International Journal of Computer Vision and Signal Processing, 1(1), 9-17(2025) ORIGINAL ARTICLE

IntegraDenoNet: A Deep Learning Based Single Cell Multiomics Integration and Cell Type Identification

Md Shaharia Hossen*, Sakib Mahmood Saad, Maria Akter Rimi, Marin Akter, Fahim Hafiz, Riasat Azim

Department of Computer Science and Engineering
United International University, Dhaka, Bangladesh

Abstract

The correlation between different phases of biological data, such as transcriptomics, metabolomics, and other omics, is important in the case of disease analysis. Multiomics aims to combine diverse omics data into a unified dataset, revealing interrelationships and their influence on complex biological processes. Although multi-omics methodologies are relatively new, their demonstrated potential to accurately uncover insights has captured the bioinformatics field. However, limited datasets and challenges in preparing unbiased models have hindered widespread application. This research introduces an innovative deep learning-based method for the seamless integration of multi-omics single-cell data, allowing for accurate classification of omics expression levels. Omics data are reconstructed using a denoising autoencoder with a learning rate scheduler, cosine annealing. Reconstructed data are integrated with labels for further downstream analysis. Our proposed method achieved minimal classification loss, approximately 0.05% compared to other recent methods. Furthermore, the proposed method achieved a consistent accuracy greater than 90% in three multi-omics datasets, beating four advanced state-of-the-art (SOTA) methods. The proposed model 'IntegraDenoNet' demonstrates improved classification accuracy and advances possibilities in precision medicine.

Contribution of the Paper: 'IntegraDenoNet' leverages deep learning to integrate multi-omics data for cell type classification, achieving 90% accuracy across three gold-standard datasets and outperforming four state-of-the-art methods.

Keywords: Multiomics, Single Cell, Denoising Autoencoder, Neural Network, Transcriptomics, Metabolomics

IJCVSP

ISSN: 2186-1390 (Online)
http://cennser.org/IJCVSP

Article History:
Received: 10.1.2025
Revised: 10.7.2025
Accepted: 22.11.2025
Published Online: 2311.2025

1. Introduction

Multiomics approaches utilize high-throughput methods to generate omics data, followed by computational analysis to extract biological insights. It is an integrative approach that combines multiple omics data to create a holistic view of biological interactions [1]. Each omics layer correlates with others and has significance in the decision of an examination, while having its limitations and constraints [2]. Single-omics data research has proven fundamentally incomprehensive, and it does not hold itself accountable for the information correlation at each molecular level. Thus, studying multiomics is far more comprehensive and expressive. Single-cell multiomics tools improve the understanding of cellular functions in both normal and pathological states [3].

The key challenges in this field are the confidential datasets and preparing unbiased models. Historically, disease prediction methods in studies initially depended on statistical approaches. These methods calculate feature values based on the detection of biological substances in the human body. A comparison of these values was conducted against predefined thresholds to analyze biomarkers from specific omics and outline disease characteristics [4]. Still, these statistical approaches for omics-disease association identification have notable limitations. Therefore, employing a multi-omics approach is essential to fully strengthen

*Corresponding author
Email addresses: mhossen201288@bscse.uiu.ac.bd (Md Shaharia Hossen), ssaad201470@bscse.uiu.ac.bd (Sakib Mahmood Saad), mrimi201428@bscse.uiu.ac.bd (Maria Akter Rimi), makter201298@bscse.uiu.ac.bd (Marin Akter), fahimhafiz@cse.uiu.ac.bd (Fahim Hafiz), riasat@cse.uiu.ac.bd (Riasat Azim)

the potential of high-dimensional data and gain in-depth insights into biological systems. With advancements in artificial intelligence, especially deep learning, multi-omics integration has gained prominence in disease classification [5].

Many researchers have emerged in this domain. K. Chaudhary, [6] utilized autoencoders to reconstruct features from high-dimensional multiomics data, clustered samples using K-means, and trained an SVM for liver cancer classification. Similarly, X. Li, [7] introduced the MoGCN model by merging features using similarity network fusion (SNF) to build a patient similarity network, and trained a GCN for cancer subtype classification. These early integration approaches integrate multiomics data before input into prediction models [8]. However, these methods often struggle with robustness and generalizability due to the high-dimensional and heterogeneous nature of multiomics data that poses challenges for machine learning models [9]. Thus, newer implementation methods gained the attention of the researchers for their flexibility across different omics data. D. Sun [10] designed a deep convolutional neural network to extract features from CNV and DNA expression data, which were subsequently combined for BRCA subtype classification. The details related to the research are discussed in Section 2.

Although late integration methods analyze individual omics features before integration, many recent studies replicate traditional models, often neglecting the possibility of high-dimensional omics data. However, the continuous evolution of deep learning offers opportunities for significant breakthroughs in multi-omics-based disease classification.

This research project aims to integrate multiomics in single-cell datasets to identify cell types with minimal losses and higher accuracy, where the loss is achieved at 0.05% and more than 90% prediction accuracy, respectively. Before going to the proposed method, this research experiments with methylation, mRNA, and miRNA data with some machine learning models. The omics datasets have been trained with labels using the ensemble classification technique with 41% accuracy; whereas normal classification models give 21%. Individual data sets processing yielded a consistent 64% loss. From the experiments of classification with lower accuracy and higher losses, this research aims to reduce the losses using the deep learning model. After going through the experimental machine learning model, deep learning models have been chosen. Considering the improvement of losses, this research proposed a novel approach for multi-omics integration and downstream analysis. Denoising autoencoders with cosine annealing and optimizers have been used for reconstructing the individual omics after normalization with minimal losses of 0.05%. Reconstructed omics data have been integrated with corresponding labels and classified using different classification models. In the scenario of classification, a deep learning model neural network gives superior results for every multi-omics dataset, which is more than 90%. This approach demonstrates an effective pipeline for multi-omics classification, indicating strong potential in omics data analysis.

2. Related Research

Several researchers have described multiomics technology. Multiomics data is exploited in a variety of ways, including statistical approaches, machine learning, and deep learning models. The associated multi-omics research study is discussed below.

T. Wang [11], introduced MOGONET which constructs graphs for each omics data using graph convolutional networks (GCNs) for omics-specified learning and integrates outputs with the View Correlation Discovery Network (VCDN) to explore cross-omics correlations. MOGONET's effectiveness was validated across various biomedical classification tasks. H. Wang [12], highlighted a trusted method HyperTMO integrate the dataset and find out the patient classification. Single-omics data is used to calculate the cosine similarity for samples, and KNN is used to construct the sample structure of the hypergraph. Hypergraph convolutional network (HGCN) finds out the high-order association and then combines the graph for multiomics representation and predicts the class. T. Athaya [13], showed the integration of extracellular miRNA with mRNA for cancer studies using the CrosePred method. To enhance the accuracy of the dataset, they have used an encoder output and a random forest classifier to find out the to classify samples as cancerous or healthy.

F. Chen [14], integrates a novel Graph Regularized Multi-view Ensemble Clustering (GRMEC-SC) model for clustering single-cell multi-omics data is a way to find out cell type. Multi mono omics data is used based on clustering and used non-negative weighted co-cluster affinity matrix learned from multi-omics data to find out the cell type. Novoloaca [15], studied emphasizes the need for integrated tools for analysis in multi-omics for precise medicine, with methodologies such as DIABLO and specific random forest models demonstrating outstanding results in biomarkers identification and illness analysis. Studies with simulated data highlight the need to select suitable approaches for multi-omics research. Lim [16], research demonstrates the precise nature of single-cell omics in detecting cell variety and unusual populations, as well as providing insights into cellular relationships and regulatory networks. Integrating multimodal omics improves applications in biology of development, oncology, and precise medicine, and provides extensive assistance in selecting appropriate approaches.

Valous [17], implemented a graph-based technique, particularly those utilizing Graph Neural Networks (GNNs), which have become critical in multi-omics research, integrating complicated biological data and improving our understanding of disease causes. This shift allows for the identification of crucial biomarkers and pathways, promoting individualized therapy. Baysoy [18], proposes that in

the field of molecular biology research, Single-cell multiomic technologies have transformed the landscape of molecular biology investigation by offering an in-depth comprehension of cellular diversity and functions. These advancements have had a profound influence on diverse fields such as cell lineage tracing, oncology, and the advancement of therapeutic strategies in tumor immunology. Kesimoglu [19], shows that, by combining multiomics data, the SUPREME model improves malignancy subtyping by achieving more precise classifications and capturing a variety of predictive signals. It finds subtypes associated with notable differences in survival, frequently revealing characteristics not found in conventional categories. SUPREME exhibits resilience and adaptability with accurate predictions spanning models and applications outside of cancer, holding promise for advancing precision medicine by providing more individualized insights into cancer treatment.

According to Flynn [20], advances in single-cell technology, such as sequencing RNA and multimodal integration, help us better comprehend the diversity of cells and function. This study investigates computational approaches for linking biological levels and pathways by combining data from various sources, including proteomic, spatial, epigenomic, and genomic information, thus paving the way for future advances in complex biological systems. Rakshit [21], described that, by combining transcriptomic, proteomic, and genomic data, multi-omics technology helps to understand how drugs work and how diseases work. Target discovery is streamlined, disease diagnosis and treatment development are transformed, and the integration of multiomics data is made easier by enhanced bioinformatics tools and data accessibility. Creighton [22], analyzed comprehensive tumor profiling has been made possible by advancements in mass spectrometry-based proteomics, with programs such as the Clinical proteome Tumor Analysis Consortium and the Worldwide Cancer Proteogenome Consortium gathering sizable proteome and multi-omics datasets relevant to cancer. By enabling the thorough examination of various cancer subtypes, these resources aid in the development of targeted diagnoses and treatments.

Marshall [23], observes that currently, cancer therapy focuses heavily on multi-omics technology for identifying precisely between individual variants and tumor-specific proteome patterns. Liquid biopsies have great opportunities as cancer biomarkers for a variety of research and therapeutic uses, while recent developments in ctDNA analysis and artificial intelligence-powered digital pathology are improving cancer diagnosis and treatment. Hasin [24], explained that multi-omics techniques interpret the complexities of disease by combining data from genomes, transcriptomics, proteomics, and metabolomics. Coordinated multi-omics studies from relevant tissues are necessary to provide comprehensive mechanistic insights; important genetic and metabolic trait assessments are provided by cohorts such as MuTher and METSIM. Liu [25], shows recent advances in single-cell multi-omics, particularly the SiaNN model, enable successful integration of scRNA-seq, scATAC-seq, and epitope data while addressing critical issues such as batch effects. Benchmark testing reveals SiaNN's high precision and adaptability in a variety of situations, with applications confirmed using pcHi-C data, making it an important tool for understanding complicated biological networks, particularly those connected to COVID-19.

In conclusion, multi-omics technologies, which combine genomes, transcriptomics, and proteomics with powerful machine learning and deep learning algorithms, have revolutionized disease classification, biomarker identification, and precision medicine.

3. Materials and Methods

The proposed model comprises three key components: 3.1 Data Availability and Review, 3.2 Reconstruction of Omics Data, and 3.3 Multiomics Integration and Downstream Analysis. Initially, the data undergoes normalization and standardization to ensure consistency and eliminate scale-related biases. After normalization, the data is further encoded using denoising autoencoders, which compress and highlight the most informative aspects of the data. The encoded features from multiple omics datasets are then concatenated into a unified multi-omics dataset, integrating diverse biological information. Finally, a Neural Network (NN), CNN, and other classification models are trained on this comprehensive multi-omics dataset to effectively predict the class. Fig. 1 shows the proposed method overview.

3.1. Data Availability and Review

To demonstrate the effectiveness of the proposed model, we performed the BRCA PAM50 subtype classification task on BRCA multi-omics data from TCGA, containing mRNA data and miRNA data for transcriptomics and DNA methylation data for epigenomics. The BRCA classification task predicts five subtypes: Normal-like, Basal-like, human epidermal growth factor receptor 2 (HER2)-enriched, Luminal A, and Luminal B. The HNSC data from TCGA and ROSMAP data were also used for this research.

Each omics is standardized and normalized. Standardization adjusts features to have a mean of zero and a standard deviation of one, removing scale biases and ensuring comparability and Normalization ensures consistency across features from different omics datasets.

3.2. Reconstruction of Omics Data

To address these challenges, this article proposes the use of denoising autoencoders to reconstruct each omics dataset, enabling a more comprehensive representation of the complex associations and heterogeneity inherent in multi-omics data.

Each omics dataset undergoes standardization and normalization to ensure uniformity and comparability across features. Noise is injected into each omic dataset. Then, a

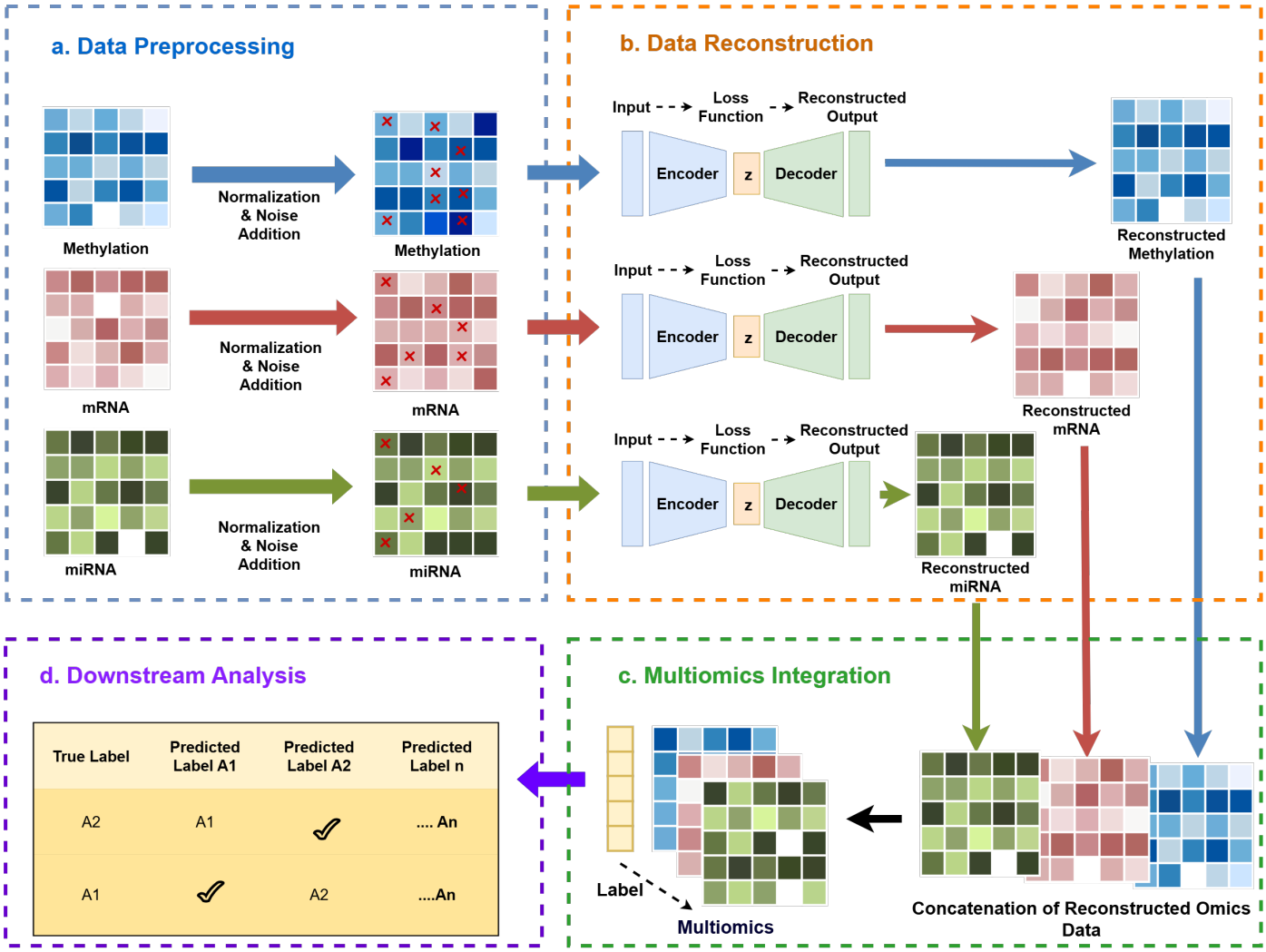

Figure 1: IntegraDenoNet: Fig (a). shows the data preprocessing, using data normalization and adding Gaussian noise for further training. A denoising autoencoder with cosine annealing, and to consider the large value difference of data, the mean squared logarithmic error loss function has been used. Reconstructed omics data have been integrated with labels, and the reconstructed multi-omics data have been used for classification. Fig. (b,c,d) shows the techniques, respectively.

denoising autoencoder is used to reconstruct the input data while filtering out added noise. Fig. 2 shows the technique.

Input of Omics Data: Accepts the standardized input $\mathbf{x} \in \mathbb{R}^n$, where n is the number of features. The Input Layer specifies the input shape of the data.

Noise Addition: Gaussian Noise adds random noise to the input to help the model learn to denoise the data. Gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$ is added to the input: $\tilde{x} \sim x + z$ where $\tilde{x}$ represents the noisy input.

Encoding of Data: Maps the noisy input $\tilde{x}$ to a latent representation h:

$$h = f(W_e \tilde{x} + b_e) \tag{1}$$

where W_e and b_e are the weights and biases, respectively, and $f(.)$ is the ReLU activation function. The encoder progressively reduces the dimensionality, employing L1/L2 regularization, dropout, and batch normalization.

Decoding of Data: Reconstructs the original input x from h:

$$\hat{x} = f(W_d h + b_d) \tag{2}$$

where $\hat{x}$ is the reconstructed output, W_d and b_d are the decoder's weights and biases, and $g(.)$ includes ReLU for intermediate layers and linear activation for the final layer.

The mean squared logarithmic error (MSLE) has been used with optimizers. In the denoising autoencoder deep learning model associated with MSLE, calculate the loss, and care about the relative scale between the noisy input and the clean input. It considers the outliers and large value differences in datasets.

$$\text{MSLE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (\log(y_i + 1) - \log(\hat{y}_i + 1))^2 \tag{3}$$

To avoid the large number of errors, MSLE helps to adapt to the large range of omics datasets. y_i denotes the actual inputs, and $\hat{y}_i$ is for predictive value addition; it uses $+1$ to avoid the issue of logarithmic 0.

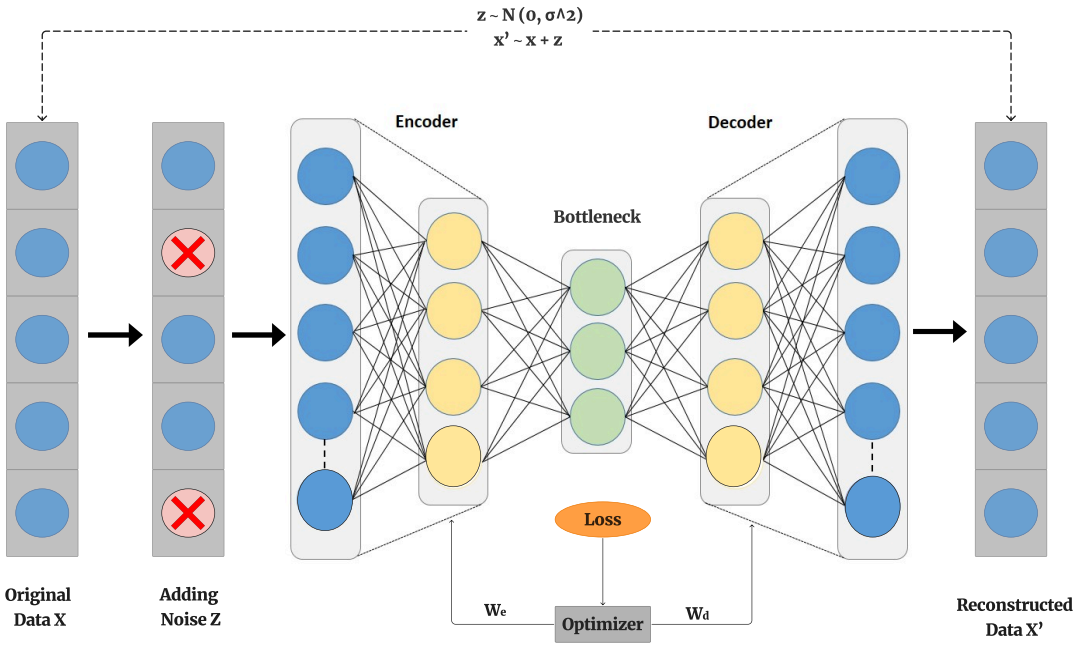

Figure 2: Denoising autoencoder with optimizer. The noise injection into the original omics data and reconstruction process.

Adaptive moment (Adam) optimization algorithms have been used to ensure more stable and faster convergence. It helps for deep deep learning models like a denoising autoencoder for stabilizing the training for noisy input. For complex and large data or parameters, it leads to better performance and accelerates convergence to lower losses in the case of reconstruction of data.

Combination of momentum and root mean square propagation (RMSP) technique of Adam helps to learn the large datasets effectively. Momentum algorithms accelerate the gradient descent algorithm with an exponentially weighted average; on the other hand, the RMSP uses an exponential moving average to improve the algorithm.

Each epoch utilizes a learning rate scheduler, cosine annealing. This technique begins with a relatively high learning rate, which is gradually reduced to a minimum value before being rapidly increased again. The cyclical pattern helps the model explore different regions of the optimization landscape, potentially avoiding local minima and improving convergence.

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right) \quad (4)$$

Where $\eta_{\min}$ and $\eta_{\max}$ are ranges for the learning rate, and T_{cur} account for how many epochs have been performed since the last restart. Each omic dataset undergoes this transformation, generating robust feature representations.

3.3. Multiomics Integration and Downstream Analysis

All the reconstructed data frames have been concatenated along the horizontal axis. Assuming a data frame of omics datasets: Methylation = A, mRNA = B, miRNA = C.

$$\text{Concatenated DataFrame} = [A \mid B \mid C]$$

where

$$A \in \mathbb{R}^{n \times m_1}, \quad B \in \mathbb{R}^{n \times m_2}, \quad C \in \mathbb{R}^{n \times m_3}$$

The above scenario results in: Multiomics(M) = $(n, m_1 + m_2 + m_3)$. For the downstream analysis label L is added with multiomics data, $[L \mid M]$. Machine learning and deep learning models have been used for classification.

4. Model Implementation and Result Analysis

Python- 3.10.12 torch-2.2.1+cu121 CUDA:0 is utilized on this platform; Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz (8 CPUs)), 1.8GHz, (Intel(R) UHD Graphics 620). Reconstructed multi-omics data with a neural network classification model gives better performance. Implementation techniques are described below:

Step 1: The individual omics data have been reconstructed using the denoising autoencoder with cosine annealing. The optimizer is Adam, the loss function is a mean squared logarithmic error, and the cosine annealing has been called back from the denoising autoencoder training step. The cosine annealing max T will increase up to 2000, and the denoising autoencoder has been trained up to 2000 epochs with a learning rate of 0.001. After that, the reconstructed omics datasets have been concatenated with labels using the concatenate function of Python.

Step 2: Features are scaled to zero mean and unit variance for better training performance. The Imbalance dataset then goes through Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset by oversampling the minority class to match the majority class. This prevents the model from being biased toward the majority class. Integer-encoded labels are converted into a one-hot encoded format required for the neural network's softmax output layer.

Step 3: The neural network model consists of an input layer accepting features of dimension, followed by two hidden layers with 512 and 512 neurons, ReLU activation, L2 regularization, and Dropout (10% and 20%, respectively), and an output layer and softmax activation for multi-class classification. It is compiled using the Adam optimizer and categorical cross-entropy loss function, tracking accuracy. Early stopping halts training if validation loss doesn't improve for 20 epochs and restores the best weights. The model runs up to 200 epochs with a batch size of 256.

Step 4: To validate the proposed model, its performance is evaluated using accuracy (ACC) and the F1 score, two widely recognized metrics for assessing classification tasks. In this setup, 80% of the data is allocated for training, while the remaining 20% is used for testing. This approach provides a reliable assessment of the model's generalization capability and minimizes the risk of overfitting.

4.1. Results Analysis

After the successful implementation of IntegraDenoNet, 3 different datasets have been trained with superior prediction results. Reconstructed data have been integrated with corresponding labels for downstream analysis. Reconstructed multi-omics data have been trained with different classification models, where neural networks give the best prediction accuracy. Random Forest, SVM, XGBoost, CNN, and Neural Network models have been used for the classification experiment. The accuracy varies in different machines for the dimension of the neuron and the dropout rate.

BRCA Multiomics Data: Table 1 shows the classification report on BRCA data. A deep learning model neural network gives the best accuracy, which is 91%. Both precision and recall for the neural network are 90%, and the harmonic mean of precision and recall, the F1 score, is also 90%, which indicates the balance of predicting true positives and false positives.

ROSMAP Multiomics Data: Compared to other classification models, a deep learning model neural network gives the best accuracy, which is 91%. Both precision and recall for the neural network are 90%, and the harmonic mean of precision and recall, the F1 score, is also 90%, which indicates the balance of predicting true positives and false positives. Table 2 indicates the classification report on ROSMAP data.

TCGA Multiomics Data: TCGA is larger compared to others datasets. The max depth of the Random Forest and XGBoost has been set to 100 to track predictions. A deep learning model neural network gives the best accuracy, which is 90%. Both precision and recall for the neural network are 90%, and the harmonic mean of precision and recall, the F1 score, is also 89%, which indicates the balance of predicting true positives and false positives. Table 3 indicates the classification report on ROSMAP data.

Table 1: Classification Report on BRCA Data

Classification Model	Metrics			
	Acc	**F1**	**P**	**R**
Neural Network	**91**% ±1.50	**90**%	**90**%	**90**%
CNN	88% ±2.80	84%	85%	84%
XGBoost	88% ±1.23	89%	89%	89%
SVM	89% ±1.50	90%	90%	90%
Random Forest	88% ±0.20	89%	89%	88%

Table 2: Classification Report on ROSMAP Data

Classification Model	Metrics			
	Acc	**F1**	**P**	**R**
Neural Network	**91**% ±1.00	**90**%	**90**%	**90**%
CNN	91% ±2.50	90%	89%	93%
XGBoost	87% ±0.50	84%	85%	84%
SVM	88% ±1.50	84%	86%	86%
Random Forest	85% ±3.50	83%	83%	83%

Table 3: Classification Report on TCGA (HNSC) Data

Classification Model	Metrics			
	Acc	**F1**	**P**	**R**
Neural Network	**90**% ±1.80	**89**%	**90**%	**90**%
CNN	82% ±3.50	81%	82%	83%
XGBoost	85% ±0.50	84%	85%	84%
SVM	83% ±3.50	81%	82%	83%
Random Forest	80% ±0.50	79%	79%	79%

The receiver operating characteristic (ROC) describes the performance of any predictions. The ROC curve of the neural network classification on multi-omics data gives significant results. The dashed diagonal line (ROC = 0.5) of the AUC curve indicates the random classifier. The ROC value of any class, 100%, means the class is perfectly identified; more than 95% indicates the predicted class is outstandingly classified. The prediction result of ROC is close to 90%, indicating the class is excellently classified. The ROC is important to measure the distinction of the predicted class of the data. For the purpose of classifying the multi-omics data, the neural network works incredibly well; Figs. 3, 4, and 5 show the results.

Fig. 3 shows the ROC of class label prediction on the

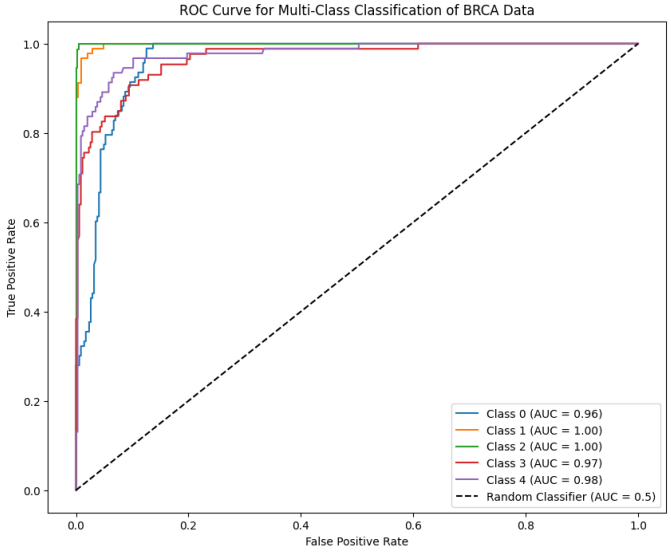

Figure 3: ROC curve for BRCA data

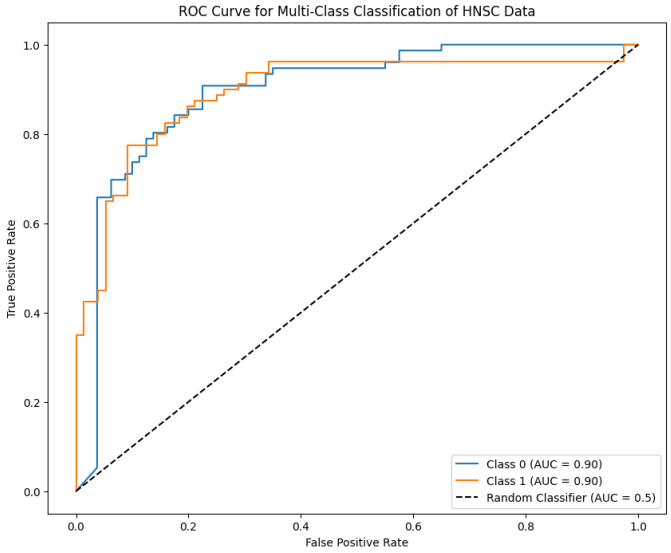

Figure 4: ROC curve for TCGA data

IntegraDenoNet model in BRCA data. On the multi-omics BRCA data, classes 1 and 2 achieved 100% AUC, which indicates that the classes are perfectly identified as true positives randomly. On the other hand, classes 0, 3, and 4 achieved more than 95% AUC, which indicated that the classes predict the true positive outstandingly.

TCGA multi-omics data predicts the true positive class label excellently, where classes 0 and class 1 achieved 90% AUC. Fig. 4 indicates the results.

Fig. 5 shows the ROC of ROSMAP multi-omics data classification. For classes 0 and 1, 95% AUC was achieved for true label prediction of ROSMAP multi-omics data, which is outstanding.

This project experiments with a neural network model on the datasets to predict performance. Methylation, mRNA, and miRNA datasets have been trained using neural networks individually to see the results. After that, the individual omics were concatenated and trained using a neural network to check the classification accuracy. The results of the experiments are poor compared to the final proposed method, IntegraDenoNet. Fig. 6 bar graphs, including the final model, show the progressive accuracy.

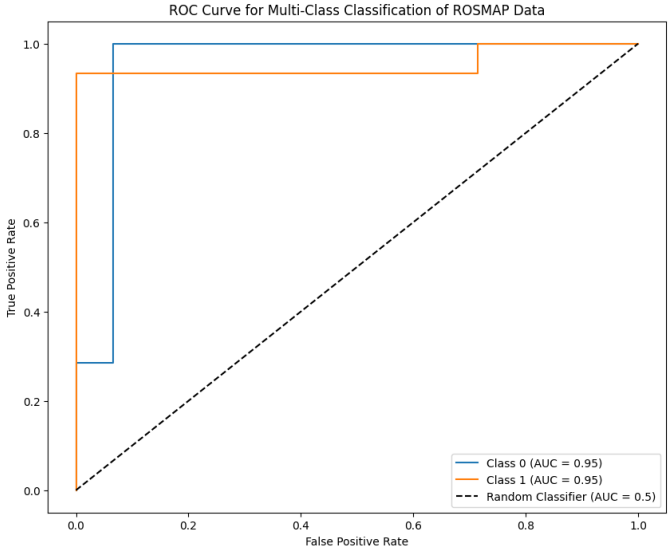

Figure 5: ROC curve for ROSMAP data

4.2. Comparison With State-of-the-art (SOTA) Methods

Compared to the similar work with the proposed IntegraDenoNet on the same datasets, it is shown in Table 4 and Table 5. On the ROSMAP data, our proposed model achieved 91% accuracy, which is better than other methods. The harmonic mean of precision and recall F1 score is achieved at 90%, which is very effective for predicting the class. The AUC for both classes on ROSMAP data is 95%, which denotes the outstanding performance on selecting the true positive class from random data. On the other hand, BRCA data also achieved 91% accuracy, which is better compared to other methods. The macro and weighted harmonic mean F1 scores are also 91% respectively, which is also better compared to others.

Table 4: Comparative Analysis of Similar Work

Related Methods	ROSMAP Data		
	Acc	**F1**	**AUC**
hyperTMO [12]	87% ±0.033	87%	90%
MOGONET [11]	85% ±0.04	86%	85%
MoGCN [7]	80% ±0.55	77%	80%
XGBoost [26]	76% ±0.04	77%	83%
IntegraDenoNet	**91**% ±1.00	**90**%	**95**%

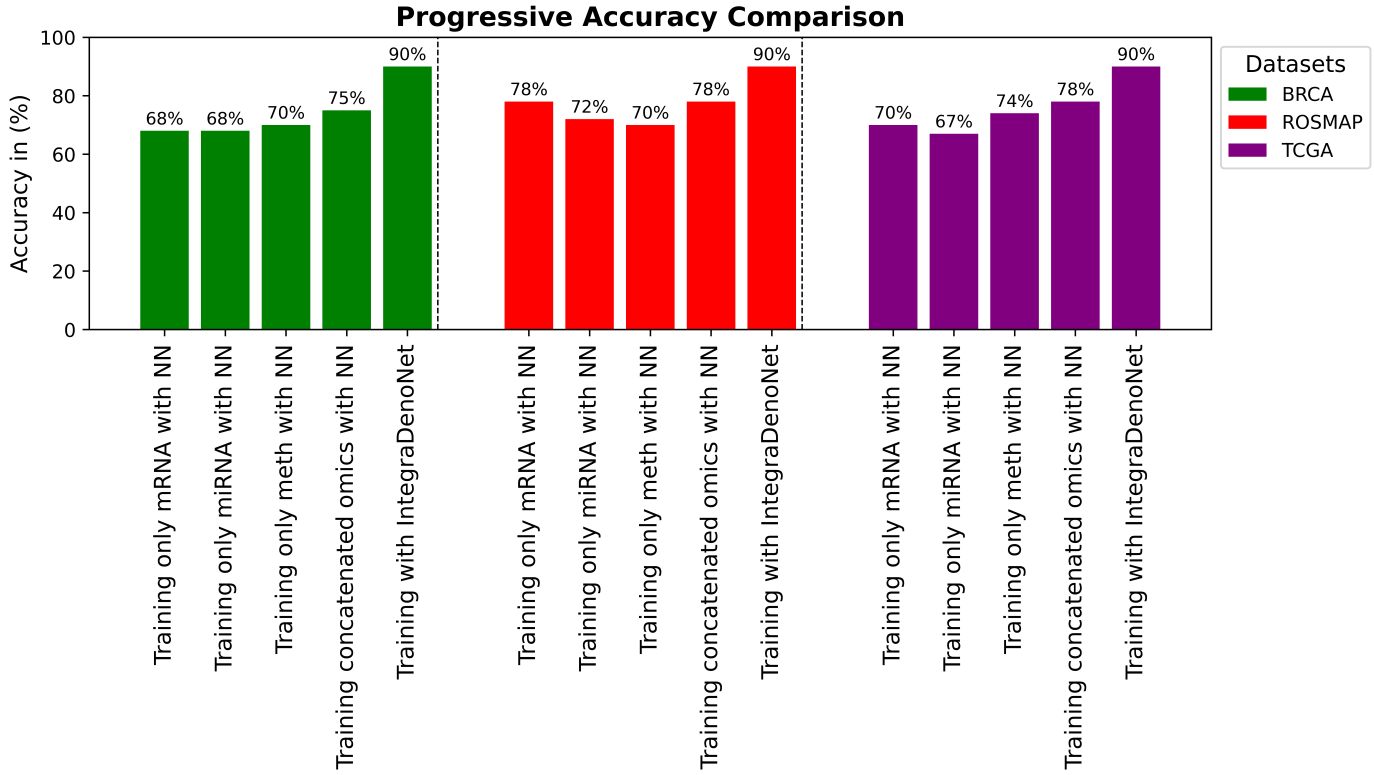

Figure 6: Accuracy for three different datasets: BRCA, ROSMAP, TCGA, using our proposed method.

Table 5: Comparative Analysis of Similar Work

Related Methods	BRCA Data		
	Acc	**F1 macro**	**F1 weighted**
hyperTMO [12]	87% ±0.023	84%	86%
MOGONET[11]	81% ±0.03	79%	81%
MoGCN [7]	80% ±0.03	76%	81%
XGBoost [26]	78% ±0.008	70%	76%
IntegraDenoNet	**91**% ±1.00	**91**%	**91**%

4.3. Scalability of The IntegraDenoNet

The proposed model, IntegraDenoNet, was developed and evaluated on a local machine equipped with an Intel Core i5-8265U CPU and 8 GB of RAM. The model was implemented using Python 3.10.11 and TensorFlow. Each omics dataset was individually trained using a denoising autoencoder, requiring approximately 35 to 40 minutes per dataset. Following this, the concatenated reconstructed data were further processed for classification, also taking approximately 35 to 40 minutes. In total, the entire pipeline takes around 140 to 160 minutes to execute end-to-end on the specified hardware.

While the model performs well on moderate-sized datasets, scaling to larger omics datasets (such as those from TCGA) or applying the model in real-time settings presents computational challenges. In such cases, the use of higher-performance computing resources such as GPUs or cloud-based platforms would be beneficial.

4.4. Limitations and Future Works

While IntegraDenoNet shows promising results for both multi-class and binary classification tasks, several limitations remain. First, although the model performs well on moderately sized datasets, its scalability to larger omics datasets (such as those from TCGA) remains a challenge due to increased computational demands. Running the model on such datasets would require higher-performance hardware, such as GPUs or cloud-based computing environments.

Future work could involve extending IntegraDenoNet for biomarker discovery and applying unsupervised classification techniques to uncover hidden patterns in multi-omics data. Additionally, developing a user-friendly software tool with adjustable model parameters could enhance the framework's accessibility, reproducibility, and applicability across various research settings.

4.5. Dataset and Code Availability

Data and code are available at the following link: Multiomics Integration.

5. Conclusion

The study demonstrates the potential of multi-omics integration for improving disease classification. Despite challenges including dataset availability, computational complexity, and framework scalability, the project provides a strong foundation for integrating single-cell omics data using advanced deep learning techniques. Future advancements include the addition of more omics data, optimized feature selection, and improved computational frameworks, which will further enhance the system's accuracy and usability. The work underscores the significance of multi-omics in understanding complex biological systems and tailoring personalized diagnostic and therapeutic strategies.

'IntegraDenoNet' performs well in predicting the classification of omics expression levels on three datasets: BRCA, TCGA, and ROSMAP. The proposed method achieved approximately 90% classification accuracy in multi-omics data with higher precision and recall, outperforming four SOTA methods.

Proposed method utilizes a denoising autoencoder in each omic dataset for reconstructing the respective omic data while removing noise from such high-dimensional biological datasets. Furthermore, a neural network architecture is employed on the concatenated multi-omics data for efficient classification. Such an integrated architecture can find the intricate pattern of multi-omics data while accurately predicting cell type.

References

[1] K. Katsos, A. Dhar, F. Moinuddin, Multiomics in precision medicine, in: The New Era of Precision Medicine, Elsevier, 2024, pp. 195–207.

[2] S. Tsimenidis, E. Vrochidou, G. A. Papakostas, Omics data and data representations for deep learning-based predictive modeling, International Journal of Molecular Sciences 23 (20) (2022) 12272.

[3] J. Lim, C. Park, M. Kim, H. Kim, J. Kim, D.-S. Lee, Advances in single-cell omics and multiomics for high-resolution molecular profiling, Experimental Molecular Medicine 56 (3) (2024) 515–526. doi:10.1038/s12276-024-01186-2.

[4] K. Miyazawa, K. Ito, M. Ito, Z. Zou, M. Kubota, S. Nomura, H. Matsunaga, S. Koyama, H. Ieki, M. Akiyama, et al., Cross-ancestry genome-wide analysis of atrial fibrillation unveils disease biology and enables cardioembolic risk prediction, Nature genetics 55 (2) (2023) 187–197.

[5] Z. Huang, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, et al., Salmon: survival analysis learning with multi-omics neural networks on breast cancer, Frontiers in genetics 10 (2019) 166.

[6] K. Chaudhary, O. B. Poirion, L. Lu, L. X. Garmire, Deep learning–based multi-omics integration robustly predicts survival in liver cancer, Clinical Cancer Research 24 (6) (2018) 1248–1259.

[7] X. Li, J. Ma, L. Leng, M. Han, M. Li, F. He, Y. Zhu, Mogcn: a multi-omics integration method based on graph convolutional network for cancer subtype analysis, Frontiers in Genetics 13 (2022) 806842.

[8] M. Picard, M.-P. Scott-Boyer, A. Bodein, O. Périn, A. Droit, Integration strategies of multi-omics data for machine learning analysis, Computational and Structural Biotechnology Journal 19 (2021) 3735–3746.

[9] P. S. Reel, S. Reel, E. Pearson, E. Trucco, E. Jefferson, Using machine learning approaches for multi-omics data analysis: A review, Biotechnology advances 49 (2021) 107739.

[10] D. Sun, M. Wang, A. Li, A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data, IEEE/ACM transactions on computational biology and bioinformatics 16 (3) (2018) 841–850.

[11] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, K. Huang, Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification, Nature communications 12 (1) (2021) 3445.

[12] H. Wang, K. Lin, Q. Zhang, J. Shi, X. Song, J. Wu, C. Zhao, K. He, HyperTMO: a trusted multi-omics integration framework based on hypergraph convolutional network for patient classification, Bioinformatics 40 (4) (2024) btae159. doi:10.1093/bioinformatics/btae159.

[13] T. Athaya, X. Li, H. Hu, A Deep Learning Method to Integrate extracelluar miRNA with mRNA for cancer studies, Bioinformatics (2024) btae653doi:10.1093/bioinformatics/btae653.

[14] F. Chen, G. Zou, Y. Wu, L. Ou-Yang, Clustering single-cell multi-omics data via graph regularized multi-view ensemble learning, Bioinformatics 40 (4) (2024) btae169. doi:10.1093/bioinformatics/btae169.

[15] A. Novoloaca, C. Broc, L. Beloeil, W.-H. Yu, J. Becker, Comparative analysis of integrative classification methods for multi-omics data, Briefings in Bioinformatics 25 (4) (2024) bbae331.

[16] J. Lim, C. Park, M. Kim, H. Kim, J. Kim, D.-S. Lee, Advances in single-cell omics and multiomics for high-resolution molecular profiling, Experimental & Molecular Medicine 56 (3) (2024) 515–526.

[17] N. A. Valous, F. Popp, I. Zörnig, D. Jäger, P. Charoentong, Graph machine learning for integrated multi-omics analysis, British Journal of Cancer (2024) 1–7.

[18] A. Baysoy, Z. Bai, R. Satija, R. Fan, The technological landscape and applications of single-cell multi-omics, Nature Reviews Molecular Cell Biology 24 (10) (2023) 695–713.

[19] Z. N. Kesimoglu, S. Bozdag, Supreme: multiomics data integration using graph convolutional networks, NAR Genomics and Bioinformatics 5 (2) (2023) lqad063.

[20] E. Flynn, A. Almonte-Loya, G. K. Fragiadakis, Single-cell multiomics, Annual review of biomedical data science 6 (1) (2023) 313–337.

[21] G. Rakshit, Komal, P. Dagur, V. Jayaprakash, Multi-omics approaches in drug discovery, in: CADD and Informatics in Drug Discovery, Springer, 2023, pp. 79–98.

[22] C. J. Creighton, Clinical proteomics towards multiomics in cancer, Mass Spectrometry Reviews (2022) e21827.

[23] J. L. Marshall, B. N. Peshkin, T. Yoshino, J. Vowinckel, H. E. Danielsen, G. Melino, I. Tsamardinos, C. Haudenschild, D. J. Kerr, C. Sampaio, et al., The essentials of multiomics, The Oncologist 27 (4) (2022) 272–284.

[24] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome biology 18 (2017) 1–15.

[25] C. Liu, L. Wang, Z. Liu, Single-cell multi-omics integration for unpaired data by a siamese network with graph-based contrastive loss, BMC bioinformatics 24 (1) (2023) 5.

[26] C. Wade, Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python, Packt Publishing, 2020.