# PREDICTING ALZHEIMER'S DISEASE FROM MIRNA SEQUENCE AND EXPRESSION DATA WITH MACHINE LEARNING

by

Sydney Monserrate

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Biology

Committee:

| | |
|---|---|
| _____ | Dr. Christopher Lockhart, Thesis Chair |
| _____ | Dr. Ancha Baranova, Committee Member |
| _____ | Dr. Iosif Vaisman, Committee Member |
| _____ | Dr. Iosif Vaisman, Director, School of Systems Biology |
| _____ | Dr. Gerald L. R. Weatherspoon, Associate Dean for Undergraduate and Graduate Affairs, College of Science |
| _____ | Dr. Fernando R. Miralles-Wilhelm, Dean, College of Science |
| Date: _____ | Spring Semester 2024<br>George Mason University<br>Fairfax, VA |

# Predicting Alzheimer's Disease from miRNA Sequence and Expression Data with Machine Learning

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Sydney Monserrate
Bachelor of Science
Virginia Commonwealth University, 2021

Director: Christopher Lockhart, Research Assistant Professor
School of Systems Biology

Spring Semester 2024
George Mason University
Fairfax, VA

## DEDICATION

I dedicate this work to all the faculty, advisors, friends, family and peers who all supported me throughout the entire process.

## ACKNOWLEDGEMENTS

I wish to thank my professors and committee chair, Christopher Lockhart, for their dedication to my future.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

Alzheimer's Disease ..... AD
Area Under the Receiver Operating Characteristic ..... AUROC
Comma-Separated Value ..... CSV
Extracellular Vesicles ..... EVs
Machine Learning ..... ML
Receiver Operator Characteristic ..... ROC

## ABSTRACT

PREDICTING ALZHEIMER'S DISEASE FROM MIRNA SEQUENCE AND EXPRESSION DATA WITH MACHINE LEARNING

Sydney Monserrate, M.S.

George Mason University, 2024

Thesis Director: Dr. Christopher Lockhart

Approximately 6.5 million people, most of whom are 65 years of age and older, have been diagnosed with Alzheimer's Disease (AD) in the United States. Diagnosing AD has notoriously been difficult because disease progression can occur before the onset of cognitive impairment, and the physiological changes in AD brains are largely only observable in post-mortem studies. AD screening has been bolstered by novel biomarkers, including expression profiles of exosomal and circulating miRNAs. Although relatively new to biological studies, these miRNAs have become a focal point due to their widespread availability in bodily fluids and potential use in disease diagnostics. The purpose of our study was to investigate the utility of machine learning (ML) to predict AD-associated outcomes with miRNA sequence and expression data. Machine learning was performed leveraging the Orange Data Mining platform, which allowed us to quickly prototype various machine learning models and assess their performance numerically and

graphically. To utilize miRNA sequence data, we employed a k-mer bag of words model to quantify subsequences within miRNAs and predict if miRNAs are involved in AD pathways. We found that a random forest model provides the best predictions with an accuracy of 0.772 and an area under the receiver operating characteristic (AUROC) of 0.813. Interestingly, out all k-mers, we found that those rich in purines are the most predictive of miRNA association with AD. As a second modelling effort, we analyzed a previously published dataset [Ludwig et al. (2019) Machine Learning to Detect Alzheimer's Disease from Circulating Non-Coding RNAs *Genom. Proteom. Bioinform*. 17(4): 430-440] that measured miRNA expression in AD and healthy patients. A random forest model produced an accuracy of 0.786 and AUROC of 0.862 approximately reproducing the published results. We explored if the likelihood for miRNAs to be associated with AD-related pathways can be used as additional selection criteria for miRNA expression profile analyses and discuss the broader applications of our machine learning models in AD diagnostics. Ultimately, we believe our machine learning models will be useful to determine for new miRNA sequences if they are likely to be involved in AD and to pre-select miRNAs as biomarkers for expression profile analysis, which could be used as a diagnostic tool.

# INTRODUCTION

## ALZHEIMER'S DISEASE

Alzheimer's disease (AD) is a neurodegenerative disorder that affects over 6 million individuals in the United States alone, accounting for the main cause of dementia [1]. AD may be characterized by late or early onset, where early onset is predominately due to gene duplication events or mutations and late onset is part of the normal aging process [1]. The amyloid cascade hypothesis proposes that AD arises due to the accumulation of Aβ peptides, which aggregate into fibrils and whose diffuse oligomers are the primary cytotoxic species [2]. However, recently the validity of this hypothesis has been called into question and focus has instead been directed to other proteins such as tau [3].

Exosomes have been shown to carry biomolecules related to the progression of AD including Aβ, tau, and miRNAs involved in AD pathways [4]. MiRNAs are small, non-coding RNAs that are approximately 18 to 25 nucleotides in length but retain cellular function, for instance as regulators [5]. The creation of mature miRNAs starts with the production of a 60 nucleotide long strand hairpin RNA referred to as a precursor miRNA (pre-miRNA) [6]. Once generated, pre-miRNAs are cleaved in the cytoplasm, and mature RNAs are then released. Within the miRNA sequence, there also exists a conserved "seed sequence." The seed sequence can start at the 5'or 3' end of mature miRNA and is eight

nucleotide bases long, and it is this region that is important for the binding of miRNA to messenger RNA [7]. Characterizing miRNAs by their seed sequences may therefore provide more accurate results, as this sequence is complementary to messenger RNA [8]. Although miRNAs are non-coding, they still play a large role in gene regulation [7]. It has been shown that miRNAs are dysregulated and differentially expressed in AD; down- or up-regulated miRNAs include miR-9, miR-107, miR-29, miR-34, miR-181, miR-106, miR-146a, and miR-155 [9]. These miRNAs have interactions with pathways involved in AD, such as pathways involving BACE1 which is responsible for cleaving the Aβ precursor protein (APP) to produce Aβ peptides.

## EXTRACELLULAR VESICLES

Extracellular vesicles (EVs) are small lipid-enclosed particles that are released into extracellular space by virtually all cell types [10]. Depending on their cellular origin, EVs contain proteins and/or several classes of nucleic acids, in addition to cellular lipids. The typical size of an EV ranges from 10 nm to 1000 μm in diameter, whereas a human cell is, on average, between 20 to 30 μm [11]. The nanoparticle size of EVs and their lipid composition permits them to readily fuse and pass through cellular membranes including the blood-brain barrier [12]. EVs exploit these properties in their role as intercellular messengers and are involved in numerous pathways including inflammation and disease [13].

There are three main subgroups of vesicles: exosomes, apoptotic bodies, and microvesicles [12]. Exosomes are the smallest with a range of 40 to 100 nm in diameter, whereas microvesicles range from 100 to 500 nm and apoptotic bodies range 500 nm to 2

μm [14]. The origin of these EVs also differ; exosomes are produced by endosomes, microvesicles are from plasma membrane budding, and apoptotic bodies are extrusions from dying cells [15]. Exosomes are known to carry biological information that is differentially expressed in cancers, cardiovascular disease, and neurological conditions [16]. However, EVs may also be a transporter of pathogens causing exacerbation and spread of disease [12].

Aside from their biological function as intercellular messengers and traffickers, there has been a growing interest in EV usage in medicine and biotechnology [17] First, the molecular cargo of EVs can be quantified for disease screening and diagnostics [18]. Second, due to the ability of EVs to bind and inject foreign material into cells, they are being exploited as potential drug delivery systems [11]. Therefore, EVs have been explored as therapeutic remedies for drug delivery systems in biotechnology and novel research. Most commonly, EVs are used as biomarkers in fluids such as blood and urine due to their biological cargo that can correlate to physiological conditions [17]. The unique characteristics of exosomes including internal cell messaging and small size gives it an advantageous edge in the scientific community. However, there are potential challenges as well, as EVs are difficult to separate and identify their tissues of origin [19].

**MiRNA Biomarkers from Exosomes**

Exosomes can be found in any biological excretion including blood, urine, and saliva, making them a rich source of biomarkers for disease since the unique molecular makeup of exosomes can be leveraged for disease screening [15]. For AD, there have been several important studies to find differentially expressed biomarkers to serve as indicators

and quantifiers of the disease [20, 5, 21, 22, 23, 24]. The benefit of these screens is that they are minimally invasive and can be performed on commonly available cell types such as those found in serum [23]. For comparison, the traditional approach to AD diagnosis has been to conduct a clinical evaluation of the patient, and/or run positron emission tomography (PET) brain scans with a radio-labelled compound to detect significant Aβ fibril deposits [25]. It should be noted that accurate diagnosis of AD is not possible until a post-mortem autopsy has been conducted [25]. New diagnostic tools are therefore critical to ease this diagnosis burden and help efficiently screen for AD in living patients.

The increasing popularity of machine learning has unveiled a new direction to explore exosome-derived biomarkers for a variety of diseases [26, 27, 28]. Indeed, recent research has utilized exosomal data in conjunction with predictive models. Holdmann et al. [26] used urine-derived miRNAs from 28 patients and a random forest model to predict prostate cancer albeit with poor model performance (area under the receiver operating characteristic curve (AUROC) < 0.7) likely due to the small sample size. Morokoff et al. [27] employed a random forest to predict glioma in 108 individuals with a >99% accuracy based on an expression signature comprised of 9 miRNAs. Similarly, a signature of 7 miRNAs was identified for AD and used a decision tree, support vector machine, and adaboost for disease prediction with an accuracy of close to 90% [28].

**PREDICTING AD WITH MIRNA DATA**

Elucidating the molecular composition of EVs such as exosomes can be the key to new health screening models as a part of standard health screenings. Such diagnostic tools are non-invasive and, with a machine learning approach, can offer predictive power.

Furthermore, the quantity of publicly available exosomal data has reached a threshold that permits rapid evaluation of machine learning models. This data can be collected and used in concert to build a robust model that is rigorously evaluated for its generalizability.

In this thesis, we performed two machine learning prediction tasks using miRNA datasets to ascertain the presence of AD. First, we utilized miRNA sequence data from miRBase [29] and indications of miRNA involvement in AD pathways from miRPathDB [30]. Foremost, our major hypothesis was that miRNA sequence data could be used to predict AD. With a k-mer bag of words model, we therefore sought to predict if a given miRNA sequence was likely involved in AD. This model will be useful to classify novel miRNA sequences as likely AD-associated or not. We show with AUROC and accuracy that our model is predictive and that miRNA sequences can be used to establish their relationship to AD pathways. Second, we utilized miRNA expression data for healthy and AD patients [21] and predicted the presence of AD based on patients' expression profiles. Our machine learning model reaffirms the previously published results but expands on the types of models and breadth of their performance. Finally, we discuss the applicability of our models as diagnostic tools and how miRNAs could be pre-selected as biomarkers.

# METHODS

In this thesis, we explored the use of machine learning algorithms to predict Alzheimer's disease (AD) from miRNA. We performed two prediction tasks. First, we predicted AD involvement from miRNA sequences to determine if miRNA subsequences can be used as predictors. Second, we predicted if a patient has AD from existing miRNA expression data [21]. The sections below describe the protocols and tools that we utilized to perform our analysis.

## DATASETS

The first dataset we used for analysis consisted of AD-labelled miRNA sequences. Human miRNA sequences were downloaded from miRBase [29]. These records were filtered for their mature 5' or 3' miRNA sequences and subsequently processed to extract the seed sequence from positions 2 to 8. To determine if miRNA sequences were associated with AD, we downloaded from miRPathDB [30] experimentally validated associations of miRNA with AD pathways. Strictly, AD pathways were annotated with the KEGG database [31]. To construct a dataset suitable for machine learning, we labelled any miRNA with an AD relationship as a positive outcome "AD-related" in our dataset and as a negative outcome otherwise.

The second dataset we used was taken from Ludwig et al. [21]. This dataset examined at the expression of 21 miRNAs in AD and healthy patients from the United States and Germany. These 21 miRNAs included let-7d-3p, let-7f-5p, miR-103a-3p, miR-107, miR-1285-5p, miR-139-5p, miR-1468-5p, miR-151-3p, miR-17-3p, miR-26a-5p,

miR-26b-5p, miR-28-3p, miR-3157-3p, miR-345-5p, miR-34a-5p, miR-361-5p, miR-4482-3p, miR-486-5p, miR-5006-3p, miR-5010-3p, and miR-532-5p. MiRNAs were selected from the group's prior research and a literature search based on their dysregulation in AD. To perform expression analysis, miRNAs were derived from the patient red blood cells. We used this dataset as provided from Ludwig et al. with the exception that missing values were replaced with 0, i.e., no differential expression. In their analysis, they utilized LightGBM, which permits missing values, whereas our modeling approaches (described below) require explicit numerical values.

### K-MER BAG OF WORDS MODEL

One of the fundamental hypotheses of our machine learning approach is that we can utilize miRNA sequences to predict if a miRNA is associated with AD pathways. Because sequence data cannot be directly used in most modeling efforts, we applied a k-mer bag of words approach [32] to quantify this data.

The k-mer bag of words approach works as follows. Each miRNA sequence is first split into consecutive words of length k. Then, these words are counted, and the counts are used as a numerical feature for machine learning. As an example, "GTAGTA" with $k = 3$ would identify the subsequences "GTA", "TAG", "AGT", and "GTA". In the next step, these words are counted: "GTA" appears twice, and both "TAG" and "AGT" appear once (Fig. 1). A k-mer approach has been used previously to analyze miRNA sequences [33] and overall has widespread use to generate feature sets for biological sequences [32]. For example, given the four nucleic acid bases, there will in total be $4^k$ independent variables

created by this model, and the number of k-mers that will be identified is the length of the sequence L – k + 1.

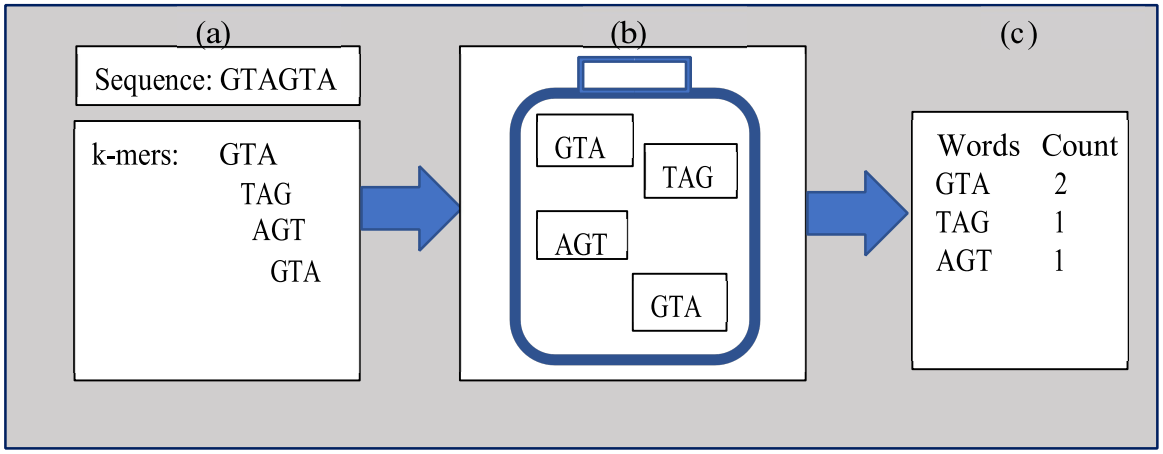


**Figure 1 k-mer bag of words model using "GTAGTA" sequences. In panel (a) all k-mers in the sequence are identified. In (b) we collect all these k-mers into a "bag" regardless of their initial sequence position. Finally, in panel (c) we count the number of each distinct k-mer.**

## ORANGE DATA MINING PLATFORM

To facilitate rapid building of machine learning models, we utilized the Orange Data Mining platform [34]. Orange is an ideal tool to use for machine learning prototyping due to its visual interface, which permits efficient testing of different machine learning algorithms, and its implementation in the Python programming language.

A representative pipeline of our machine learning analysis is presented in Fig. 2. The Orange Data Mining platform operates through widgets, each of which performs a function, and these widgets can be strung together to produce a cohesive analysis. We present two pipelines. In Fig. 2a, this explores use of a k-mer bag of words model, whereas in Fig. 2b our pipeline implements a standard machine learning approach as expected on

a tabular dataset which does not require additional processing. In Orange, the first step is to import data using the "CSV File Import" widget. To construct a k-mer bag of words model as in Fig. 2a, this is followed by use of "Corpus," which specifies textual elements (sequences) within the datasets, "Preprocess Text," which splits sequences into k-mers, and "Bag of Words," which counts the number of distinct k-mers in each sequence. Once the input data has been appropriately processed, "Select Columns" is used to specify the independent and dependent variables. This is then fed into a machine learning model, "Random Forest" as an example, and "Test and Score." In this case, "Random Forest" contains all the parameters necessary to run a random forest model. "Test and Score" specifies how the dataset should be divided into training and testing subsets or if cross-validation should be used. This module also displays the performance of the model.

**(a)**


**(b)**

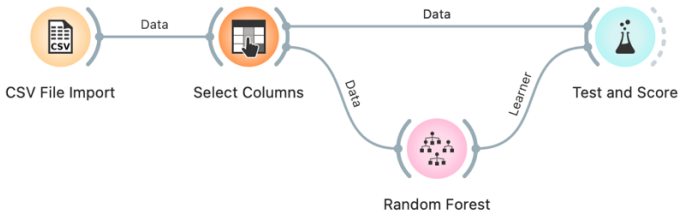

**Figure 2 Example model pipeline with Orange using a (a) k-mer bag of words model or (b) generic tabular dataset.**

## MACHINE LEARNING MODELS

We examined several machine learning models with varying levels of complexity, all of which were included as part of the Orange Data Mining platform [34]. The models we tested were k-nearest neighbors, naïve Bayes, logistic regression, support vector machine, decision tree, random forest, adaboost, and xgboost. Unless otherwise noted, all models were used with their default settings within Orange. As evidence of their applicability, previous machine learning studies focusing on miRNA sequences have also implemented these models [35]. To evaluate the performance of our model, we utilized k-fold cross validation with k = 5. The success of our models was evaluated on accuracy, precision, recall, and the area under the receiver operating characteristic (AUROC).

## RESULTS & DISCUSSION

Utilizing a machine learning approach with the Orange Data Mining platform [34], we studied miRNA involved in Alzheimer's disease (AD). Our efforts are divided into two tasks. First, we used miRNA sequences and information regarding their association with AD-related pathways to determine the predictive ability of miRNA sequences in predicting AD. In lieu of pathway enrichment studies, the results of this first model will be directly relevant to not only provide annotations for novel miRNA but also provide a quantitative ranking that indicates the likelihood of their relationship to AD. Second, we used a previously published dataset from Ludwig et al. [21] that explored use of circulating miRNAs as biomarkers for AD. We replicated their study, connected their results to those of our first model, and further discuss how our models can be used to improve miRNA selection as biomarkers for AD.

### PREDICTING ALZHEIMER'S DISEASE INVOLVEMENT OF MIRNAS

In this study we leveraged miRNAs from publicly available datasets to detect their involvement in the AD pathway. Specifically, mature miRNA sequences were downloaded from miRBase [29]. These sequences were further processed into their seed sequences, which is taken from positions 2 to 8 of the mature miRNA sequence and represents the canonical sequence where it binds to and affects protein messenger RNA [6]. Sequences were labelled as AD-related if they were found within the miRPathDB (30-Backes et al., 2017) experimentally validated dataset for AD, and AD pathways were annotated by

KEGG [31]. In total, our dataset contained 2,461 miRNA sequences, and 833 of these were marked as associated with AD.

**Model Preparation and Selection**

Several considerations must be made for machine learning evaluation of miRNA sequences. First, should mature miRNA sequences be used with the k-mer bag of words model, or should miRNA be restricted to their seed sequences? Second, what value of k produces the best results? Because high values of k results in a combinatorial explosion of independent variables and sparsity, we limit our exploration of k values from 3 to 5 (Fig. 3). Furthermore, we report only the results from the best model out of all those introduced in Methods.

**(a)**

**(b)**

**(c)**

**Figure 3 Word clouds showing the frequency of k-mers in the miRBase dataset when (a) k = 3, (b) k = 4, or (c) k = 5.**

As shown in Table 1, use of seed sequences consistently results in higher model performance. Because it is the seed sequences that directly interact and regulate messenger

RNA [7], our results suggest that focusing on these sequences is required for adequate model performance and that the rest of the mature miRNA provides background noise. The seed region is also commonly used to predict interactions of miRNA with targets due to its high degree of conservation [36]. It should be noted that, for these models, we considered a sample as a miRNA based on its designator from miRBase. When considering mature miRNA sequences, over 99% of these sequences were unique. Restricting these to their seed regions results in 79% unique sequences, which might result in some classification error in our prediction estimates.

**Table 1 Model performance versus k and sequence type.**

| k | Sequence Type | Best Model[a] | AUROC | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|
| 3 | Mature | Logistic Regression | 0.587 | 0.659 | 0.481 | 0.109 |
| 4 | Mature | XGBoost | 0.614 | 0.649 | 0.472 | 0.327 |
| 5 | Mature | Naïve Bayes | 0.605 | 0.588 | 0.419 | 0.561 |
| 3 | Seed | Random Forest | 0.785 | 0.757 | 0.688 | 0.516 |
| 4 | Seed | Random Forest | 0.794 | 0.760 | 0.682 | 0.544 |
| 5 | Seed | Random Forest | 0.770 | 0.734 | 0.623 | 0.540 |

[a]Best model was chosen based on AUROC.

The choice of k in the k-mer bag of words model must also be rigorously considered. For short sequences, as with miRNA seed sequences, large values of k will result in the creation of many independent variables which will sparsely be covered for any particular sample (Fig. 3). At the same time, larger values of k results in independent

variables that contain more information because these variables are more representative of the specific sequence being considered. Limiting our investigation to k of 3, 4, and 5, we found that k = 4 resulted in the best model performance. Taking these results together, our initial modeling efforts therefore indicate that seed sequences should be used and the k-mer bag of word model should be used with k = 4. Interestingly, the seed sequences considered are short enough that, instead of a k-mer bag of words model, we could implement a position-specific model and one-hot encode the nucleotide type for each position. However, based on our tests (not shown) this approach does not yield markedly different results than those presented above.

Using a k-mer bag of words model to construct our independent variables, how does model performance change for different machine learning algorithms? We focused only on the models introduced in Methods and their performance based on accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUROC). These results, presented in Table 2, indicate that overall, it is a random forest model that performs best, with AUROC of 0.794, accuracy of 0.760, precision of 0.682, and recall of 0.544. Other tree-based ensemble models, adaboost and xgboost, were also highly performative.

**Table 2 Performance of machine learning models with k = 4 using miRNA seed sequences**

| Model | AUROC | Accuracy | Precision | Recall |
|---|---|---|---|---|
| Random Forest | 0.794 | 0.760 | 0.682 | 0.544 |
| AdaBoost | 0.789 | 0.767 | 0.666 | 0.627 |
| XGBoost | 0.767 | 0.741 | 0.660 | 0.483 |
| k-Nearest Neighbors | 0.751 | 0.722 | 0.600 | 0.534 |
| Decision Tree | 0.731 | 0.737 | 0.643 | 0.501 |
| Naïve Bayes | 0.725 | 0.684 | 0.532 | 0.558 |
| Logistic Regression | 0.719 | 0.704 | 0.593 | 0.401 |
| Support Vector Machine | 0.602 | 0.539 | 0.390 | 0.643 |

To support AUROC in Table 2, we plot in Fig. 4 the receiver operating characteristic (ROC) curve. ROC plots the true positive rate against the false positive rate, and the area under this curve (AUROC) can be expressed as the probability that a model can take a random negative and random positive sample and correctly order them. As seen in Fig. 4, tree-based ensemble models (random forest, adaboost, and xgboost) all perform best, whereas other models exhibit distributions closer to the diagonal line, which is expected if the model behaves randomly.

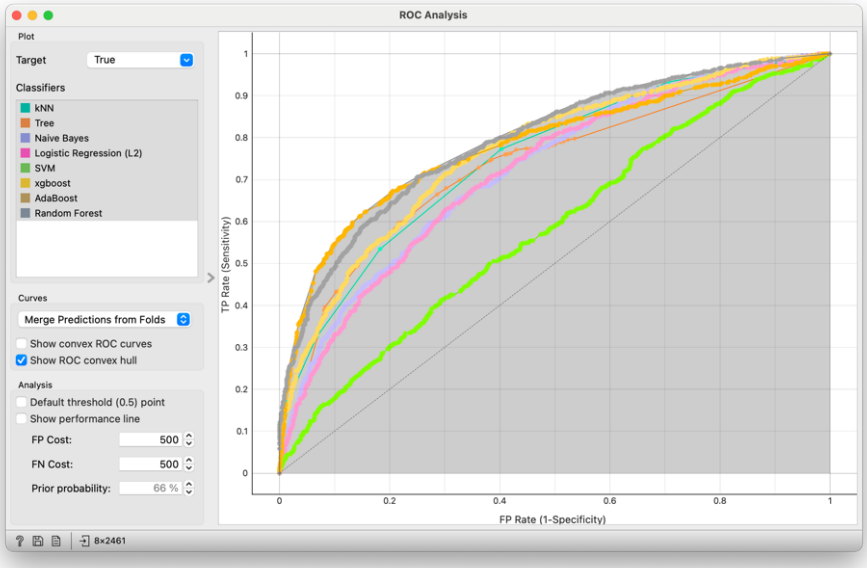


**Figure 4 ROC performance for machine models with k = 4 using miRNA seed sequences. ROC shows the true positive rate vs the false positive rate. The machine learning models corresponding to the lines are identified on the left-hand panel.**

## Model Performance

Our initial investigation in Table 2 and Fig. 4 indicate that random forests perform best for our dataset, which utilizes seed sequences and a k-mer bag of words model with k = 4. Because this initial investigation used the default random forest parameters as defined by the Orange Data Mining Platform [34], we further explored the impact of changing model hyperparameters. We found that, by increasing the number of trees from the default (10) to 100 and by setting the number of features considered by each tree as the square root of the total number of features (16), we marginally improved performance bringing accuracy from 0.760 to 0.772 and AUROC from 0.794 to 0.813.

However, even though accuracy and AUROC demonstrate adequate model performance, the recall of our model remains low (0.534). (For comparison, precision is

0.720.) Recall is computed by considering the number of true positives divided by the number of actual positives and expresses the probability associated with correctly predicting a positive sample. With a recall of roughly 0.5, this indicates that only approximately half of our predictions for actual positives are accurate.

To probe this issue, we present in Table 3 the confusion matrix that shows the number and percent of samples that were actually AD-related and predicted as AD-related (true positives), AD-related but not predicted as AD-related (false negatives), not AD-related but predicted as AD-related (false positives), and not AD-related and not predicted as AD-related (true negatives). The majority of our dataset (59%) is comprised of true negatives and a smaller portion (18%) comprises true positives. The proportion of false negatives (16%) exceeds false positives (7%) and approximately matches the proportion of true positives, which agrees with the reported recall.

**Table 3 Performance of machine learning models with k = 4 using miRNA seed sequences**

| | | **Predicted** | |
|---|---|---|---|
| | | **Not AD** | **AD** |
| **Actual** | **Not AD** | 1455 (59%) | 173 (7%) |
| | **AD** | 388 (16%) | 445 (18%) |

The misclassification seen in false positives or false negatives must be driven by the information collected in our k-mer bag of words model. Table 4 explores this possibility by considering the occurrence of specific k-mers in the true positive, true negative, false

positive, and false negative populations. The correlation of false positive k-mer occurrences with true positive occurrences is 0.77, whereas its correlation with true negative occurrences is 0.57. This indicates that false positives are largely driven by strong predictors of true positives, for instance the k-mer “GAGG”. For comparison, false negatives have correlations of 0.27 and 0.69 with true positives and true negatives, respectively, indicating that there are strong predictors of true negatives like the k-mer “GGGG”. The low correlation of false negatives with true positives also demonstrates that, for these sequences, there is insufficient information within existing true positives to result in adequate classification.

**Table 4 Occurrence of k-mers present in true positive, true negative, false positive, and false negative samples**

| k-mer | True Positives | True Negatives | False Positives | False Negatives |
|---|---|---|---|---|
| GAGG | 14.2% | 1.9% | 9.2% | 3.4% |
| GGAG | 13.9% | 3.0% | 6.9% | 3.4% |
| GGGA | 12.4% | 4.5% | 5.2% | 4.4% |
| AGUG | 8.8% | 3.0% | 6.9% | 3.9% |
| UGGG | 7.6% | 5.1% | 4.6% | 5.2% |
| GCAG | 7.0% | 2.1% | 7.5% | 4.4% |
| GGGG | 6.7% | 7.1% | 3.5% | 7.2% |
| AGGC | 6.7% | 1.4% | 6.9% | 3.4% |
| AGAG | 6.5% | 0.8% | 6.4% | 1.8% |
| GGCA | 5.8% | 2.3% | 5.8% | 3.6% |
| AGCA | 5.4% | 1.2% | 8.7% | 2.6% |
| GCAC | 5.2% | 0.3% | 5.8% | 2.1% |
| AAAC | 4.5% | 1.2% | 5.2% | 1.3% |
| GAGA | 4.5% | 1.6% | 5.8% | 3.6% |
| CCUC | 4.0% | 1.7% | 5.8% | 2.3% |
| CAGC | 4.0% | 1.3% | 6.4% | 3.4% |
| CCUG | 3.8% | 3.1% | 5.2% | 4.9% |
| CCCU | 3.4% | 3.0% | 5.2% | 5.2% |
| CCAG | 3.1% | 1.4% | 5.2% | 3.4% |
| CUCC | 3.1% | 0.8% | 5.2% | 2.3% |
| CAGU | 2.2% | 2.4% | 5.8% | 4.1% |
| GUAG | 2.2% | 1.5% | 5.8% | 2.8% |
| CUGC | 1.3% | 2.2% | 3.5% | 5.2% |
| GUGU | 0.0% | 1.6% | 5.2% | 3.4% |

Only k-mers with >5% false positives or false negatives are shown

How can the performance of our model be improved? The results above indicate that the information contained with the k-mer bag of words model is not sufficient to eliminate false positives and, particularly, false negatives. Following other predictive models based on miRNAs [36], we experimented with the addition of miRNA energetic

information to assess if it improves model performance. Using the RNAcofold program of the ViennaRNA package [37], we computed the interaction energy between the miRNA seed sequence and its reverse complement. The impact to our modeling results were subtle. The AUROC remained approximately the same (0.813 vs 0.812), but slight changes were seen for accuracy (0.772 vs 0.779), precision (0.720 vs 0.735), and recall (0.534 vs 0.541). Although these results are modest, they indicate that including additional independent variables derived from the miRNA sequences may improve or stabilize model performance.

**Interpretation of Model Results**

Our model predicts if miRNAs are likely to be associated with AD based on their subsequences. We can utilize the model results to identify those subsequences (k-mers) that are predictive using the Orange Data Mining platform's "Feature Importance" widget [34]. In Fig. 5, we show the top 10 k-mers ranked by our model, where the rank is determined from the ability of that k-mer to decrease model performance as measured by AUROC. The results in Fig. 5 strongly suggest that purines, as opposed to pyrimidines, are predictive of miRNA association with AD-related proteins. Seven of the top 10 k-mers contain only purines, and it is "GAGG" that produces the greatest effect. Interestingly, a recent study of miRNA sequences in humans revealed that the distribution of purines and pyrimidines is non-random with an abundance of pyrimidines [38]. These findings conflict with our results which shows, for AD-associated miRNAs, that there is a greater likelihood of purines in their sequences. The presence of purines in AD-related miRNAs may therefore be a target for future AD-related diagnostics or therapies. Interestingly, purine-

rich (specifically, G-rich) miRNAs are also more likely to be found in exosomes [39]. The "GAGG" motif, which we recognized as the most important k-mer for predicting if a miRNA was involved in AD pathways, has also been associated with miRNAs that bind to transcription factors [40].

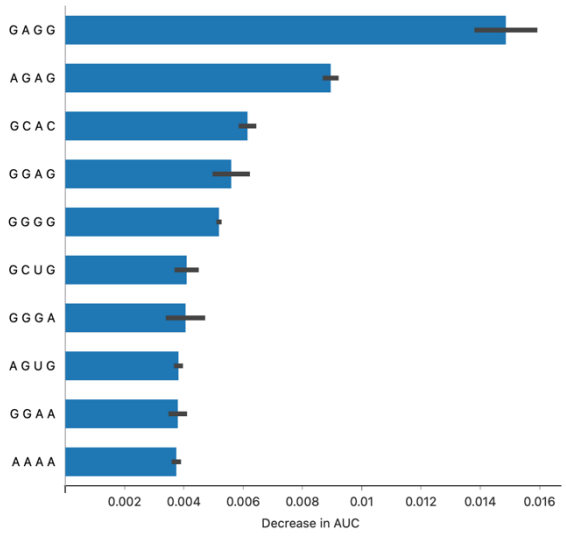


**Figure 5 Decrease in AUROC from specific k-mers indicating their model importance.**

## PREDICTING ALZHEIMER'S DISEASE FROM MIRNA EXPRESSION DATA

As a second modeling pursuit, we analyzed data from Ludwig et al. [21] which measured the expression profile of 21 miRNAs among 145 AD and 214 healthy patients located in the United States and Germany. The miRNA analyzed was circulating and derived from patient blood cells. The authors also performed an enrichment analysis and found that up-regulated miRNAs were largely expressed in serum, exosomes, t-cells, and

B-cells, whereas down-regulated miRNAs were expressed in monocytes and T-cells. This dataset represents a look at how miRNA can be used as a clinical basis to diagnosis AD, and we can directly investigate our ability to construct predictive machine learning models.

Table 5 shows the performances of our machine learning models against the Ludwig dataset. Random forest was best with an AUROC of 0.862 and accuracy of 0.786. Other models such as logistic regression, xgboost, support vector machine, and naïve Bayes also showed reasonable performance with AUROC > 0.8. To put these results into context, Ludwig et al. [21] utilized LightGBM to predict AD from their expression data with an AUROC of 0.876, only marginally higher than our best model and likely due either to their model choice, treatment of missing values (in our case we assumed missing values were 0), or cross-validation strategy.

**Table 5 Performance of machine learning models for Ludwig et al. dataset**

| Model | AUROC | Accuracy | Precision | Recall |
|---|---|---|---|---|
| Random Forest | 0.862 | 0.786 | 0.785 | 0.786 |
| Logistic Regression | 0.846 | 0.780 | 0.779 | 0.780 |
| XGBoost | 0.840 | 0.760 | 0.759 | 0.760 |
| Support Vector Machine | 0.832 | 0.772 | 0.770 | 0.772 |
| Naïve Bayes | 0.817 | 0.738 | 0.748 | 0.738 |
| k-Nearest Neighbors | 0.737 | 0.716 | 0.713 | 0.716 |
| AdaBoost | 0.710 | 0.719 | 0.720 | 0.719 |
| Decision Tree | 0.660 | 0.721 | 0.721 | 0.721 |

Fig. 6 shows the ROC curve for all models considered. This graph indicates the poor performance of models such as k-nearest neighbors, adaboost, and decision tree. All

other models perform reasonably well and demonstrate ROC curves without any significant defects.

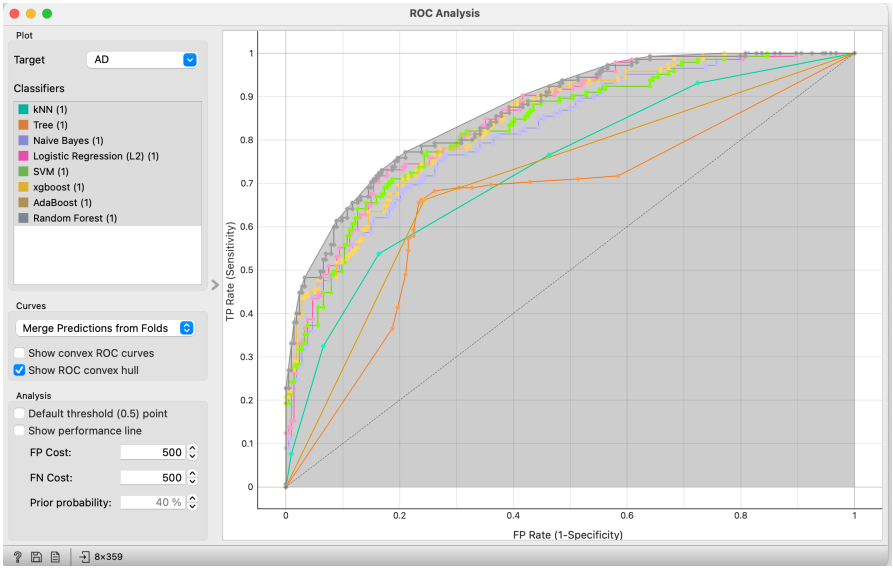


**Figure 6 ROC performance for Ludwig dataset. ROC shows the true positive rate vs the false positive rate. The machine learning models corresponding to the lines are identified on the left-hand panel.**

Ultimately, we sought to understand if we could improve model results for the Ludwig dataset using information from our first model based on miRNA sequences. Ludwig et al. [21] chose miRNAs for analysis based on their prior work and literature analysis. Because these miRNAs were either up- or down-regulated in AD compared to healthy controls, there is no correlation (0.02) between the posterior probability of miRNA involvement in AD pathways computed from our first modeling effort and miRNA importance when applying machine learning to the Ludwig dataset. This is to be expected because of the inclusion of down-regulated miRNAs in the Ludwig dataset, which strictly

cannot be considered in our first model which seeks to establish a positive association between miRNA and AD pathways. For comparison, the posteriors from our first model does show minor correlation (0.23) with the importances from exclusively up-regulated miRNAs.

The Ludwig dataset identified 8 up-regulated miRNAs: miR-107, miR-17-3p, miR-26a-5p, miR-26b-5p, miR-345-5p, miR-486-5p, miR-5006-3p, and miR-532-5p. Interestingly, if the Ludwig dataset is filtered to consider only these up-regulated miRNA, the predictive performance is high (AUROC of 0.852 and accuracy of 0.760 using a random forest model). We can therefore conclude that, within the Ludwig dataset, that up-regulated miRNAs are predictive of AD and the inclusion of down-regulated miRNAs is not strictly necessary.

Given that our first modeling effort seeks to exclusively predict positive association between miRNAs and AD, can those results be used to establish which miRNAs should be considered for study? We again filtered the Ludwig dataset but now considered only those miRNAs that were predicted as AD-associated from our first model (let-7f-5p, miR-103a-3p, miR-107, miR-1285-5p, miR-26a-5p, miR-26b-5p, and miR-5006-3p). The model has marginally diminished performance (AUROC of 0.827 and accuracy of 0.758 using a random forest model). However, these results and whether this approach can be useful is inconclusive because, for comparison, the remaining miRNAs that were predicted by our first effort to not be relevant to AD pathways still result in good performance (AUROC of 0.847 and accuracy 0.780). The miRNAs selected by Ludwig et al. all offer prediction of AD, and therefore we cannot conclude if we can reliably pre-select miRNA for study based

on these initial results. Despite the lack of an appropriate control, we believe that our models can still yield valuable insights into which miRNAs should be further explored as AD biomarkers.

# CONCLUSIONS

Efficient and non-invasive diagnosis of Alzheimer's disease (AD) remains an unsolved issue, and to this day AD can only truly be diagnosed with 100% certainty through post-mortem studies [41]. Ideally, diagnosis should proceed by collecting easily accessible patient samples like serum and, less so, cerebrospinal fluid and measuring specific biomarkers that are correlated with disease progression. For AD, circulating and exosomal miRNAs are a promising source of biomarkers that have shown dysregulation between healthy and AD patients [5, 21, 20]. When coupled with machine learning, expression profiles of these biomarkers can be a useful tool for diagnosis. Although the initial results of these models have been met with enthusiasm, there is currently no consensus about which biomarkers are adequate as a minimal set for AD diagnosis.

In this work, we sought to establish the relationship between miRNAs and AD to greater detail using machine learning and the Orange Data Mining platform [34]. Our main pursuit was to predict the likelihood of miRNA interaction with AD pathways from miRNA sequences, and we hypothesized that miRNA sequence information is sufficient to predict this pathway relationship. Secondly, we utilized previously published miRNA expression profiles of patients to predict if those patients had AD. Our goal with this second model was to survey machine learning models that offered prediction and evaluate if our sequence-based model could be used to assist in the pre-selection of miRNAs that are useful for analysis.

We therefore considered two separate approaches, both of which utilized machine learning to predict AD. Our first model focused on human miRNAs whose sequences were extracted from miRBase [29] and were annotated as AD-related if they were involved in AD pathways as curated in miRPathDB [30]. A k-mer bag of words model was applied to miRNAs to produce a quantitative dataset. We assessed if a full or seed sequences produce better modeling results and the impact of the size of k in the k-mer bag of words model on these results. We found that a random forest model utilizing seed sequences and setting k = 4 produces the best outcome, as measured by the area under the receiver operating characteristic (AUROC) and accuracy of 0.813 and 0.772, respectively. These results unambiguously demonstrate that miRNA sequences are indeed predictive of miRNA interaction with AD pathways. However, this model still could be improved due to its high rate of false negatives. We explored if additional information sources such as miRNA energetic information may aid in prediction, but this extra feature only provided a marginal improvement. More features, for instance those available through iLearn [42] could be useful to improve our model performance further. In addition, information regarding the location of miRNA expression (e.g., serum or cerebrospinal fluid) could provide discriminating information.

As a second approach, we utilized the miRNA expression dataset for healthy and AD patients from Ludwig et al. [21]. Using a random forest model, we were largely able to replicate the AUROC of the original authors (0.876 vs 0.862 from our model). We hypothesized that slight difference in model performance could be due to the different model framework and our treatment of missing values in the dataset. Whereas the authors'

LightGBM model permits missing data elements, the models available through the Orange Data Mining platform [34] require numerical features. Interestingly, we also demonstrated that several other machine learning models such as logistic regression, xgboost, support vector machine, and naïve Bayes produce acceptable predictive outcomes (AUROC > 0.8).

Given that our models have demonstrated their respective predictive abilities, what is the practical utility of these models? First, our sequence-based model is designed to predict if a given miRNA sequence is associated with AD-related pathways. Therefore, for novel miRNA sequences, we can predict if they are likely involved in AD. This is directly useful as a first approximation for novel miRNAs and to establish if they should be investigated further for their relationship to AD. Our modeling results pointed to purine-rich miRNA seed sequences as potential drivers of this relationship to AD, so these results can be further leveraged as a basis to understand the mechanism of miRNA involvement in AD.

Second, there has not yet been an established consensus regarding which miRNAs can act as the minimal set for AD diagnosis. Our sequenced-based model can be used as part of the criteria to help pre-select miRNAs for AD-related studies. Utilizing the Ludwig et al. dataset [21], we pre-selected miRNAs and observed overall satisfactory model performance. Because this dataset had already undergone miRNA pre-selection and lacks an appropriate control, these results are only tentative and should be replicated with additional datasets to determine if the approach is valid.

Third, ultimately our models are designed for AD diagnosis. Whether or not we perform pre-selection of miRNA sequences using our sequence-based model, we have

demonstrated that machine learning can predict AD given patient miRNA expression profiles. This further establishes the relevance of machine learning and modeling for disease diagnosis and indicates that miRNAs can act as suitable biomarkers for AD.

Overall, our approach has provided an exploration of the use of machine learning and Orange to understand and predict the involvement of miRNAs in AD utilizing sequence and expression data. Our initial results are encouraging and based on these results further establish the suitability of miRNAs as biomarkers for AD.

## REFERENCES


1. 2022 Alzheimer's disease facts and figures. (2022). *Alzheimer's & Dementia: the Journal of the Alzheimer's Association*, *18*(4), 700–789. https://doi.org/10.1002/alz.12638

2. Haass, C., & Selkoe, D. J. (2007). Soluble protein oligomers in neurodegeneration: Lessons from the Alzheimer's amyloid β-peptide. *Nature Reviews Molecular Cell Biology*, *8*(2), 101–112. https://doi.org/10.1038/nrm2101

3. Young-Pearse, T. L., Lee, H., Hsieh, Y.-C., Chou, V., & Selkoe, D. J. (2023). Moving beyond amyloid and tau to capture the biological heterogeneity of Alzheimer's disease. *Trends in Neurosciences*, *46*(6), 426–444. https://doi.org/10.1016/j.tins.2023.03.005

4. Shetty, A. K., & Upadhya, R. (2021). Extracellular vesicles in health and disease. *Aging and Disease*, *12*(6), 1358–1362. https://doi.org/10.14336/AD.2021.0827

5. Femminella, G. D., Ferrara, N., & Rengo, G. (2015). The emerging role of microRNAs in Alzheimer's disease. *Frontiers in Physiology*, *6*, 40. https://doi.org/10.3389/fphys.2015.00040

6. Zeng, Y. (2006). Principles of micro-RNA production and maturation. *Oncogene*, *25*(46), 6156–6162. https://doi.org/10.1038/sj.onc.1209908

7. Kehl, T., Backes, C., Kern, F., Fehlmann, T., Ludwig, N., Meese, E., Lenhof, H.-P., & Keller, A. (2017). About miRNAs, miRNA seeds, target genes and target pathways. *Oncotarget*, *8*(63), 107167–107175. https://doi.org/10.18632/oncotarget.22363

8. Zhang, H., Artiles, K. L., & Fire, A. Z. (2015). Functional relevance of "seed" and "non-seed" sequences in microRNA-mediated promotion of C. elegans developmental progression. *RNA*, *21*(11), 1980–1992. https://doi.org/10.1261/rna.053793.115

9. Liu, S., Fan, M., Zheng, Q., Hao, S., Yang, L., Xia, Q., Qi, C., & Ge, J. (2022). MicroRNAs in Alzheimer's disease: Potential diagnostic markers and therapeutic targets. *Biomedicine & Pharmacotherapy*, *148*, 112681. https://doi.org/10.1016/j.biopha.2022.112681

10. Han, C., Yang, J., Sun, J., & Qin, G. (2022). Extracellular vesicles in cardiovascular disease: Biological functions and therapeutic implications. *Pharmacology & Therapeutics*, *233*, 108025. https://doi.org/10.1016/j.pharmthera.2021.108025

11. Zhang, L., & Yu, D. (2019). Exosomes in cancer development, metastasis, and immunity. *Biochimica et Biophysica Acta. Reviews on Cancer*, *1871*(2), 455–468. https://doi.org/10.1016/j.bbcan.2019.04.004

12. Kalluri, R., & LeBleu, V. S. (2020). The biology, function, and biomedical applications of exosomes. *Science (New York, N.Y.)*, *367*(6478), eaau6977. https://doi.org/10.1126/science.aau6977

13. Zhang, Y., Liu, Y., Liu, H., & Tang, W. H. (2019). Exosomes: Biogenesis, biologic function and clinical potential. *Cell & Bioscience*, *9*, 19. https://doi.org/10.1186/s13578-019-0282-2

14. Battistelli, M., & Falcieri, E. (2020). Apoptotic bodies: Particular extracellular vesicles involved in intercellular communication. *Biology*, *9*(1), 21. https://doi.org/10.3390/biology9010021

15. Doyle, L. M., & Wang, M. Z. (2019). Overview of extracellular vesicles, their origin, composition, purpose, and methods for exosome isolation and analysis. *Cells*, *8*(7), 727. https://doi.org/10.3390/cells8070727

16. Mosquera-Heredia, M. I., Morales, L. C., Vidal, O. M., Barceló, E., Silvera-Redondo, C., Vélez, J. I., & Garavito-Galofre, P. (2021). Exosomes: Potential disease biomarkers and new therapeutic targets. *Biomedicines*, *9*(8), 1061. https://doi.org/10.3390/biomedicines9081061

17. Song, Y., Kim, Y., Ha, S., Sheller-Miller, S., Yoo, J., Choi, C., & Park, C. H. (2021). The emerging role of exosomes as novel therapeutics: Biology, technologies, clinical applications, and the next. *American Journal of Reproductive Immunology (New York, N.Y.: 1989)*, *85*(2), e13329. https://doi.org/10.1111/aji.13329

18. Silva, A. M., Lázaro-Ibáñez, E., Gunnarsson, A., Dhande, A., Daaboul, G., Peacock, B., Osteikoetxea, X., Salmond, N., Friis, K. P., Shatnyeva, O., & Dekker, N. (2021). Quantification of protein cargo loading into engineered extracellular vesicles at single-vesicle and single-molecule resolution. *Journal of Extracellular Vesicles*, *10*(10), e12130. https://doi.org/10.1002/jev2.12130

19. Li, X., Corbett, A. L., Taatizadeh, E., Tasnim, N., Little, J. P., Garnis, C., Daugaard, M., Guns, E., Hoorfar, M., & Li, I. T. S. (2019). Challenges and opportunities in exosome research-Perspectives from biology, engineering, and cancer therapy. *APL Bioengineering*, *3*(1), 011503. https://doi.org/10.1063/1.5087122

20. Sproviero, D., Gagliardi, S., Zucca, S., Arigoni, M., Giannini, M., Garofalo, M., Fantini, V., Pansarasa, O., Avenali, M., Ramusino, M. C., Diamanti, L., Minafra, B., Perini, G., Zangaglia, R., Costa, A., Ceroni, M., Calogero, R. A., & Cereda, C. (2022). Extracellular vesicles derived from plasma of patients with neurodegenerative disease have common transcriptomic profiling. *Frontiers in Aging Neuroscience*, *14*, 785741. https://doi.org/10.3389/fnagi.2022.785741

21. Ludwig, N., Fehlmann, T., Gogol, M., Maetzler, W., Deutscher, S., Gurlit, S., Schulte, C., Von Thaler, A.-K., Deuschle, C., Metzger, F., Berg, D., Suenkel, U., Keller, V., Backes, C., Lenhof, H.-P., Meese, E., & Keller, A. (2019). Machine learning to detect Alzheimer's disease from circulating non-coding RNAs. https://doi.org/10.1016/j.gpb.2019.09.004

22. Batabyal, R. A., Bansal, A., Cechinel, L. R., Authelet, K., Goldberg, M., Nadler, E., Keene, C. D., Jayadev, S., Domoto-Reilly, K., Li, G., Peskind, E., Hashimoto-Torii, K., Buchwald, D., & Freishtat, R. J. (2023). Adipocyte-derived small extracellular vesicles from patients with Alzheimer disease carry miRNAs predicted to target the CREB signaling pathway in neurons. *International Journal of Molecular Sciences*, *24*(18), 14024. https://doi.org/10.3390/ijms241814024

23. Kumar, P., Dezso, Z., MacKenzie, C., Oestreicher, J., Agoulnik, S., Byrne, M., Bernier, F., Yanagimachi, M., Aoshima, K., & Oda, Y. (2013). Circulating miRNA biomarkers for Alzheimer's disease. *PLoS ONE*, *8*(7), e69807. https://doi.org/10.1371/journal.pone.0069807

24. Swarbrick, S., Wragg, N., Ghosh, S., & Stolzing, A. (2019). Systematic review of miRNA as biomarkers in Alzheimer's disease. *Molecular Neurobiology*, *56*(9), 6156–6167. https://doi.org/10.1007/s12035-019-1500-y

25. DeTure, M. A., & Dickson, D. W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration*, *14*(1), 32. https://doi.org/10.1186/s13024-019-0333-5

26. Holdmann, J., Markert, L., Klinger, C., Kaufmann, M., Schork, K., Turewicz, M., Eisenacher, M., Degener, S., Dreger, N. M., Roth, S., & Savelsbergh, A. (2022). MicroRNAs from urinary exosomes as alternative biomarkers in the differentiation of benign and malignant prostate diseases. *Journal of Circulating Biomarkers*, *11*, 5–13. https://doi.org/10.33393/jcb.2022.2317

27. Morokoff, A., Jones, J., Nguyen, H., Ma, C., Lasocki, A., Gaillard, F., Bennett, I., Luwor, R., Stylli, S., Paradiso, L., Koldej, R., Paldor, I., Molania, R., Speed, T. P., Webb, A., Infusini, G., Li, J., Malpas, C., Kalincik, T., …, Kaye, A. H. (2020). Serum microRNA is a biomarker for post-operative monitoring in glioma. *Journal of Neuro-Oncology*, *149*(3), 391–400. https://doi.org/10.1007/s11060-020-03566-w

28. Lugli, G., Cohen, A. M., Bennett, D. A., Shah, R. C., Fields, C. J., Hernandez, A. G., & Smalheiser, N. R. (2015). Plasma exosomal miRNAs in persons with and without Alzheimer disease: Altered expression and prospects for biomarkers. *PLoS ONE*, *10*(10), e0139233. https://doi.org/10.1371/journal.pone.0139233

29. Griffiths-Jones, S. (2006). miRBase: The microRNA sequence database. In Y. Shao-Yao, *MicroRNA Protocols* (Vol. 342, pp. 129–138). Humana Press. https://doi.org/10.1385/1-59745-123-1:129

30. Backes, C., Kehl, T., Stöckel, D., Fehlmann, T., Schneider, L., Meese, E., Lenhof, H.-P., & Keller, A. (2017). miRPathDB: A new dictionary on microRNAs and target pathways. *Nucleic Acids Research*, *45*(D1), D90–D96. https://doi.org/10.1093/nar/gkw926

31. Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

32. Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19, 1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022

33. Yousef, M., Khalifa, W., Acar, İ. E., & Allmer, J. (2017). MicroRNA categorization using sequence motifs and k-mers. *BMC Bioinformatics*, 18(1). https://doi.org/10.1186/s12859-017-1584-1

34. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research*, *14*(71), 2349–2353.

35. Xu, A., Kouznetsova, V. L., & Tsigelny, I. F. (2022). Alzheimer's disease diagnostics using miRNA biomarkers and machine learning. *Journal of Alzheimer's Disease*, *86*(2), 841–859. https://doi.org/10.3233/JAD-215502

36. Riffo-Campos, Á. L., Riquelme, I., & Brebi-Mieville, P. (2016). Tools for sequence-based miRNA target prediction: What to choose? *International Journal of Molecular Sciences*, *17*(12), 1987. https://doi.org/10.3390/ijms17121987

37. Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26. https://doi.org/10.1186/1748-7188-6-26

38. Das, J. K., Choudhury, P. P., Chaudhuri, A., Hassan, S. S., & Basu, P. (2018). Analysis of purines and pyrimidines distribution over miRNAs of human, gorilla, chimpanzee, mouse and rat. *Scientific Reports*, *8*(1), 9974. https://doi.org/10.1038/s41598-018-28289-x

39. Momose, F., Seo, N., Akahori, Y., Sawada, S., Harada, N., Ogura, T., Akiyoshi, K., & Shiku, H. (2016). Guanine-rich sequences are a dominant feature of exosomal microRNAs across the mammalian species and cell types. *PLoS ONE*, *11*(4), e0154134. https://doi.org/10.1371/journal.pone.0154134

40. Goldie, B. J., Fitzsimmons, C., Weidenhofer, J., Atkins, J. R., Wang, D. O., & Cairns, M. J. (2017). MiRNA enriched in human neuroblast nuclei bind the MAZ transcription factor and their precursors contain the MAZ consensus motif. *Frontiers in Molecular Neuroscience*, *10*, 259. https://doi.org/10.3389/fnmol.2017.00259

41. Rocca, W. A., Amaducci, L. A., & Schoenberg, B. S. (1986). Epidemiology of clinically diagnosed Alzheimer's disease. *Annals of Neurology*, *19*(5), 415–424. https://doi.org/10.1002/ana.410190502

42. Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., Zhu, Y., Powell, D. R., Akutsu, T., Webb, G. I., Chou, K.-C., Smith, A. I., Daly, R. J., Li, J., & Song, J. (2020). iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings in Bioinformatics*, *21*(3), 1047–1057. https://doi.org/10.1093/bib/bbz041

## BIOGRAPHY

Sydney Monserrate received her Bachelor of Science from Virginia Commonwealth University; she then went on to George Mason University to complete her Master of Science.