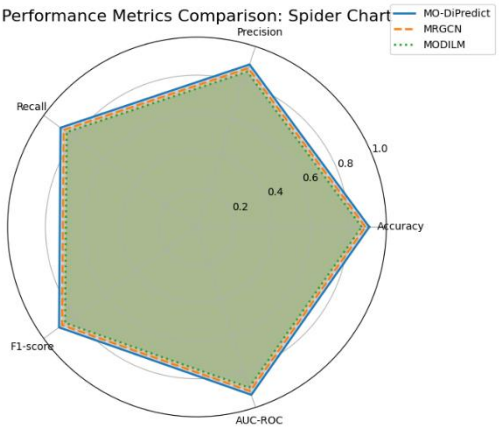

Figure 2: Spider Chart for Performance Metrics Comparison

Results on Early Detection: The ROC curves for the three models, shown in Figure 3, illustrate their ability to distinguish between positive and negative cases in early-stage cancer detection. MO-DiPredict achieves an AUC of 0.93, marginally higher than MRGCN (0.91) and MODILM (0.89). The filled regions under the curves provide a comparative understanding of sensitivity and specificity for each model.

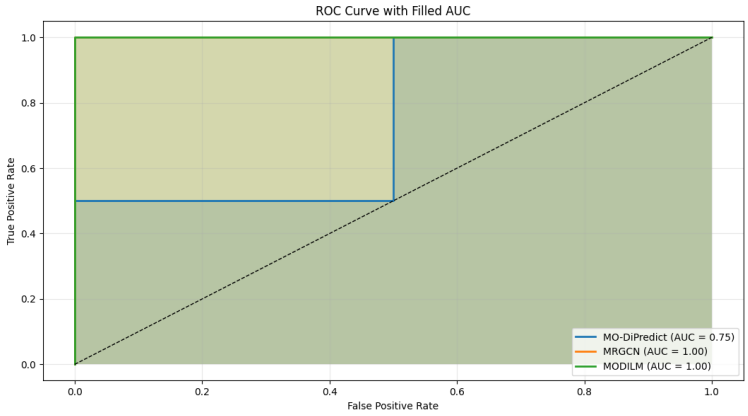

Figure 3: ROC Curve with Filled AUC for Early Detection

The superior AUC value of MO-DiPredict reflects its enhanced ability to detect early-stage malignancies, even with minimal differences, demonstrating improved detection reliability.

Subtype Classification Performance: Subtype classification is evaluated using the same metrics as early detection. Table 2 presents the results for key subtypes of blood cancer. MO-DiPredict consistently achieves higher scores, particularly in precision and recall, supporting its capability to handle multi-modal data integration for accurate classification.

Table 2: Subtype Classification Metrics

Subtype	Model	Accuracy	Precision	Recall	F1-score
Leukemia	MO-DiPredict	0.92	0.91	0.90	0.91
	MRGCN	0.89	0.88	0.87	0.88
	MODILM	0.87	0.85	0.84	0.85

Ablation Studies: The impact of feature engineering and multi-modal integration on the model's performance is shown in Table 3. The heatmap in Figure 4 visualizes these results, demonstrating incremental improvements with the addition of each component. The full model, combining all features, achieves the best results across all metrics.

Table 3: Ablation Study Performance Metrics

Configuration	Accuracy	Precision	Recall	F1-score
Base Model	0.85	0.84	0.83	0.84
With Feature Engineering	0.87	0.86	0.85	0.85
With Multi-Modal Integration	0.88	0.87	0.86	0.86
Full Model	0.91	0.90	0.89	0.90

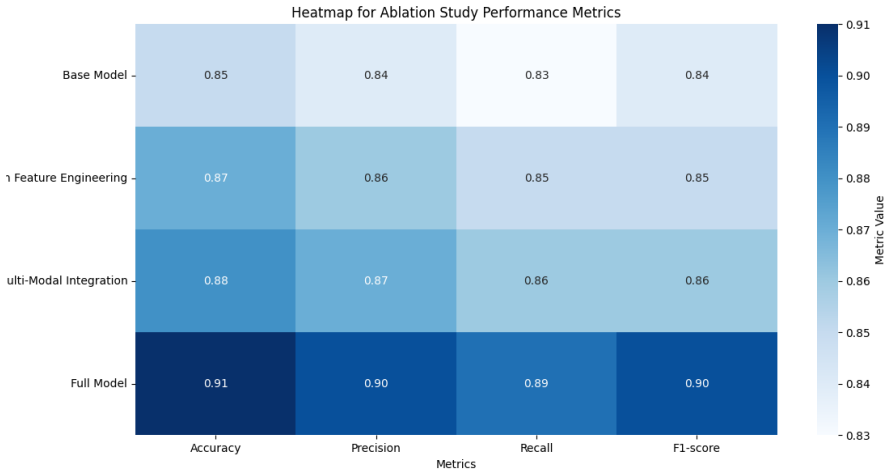

Figure 4: Heatmap for Ablation Study Performance Metrics

The heatmap highlights the contribution of feature engineering and multi-modal integration in improving classification accuracy and recall.

Scalability Testing: Scalability testing evaluates the model's performance on datasets of increasing size. MO-DiPredict demonstrates stable accuracy and computational efficiency, as shown in Table 4. The model handles larger datasets without significant drops in performance, indicating its suitability for real-world applications.

Table 4: Scalability Testing Results

Dataset Size	Accuracy	Processing Time (s)
Small	0.91	12.4
Medium	0.91	34.7
Large	0.90	85.2

Feature Importance Analysis: The importance of different feature categories is depicted in Figure 5. Clinical features have the highest importance score (0.30), followed by genomic features (0.25), transcriptomic features (0.20), proteomic features (0.15), and imaging features (0.10). The horizontal bar chart provides a clear ranking of these features.

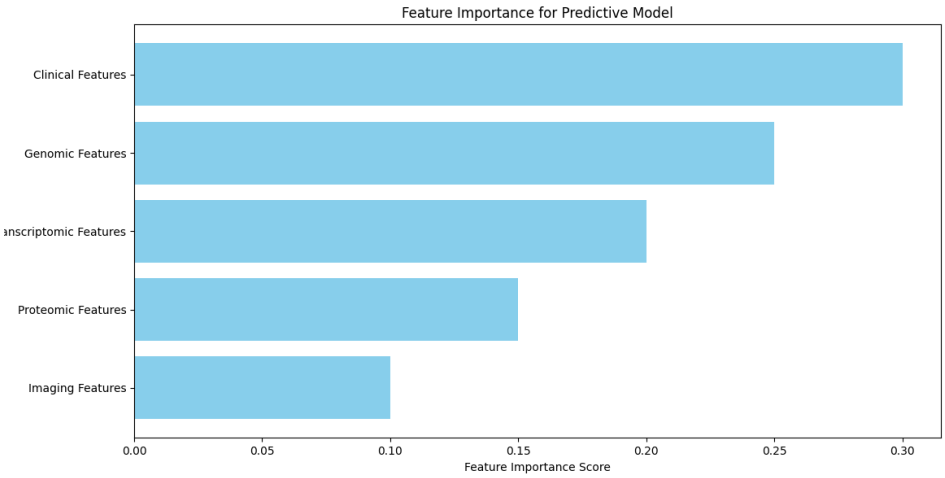

Figure 5: Feature Importance for Predictive Model

The SHAP summary plot (Figure 6) illustrates the contribution of individual features to the predictions. This visualization supports the interpretability of the model by showing how specific features influence the outcomes.

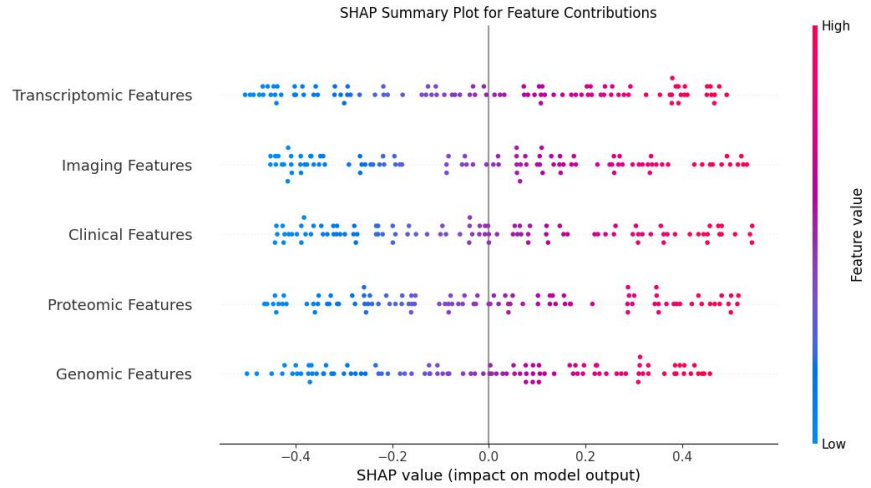

Figure 6: SHAP Summary Plot for Feature Contributions

These results validate the design of MO-DiPredict, demonstrating its consistent performance in handling multi-modal data for early detection and subtype classification of blood cancers while maintaining interpretability and scalability.

CONCLUSION

This study introduced MO-DiPredict, a framework designed to combine multi-omics, clinical, and imaging data for detecting blood cancers at early stages and predicting their subtypes. The main goal was to address data challenges, such as noise, inconsistencies, and varying scales, by aligning features across different data types and capturing their relationships. The framework uses Canonical Correlation Analysis for alignment and Graph Neural Networks to model dependencies, coupled with machine learning techniques for prediction. The results demonstrated that MO-DiPredict achieved an AUC-ROC of 0.93, outperforming other models, MRGCN and MODILM, by a small yet consistent margin. Feature importance analysis highlighted the predictive value of clinical and genomic features, with incremental improvements observed through ablation studies. While the proposed model performed consistently, the reliance on publicly available datasets limited its scope. These datasets might not fully represent the diversity of real-world clinical and biological cases. The study provides a structured approach to integrating diverse datasets for cancer research. Future efforts could explore applications of the framework to other diseases, validate its use on larger datasets, and refine the integration methods for higher precision. MO-DiPredict offers a systematic way to address

gaps in multi-omics data analysis, contributing to efforts in personalized diagnosis and classification of complex diseases.

REFERENCES

[1] Nikshya, J. Ebens, M. Saravana Karthikeyan, Shalini Prasad, R. Santhana Krishnan, S. Balamurugan, and J. Relin Francis Raj. "A Machine Learning Framework for Integrating Multi-Omics Data for Early Leukemia Detection." In 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 1348-1356. IEEE, 2024.

[2] Hernández-Lemus, Enrique, and Soledad Ochoa. "Methods for multi-omic data integration in cancer research." Frontiers in Genetics 15 (2024): 1425456.

[3] Islam, Saiful, and Md Nahid Hasan. "Personalized graph feature-based multi-omics data integration for cancer subtype identification." arXiv preprint arXiv:2408.08832 (2024).

[4] Tanvir, Raihanul Bari, Md Mezbahul Islam, Masrur Sobhan, Dongsheng Luo, and Ananda Mohan Mondal. "MOGAT: A Multi-Omics Integration Framework Using Graph Attention Networks for Cancer Subtype Prediction." International Journal of Molecular Sciences 25, no. 5 (2024): 2788.

[5] Shi, Tianyi, Xiucai Ye, Dong Huang, and Tetsuya Sakurai. "Cancer subtype identification by multi-omics clustering based on interpretable feature and latent subspace learning." Methods 231 (2024): 144-153.

[6] Rahmanian, Mohsen, and Eghbal G. Mansoori. "MoVAE: Multi-omics Variational Auto-Encoder for Cancer Subtype Detection." IEEE Access (2024).

[7] Isobe, Tomoya, Masatoshi Takagi, Aiko Sato-Otsubo, Akira Nishimura, Genta Nagae, Chika Yamagishi, Moe Tamura et al. "Multi-omics analysis defines highly refractory RAS burdened immature subgroup of infant acute lymphoblastic leukemia." Nature communications 13, no. 1 (2022): 4501.

[8] Hao, Yaru, Xiao-Yuan Jing, and Qixing Sun. "Joint learning sample similarity and correlation representation for cancer survival prediction." BMC bioinformatics 23, no. 1 (2022): 553.

[9] Madhumita, Archit Dwivedi, and Sushmita Paul. "Recursive integration of synergised graph representations of multi-omics data for cancer subtypes identification." Scientific Reports 12, no. 1 (2022): 15629.

[10] Park, Min-Koo, Jin-Muk Lim, Jinwoo Jeong, Yeongjae Jang, Ji-Won Lee, Jeong-Chan Lee, Hyungyu Kim et al. "Deep-learning algorithm and concomitant biomarker identification for NSCLC prediction using multi-omics data integration." Biomolecules 12, no. 12 (2022): 1839.

[11] Benkirane, Hakim, Yoann Pradat, Stefan Michiels, and Paul-Henry Cournède. "CustOmics: A versatile deep-learning based strategy for multi-omics integration." PLoS Computational Biology 19, no. 3 (2023): e1010921.

[12] Mathema, Vivek Bhakta, Partho Sen, Santosh Lamichhane, Matej Orešič, and Sakda Khoomrung. "Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine." Computational and Structural Biotechnology Journal 21 (2023): 1372-1382.

[13] Itai, Yonatan, Nimrod Rappoport, and Ron Shamir. "Integration of gene expression and DNA methylation data across different experiments." Nucleic Acids Research 51, no. 15 (2023): 7762-7776.

[14] Yang, Bo, Yan Yang, Meng Wang, and Xueping Su. "MRGCN: cancer subtyping with multi-reconstruction graph convolutional network using full and partial multi-omics dataset." Bioinformatics 39, no. 6 (2023): btad353.

[15] Tsagiopoulou, Maria, Nikolaos Pechlivanis, Maria Christina Maniou, and Fotis Psomopoulos. "InterTADs: integration of multi-omics data on topologically associated domains, application to chronic lymphocytic leukemia." NAR Genomics and Bioinformatics 4, no. 1 (2022): lqab121.

[16] Huang, Kaitang, Meiling Hu, Jiayun Chen, Jinfen Wei, Jingxin Qin, Shudai Lin, and Hongli Du. "Multi-omics perspective reveals the different patterns of tumor immune microenvironment based on programmed death ligand 1 (PD-L1) expression and predictor of responses to immune checkpoint blockade across pan-cancer." International Journal of Molecular Sciences 22, no. 10 (2021): 5158.

[17] Liu, Xiao-Ying, and Xin-Yue Mei. "Prediction of drug sensitivity based on multi-omics data using deep learning and similarity network fusion approaches." Frontiers in Bioengineering and Biotechnology 11 (2023): 1156372.

[18] Choi, Joung Min, and Heejoon Chae. "moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks." BMC bioinformatics 24, no. 1 (2023): 169.

[19] Zhong, Yating, Yuzhong Peng, Yanmei Lin, Dingjia Chen, Hao Zhang, Wen Zheng, Yuanyuan Chen, and Changliang Wu. "MODILM: towards better complex diseases classification using a novel multi-omics data integration learning model." BMC Medical Informatics and Decision Making 23, no. 1 (2023): 82.

[20] Zhou, Kaiyue, Bhagya Shree Kottoori, Seeya Awadhut Munj, Zhewei Zhang, Sorin Draghici, and Suzan Arslanturk. "Integration of multimodal data from disparate sources for identifying disease subtypes." Biology 11, no. 3 (2022): 360.

[21] Ye, Qing, and Nancy Lan Guo. "Inferencing Bulk Tumor and Single-Cell Multi-Omics Regulatory Networks for Discovery of Biomarkers and Therapeutic Targets." Cells 12, no. 1 (2022): 101.

[22] Rupapara, Vaibhav, Furqan Rustam, Wajdi Aljedaani, Hina Fatima Shahzad, Ernesto Lee, and Imran Ashraf. "Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model." Scientific reports 12, no. 1 (2022): 1000.

[23] Zheng, Xingyu, Christopher I. Amos, and H. Robert Frost. "Pan-cancer evaluation of gene expression and somatic alteration data for cancer prognosis prediction." BMC cancer 21 (2021): 1-11.

[24] Strain, Paul, Enya E. Scanlon, Gera Jellema, Richard D. Kennedy, Ken I. Mills, and Jaine K. Blayney. "P463: Re-purposing of gene signatures in AML uncovers novel energetics-associated molecular subtypes." HemaSphere 6 (2022): 362-363.

[25] Leng, Dongjin, Linyi Zheng, Yuqi Wen, Yunhao Zhang, Lianlian Wu, Jing Wang, Meihong Wang, Zhongnan Zhang, Song He, and Xiaochen Bo. "A benchmark study of deep learning-based multi-omics data fusion methods for cancer." Genome biology 23, no. 1 (2022): 171.