

# A Secure AI-enabled Platform to Support the Automated Alzheimer's Disease Diagnosis

Claudio A. Ardagna<sup>1,3</sup>, Veronica Buttarò<sup>2,†</sup>, Michelangelo Ceci<sup>2,3,4</sup>, Marco Luzzara<sup>1</sup>, Antonio Pellicani<sup>2,3,\*,†</sup>, Gianvito Pio<sup>2,3</sup> and Antongiacomo Polimeno<sup>1</sup>

<sup>1</sup>Dept. of Computer Science, University of Milan, Via Celoria 18, Milan, 20133, Italy

<sup>2</sup>Dept. of Computer Science, University of Bari, Via Orabona 4, Bari, 70125, Italy

<sup>3</sup>Data Science Laboratory, National Interuniversity Consortium for Informatics (CINI), Via Volturno 58, Roma, 00185, Italy

<sup>4</sup>Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, Ljubljana, 1000, Slovenia

## Abstract

This paper presents the BA-PHERD platform, a secure AI-enabled platform for automated Alzheimer's disease diagnosis using microRNA expression data. The BA-PHERD platform employs a cloud-edge architecture with secure data ingestion, governance-compliant storage, and role-based access control to ensure privacy protection. From a machine learning viewpoint, the platform also introduces a novel patient embedding method that captures regulatory relationships between miRNAs through a correlation network, integrating expression patterns with RNABERT sequence features and miRNA expression values. Unlike traditional approaches treating miRNAs independently, our method constructs a correlation network and applies a GraphSAGE-based approach to generate comprehensive patient representations. Experimental validation on a dataset comprising 1,256 subjects (300 controls, 115 MCI, and 841 AD patients) demonstrates significant improvements, achieving F1-score of 0.66 and an accuracy value of 0.76, over competitor methods.

## Keywords

Machine learning, Secure platform, Alzheimer's Disease diagnosis

## 1. Introduction

Machine Learning (ML) techniques have been increasingly applied to analyze biological data, and, in particular, microRNA (miRNA) expression values. These small non-coding RNA molecules regulate post-transcriptional gene expression [1] and have emerged as potential disease biomarkers due to their stability in biological fluids and distinctive expression patterns in conditions such as Alzheimer's disease [2], Parkinson's disease [3], and various cancers [4, 5]. The advancement of high-throughput sequencing technologies [6] has significantly accelerated miRNA discovery, with findings systematically documented in public repositories, such as miRBase [7]. This huge amount of data enabled researchers to employ ML algorithms to identify complex expression patterns for disease diagnosis, and to generate more accurate biological hypotheses that can be validated through in-vitro experiments [8, 9].

Despite these advances, applying computational approaches to miRNA data raises significant challenges: *i*) the limited availability of high-quality data often results in small sample sizes, leading to overfitting issues and poor generalization capabilities; *ii*) most datasets exhibit the "curse of dimensionality", since they collect expression values of numerous miRNAs (features) compared with relatively few patients (instances), that introduces the need of feature extraction/reduction techniques [10]; *iii*) traditional ML approaches [11, 12] frequently treat miRNA expression levels as independent vari-

ITADATA2025: The 4<sup>th</sup> Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

\*Corresponding author.

<sup>†</sup>These authors contributed equally.

✉ claudio.ardagna@unimi.it (C. A. Ardagna); veronica.buttaro@uniba.it (V. Buttarò); michelangelo.ceci@uniba.it (M. Ceci); marco.luzzara@unimi.it (M. Luzzara); antonio.pellicani@uniba.it (A. Pellicani); gianvito.pio@uniba.it (G. Pio); antongiacomo.polimeno@unimi.it (A. Polimeno)

ORCID: 0000-0001-7426-4795 (C. A. Ardagna); 0009-0007-9910-8538 (V. Buttarò); 0000-0002-6690-7583 (M. Ceci); 0000-0002-6414-3727 (M. Luzzara); 0000-0002-4193-3486 (A. Pellicani); 0000-0003-2520-3616 (G. Pio); 0009-0003-1197-568X (A. Polimeno)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ables, overlooking complex biological interrelationships that could potentially enhance both predictive accuracy and biological interpretability of the resulting models.

Beyond methodological aspects, the development of systems to effectively gather and manage miRNA expression data introduces additional challenges, mostly related to data governance. Indeed, ensuring both data quality and patient privacy protection requires frameworks that comply with specific regulations, while maintaining the integrity of the collected data. The sensitive nature of patient-derived miRNA data requires stringent data protection protocols, particularly when integrated with other clinical information for disease profiling. These governance considerations become increasingly important as miRNA-based diagnostic approaches move toward clinical implementation.

To address these challenges, the project PRIN 2022 - BA-PHERD (Big Data Analytics Pipeline for the Identification of Heterogeneous Extracellular non-coding RNAs as Disease Biomarkers) proposes a comprehensive pipeline that enables healthcare practitioners to securely collect, manage, and analyze patient miRNA data while providing automated diagnostic insights to support clinical decision-making.

The BA-PHERD pipeline provides end-to-end functionality, from secure data collection and compliant data storage, to advanced analytical capabilities. The BA-PHERD pipeline includes an AI-based diagnosis module based on a novel patient embedding method that creates patient representations by modeling their miRNA expression profiles as a correlation network, leveraging the underlying biological relationships between these molecules. This network-based approach captures both individual miRNA expression patterns and inter-miRNA interactions to characterize each patient within a unified representation. The proposed embedding approach integrates data-driven statistical measures with structural information derived from the pre-trained model RNABERT [13], providing a comprehensive characterization of patient-specific miRNA activities in disease contexts. The constructed patient representations can be subsequently adopted for any downstream tasks, including disease diagnosis and biomarker discovery.

We evaluate the proposed approach for diagnosis purposes, using clinical miRNA expression datasets from the GEO repository, with a particular focus on Alzheimer’s disease (AD). This neurodegenerative disease exhibits distinct diagnostic challenges for physicians, including the fact that initial symptoms are often mistakenly attributed to normal cognitive aging, and the limitations of existing diagnostic procedures, that are invasive, costly, and labor-intensive [14]. The considered dataset includes subjects with Alzheimer’s disease (AD), Mild Cognitive Impairment (MCI), and healthy controls (CN). Particular attention is given to the MCI category, as it represents a critical transitional phase that could provide key insights into early detection and intervention [15].

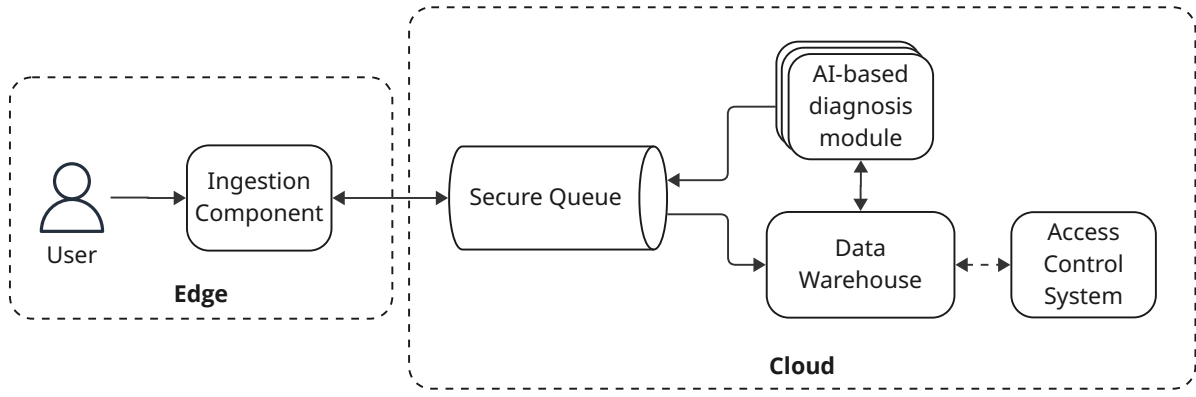
The remainder of the paper is organized as follows: Section 2 describes the BA-PHERD platform; Section 3 presents the experimental setup and discusses the results; Section 4 draws some conclusions and highlight possible future work.

## **2. The BA-PHERD platform**

In this section, we introduce the BA-PHERD platform. In particular, we first provide a brief overview of the technological architecture. Subsequently, we describe our AI-based diagnosis module, including the novel patient embedding strategy.

### **2.1. Architecture Overview**

In the context under consideration, it is important to note that clinicians and researchers may operate at different paces: clinicians mostly focus on immediate patient care, while researchers work on an extended timeline, requiring longer periods for data analysis and model development. In addition, healthcare facilities often lack the computational infrastructure necessary for complex miRNA analysis, requiring data to be processed in external data centers or specialized facilities. This scenario creates a critical need for secure data transfer and processing that supports strict privacy requirements, as sensitive patient information possibly need to cross organizational boundaries.



**Figure 1:** High level architecture overview of the BA-PHERD platform.

To address these challenges, we propose a digital platform that enables secure data sharing between clinical sites and research teams. The platform employs a flexible anonymization approach to safeguard patient privacy throughout the entire pipeline and data life cycle. Initially, clinical sites anonymize patients’ data, for instance, by replacing patient identifiers with persistent unique codes, ensuring that sensitive information remains under the control of the healthcare provider. Subsequently, each access to patients’ data is mediated by access control mechanisms that enforce anonymization policies based on user roles and permissions. For instance, when a researcher requests access to sensitive data, the system applies additional anonymization measures tailored to the researcher’s access rights.

Figure 1 shows a high-level overview of the platform architecture, illustrating the division of responsibilities between the edge and cloud components. The proposed solution is structured around five essential components, designed to address specific challenges of analyzing miRNA expression data:

- i) *Ingestion component.* It enables the ingestion of miRNA expression data from clinical sources.
- ii) *Secure queue.* It serves as the communication backbone, ensuring reliable transfer of data and intermediate processing results while maintaining workflow coordination. Its event-driven architecture allows the edge and cloud components to independently scale and evolve.
- iii) *Data Warehouse.* It provides governance-compliant management of raw and processed data, supporting storage and structured queries for data retrieval and analysis.
- iv) *Access control system.* It regulates data access through preconfigured roles and permissions, using a dual approach: resource-based access control (RBAC) that leverages database table structures, and attribute-based access control (ABAC) that uses metadata tags to enforce detailed access policies.
- v) *AI-based diagnosis module.* It analyzes miRNA expression profiles exploiting our novel patient embedding strategy for diagnosis purposes (see Section 2.2 for a detailed description).

The diagram in Figure 1 emphasizes the interaction flow among the five core components of the system. It highlights how these components collaborate to support machine learning tasks, from secure data ingestion at the edge to scalable processing and controlled access in the cloud environment.

The platform is designed to accommodate distinct user roles, each with specific responsibilities and permissions. In a representative real-world usage scenario, specialist doctors and researchers interact within the platform as follows.

Specialist doctors are responsible for uploading patients’ data to the platform, that will be ingested and stored in the *Data Warehouse*. During this process, they specify the intended scope of the data, either for *training* or *predictive* purposes. The former builds or updates a model using newly acquired data, while the latter generates diagnostic predictions based on the latest version of the model. Data

accessed from these tasks are filtered and modified by data protection policies, enforced through the *Access Control system*. The *Data Warehouse* automatically applies anonymization strategies, based on the roles/groups of the user who initiated the task. Two main mechanisms are employed: *row-level filtering*, which ensures that users view only records associated with their own group, and *column-level masking*, which anonymizes specific fields depending on the user's role. For example, the *Access Control system* masks personally identifiable information (PII) columns when accessed by a researcher, whereas it grants full data access to a specialist doctor.

The results of both types of task are accessible through the platform for a predefined duration, ensuring timely availability while adhering to data retention policies.

The proposed approach is based on containerized software architecture, making it compatible with complex orchestration systems such as Kubernetes, which enables scalable computational power to handle intensive machine learning workloads. This integrated platform bridges the gap between clinical data collection and research analysis, while ensuring appropriate privacy safeguards and supporting the distinct workflows of both clinical and research environments.

## 2.2. AI-based diagnosis module

After acquiring and managing the patient data, the BA-PHERD pipeline proceeds to perform automatic disease diagnosis exploiting a novel patient embedding approach that leverages miRNA correlation networks and machine learning techniques.

The first step involves the construction of a correlation network that captures the regulatory relationships among miRNAs. This network representation serves as the foundation for deriving meaningful embeddings that reflect both individual miRNA expression patterns and their functional associations. More specifically, given a dataset of  $p$  patients and  $n$  miRNAs represented as a matrix  $X \in \mathbb{R}^{p \times n}$ , we compute a correlation matrix  $C \in \mathbb{R}^{n \times n}$  where each element  $C_{ij}$  represents the Pearson correlation coefficient between miRNAs  $i$  and  $j$ . We then assess the statistical significance of each correlation using the Pearson correlation test, transforming  $C$  into an adjacency matrix  $B$  where only statistically significant correlations (p-value < 0.05) are retained. The resulting adjacency matrix  $B$  defines our miRNA correlation network  $G = (V, E)$ , where each vertex represents a miRNA and edges represent significant correlations between pairs of miRNAs. Furthermore, to enrich each node with biological features, we utilize RNABERT [13], a pre-trained transformer model designed for RNA sequence embedding. Specifically, for each miRNA  $i \in V$ , we adopt RNABERT to generate a feature vector  $\mathbf{f}_i$  that captures complex nucleotide patterns and contextual information within the miRNA sequence.

Focusing on the obtained correlation network, it is worth noting that we can find negatively correlated miRNA pairs. These negative correlations are biologically meaningful, and may provide biological insights into miRNA regulatory mechanisms. To better handle these correlations, we implement a dedicated transformation phase. Specifically, we define a new matrix  $A$  starting from  $B$  and adding, for each pair  $(i, j)$  with significant negative correlation  $A_{ij} < 0$ , complementary nodes  $\tilde{i}$  and  $\tilde{j}$  with feature vectors located, in the embedding space, specularly with respect to the center, namely:

$$\begin{aligned}\mathbf{f}_{\tilde{i}} &= \mathbf{1} - \mathbf{f}_i \\ \mathbf{f}_{\tilde{j}} &= \mathbf{1} - \mathbf{f}_j\end{aligned}\tag{1}$$

We then establish two new positive connections in the adjacency matrix  $A$ , specifically:

$$\begin{aligned}A_{i,\tilde{j}} &= |B_{i,j}| \\ A_{\tilde{i},j} &= |B_{i,j}|\end{aligned}\tag{2}$$

Notably, this transformation preserves negative relationships while maintaining a coherent network structure that can be processed by downstream graph-based algorithms.

After creating the miRNA correlation network, a key step is to obtain a suitable patient representation that can be used for automated diagnosis. For this purpose, we start by generating informative embeddings for each miRNA in the correlation network, employing a variant of GraphSAGE [16], an

inductive learning algorithm for node embeddings. Specifically, since we lack node-level labels for miRNAs, we adopt a self-supervised training approach that aims to predict the existing links in the network, formulating the objective function as:

$$\mathcal{L} = - \sum_{(u,v) \in \mathcal{D}_{\text{pos}}} \log(\sigma(h_u^T \cdot h_v)) - \sum_{(u,v) \in \mathcal{D}_{\text{neg}}} \log(1 - \sigma(h_u^T \cdot h_v)) \quad (3)$$

where  $\mathcal{D}_{\text{pos}}$  is the set of nodes that are connected by edges in  $E$  or co-occur within a fixed-length random walk,  $\mathcal{D}_{\text{neg}}$  is a set of randomly sampled disconnected nodes,  $h_u$  and  $h_v$  are the embeddings for nodes  $u$  and  $v$ , respectively, and  $\sigma$  is the sigmoid function.

GraphSAGE generates node embeddings through iterative neighborhood aggregation, where each node's representation is updated by combining information from its local neighborhood structure. For each node  $v \in V$ , the  $k$ -th layer update is computed as:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot \text{AGG}^{(k)}\left(\{h_v^{(k-1)}\} \cup \{h_u^{(k-1)} \mid u \in \mathcal{N}(v)\}\right)\right) \quad (4)$$

where  $W^{(k)}$  is the learnable weight matrix at layer  $k$ ,  $\sigma$  is an activation function, and  $\mathcal{N}(v)$  denotes the neighborhood of node  $v$ . To effectively capture the varying importance of correlations in our miRNA network, we implement a weighted mean aggregator that accounts for correlation strengths rather than treating all neighbors equally:

$$\text{AGG}^{(k)}(v) = \frac{h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} A_{vu} \cdot h_u^{(k-1)}}{1 + \sum_{u \in \mathcal{N}(v)} A_{vu}} \quad (5)$$

where  $A_{vu}$  represents the correlation strength between miRNAs  $v$  and  $u$ , ensuring that strongly correlated miRNAs contribute more significantly to the embedding update.

Starting with RNABERT features as initial node representations ( $h_v^{(0)} = \mathbf{f}_v$ ), we apply two GraphSAGE layers to progressively capture both direct neighborhood correlations and extended two-hop miRNA relationships within the correlation network. Then, after completing the self-supervised training process, we systematically discard the complementary nodes  $\tilde{i}$  and  $\tilde{j}$  that were introduced to handle negative correlations, while their learned representations have already influenced the embeddings of the original miRNA nodes through the neighborhood aggregation process. This approach ensures that the biological insights from negative correlations are preserved in the final embeddings without artificially expanding the feature space. This process results in an embedding matrix  $H \in \mathbb{R}^{n \times d}$ , where each row  $h_i$  represents the learned embedding for miRNA  $i$ , and  $d$  is the embedding dimension that captures both sequence-based features and network topology information.

Finally, to generate patient-level embeddings that effectively capture individual disease signatures, we first normalize each row of the original expression matrix  $X$  to account for varying sequencing depths and total RNA content across patients:

$$\tilde{X}_{ij} = \frac{X_{ij}}{\sum_{k=1}^n X_{ik}} \quad (6)$$

This row-wise normalization ensures that each patient's expression profile sums to unity, effectively controlling for technical variations in total expression magnitude. Subsequently, we project each patient into the learned miRNA embedding space through a weighted linear combination that leverages both the patient's expression profile and the network-informed miRNA representations:

$$P = \tilde{X} \cdot H \quad (7)$$

where each element  $P_{ij}$  represents the contribution of miRNA embedding dimension  $j$  to patient  $i$ 's representation, weighted by the patient's normalized expression levels.

The resulting patient embedding matrix  $P \in \mathbb{R}^{p \times d}$  contains representations for all  $p$  patients, effectively integrating three key sources of information: individual patient expression patterns, miRNA

sequence characteristics from RNABERT, and topological relationships captured from the correlation network structure. Furthermore, to enhance diagnostic accuracy and create a more comprehensive patient representation, we integrate these embeddings with relevant clinical variables, including demographic data (age and sex) and Apolipoprotein E (ApoE) genotype values. ApoE variants are well-established biomarkers associated with various neurodegenerative and cardiovascular conditions, making them particularly valuable for complete disease characterization.

For the final diagnostic classification, we employ a Random Forest (RF) classifier [17] with the Gini impurity heuristic. RF can naturally handle the integration of heterogeneous features (through concatenation) from both embeddings and clinical variables, without requiring extensive preprocessing or feature scaling steps. This ensemble method is particularly well-suited, as it can effectively capture complex interactions between miRNA-derived features and clinical indicators.

In summary, this module integrated in the BA-PHERD platform offers several key advantages over conventional miRNA-based diagnostic approaches: (1) it captures the complex regulatory relationships between miRNAs, rather than treating them as independent features, (2) it integrates expression-based and sequence-derived information, providing a more comprehensive representation of miRNA biology, and (3) it seamlessly combines molecular data with clinical variables for enhanced diagnostic accuracy.

### 3. Experiments

To evaluate the performance of the BA-PHERD platform, we conducted some experiments using a dataset obtained from NCBI Gene Expression Omnibus (GEO)<sup>1</sup>, a comprehensive repository for microarray and RNA-seq experimental data. The considered dataset contains different AD studies, namely GSE120584, GSE150693 and GSE242923. The selected studies include 300 control subjects (CN), 115 patients with mild cognitive impairment (MCI), and 841 patients with Alzheimer’s disease (AD). Inconsistencies in microRNA identifiers across different studies needed some standardization steps to guarantee a reliable analysis: *i*) eliminating duplicate identifiers within each dataset to ensure unique representation of microRNAs, *ii*) removing whitespaces to maintain consistent formatting, *iii*) truncating version-specific suffixes (e.g., v2 from miR-123a-3p-v2); and *iv*) mapping all identifiers to the latest version as reported in the miRBase repository.

We evaluated model performance through a stratified 5-fold cross-validation on the integrated dataset, ensuring that each fold preserves the original class proportions of the diagnostic categories (CN, MCI, AD). For each fold, we collected precision, recall, and F1-score measures for individual classes, then computed the overall performance by averaging these metrics across folds and classes. To ensure equal importance is given to all diagnostic categories regardless of class size, we report macro-averaged results, which prevents the larger AD group from dominating the performance evaluation.

For the automated disease diagnosis, we employed a Random Forest classifier with 100 trees using the scikit-learn 1.4 implementation. To comprehensively evaluate the contribution of each view of the patients, we performed experiments across multiple combinations of the following features sets:

- the original miRNA expression data (henceforth denoted as **expr**);
- patient metadata, that contain clinical information, namely, age, sex and ApoE (henceforth denoted as **meta**);
- the features identified through the embedding approach we propose in Section 2.2 - Equation (7) (henceforth denoted as **emb**).

All the considered combinations include the **emb** view, since it is the core of the proposed method. On the other hand, we compare the obtained results with those achieved by the approach proposed in [18], which is mainly based on the adoption of a Random Forest (RF) or a Multi-Layer Perceptron (MLP) on the concatenation of the views **expr** and **meta**.

---

<sup>1</sup><https://www.ncbi.nlm.nih.gov/geo/>



Method	Views	Prec.	Rec.	F1	Acc.
RF [18]	expr	0.78	0.45	0.48	0.72
	meta	0.57	0.47	0.47	0.72
	expr-meta	0.78	0.44	0.47	0.71
MLP [18]	expr	0.62	0.50	0.45	0.55
	meta	0.46	0.40	0.38	0.70
	expr-meta	0.69	0.49	0.48	0.64
BA-PHERD	emb	0.76	0.57	0.61	0.73
	expr-emb	0.82	0.58	0.63	0.75
	meta-emb	0.80	0.60	0.65	0.75
	expr-meta-emb	0.84	0.59	0.64	0.76

**Table 1**

Results obtained by the BA-PHERD platform and its competitors, through stratified 5-fold cross validation.

### 3.1. Results and Discussion

In Table 1 we show the obtained results. From the table, we can observe that the competitor based on RF achieves a maximum F1-score of 0.48 with the **expr** view, while the MLP variant obtains the same result when also including the **meta** view. Both variants suffer from low average recall (0.44-0.47 for RF, 0.40-0.50 for MLP). On the other hand, the embedding strategy implemented in the BA-PHERD platform demonstrates substantial improvements over the competitors. Indeed, the **emb** view alone significantly outperforms all the results obtained by competitors, achieving an F1-score of 0.61 and an accuracy value of 0.73, representing a 27% improvement in F1-score compared to the best result obtained by competitors (0.48). The combination of multiple views in our approach leads to even superior performance, with the **expr-meta-emb** combination achieving the highest F1-score of 0.64 and an accuracy value of 0.76, outperforming the F1-score of the best competitor by 33%. The **meta-emb** combination (F1: 0.65, Acc: 0.75) demonstrates that clinical metadata and the features extracted through our embedding strategy can already reach the best performances in the diagnosis of AD. This is somehow expected because expression data are exploited in the computation of the final feature set (see Equation (7)). It is important to note that, with these two combinations (**meta-emb** and **expr-meta-emb**), performances in terms of precision and recall appear balanced (precision in the interval 0.80-0.84, recall in the interval 0.59-0.60), suggesting potential clinical utility.

The consistent improvement observed when including the **emb** view highlights the value of the proposed patient embedding approach that captures expression patterns through *i)* the original miRNA expression values, *ii)* regulatory relationships via correlation-based miRNA interactions, and *iii)* sequence-level information through the features extracted by RNABERT.

## 4. Conclusions

In this paper, we presented the BA-PHERD platform, a result of the PRIN 2022 project BA-PHERD - Big Data Analytics Pipeline for the Identification of Heterogeneous Extracellular non-coding RNAs as Disease Biomarkers. This framework addresses the critical need for accurate and early diagnosis of Alzheimer’s disease through the integration of biological data, a novel patient embedding approach and machine learning techniques.

Our approach uniquely captures expression patterns through the original miRNA expression data, regulatory relationships via correlation-based miRNA interactions, and sequence-level information through the RNABERT method. The proposed methodology achieves significant performance improvements, demonstrating a substantial enhancement in F1-score compared to existing methods.

The complete framework encompasses end-to-end capabilities for gathering, managing, processing, and analyzing both miRNA expression and clinical data, while maintaining strict compliance with

patient privacy criteria and regulatory requirements. We are currently working on validating the whole platform, involving actual clinicians, researchers and new patients, whose data are collected and processed directly in the hospitals. For future work, we plan to incorporate explainability approaches to identify candidate miRNA biomarkers that contribute most significantly to Alzheimer's disease diagnosis, that will be subsequently validated through in-lab experiments.

## Acknowledgments

This work has been partially supported by the European Union - NextGenerationEU through the Italian Ministry of University and Research, Projects PRIN 2022 "BA-PHERD: Big Data Analytics Pipeline for the Identification of Heterogeneous Extracellular non-coding RNAs as Disease Biomarkers", grant n. 2022XABBMA, CUP: H53D23003690006, CUP: G53D23002910006, and CN3 RNA - "National Center for Gene Therapy and Drugs based on RNA Technology", CUP: H93C22000430007.

## Declaration on Generative AI

The authors used Grammarly for grammar and spelling check. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

- [1] J. O'Brien, H. Hayder, Y. Zayed, C. Peng, Overview of microRNA biogenesis, mechanisms of actions, and circulation, *Frontiers in Endocrinology* 9 (2018).
- [2] X. Chen, D. Xie, Q. Zhao, Z.-H. You, MicroRNAs and complex diseases: from experimental results to computational models, *Briefings in bioinformatics* 20 (2019) 515–539.
- [3] H. Ding, Z. Huang, M. Chen, C. Wang, X. Chen, J. Chen, J. Zhang, Identification of a panel of five serum mirnas as a biomarker for parkinson's disease, *Parkinsonism & Related Disorders* 22 (2016) 68–73.
- [4] J. V. Carter, N. J. Galbraith, D. Yang, J. F. Burton, S. P. Walker, S. Galandiuk, Blood-based microRNAs as biomarkers for the diagnosis of colorectal cancer: a systematic review and meta-analysis, *British journal of cancer* 116 (2017) 762–774.
- [5] M. G. Schrauder, R. Strick, R. Schulz-Wendtland, P. L. Strissel, L. Kahmann, C. R. Loehberg, M. P. Lux, S. M. Jud, A. Hartmann, A. Hein, et al., Circulating micro-rnas as potential blood-based markers for early stage breast cancer detection, *PloS one* 7 (2012) e29770.
- [6] Y. Hu, W. Lan, D. Miller, Next-generation sequencing for microRNA expression profile, *Bioinformatics in microRNA research* (2017) 169–177.
- [7] S. Griffiths-Jones, mirbase: the microRNA sequence database, *MicroRNA Protocols* (2006) 129–138.
- [8] A. Vishnoi, S. Rani, mirna biogenesis and regulation of diseases: an updated overview, *MicroRNA profiling: methods and protocols* (2022) 1–12.
- [9] G. Pio, M. Ceci, C. Loglisci, D. D'Elia, D. Malerba, Hierarchical and Overlapping Co-Clustering of mRNA: miRNA Interactions, in: *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track*, Montpellier, France, August 27-31, 2012, volume 242 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2012, pp. 654–659.
- [10] H. Wirth, M. V. Çakir, L. Hopp, H. Binder, Analysis of microRNA expression using machine learning, *miRNomics: MicroRNA Biology and Computational Analysis* (2014) 257–278.
- [11] H. Azari, E. Nazari, R. Mohit, A. Asadnia, M. Maftooh, M. Nassiri, S. M. Hassanian, M. Ghayour-Mobarhan, S. Shahidsales, M. Khazaei, et al., Machine learning algorithms reveal potential mirnas biomarkers in gastric cancer, *Scientific reports* 13 (2023) 6147.



- [12] N. Gilani, R. Arabi Belaghi, Y. Aftabi, E. Faramarzi, T. Edgünlü, M. H. Somi, Identifying potential miRNA biomarkers for gastric cancer diagnosis using machine learning variable selection approach, *Frontiers in genetics* 12 (2022) 779455.
- [13] M. Akiyama, Y. Sakakibara, Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning, *NAR genomics and bioinformatics* 4 (2022) lqac012.
- [14] B. P. Leifer, Early diagnosis of alzheimer's disease: clinical and economic benefits, *Journal of the American Geriatrics Society* 51 (2003) S281–S288.
- [15] M. N. Sabbagh, M. Boada, S. Borson, M. Chilukuri, B. Dubois, J. Ingram, A. Iwata, A. Porsteinsson, K. Possin, G. Rabinovici, et al., Early detection of mild cognitive impairment (MCI) in primary care, *The Journal of prevention of Alzheimer's disease* 7 (2020) 165–170.
- [16] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, *Advances in neural information processing systems* 30 (2017).
- [17] X. Chen, H. Ishwaran, Random forests for genomic data analysis, *Genomics* 99 (2012) 323–329.
- [18] D. Rosa, A. Pellicani, G. Pio, D. D'Elia, M. Ceci, Exploiting microRNA Expression Data for the Diagnosis of Disease Conditions and the Discovery of Novel Biomarkers, in: *International Symposium on Methodologies for Intelligent Systems*, Springer, 2024, pp. 77–86.