MiRNA to Feature Correlation: A Strategy to Enhance Explainability in Alzheimer's Diagnosis

Emanuele Tanzi, matr. 807406

Table of contents

Abstract

In recent years, microRNAs (miRNAs) have emerged as promising non-invasive biomarkers for the early diagnosis of Alzheimer's disease (AD). However, machine learning models built on complex genomic embeddings often lack interpretability, limiting their clinical utility. This study proposes a methodology to bridge this gap by identifying a mapping between miRNA expression profiles and the genetic features used in classification models. Using Lasso regression, we derive a sparse linear transformation that links miRNA expressions to the embedding features of a previously trained Random Forest model. We then employ SHAP (SHapley Additive exPlanations) to interpret the model's predictions and trace feature importance back to specific miRNA expressions. This dual approach enables a biologically grounded and interpretable classification of patients into AD, Mild Cognitive Impairment (MCI), and Normal Control (NC) groups. The results highlight several miRNAs with consistent predictive influence across correctly classified cases, suggesting their potential role as stage-specific biomarkers. More importantly, the study introduces a general framework to enhance the explainability of genomic classifiers in neurodegenerative disease diagnostics, paving the way for more transparent and trustworthy AI applications in medicine.

1. Introduction

1.1. Genetic Analysis in medical diagnosis

The incorporation of genetic data into medical diagnostics has significantly advanced the early detection and mechanistic understanding of a wide range of diseases. High-resolution analyses of molecular-level alterations—such as gene expression profiles, sequence mutations, and regulatory elements including microRNAs (miRNAs)—enable the identification of pathological signatures at early stages of disease development.

Biomarkers, defined as quantifiable indicators of biological processes or pathological states, play a critical role in disease diagnosis, progression monitoring, and therapeutic evaluation. In the context of neurodegenerative disorders, ideal biomarkers are expected to be non-invasive, highly sensitive to early pathological changes, and mechanistically linked to disease etiology[1].

MicroRNAs (miRNAs) are short (~22 nucleotides), non-coding RNA molecules that modulate gene expression post-transcriptionally by binding to target messenger RNAs (mRNAs). Due to their stability in biological fluids such as blood, cerebrospinal fluid, and saliva, and their responsiveness to disease-related molecular alterations, miRNAs have emerged as promising candidates for non-invasive biomarker discovery. In Alzheimer's disease, distinct miRNA expression patterns have been implicated in key pathological mechanisms, including neuroinflammation, synaptic dysfunction, and amyloid-beta metabolism, thereby supporting

[1] Etiology refers to the cause or origin of a disease or a condition.

their diagnostic potential.

1.2. Genetic Analysis in Alzheimer's diagnosis

In neurodegenerative diseases such as Alzheimer's, early diagnosis is crucial for effective patient management and clinical trial stratification. Genetic and transcriptomic biomarkers, including miRNA expression profiles, have shown strong predictive power in identifying mild cognitive impairment (MCI) and distinguishing it from both healthy aging and advanced AD.

Numerous studies have demonstrated the value of circulating miRNAs as non-invasive biomarkers for early and differential diagnosis. Petracci et al. [1] highlighted the diagnostic relevance of blood-based miRNAs, particularly those in canonical regulatory pathways. Other bioinformatics studies identified miRNA–gene networks linked to AD progression: for example, Ninanajan et al. [2]. pinpointed hub genes and associated miRNAs such as miR-29a and miR-132, while Zhou and Pang [3] focused on inflammatory markers like miR-146a and miR-21.

Advances in AI and multi-omics integration have accelerated biomarker discovery. Latifi-Navid et al. [4]. identified a panel of 9 miRNAs and 8 genes predictive of AD onset and severity, while Subasinghe et al. [5] examined the role of miRNAs regulating mitochondrial function in early-stage AD.

Comprehensive reviews, such as Afrin et al. [6], advocate for multi-modal biomarker panels combining miRNAs, lipids, and imaging to enhance diagnostic precision and therapeutic decision-making.

1.3. Introduction of the applied methodology

While numerous studies have underscored the diagnostic potential of miRNA expression profiles in neurodegenerative diseases, their integration into predictive models often lacks interpretability—an essential requirement in clinical settings. In a previous study, a dimensionality reduction approach was applied to miRNA expression data, producing 128-dimensional embedding vectors for each patient. These embeddings served as input to a Random Forest classifier capable of predicting disease status across three classes: Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), and Normal Controls (NC). Although the model achieved reasonable classification performance, the black-box nature of the embeddings hindered any biologically grounded interpretation of the predictions.

The present study addresses this limitation by introducing a methodological framework aimed at enhancing model explainability. Specifically, we propose a strategy to trace back the predictive power of the learned features to the original miRNA expression profiles. This approach enables the identification of potentially relevant miRNA biomarkers while retaining the predictive capabilities of the existing model. By combining feature-to-miRNA mapping with model interpretation tools such as SHAP, we seek to bridge the gap between

accurate classification and biological insight, ultimately contributing to the development of interpretable diagnostic tools in Alzheimer's research.

2. Methodology

2.1. Dataset Description

The dataset used in this study was derived from the work described in the introduction and is distributed across the following files, each of which is briefly described below:

- *subj_embeddings_train.csv*: contains, for each patient (in total, 1256), the values associated with the 128 genetic features extracted in the previous study.
- *df_concat_final.csv*: includes the original measurements of numerous miRNA expressions for each patient. In addition to miRNA-related data (corresponding to 2558 columns), this file also contains demographic and clinical information such as country of origin, age, gender, and the measured *APOE4* allele count. It also provides the ground-truth label for each patient, corresponding to one of three diagnostic classes:
  - ***AD**: Alzheimer's disease*
  - ***MCI**: Mild Cognitive Impairment*
  - ***NC**: Normal Control*
- *graph_embeddings.csv*: stores the miRNA-based embeddings computed in the previous study.
- *feature_importance_prod_final.csv*: contains feature importance scores derived from model training, providing a quantitative assessment of each feature's contribution to classification.

In the following sections, we describe the pipeline developed to utilize the mapping between miRNA expressions and feature embeddings, with the aim of improving model explainability and identifying miRNA biomarkers relevant to disease classification. We begin by outlining the preprocessing steps applied to the data extracted from these files.

2.2. Pipeline Description

2.2.1. Data Cleaning

The starting point, based on the conclusions of the study described above, is as follows:

$$P = X \cdot H$$

Where:

- P represents the matrix extracted from *subj_embeddings_train.csv*.
- X represents the matrix extracted from *df_concat_final.csv.*

- H represents the matrix extracted from *graph_embeddings.csv*.

So, as a starting point, we correctly extracted the matrices, taking care not to lose the multiplicity between matrices X and H, verifying that we obtained matrix P as a result, thus avoiding losing the order on both patients and features.

Eventually, we obtained:

- $P \in \mathbb{R}^{1256\times128}$
- $X \in \mathbb{R}^{1256\times2558}$
- $H \in \mathbb{R}^{2558\times128}$

and verified that the product above described is correct (after applying a standardization on X and P).

2.2.2. Computing mapping between miRNA expressions and embeddings

The primary objective of this phase was to determine, for each extracted feature, which miRNA expressions were most influential. This was achieved by leveraging two datasets: one associating each patient with their genetic features, and another providing their miRNA expression profiles.

Since the transformation matrix **H**—originally used to project miRNA expressions into the embedding space—is assumed to be unavailable, our goal was to construct a comparable structure that could support downstream explainability. Unlike the original transformation, however, the desired mapping should allow for a high degree of sparsity, enabling us to distinguish relevant miRNAs from non-informative ones. This filtering is crucial to reduce the dimensionality of the problem and improve the interpretability of model predictions, while minimizing information loss.

To this end, we adopted **Lasso** (Least Absolute Shrinkage and Selection Operator), a linear regression technique with L1 regularization. Lasso is well suited for feature selection, as it tends to shrink the coefficients of less relevant variables to zero, effectively retaining only the most informative inputs. Applied in our context, Lasso identifies, for each genetic feature, a sparse set of miRNA expressions that best explains its variability. The result is a linear mapping, represented as a weight matrix, that quantifies the influence of each miRNA on each genetic feature.

Lasso finds, for each genetic feature, the minimum set of miRNAs that best explains its variability, allowing me to obtain a weight vector. The mapping obtained is a linear mapping that associates miRNA expressions with the genetic characteristics of patients.

The choice of Lasso was driven by two main considerations:

- The need to reduce the dataset to the most relevant miRNA expressions, thereby identifying the subset of variables most predictive of disease status in Alzheimer's patients.

- The knowledge that a functional transformation exists from miRNA space to feature space (as demonstrated in the original embedding process), even if it is not explicitly accessible. Therefore, rather than reconstructing miRNA expressions from features, our aim was to estimate an interpretable approximation of this mapping—favoring sparsity and interpretability over exact reconstruction.

$$Lasso\ Objective\ Function: \min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

- X: standardized miRNA expressions
- y: standardized target genetic feature
- β: weight vector (coefficients)
- λ: regularization parameter (automatically tuned via cross-validation)

Experimental settings:

- Model: lassoCV (with automatic λ selection)
- Cross validation: 5-fold
- Maximum iterations: 10,000
- Standardization: *StandardScaler*, applied to bot X and P

In the following subsections, we describe in detail the procedure used to compute the weight matrix and refine the set of miRNA expressions for subsequent analysis.

Analysis of the most influential features

To guide the computation of the miRNA-to-feature mapping, we first identified the most relevant features for classification. The *feature_importance* dataframe—obtained from a previously trained model—was sorted according to feature relevance in patient classification tasks. At this stage, Lasso regression was applied exclusively to this subset of top-ranked features.

The goal was to learn a mapping capable of reconstructing the feature embeddings from miRNA expression vectors. The resulting weight matrix captures, for each feature, the contribution of each miRNA expression. This mapping serves as the foundation for the subsequent explainability steps, as each weight can be interpreted as a measure of importance.

First reduction in dataset size

To reduce computational complexity and enhance focus on the most informative signals, we performed an initial dimensionality reduction on the weight matrix obtained from Lasso. Specifically, for each miRNA, we computed the sum of its absolute weights across all features and retained only the top 300 miRNAs with the highest cumulative influence.

This yielded a reduced version of the matrix **X**, of dimensions 1256 × 300, compared to the original 1256 × 2558 matrix.

Training Lasso on the entire dataset

Following the initial reduction, we retrained Lasso on the complete set of features using the 300 selected miRNAs. This training phase was computationally efficient due to the lower dimensionality.

Importantly, the reconstruction accuracy of the original feature matrix **P** was preserved, with excellent performance metrics:

- $R^2 = 0.99$
- $MSE = 0.01$

The resulting weight matrix, representing the linear mapping from miRNA expressions to genetic features, has dimensions 128 × 300.

Given the strong reconstruction performance and the still relatively large number of miRNAs retained, we proceeded with a second dimensionality reduction.

Second reduction in dataset size

To further sparsify the model, we applied a thresholding strategy on the weight matrix. For each feature (row), we computed a threshold based on the maximum absolute value in that row and zeroed out all weights below this threshold. Subsequently, we removed all miRNA expressions (columns) whose weights were null across all features.

This process yielded a final weight matrix of dimensions 128 × 93.

Evaluating P reconstruction

As a final validation step, we assessed the ability of the reduced mapping to reconstruct the original feature matrix **P**. Despite the aggressive reduction in dimensionality, the reconstruction remained highly accurate:

- $R^2 = 0.98$
- $MSE = 0.02$

These results confirm that the mapping retains essential information while significantly reducing the number of miRNA expressions considered, enabling more efficient and interpretable downstream analysis.

2.2.3. Training a Random Forest model

To establish a reference for downstream explainability analysis, we trained a Random Forest classifier on the original dataset—specifically, the matrix **P** containing the extracted genetic features. While Random Forest models are known for their predictive power, they are not inherently interpretable; this made them a suitable baseline for testing the effectiveness of our miRNA-based explainability strategy.

In addition to the genetic features, we incorporated relevant patient-level metadata during training. In particular, **age** and **APOE4** status were retained due to their well-documented

association with Alzheimer's progression. Other variables, such as country of origin, were excluded due to low variability and limited predictive relevance.

The model training process followed these steps:

- Exclusion of all patients with missing values in personal data.
- An 80/20 train-test split.
- Hyperparameter tuning via grid search.
- Optimization based on the **macro-averaged F1 score (F1-macro)**, to account for the class imbalance (67% AD, 24% NC, 9% MCI).

Once the optimal model configuration was identified, it was evaluated on the test set. The results were overall acceptable, though imbalanced performance persisted across the three diagnostic classes. As expected, the model demonstrated strong performance in identifying Alzheimer's disease cases, while classification of MCI and NC was less accurate.

The evaluation metrics for the best-performing model are shown in the table below.

	Precision	Recall	F1-score	Support
AD	0.787	0.897	0.839	165
MCI	1.000	0.565	0.722	23
NC	0.600	0.764	0.754	58

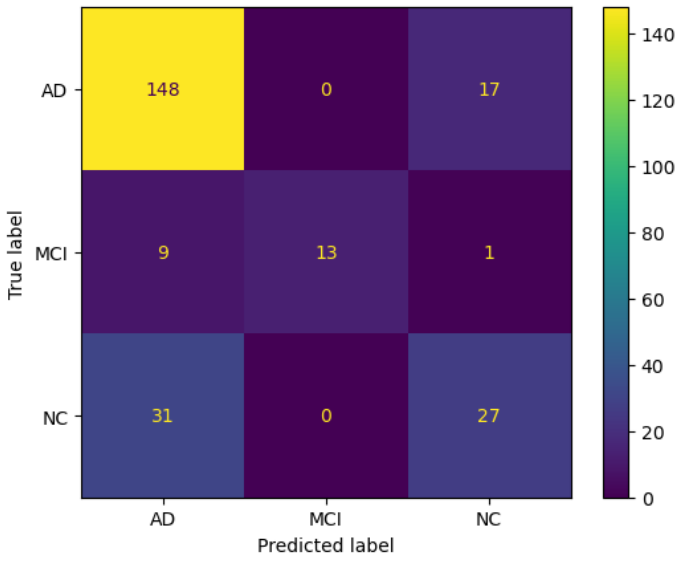

Figure 1: confusion matrix

Notably, the model tended to overclassify patients as AD, including several MCI cases. While some misclassifications may be informative to domain experts—e.g., MCI cases consistently predicted as AD—the goal of this study is not to optimize classification accuracy. Rather, our focus is on leveraging the model's behavior to develop an explainability pipeline that can be generalized to more complex, black-box classifiers.

2.2.4. SHAP for interpreting classification

The second part of the analysis focused on applying SHAP (SHapley Additive exPlanations) to interpret the contribution of individual features to the classification outcomes produced by the trained model. SHAP provides a principled method to quantify how each feature influences a specific prediction and in what direction.

To operationalize this, we implemented a function that, given a test instance, identifies both the predicted class and the most influential features contributing to that prediction. The function returns a ranked list of features sorted by their SHAP values (in absolute terms), allowing for the identification of both supporting and opposing evidence in the model's decision-making. Alongside each SHAP value, the corresponding raw feature value for the specific instance is also returned, facilitating further interpretation and the detection of potential inconsistencies in the model's reasoning.

2.2.5. Mapping + SHAP: explainability based on miRNA

The final step of the pipeline aimed to trace the influence of miRNA expressions on individual predictions by leveraging the SHAP-derived feature importances and the miRNA-to-feature mapping computed earlier.

To achieve this, we developed the miRNA_explain function. This function first filters out all features unrelated to miRNA-derived embeddings. It then identifies the most influential features for a given prediction and maps them back to the original miRNA expressions using the weight matrix. For each relevant miRNA, the function aggregates both its associated weights and raw expression values. The result is a ranked list of miRNAs for each test instance, ordered by their cumulative influence (*aggregated_weight*).

To compute the SHAP value associated to each miRNA, the following formula has been used:

$$agg\_weight\ (m) = \sum_{f \in \mathcal{F}_m} W_{f,m} \cdot SHAP(f)$$

While this second formula has been used to obtain the overall value of the observations for miRNA expressions w.r.t. the average population

$$agg\_raw\ (m) = \sum_{f \in \mathcal{F}_m} W_{f,m} \cdot RAW(f)$$

Additionally, to identify consistent miRNA patterns across correct classifications, we implemented a second function capable of extracting summary statistics for each miRNA across the dataset. The metrics include:

- ***mean_weight***: indicates whether the miRNA tends to support or oppose the classification of a given class.
- ***Std_weight***: reflects the variability of the miRNA's influence, helping assess the reliability of *mean_weight*.

- ***level_vs_avg***: measures how the miRNA's expression level compares to the population average.
- ***positive_ratio***: quantifies how often the miRNA contributed positively to the classification of the target class.
    - $\frac{1}{N}\sum_{i=1}^{N} 1(aggr\_weight_i(m) > 0)$
- ***mean_raw***: provides the average raw expression level of the miRNA in the examples under analysis.

These aggregated statistics support the identification of miRNAs that consistently influence classification decisions, offering a foundation for interpreting predictions in a biologically meaningful way.

3. Results

Beyond achieving correct classifications, the overarching objectives of this work were twofold:

- To associate each prediction with a biologically plausible explanation based on miRNA biomarkers, enabling domain experts to understand the rationale behind individual classifications.
- To identify, from the available data, miRNA biomarkers that may be relevant for predicting each of the three diagnostic classes (AD, MCI, NC).

As described in Section 2.2.4, SHAP was used to quantify the contribution of each feature to the model's predictions. A consistent finding across predictions is that **age** and **APOE4** status emerge as the two most influential features:

These findings align with existing literature and with the empirical observations from our model, confirming the importance of including these variables in the classification process. An illustrative example of SHAP-based feature importance is shown in *Figure 2*.

	feature	shap_value	raw_value	direction (+ ⇒ pro-class)
0	num__age	0.11807506435215145	0.7014892076044722	+
1	num__apoe4	-0.03047287971166497	-0.7684837404432598	-
2	num__49	-0.011635854921354393	0.06833541148984061	-
3	num__109	-0.007178682055379113	-0.045846352374075514	-
4	num__32	0.006494414656083336	0.019938056881912882	+
5	num__24	0.0060228473392271555	-0.09323352282947651	+
6	num__29	0.005978601435954261	0.0668188105949285	+
7	num__30	0.005455914735912586	0.03198035339015115	+
8	num__112	0.0050926949311178756	-0.10139664498823846	+
9	num__74	0.005021357545629144	0.12856700535843185	+

Figure 2: example of the most important features in a classification

We can interpret the results expressed in Figure 2 as follows:

- The observation of the age feature, which is significantly above the population average, as indicated by the row_value, pushes towards the predicted class (in the example, *AD*) since the positive shap_value.
- The observation of the feature apoe4, which is significantly below the population average, pushes against the predicted class, observing a negative shap_value.

The results are ordered based on the maximum absolute value of shap_value.

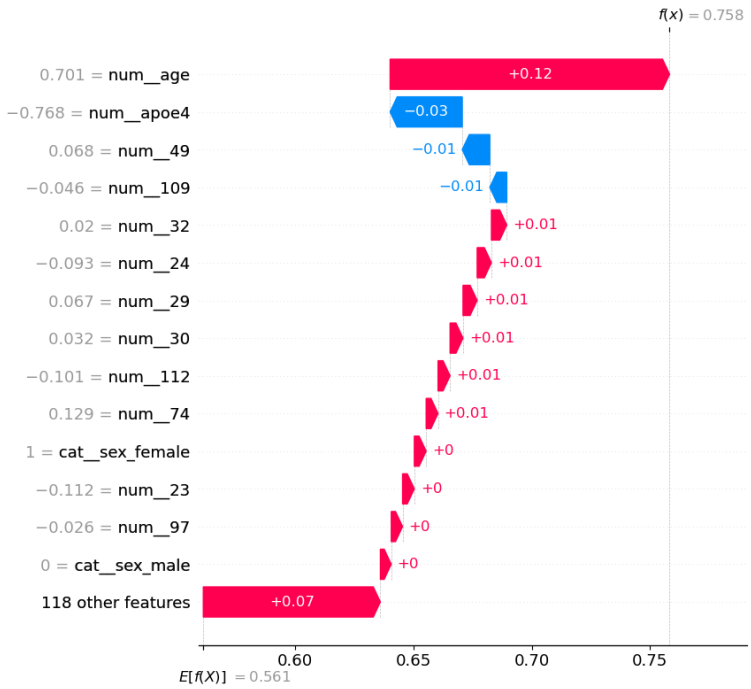

Figure 3: waterfall plot, graphical representation of SHAP values for a classification

While informative, these results represent only an intermediate step. Our primary goal is to trace predictive relevance back to specific **miRNA expressions**, thus identifying potential biomarkers. For this reason, we applied the mapping strategy described in Section 2.2.5 to translate SHAP feature importances into miRNA-level explanations.

The structure of the explanation remains similar: for each test instance, we derive a ranked list of miRNA expressions based on their aggregated contribution (aggregated_weight) to the most influential features. These are presented along with information on whether each miRNA's expression level is above or below the population average (agg_raw). An example of such output is shown in *Figure 4*.

	miRNA	aggregated_weight	agg_raw
0	hsa-miR-2861	0.011440480264205436	-0.11504813784851482
1	hsa-miR-6729-5p	0.008664085279676197	-0.3790482808194421
2	hsa-miR-1908-5p	0.005742331446907871	0.6307945470772633
3	hsa-miR-5787	0.0055785196181378635	0.31320730732942126
4	hsa-miR-150-5p	0.005062763113042427	0.3731062677960225
5	hsa-miR-1281	0.0042825065667354625	-0.008208151767773142
6	hsa-miR-6126	0.004256624269463154	0.20265186447676567
7	hsa-miR-6786-5p	0.004151927212852016	-0.028433771966038867
8	hsa-miR-92a-3p	0.0032325690591932512	-0.02422571122226973
9	hsa-miR-4306	0.003042857443553049	-0.012078345850617544

Figure 4: example representing miRNAs influence in a classification

Interpretation of these results can be exemplified as follows:

- **hsa-miR-2861** positively influenced the classification, being under-expressed relative to the population average.
- **hsa-miR-1908-5p** also contributed positively to the classification, but in this case through over-expression.

To further explore these findings, we extended the analysis across all correctly classified samples. For each class (AD, MCI, NC), we aggregated miRNA importance scores to identify consistently influential biomarkers. The results are summarized in *Figures 5*, *6*, and *7*.

	miRNA	mean_weight	std_weight	positive_ratio	mean_raw	direction	level_vs_avg
8	hsa-miR-1908-5p	0.004239526253651985	0.008624096270559 7	0.7094594594594594	0.9970062496696276	pro-class	high
87	hsa-miR-7704	-0.004204262707793654	0.009632497183149897	0.2635135135135135	-1.342058296125548	anti-class	low
69	hsa-miR-6780a-3p	0.003112654508209 2933	0.006421297787522 0075	0.6891891891891891	1.0975439195919257	pro-class	high
61	hsa-miR-6729-5p	-0.002360817431538 4025	0.011207023553603672	0.4054054054054054 3	-0.5714663604836641	anti-class	low
89	hsa-miR-8072	-0.001842790433129 1016	0.003550289512466748	0.3581081081081081	-0.5561753363800935	anti-class	low
26	hsa-miR-3937	-0.001697787477717 2684	0.003180574855863074 7	0.2432432432432432 6	-0.5905679670898752	anti-class	low
48	hsa-miR-5196-5p	0.001688816186058 6584	0.004318760186137 7125	0.6148648648648649	1.009707223726697	pro-class	high
40	hsa-miR-4734	-0.001614511606654 2701	0.005768183915825981	0.3513513513513513 7	-1.2382792355219963	anti-class	low
51	hsa-miR-5787	0.001504480561235 0525	0.005955225501346829	0.6216216216216216	0.3034530367679310 6	pro-class	high
50	hsa-miR-557	-0.001377635672877 7092	0.002111835299058 6184	0.25	-0.2320219732329579 2	anti-class	low

Figure 5: influence of miRNAs for class 'AD'

	miRNA	mean_weight	std_weight	positive_ratio	mean_raw	direction	level_vs_avg
87	hsa-miR-7704	-0.009493698908527621	0.008630809917318813	0.14814814814814814	-1.0860131564950821	anti-class	low
61	hsa-miR-6729-5p	-0.006854652565146517	0.011401637817560284	0.2962962962962963	-0.2580471864952984	anti-class	low
18	hsa-miR-3187-5p	0.005383880564298723	0.003648241780970211 2	0.9629629629629629	-0.037380463981861 84	pro-class	low
8	hsa-miR-1908-5p	0.005309776549170138	0.007514828941373 99	0.8518518518518519	0.5037096746740907	pro-class	high
24	hsa-miR-371a-5p	0.005295211472061 74	0.007580141587175979	0.7777777777777778	0.4119521900557422 4	pro-class	high
40	hsa-miR-4734	-0.004303543186370 12	0.006527343470696605 5	0.2592592592592592 4	-0.9968942629297947	anti-class	low
51	hsa-miR-5787	0.003693659659411346	0.006396170650358696	0.6666666666666666	0.1319849317735890 8	pro-class	high
1	hsa-miR-1237-5p	0.003651341778599176	0.005525834462808761	0.8518518518518519	-0.2998487778427520 5	pro-class	low
22	hsa-miR-3663-3p	-0.003441508304852742	0.004330847875828473	0.2592592592592592 4	-0.2774664146901906	anti-class	low
89	hsa-miR-8072	-0.003004054421878906 3	0.004169926870930897	0.2592592592592592 4	-0.4372135229424948	anti-class	low

Figure 6: influence of miRNAs for class 'NC'

	miRNA	mean_weight	std_weight	positive_ratio	mean_raw	direction	level_vs_avg
8	hsa-miR-1908-5p	0.004239526253651985	0.008624096270559 7	0.7094594594594594	0.9970062496696276	pro-class	high
87	hsa-miR-7704	-0.004204262707793654	0.009632497183149897	0.2635135135135135	-1.342058296125548	anti-class	low
69	hsa-miR-6780a-3p	0.003112654508209 2933	0.006421297787522 0075	0.6891891891891891	1.0975439195919257	pro-class	high
61	hsa-miR-6729-5p	-0.002360817431538 4025	0.011207023553603672	0.4054054054054054 3	-0.5714663604836641	anti-class	low
89	hsa-miR-8072	-0.001842790433129 1016	0.003550289512466748	0.3581081081081081	-0.5561753363800935	anti-class	low
26	hsa-miR-3937	-0.001697787477717 2684	0.003180574855863074 7	0.2432432432432432 6	-0.5905679670898752	anti-class	low
48	hsa-miR-5196-5p	0.001688816186058 6584	0.004318760186137 7125	0.6148648648648649	1.009707223726697	pro-class	high
40	hsa-miR-4734	-0.001614511606654 2701	0.005768183915825981	0.3513513513513513 7	-1.2382792355219963	anti-class	low
51	hsa-miR-5787	0.001504480561235 0525	0.005955225501346829	0.6216216216216216	0.3034530367679310 6	pro-class	high
50	hsa-miR-557	-0.001377635672877 7092	0.002111835299058 6184	0.25	-0.2320219732329579 2	anti-class	low

Figure 7: influence of miRNAs for class 'MCI'

These aggregated outcomes provide preliminary evidence of distinct expression patterns associated with specific disease stages. Notable examples include:

- **miR-1908-5p** and **miR-6780a-3p** consistently act as **pro-AD** markers, characterized by a positive mean weight and a high positive ratio (*mean_weight > 0.004*, *positive_ratio > 0.7*).
- **miR-3187-5p** emerges as a strong **pro-NC** marker (*positive_ratio = 0.96*), with a neutral role in the other two classes.
- **miR-7704** shows a **negative association with both AD and NC**, with a consistently negative mean weight (*mean_weight < -0.01*), suggesting a potential anti-class effect.

These findings demonstrate the potential of the proposed methodology to identify interpretable, miRNA-level biomarkers linked to model decisions. While these results remain exploratory, they suggest promising directions for future validation and refinement using larger datasets and more robust classifiers.

4. Conclusions and Future Work

4.1. Conclusions

This study set out to enhance the explainability of machine learning models used for Alzheimer's disease classification by establishing a biologically meaningful link between predictive features and miRNA expression profiles.

Starting from a previously developed dataset, which included both raw miRNA expression data and corresponding embedding vectors used in classification tasks, we proposed a method to reconstruct the relationship between miRNAs and learned features. By applying Lasso regression, first to a restricted set of highly relevant features and later to the full set, we derived a sparse and interpretable weight matrix that quantifies the influence of each miRNA on the model's internal representation.

Due to the unavailability of the original predictive model, we trained a new Random Forest classifier on the same feature space. The trained model served as a reference for explainability. SHAP was then employed to identify the most influential features contributing to each prediction. By leveraging the previously computed miRNA-to-feature mapping, we translated SHAP values into miRNA-level attributions, thereby offering insight into the molecular underpinnings of each classification.

This approach enabled us not only to provide a traceable explanation for individual predictions but also to identify recurring miRNA signatures associated with specific disease classes. While the model's predictive performance is not the focus of this work, the proposed pipeline offers a general framework for integrating explainability into genomics-based diagnostic models.

In summary, the main contribution of this study lies in the design of a reproducible and interpretable methodology that allows classification models to be made more transparent, while at the same time uncovering potential miRNA biomarkers for further biological investigation.

4.2. Future Works

The methodology developed in this study is applicable regardless of the specific predictive capacity of the underlying black-box model, and remains valid across different strategies aimed at reducing the dimensionality of miRNA expression data.

Future efforts should focus on extending the analysis to larger and more diverse datasets, which would likely improve classification performance and strengthen the reliability of identified biomarkers. In particular, training more robust models—whether based on Random Forests or alternative architectures—could enhance the detection of early-stage neurodegenerative patterns and further support model interpretability.

Additionally, given the extensive body of literature linking miRNA signatures to neurodegenerative disease mechanisms, future work may benefit from integrating prior biological knowledge to guide the selection of miRNAs used in the mapping. Focusing on the most biologically plausible and statistically influential expressions could yield more interpretable and clinically relevant diagnostic tools.

References

[1] Petracci et al., «Exploring the Role of microRNAs as Blood Biomarkers in Alzheimer's Disease and Frontotemporal Dementia,» *International Journal of Molecular Sciences,* 2025.

[2] V. Niranajan et al., «Integrated Bioinformatics Approach Reveals Key Genetic Biomarkers and Therapeutic Targets in Alzheimer's Disease,» p. 42, 2025.

[3] X. Pang, M. Zhou, «Polyphenols and miRNA interplay: a novel approach to combat apoptosis and inflammation in Alzheimer's disease,» 2025.

[4] H. Latifi-Navid et al., «AI-assisted multi-OMICS analysis reveals new markers for the prediction of AD,» *BBA Molecular Basis of Disease,* 2025.

[5] K. Subasinghe, «miRNA mediated mitochondrial function and gene regulation associated with Alzheimer's disease,» 2025.

[6] M. R. Afrin et al., «Advanced Biomarkers: Beyond amyloid and tau: Emerging non-traditional biomarkers for Alzheimer's diagnosis and progression,» *Ageing Research Reviews,* 2025.

Acknowledgements

I would like to thank Dr. Veronica Buttaro for her valuable support.