EARLY ALZHEIMER'S DIAGNOSIS USING A MULTIVIEW APPROACH ON GENETIC AND CLINICAL DATA

Rosa Vicenti
Department of Computer Science
University of Bari

ABSTRACT

Early diagnosis of Alzheimer's disease remains one of the major challenges in the medical field, mainly due to the limited sensitivity of traditional methods during the initial stages of the condition. In this study, we propose a multiview deep learning approach that integrates genetic data (miRNA expression) and clinical metadata using dedicated autoencoders. Each data type is processed through a separate autoencoder to extract meaningful latent representations, which are then concatenated and used for classification via Random Forest and Multilayer Perceptron (MLP) models.

We conducted experiments on two datasets acquired through different technologies (microarray and high-throughput sequencing) and evaluated the model's generalization both through cross-validation and in cross-dataset settings. Results show that the multiview approach improves predictive performance compared to using only miRNA features, particularly when using the Random Forest classifier.

This work highlights the potential of multimodal autoencoder-based methods for early and accurate diagnosis of Alzheimer's disease, even when working with small and heterogeneous datasets.

1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that primarily affects cognitive function, leading to memory loss and cognitive decline. Early diagnosis is critical for effective intervention and management, yet traditional diagnostic approaches relying on neuroimaging, cognitive tests, and genetic biomarkers often lack the sensitivity needed for early-stage detection.

With advancements in artificial intelligence, deep learning models have demonstrated their potential in identifying patterns within complex, high-dimensional biomedical data. Among these, Graph Neural Networks (GNNs) and other machine learning approaches have gained prominence in analyzing relationships among various biological entities, such as genes, proteins, and imaging biomarkers. By leveraging graph-based models, it becomes possible to capture intricate associations between different features of AD, improving diagnostic accuracy.

This study explores the application of multimodal deep learning techniques, including autoencoders, and ensemble learning models such as Random Forest and Multilayer Perceptron (MLP), for the early diagnosis of Alzheimer's Disease. The research incorporates both genetic (miRNA data) and metadata to enhance predictive performance.

The approach involves the integration of feature extraction via autoencoders and traditional classification models to achieve robust and scalable AD detection. By comparing different configurations and preprocessing strategies, we aim to determine the most effective pipeline for analyzing multimodal data, ultimately contributing to more accurate and generalizable early detection methods for Alzheimer's Disease.

2 Related Work

In recent years, graph embedding techniques have been widely used in the biomedical field. A comprehensive review of the latest methodologies was conducted in [15], highlighting advances in integrating complex networks into the biomedical sector.
The work of [17] analyzed 11 representative methods across seven datasets from public sources, showing that matrix factorization-based approaches (e.g., HOPE, GraRep) are particularly effective for link prediction, while random walk-based methods (e.g., Struc2vec, node2vec [4], DeepWalk) achieve better results in node classification. At the same time, neural network-based methods such as LINE and SDNE have demonstrated competitive performance across different datasets, with GAE being particularly effective for large-scale graphs.
A further advancement is represented by the MRMTI model [7], which combines similarity networks and a bipartite miRNA-gene network into a heterogeneous graph, using multi-relation graph convolutions to capture structural and semantic information. Another relevant approach is DWLMI [16], which leverages DeepWalk to infer associations between lncRNA and miRNA in a heterogeneous graph comprising relationships with diseases, proteins, and drugs. The DWLMI model achieved an accuracy of 95.22% and an AUC of 98.56%, demonstrating the effectiveness of the approach.
Finally, the LINE (Large-scale Information Network Embedding) method [13] provides an efficient solution for large-scale networks, preserving both local and global structures through specific objective function optimization.

Several methods exist to learn node embeddings in graphs. GraphSAGE [5] can generalize to new nodes by aggregating neighbor features. Struc2vec [10] captures node roles based on structural similarity. HyperSAGE [2] handles complex relationships in hypergraphs using message passing. Role2Vec [1] learns embeddings based on node roles, using less memory and generalizing well.

METHOD	FEATURE-AWARE	TRANSDUCTIVE / INDUCTIVE
GraphSAGE [5]	Yes	Inductive
Struc2vec [10]	No	Transductive
HyperSAGE[2]	Yes	Inductive
Role2vec[1]	Yes	Transductive/ Inductive

2.1 Autoencoder

A relevant study by Venugopalan et al. [14] proposes a multimodal deep learning architecture for predicting the stage of Alzheimer's disease by integrating multiple data types, including MRI images, SNPs, and clinical metadata (EHR). In their approach, each data modality is processed by a dedicated autoencoder to extract intermediate feature representations. These representations are then concatenated and used as input to a classifier (e.g., SVM or Random Forest). Their results demonstrate that combining latent features from different modalities can improve prediction performance, especially in complex biomedical contexts.

2.1.1 Graph Autoencoder (GAE)

In recent years, the integration of bioinformatics and genetic data analysis has allowed us to search for new predictive biomarkers. In this context, Graph Autoencoders (GAEs) emerge, unsupervised learning models that extract latent representations from graph-structured data. A GAE typically learns an embedding for each node by reconstructing the adjacency matrix of the original graph. [11]
In [12] the authors proposed MGATE, a multi-channel graph attention autoencoder designed to predict novel lncRNA-disease associations (LDA). The goal is to enhance the identification of disease-related lncRNAs by leveraging a complex multi-layer graph representation that integrates rich biological data. A triple-layer complex graph is built, incorporating the relationships and similarities between lncRNAs, miRNAs, and diseases. A multi-channel graph autoencoder (GAE) captures both inter-graph (between different entities) and intra-graph (within the same entity) relationships. A graph-level attention mechanism and optimization strategies are applied to effectively merge different types of representations. Random Forest is used to predict new disease-related lncRNA candidates. MGATE outperforms seven state-of-the-art methods for LDA prediction. It achieves higher AUC and AUPR scores, indicating superior accuracy in detecting true lncRNA-disease associations. Case studies on three types of cancer validate its effectiveness in ranking high-confidence lncRNA candidates. The model can be extended to other association prediction tasks, including miRNA-disease associations, drug-disease associations, and drug-target interactions.

This study [8] introduces HeMDAP, a novel approach designed to enhance the prediction of miRNA-disease associations (MDA). The method leverages a heterogeneous graph connecting miRNAs, genes, and diseases, integrating multi-view learning with contrastive learning strategies to accurately identify these associations.

The study conducts both structural and semantic analyses of the graph using two distinct learning perspectives: network structure view and meta-path view.

Experimental results demonstrate that HeMDAP outperforms existing methods in predicting miRNA-disease associations, proving particularly effective in analyzing complex biological networks.

Despite its promising results, there is room for further improvement. HeMDAP marks a significant step forward in AI-driven diagnostics, providing an efficient and adaptable method for studying intricate biological networks.

3 Proposed Approach

This work follows a multimodal learning strategy inspired by the architecture proposed by Venugopalan et al. [14], where different types of biomedical data are processed through separate autoencoders to extract meaningful latent features. In our case, we adopt a similar method by applying dedicated autoencoders to miRNA data and clinical metadata, obtaining intermediate feature representations from each modality. These representations are then concatenated and used as input to a classification model.

Two types of classifiers were explored in this study: Random Forest [6], a robust ensemble method particularly suited for tabular and high-dimensional data, and Multilayer Perceptron (MLP) [9], a neural network architecture capable of modeling complex non-linear relationships.

4 Data processing and cleaning

In this study, two distinct datasets were used to evaluate the generalizability of the proposed approach across different data acquisition techniques.

The first dataset is a concatenation of three publicly available datasets retrieved from the Gene Expression Omnibus (GEO) repository [3]. All three datasets were generated using the **microarray technique** for miRNA expression profiling. After preprocessing and harmonization, the combined dataset provides a larger and more diverse set of samples, enabling more robust training and evaluation under standard cross-validation settings.

The second dataset was obtained using a different technology, namely **high-throughput sequencing** (HTS). This technique offers higher sensitivity and resolution compared to microarrays and is considered more modern. One of the key experiments in this study aims to address whether a model trained on microarray-based data can effectively generalize to data obtained through high-throughput sequencing. To test this, we train the model exclusively on the microarray dataset and evaluate its performance on the sequencing-based dataset. This cross-platform validation allows us to assess the compatibility between the two technologies and understand the limitations and potential of transferability in the context of Alzheimer's disease diagnosis using miRNA expression profiles.

Therefore, in the rest of this work, we will use the name **df_microarray** to refer to the dataset built from the concatenation of microarray-based data, and **df_hts** to refer to the dataset obtained through high-throughput sequencing.

Below, we provide details about their structure and class distributions.

4.1 *df_microarray* Dataset

The dataset **df_microarray** consists of **1256 rows** and **2610 variables**, including both miRNA expression data and metadata.

The dataset has the following class distribution:

Class	Number of Samples
AD	841
NC	300
MCI	115

Five-Fold Cross Validation

To evaluate the generalization ability of the model, a five-fold cross-validation strategy was applied. This configuration was used exclusively on the `df_microarray` dataset.

The procedure was structured as follows:

- The `df_microarray` dataset was split into five training folds and five corresponding test folds.
- For each of the five pairs:
  - An autoencoder was trained on the training fold.
  - The trained autoencoder was then used to extract compressed features from both the training and the corresponding test fold.
  - These features were given as input to a classifier (Random Forest or MLP), which was trained on the training fold and evaluated on the corresponding test fold.
- At the end of the process, the classification metrics (precision, recall, F1-score) obtained on each fold were averaged to compute the final performance.

This configuration allows for a balanced use of the entire dataset, improving the reliability and stability of the evaluation.

4.2 *df_hts* Dataset

The dataset **df_hts** contains **81 rows** and **2610 variables**.

The class distribution for **df_hts** is as follows:

Class	Number of Samples
AD	28
NC	21
MCI	32

4.3 Dataset Preprocessing

To enhance data quality, the following preprocessing steps were applied to the original dataset:

- **Handling missing data**: Removed MiRNA with a missing rate greater than 5%.
- **Null value replacement**: Replaced all null values with zero.
- **Column removal**: Deleted the column ‘Unnamed: 0’ , which held no informative value.

4.3.1 Metadata Preprocessing:

- Selected metadata features: age, sex, and apoe4.
- Normalized the age variable to a range of [1,2] using MinMaxScaler.
- Encoded categorical variables:
- Sex: Mapped female $\rightarrow$ 0, male $\rightarrow$ 1, with missing values assigned to 3.

The features (train set) and target variable "disease’ were separated.
The target variable was converted into numerical values using label encoding.

5 Methodology

Two main experimental configurations were implemented:

- **MiRNA-only:** Feature extraction and classification were performed using only MiRNA expression data.
- **Multiview (MiRNA + Metadata):** Separate autoencoders were trained on MiRNA and metadata features. Their compressed outputs were then concatenated and used for classification.

When the training and test sets had different feature sets, only the common features were retained. Any feature present in the training set but missing in the test set (or vice versa) was removed to ensure consistency.

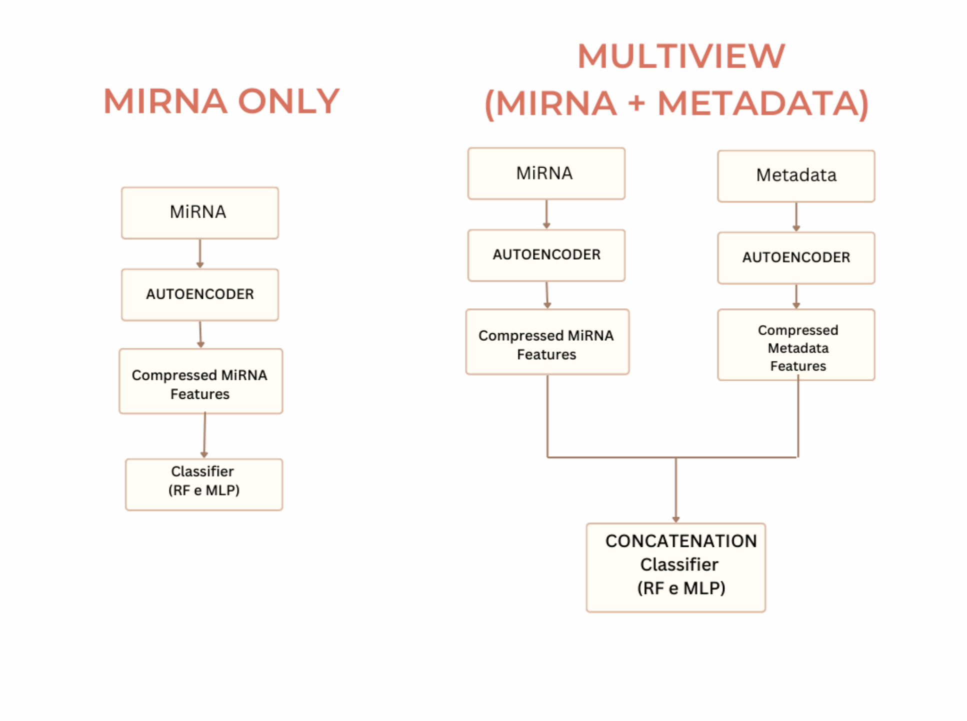
MIRNA ONLY
MiRNA
AUTOENCODER
Compressed MiRNA Features
Classifier (RF e MLP)
MULTIVIEW (MIRNA + METADATA)
MiRNA
AUTOENCODER
Compressed MiRNA Features
Metadata
AUTOENCODER
Compressed Metadata Features
CONCATENATION Classifier (RF e MLP)

Figure 1: Flowchart of the feature extraction and classification pipeline

5.1 Autoencoder Architecture

The proposed autoencoder is a two-level model that uses L2 regularization and dropout to prevent overfitting and improve generalization. Below is a description of the components and parameters:

1. Input and Initial Dimension

Input Layer: Receives a matrix with a number of features equal to the input dimension (input_dim), corresponding to the number of columns in the training set.

2. Encoder

The encoder compresses the input into two hidden layers:

- **First Hidden Layer:**
  - Neurons: 256 (parameter hidden_dim1).
  - Activation: ReLU.
  - Regularization: L2 with coefficient REG_COEFF = 0.03.
  - Dropout: A dropout with a rate of DROPOUT_RATE = 0.6 is applied after this layer.
- **Second Hidden Layer (Bottleneck):**
  - Neurons: 128 (parameter hidden_dim2), which is the intermediate encoding.
  - Activation: ReLU.
  - Regularization: L2 with coefficient 0.03.
  - Dropout: A dropout with a rate of 0.6 is applied.

3. Decoder

The decoder reconstructs the original input:

- **Decoding Layer:**
  - Neurons: 256, gradually restoring the original feature dimension.
  - Activation: ReLU.
  - Regularization: L2 with coefficient 0.03.
  - Dropout: A dropout with a rate of 0.6 is applied.
- **Output Layer:**
  - Neurons: Equal to input_dim, to reconstruct the original input.
  - Activation: Sigmoid, useful for outputs normalized between 0 and 1.
  - Regularization: L2 with coefficient 0.03.

4. Compilation and Training

- **Optimizer:** Adam.
- **Loss Function:** Mean Squared Error (MSE).
- **Training Parameters:**
  - Epochs: 100.
  - Batch Size: Experiments are conducted with both batch sizes 3 and 256.
  - The model is validated using test data during training to check its generalization.

5. Extraction of Intermediate Features

After training, a separate encoder model is created to extract the 128-dimensional intermediate features from both the training and test sets. This compact representation can be used for classification tasks (5.2).

5.2 Feature Classification with RF and MLP

The intermediate features were then used as input for two classifiers: Random Forest and Multilayer Perceptron (MLP).
For both Classifiers, the following hyperparameters were chosen:

- **Random Forest:**
  - Number of decision trees: 100
  - Random state: 42
- **MLP:**
  - Maximum iterations: 100
  - Random state: 42

6 Results and discussion

Configurations ID:

1. Five-Fold Cross Validation on *df_microarray*
2. Train: df_microarray, Test: df_hts
3. Train/Test: df_hts (50/50 split)

ID	Batch Size	Model	F1 (macro) only mirna	F1 (macro) multiview
1	3	Random Forest	0.55	**0.59**
1	3	MLP	0.42	**0.49**
1	256	Random Forest	0.46	**0.62**
1	256	MLP	0.42	0.42
2	3	Random Forest	0.17	0.14
2	3	MLP	0.17	0.17
2	256	Random Forest	0.14	0.14
2	256	MLP	0.14	**0.17**
3	3	Random Forest	0.35	0.25
3	3	MLP	0.13	**0.16**
3	256	Random Forest	0.22	**0.27**
3	256	MLP	0.13	**0.2**

Table 1: F1-Score Results

Table 1 clearly shows that the **multiview approach**, which combines MiRNA data with metadata, provides an advantage over using MiRNA data alone.

ID	Batch Size	Model	Precision (macro) only mirna	Precision (macro) multiview
1	3	Random Forest	0.72	**0.73**
1	3	MLP	0.39	0.39
1	256	Random Forest	0.5	**0.68**
1	256	MLP	0.45	**0.53**
2	3	Random Forest	0.12	0.12
2	3	MLP	0.12	0.09
2	256	Random Forest	0.12	0.09
2	256	MLP	0.12	0.12
3	3	Random Forest	0.37	0.26
3	3	MLP	0.08	**0.11**
3	256	Random Forest	0.16	**0.28**
3	256	MLP	0.11	0.08

Table 2: Precision Results

ID	Batch Size	Model	Recall (macro) only mirna	Recall (macro) multiview
1	3	Random Forest	0.52	**0.54**
1	3	MLP	0.47	0.47
1	256	Random Forest	0.44	**0.55**
1	256	MLP	0.51	0.51
2	3	Random Forest	0.33	0.33
2	3	MLP	0.33	0.33
2	256	Random Forest	0.33	0.33
2	256	MLP	0.33	0.33
3	3	Random Forest	0.35	0.25
3	3	MLP	0.33	0.33
3	256	Random Forest	0.19	**0.28**
3	256	MLP	0.33	0.33

Table 3: Recall Results

Experiment 1 – Five-Fold Cross Validation on df_microarray

In this first setting, the multiview pipeline achieved the best overall performance:

- With batch size 3, the **Random Forest** model improved from 0.55 (MiRNA-only) to 0.59 (multiview).
- For **MLP**, the F1-score increased from 0.42 to 0.49.
- With batch size 256, the gain was even more significant: **Random Forest** jumped from 0.46 to 0.62.

These results confirm that the inclusion of metadata helps the classifier better distinguish between classes, especially when the model is more stable and generalizable.

Experiment 2 – Train on *df_microarray*, Test on *df_hts*

In this cross-dataset scenario, performance was generally lower, as expected, since the two datasets were acquired using different technologies (microarray vs. high-throughput sequencing). This introduces a domain shift that makes generalization more challenging. Nonetheless, the multiview approach still performed slightly better or on par with the MiRNA-only setup:

- For instance, with batch size 3 and the **MLP** model, both approaches reached 0.17.
- For **Random Forest**, multiview improved the F1-score from 0.14 to 0.17.

Experiment 3 – Train/Test on *df_hts* (50/50 split)

In this case, where training and testing data come from the same distribution, the multiview configuration led to more consistent improvements:

- **Random Forest** improved from 0.22 (MiRNA-only) to 0.27 (multiview).
- **MLP** improved from 0.13 to 0.20.

These results suggest that metadata contains relevant and complementary information that enhances the learned representation.

Batch Size Analysis

In the comparative article [14] used as a reference, a batch size of 3 was applied during training. In this work, we also tested a larger batch size of 256 to assess its effect on performance. The results show that such a small batch size (3) does not consistently improve performance. In some cases, it even led to worse results, particularly with cross-dataset generalization. Moreover, training with a smaller batch size required significantly more time due to the increased number of updates per epoch.

7 Conclusion

Integrating metadata with MiRNA data through the multiview approach proved effective in most scenarios.

The improvement was particularly evident for **Random Forest**.

Overall, multiview representation provides richer information, leading to better classification performance across different settings.

Additionally, our experiments suggest that a larger batch size like 256 can yield better results and reduce training time, making it a more practical and efficient choice for this type of data.

8 Graphical Summary of Results

In this section, we provide a set of visualizations comparing macro-level performance metrics (Precision, Recall, and F1-score) across the different experimental configurations and batch sizes. The comparisons include both Random Forest and MLP classifiers, evaluated under two data modalities: MiRNA-only and Multiview (MiRNA + Metadata).

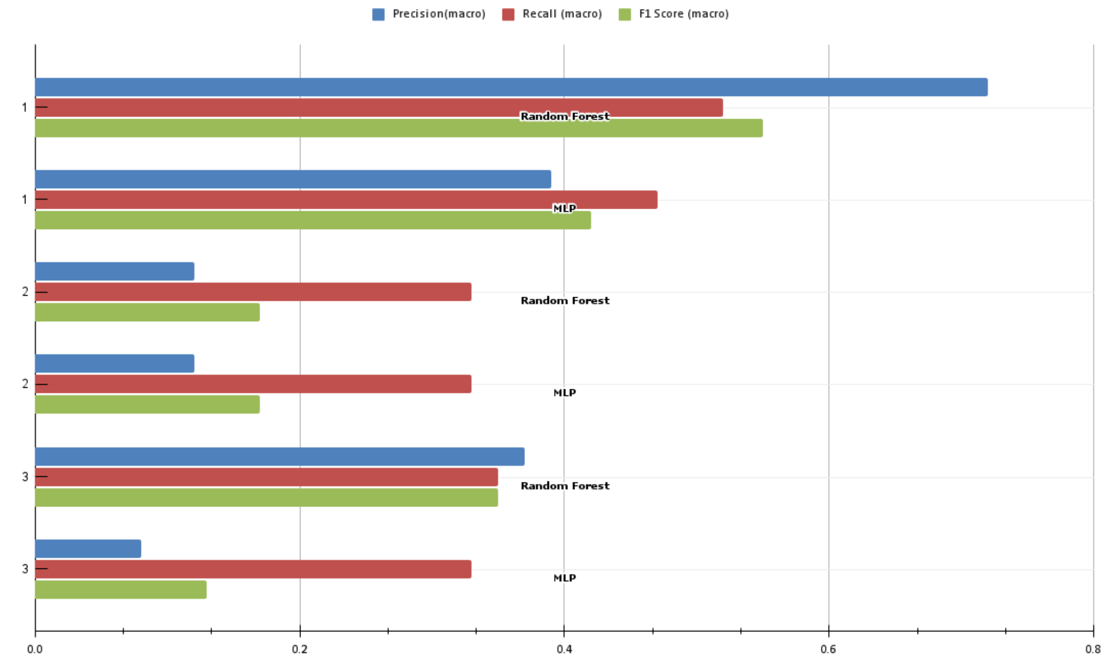

Figure 2: Results of macro metrics on Mirna only and batch size of 3 configuration

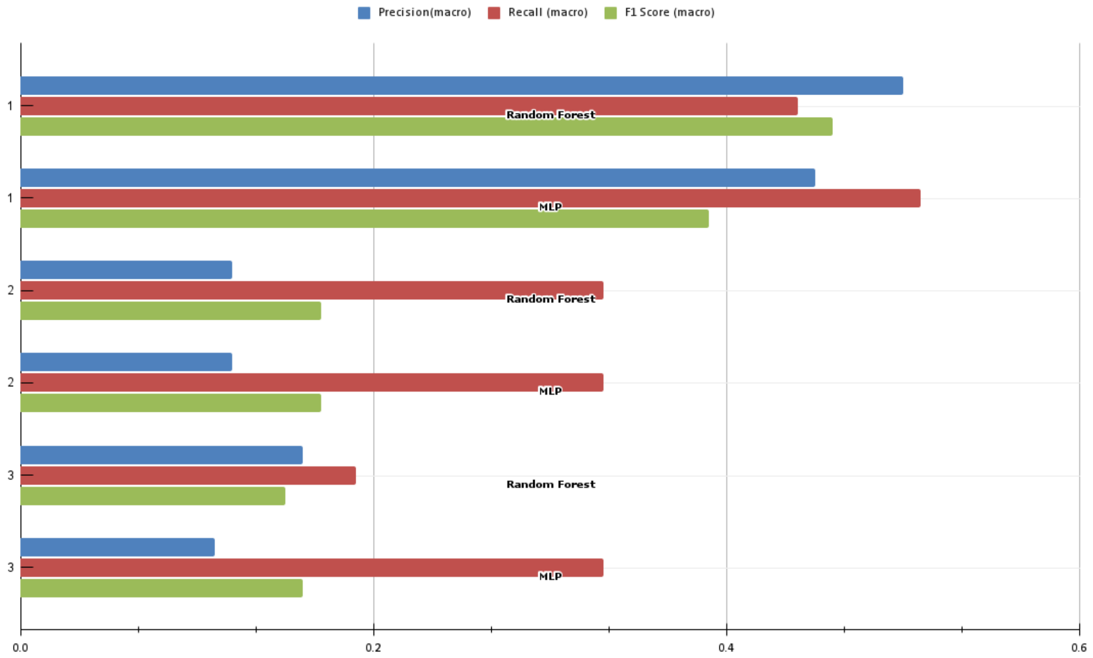

Figure 3: Results of macro metrics on Mirna only and batch size of 256 configuration

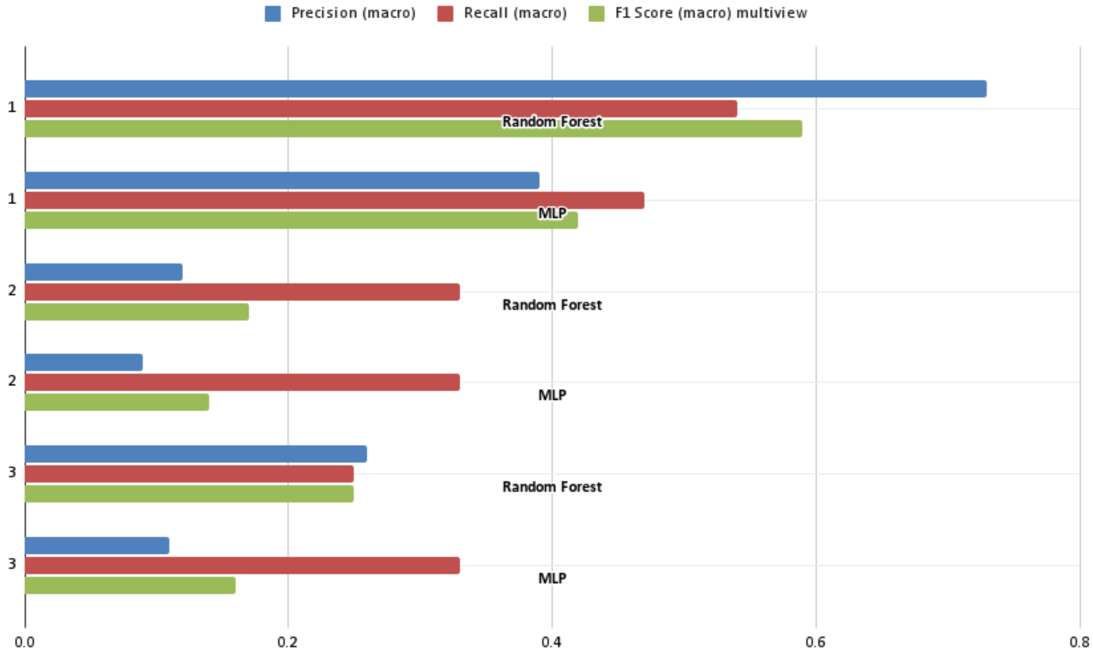

Figure 4: Results of macro metrics on Multiview and batch size of 3 configuration

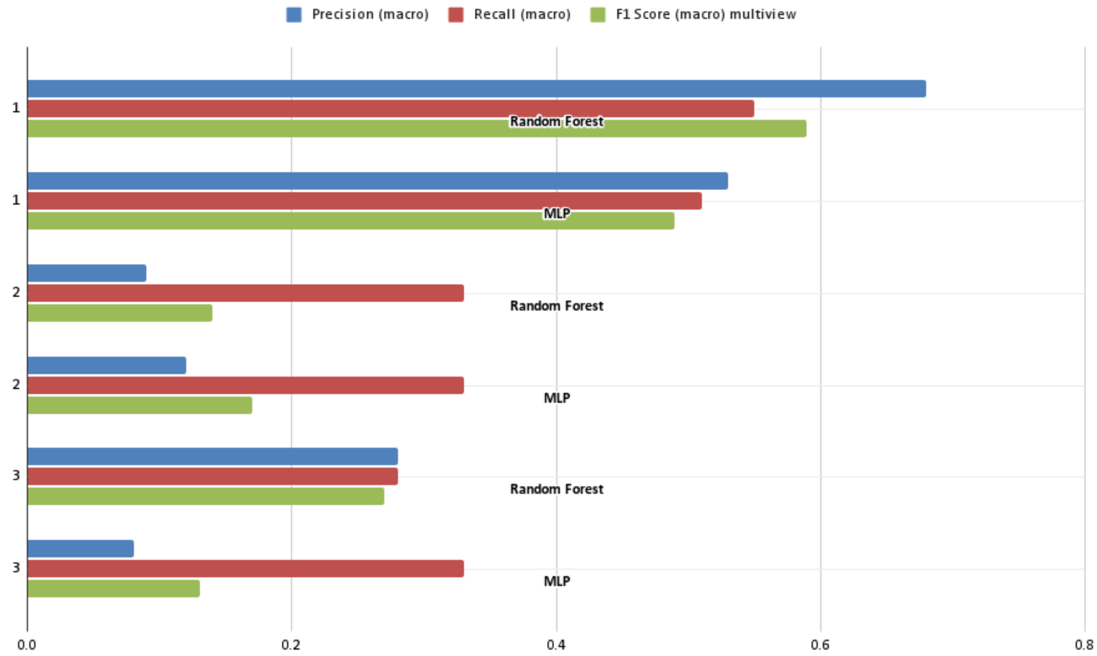

Figure 5: Results of macro metrics on Multiview and batch size of 256 configuration

References

[1] Nesreen K Ahmed, Ryan Rossi, John Boaz Lee, Theodore L Willke, Rong Zhou, Xiangnan Kong, and Hoda Eldardiry. Learning role-based graph embeddings. *arXiv preprint arXiv:1802.02896*, 2018.

[2] Devanshu Arya, Deepak K Gupta, Stevan Rudinac, and Marcel Worring. Hypersage: Generalizing inductive representation learning on hypergraphs. *arXiv preprint arXiv:2010.04558*, 2020.

[3] Emily Clough and Tanya Barrett. The gene expression omnibus database. *Statistical Genomics: Methods and Protocols*, pages 93–110, 2016.

[4] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[5] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[6] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[7] Jiawei Luo, Wenjue Ouyang, Cong Shen, and Jie Cai. Multi-relation graph embedding for predicting mirna-target gene interactions by integrating gene sequence information. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 04 2022.

[8] Yunjie Ma, Fei Wang, Qiyu Feng, Zuocheng Wang, and Luyu Xie. Hemdap: Heterogeneous graph self-supervised learning for mirna-disease association prediction. *IEEE Transactions on Computational Biology and Bioinformatics*, pages 1–12, 2025.

[9] Sankar K Pal and Sushmita Mitra. Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on neural networks*, 3(5):683–697, 1992.

[10] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 385–394, 2017.

[11] Julian Carvajal Rico, Adel Alaeddini, Syed Hasib Akhter Faruqui, Susan P Fisher-Hoch, and Joseph B Mccormick. A generative framework for predictive modeling of multiple chronic conditions using graph variational autoencoder and bandit-optimized graph neural network. *arXiv preprint arXiv:2409.13671*, 2024.

[12] Nan Sheng, Lan Huang, Yan Wang, Jing Zhao, Ping Xuan, Ling Gao, and Yangkun Cao. Multi-channel graph attention autoencoders for disease-related lncrnas prediction. *Briefings in Bioinformatics*, 23(2):bbab604, 02 2022.

[13] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *WWW*. ACM, 2015.

[14] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific Reports*, 11:3254, 02 2021.

[15] Yaozu Wu, Yankai Chen, Zhishuai Yin, Weiping Ding, and Irwin King. A survey on graph embedding techniques for biomedical data: Methods and applications. *Information Fusion*, 100:101909, 2023.

[16] Long Yang, Liping Li, and Hai-Cheng Yi. Deepwalk based method to predict lncrna-mirna associations via lncrna-mirna-disease-protein-drug graph. *BMC Bioinformatics*, 22, 02 2022.

[17] Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M. Lin, Wen Zhang, Ping Zhang, and Huan Sun. Graph embedding on biomedical networks: Methods, applications, and evaluations. *CoRR*, abs/1906.05017, 2019.