Research Article

Exploration of Potential Biomarkers for Diagnosing Dementia through Machine Learning Techniques

Muchen Xu

Beijing Academy International Department, No. 10, Kangyuan Road,
Dongba, Chaoyang District, Beijing, 100018, China

ABSTRACT

Dementia, marked by cognitive decline and neuropsychiatric symptoms, significantly impacts individuals and society, especially with an aging global population. Despite the need for early diagnosis and intervention, current diagnostic methods are costly and invasive. This study investigates blood-based microRNAs (miRNAs) as non-invasive, cost-effective biomarkers for early dementia diagnosis and subtype differentiation. Using machine learning techniques, we analyzed serum miRNA expression profiles from the Gene Expression Omnibus database (GSE120584), which includes samples from dementia and cognitively normal controls. Our approach involved Support Vector Machine (SVM), Random Forest (RF), Recursive Feature Elimination (RFE), and Neural Networks for feature selection. Logistic Regression was used for classification. Pathway analysis was further performed on the target genes of the identified miRNA biomarkers to explore biological insights behind these biomarkers. We identified miRNAs such as miR-6777-3p, miR-1471, and miR-6806-5p as potential biomarkers for dementia diagnosis and miRNAs like miR-4290 and miR-3184-3p for subtype differentiation. Among the miRNA biomarkers, miR-371b-3p, miR-1539, and miR-4290 are newly discovered biomarkers that have not been mentioned in any studies before. Additionally, this study demonstrates the power of integrating deep learning with traditional machine learning techniques to find new outcomes. This study also reveals the connection between dementia and infectious diseases on a molecular level, providing new therapeutic insights in dementia.

Keywords: Dementia, Biomarkers, Machine learning, Deep learning, Predictive modeling, Pathway analysis, Infection

INTRODUCTION

Dementia is a syndrome of cognition decline in multiple etiologies, often associated with a range of neuropsychiatric features including emotional lability, impulse control disorders, and depressive symptoms. Those with dementia typically lose their independence and require the care of others to survive. Approximately 6% of individuals aged 65 and above are estimated to be affected by dementia (1). Due to the rapid aging of the global population, dementia has become a global concern, placing a huge burden on health and social care. In 2024, it is estimated that around 60 million people will be diagnosed with dementia worldwide, and by 2050, the number is expected to reach 135 million. The global cost of dementia care was estimated at $604 billion in 2010 and

Corresponding author: Muchen Xu, E-mail: xu.muchen@163.com.

Received July 29, 2024**; Accepted** August 28, 2024
https://doi.org/10.70251/HYJR2348.231324

is expected to rise to $1 trillion by 2030. Despite the large number of people suffering from dementia and the fact that the disease is widely researched by medical and biological scientists, a cure for dementia is still in absence (2).

There are more and more studies hinting that the progression of symptoms can be slowed and patients' cognitive function can be improved (3). This has led to international attention focusing on the early diagnosis and intervention of dementia. Imaging techniques and identifications of biomarkers are currently the mainstream methods for diagnosing dementia (2). However, commonly used imaging techniques such as positron emission tomography (PET) and magnetic resonance imaging (MRI) are usually costly and scarce (4). Traditional methods using β-amyloid and tau proteins in cerebrospinal fluid as biomarkers require invasive procedures like lumbar punctures, which have high rates of procedural failures and infections (5).

In light of these impediments, there arises an urgent need for less invasive and more universally applicable diagnostic approaches (6). Different research has investigated surrogate biomarkers for dementia, where blood-based markers have shown great potential (7). Additionally, various studies have shown that miRNA expression levels are altered in dementia patients compared to healthy populations (8). Other studies also demonstrate the possible value of using miRNA to differentiate between dementia subtypes (9). These findings suggest that blood-based miRNAs have the potential to be used as biomarkers for early diagnosis and subtype differentiation of dementia.

Advances in computational technology in the late 20th and early 21st centuries have dramatically improved the practice of machine learning, enabling it to significantly outperform traditional statistical approaches (10, 11). Now, the dominance of ever-growing big data processing tasks and high-performance computational capabilities continues to drive the requirements of precision models in modern biological research. Biological datasets frequently contain high-dimensional data, noise, and missing values. For these reasons, machine-learning techniques are commonly used in biological studies, particularly within genetics and proteomics (12, 13).

Recognizing the value blood-based miRNAs and the superiority of machine learning in building complex models, several studies have used machine learning methods to identify potential biomarkers for dementia (14, 15). At the same time, with the rapid development of artificial intelligence technology in recent years, the use of deep learning to analyze datasets has become a new machine learning method (16). The use of deep learning to build predictive models may achieve different results from traditional machine learning, and may achieve higher accuracy, which is of great research value (17).

This paper builds on previous research to conduct more in-depth machine learning analysis, which includes the traditional machine learning methods and the latest deep learning methods, to analyze biological features that may have potential associations with dementia, especially miRNA expression profiles in blood. By doing so, this research aims to explore their potential as biomarkers for early diagnosis and subtype differentiation of dementia. Additionally, this study endeavors to explore the potential value of using deep learning technologies in biological research and the additional results deep learning can provide compared to traditional machine learning methods. This study further conducts pathway analysis on the target genes of the miRNA biomarkers identified to find connections between dementia and other diseases at a molecular level, revealing new biological insights behind the pathways identified. This approach is designed not only to improve diagnostics capabilities but also to increase our knowledge of the genetic basis of various forms of dementia.

The core hypothesis of this study is that certain miRNAs are differentially expressed in different dementias and could be utilized as biomarkers for the early detection and categorization of dementia subtypes. Additionally, these miRNA biomarkers can be used in predictive areas of clinical practice.

METHODS

Data properties

The miRNA expression profiles were obtained from the Gene Expression Omnibus (GEO) database with the accession code GSE120584. The dataset includes 1,601 serum samples: 1,021 of Alzheimer's disease (AD), 91 of vascular dementia (VaD), 169 of dementia with Lewy bodies (DLB), 32 of mild cognitive impairment (MCI), and 288 of normal cognitive function (NC). A total of 2,547 miRNAs are identified in the dataset.

The background information on the genetic and demographic profile of the participants in the study is also included in the dataset, such as age, sex, and presence of genotype apoe ε4 allele (APOE4), which has been shown to be strongly associated with dementia (18-22).

Data pre-processing

After initial processing, our dataset ended up containing five biological feature categories: diagnosis type (dependent variable), age, sex, APOE4, and miRNA expression profiles (independent variable). In order to explore the biomarkers for dementia subtype differentiation, the original dataset was divided into different subsets according to the diagnostic type of dementia subtype. Since the dataset contains only 32 MCI samples, we excluded this category from our analysis, The remaining subsets included: AD cases and NC cases, VaD cases and NC cases, DLB cases and NC cases, and cases with dementia and NC cases.

We further used the One-Hot Encoding method to process the categorical data. For diagnosis type in each subset, we record dementia as 1 and non-dementia as 0. Studies have shown that women have a higher risk of developing dementia (23), so we recorded female as 1 and male as 0. Then we normalized the continuous data using the Z-score method (24).

Feature selection

Support vector machine. Feature selection has the ability to improve model performance, reduce computational costs, and yield better output results (25). The embedded method was chosen for feature selection in this study because it possesses lower computational costs than the wrapper method while also offering higher accuracy than the filter method, achieving a balance.

As a supervised learning method for classification and regression tasks, Support Vector Machine (SVM) has been shown to perform well with high-dimensional data due to its robustness and ability to handle nonlinear classification problems through kernel techniques, including biological datasets (26, 27).

We performed SVM feature selection for each subset and retained the top 100 most important features for further analysis.

RF and RFE. Random Forest (RF) is an integrated learning-based algorithm that improves the accuracy of the constructed model by constructing multiple decision trees and combining the predictions of each decision tree. However, due to the complexity of its algorithms, the general training and prediction process for RF is relatively long (28).

Recursive Feature Elimination (RFE) is a feature selection method that fits a model and then recursively trains the model, evaluates the importance of the features, and progressively eliminates the least important features until a subset with the optimal features is selected (11). RFE has been shown to improve classification performance by removing the irrelevant features (29).

Therefore, combining RFE with RF can help to remove redundant features, reducing the risk of overfitting and the model's complexity, which ultimately improves the model performance and prediction accuracy. Moreover, by including RFE, we can make the RF model more computationally efficient. In biological research, there are often precedents for combining RFE and random forests (27).

We implemented RF and RFE in each subset to filter out the 20 most significant biological features from the 100 biological features identified through SVM.

Deep learning. Deep learning (DL) simulates the human brain's processing of data by constructing and training neural networks. It is particularly suitable for processing large and complex data sets (15).

We utilized the torch Python library for the DL Method. TabTransformer was used as the neural network. Through repeated testing, we set the number of hidden units to 512, the number of layers to 3, the number of attention heads to 8, and the dropout rate to 0.3. The model reached relatively high accuracies with these hyperparameters. We then used Cross-entropy loss as the loss function and Adam optimizer for training to further initialize the model. The early stopping was incorporated inside the model to prevent overfitting. We set the numbers of epochs for the training loop to 50 epochs and trained the model.

We implemented DL in each subset to filter out the 20 most significant biological features from the 100 biological features identified through SVM.

Logistic regression

Logistic Regression (LR) is a supervised learning algorithm primarily used for binary classification tasks. LR model assumes a logistic function relationship between the outcome variable (Y) and the predictor (X), where (P(Y=1|X)) represents the conditional probability of the event (Y) taking the value 1 given (X) (54). The model transforms linear combinations of predictors into probabilities using a logistic function, with model parameters β estimated using Maximum Likelihood Estimation (MLE).

$$P(Y=1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}$$

In decision-making, a classification threshold is applied and if P(Y=1|X) is greater than the threshold, Y is predicted as 1, otherwise, it is 0 (30). LR is used to calculate the odds ratios (value = P(Y=1|X) / P(Y=0|X)) for each of the last 20 biological features retained to assess their influence on the probability of dementia. An odds ratio greater than 1 reveals a risk factor, while an odds ratio less than 1 indicates a protective factor. The p-value (31) of each feature was also calculated.

Model validation

Datasets used in each model were divided into training and test sets in a ratio of 80%:20%. The models were trained on the training set and were validated on the test set. The accuracy score of SVM and RF and RFE were calculated by constructing a random forest classifier using the 100 and 20 features selected respectively. The classifier then used the selected features from the test set to make predictions. By comparing the actual labels with the predicted labels, we were able to calculate out the accuracy score of each model. For the deep learning algorithms, the model's performance was evaluated using the trained PyTorch model on the test set.

Pathway analysis

In order to understand the biological meanings behind the miRNA biomarkers identified in this study, we further implemented pathway analysis on the target genes of these miRNAs. The target genes of each miRNA were found using miRTarBase. Among the 710 target genes identified, 2 genes are shared by 4 miRNAs, 10 genes are shared by 3 miRNAs, and 196 are shared by 2 miRNAs. Pathway analysis using DAVID was then implemented on the 208 target genes that are shared among different miRNA biomarkers (32).

Pathway analysis was further performed on the target genes of the newly discovered miRNA biomarkers that were not mentioned in previous studies to validate these findings.

Code availability

All machine learning and graph generating procedures were done using python. The libraries used inside this study included: pandas, numpy, torch, sklearn, matplotlib, and scipy. The codes are available on https://gist.github.com/isospin1/fbfa50e6e67c40387ce3b60f480ff2c1.

RESULTS

Model accuracies

This dataset was also previously used for the prediction models developed for search of miRNA biomarker of dementia, where it demonstrated high performance in the predictive models developed (14). The results obtained from experiments underscore the value of this dataset in progressing the identification of dementia biomarkers.

The baseline accuracies and accuracy scores of each model in each subset are shown in Figure 1.

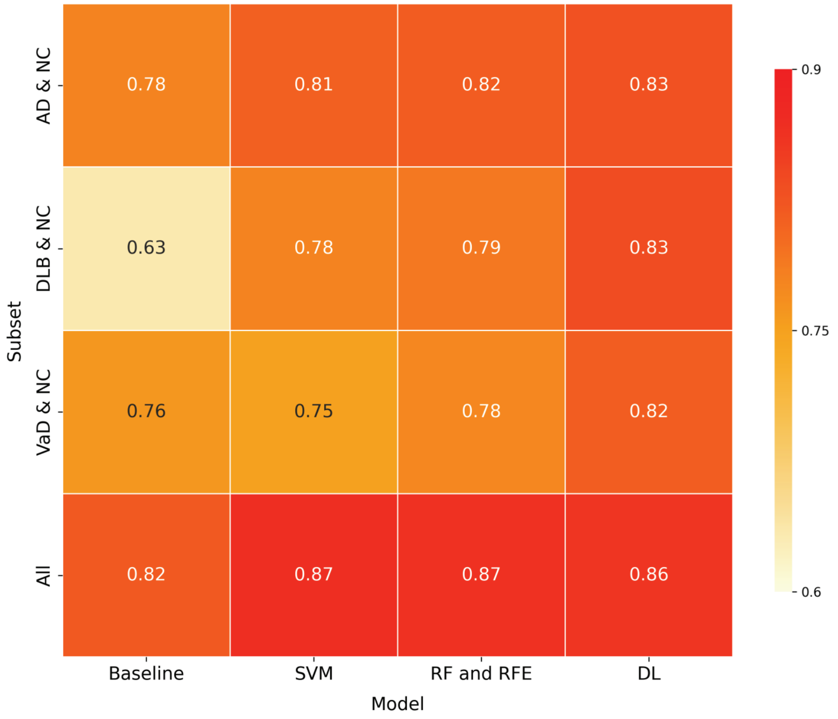

Figure 1. The accuracy score of each model. The accuracy score of different models in each subset is calculated through the procedures in the model validation method. The results are presented above. The row represents which model the accuracy score refers to, and the column represents which subset the model is performed in. The darker the color, the higher the accuracy score that box represents.

Feature selection results

Features were sorted based on the importance parameter obtained through SVM, and the top 100 biological features that have maximum influence on the dependent variable were identified. From the results generated, we find that miR-1471 and miR-4448 appears in three subsets, and miR-1233-3p, miR-6836-3p, miR-4701-3p, miR-6881-5p, miR-6772-3p, miR-4697-3p, and miR-6764-3p appears in two subsets. Their recurrence in the results of different subsets indicates their potential as biomarkers for dementia.

Table 1 shows the biological features that appeared in at least two subsets in the list of top 20 features identified by RF and RFE and DL. From the table, we can see that miR-6777-3p appears in all 4 subsets, followed by miR-1471 and miR-6806-5p, miR-4419a, and miR-208a-5p appearing in 3 subsets. Next, we have 21 miRNAs such as miR-6088 appearing in 2 subsets. It is worth noting that miR-1471 appears in the results of both algorithms.

For the miRNAs do not share biological features across different subsets, we store those with the highest importance values in each subset as shown in Table 2.

Table 1. The shared features

RF and RFE		Deep Learning	
Feature	**Subset**	**Feature**	**Subset**
miR-6777-3p	AD and NC, DLB and NC, VaD and NC, All and NC	miR-4419a	DLB and NC, VaD and NC, All and NC
miR-6806-5p	AD and NC, VaD and NC, All and NC	miR-208a-5p	AD and NC, DLB and NC, All and NC
miR-1471	AD and NC, VaD and NC, All and NC	miR-6806-5p	VaD and NC, All and NC
miR-6088	DLB and NC, VaD and NC	miR-1202	DLB and NC, VaD and NC
miR-4697-3p	DLB and NC, VaD and NC	miR-4488	DLB and NC, All and NC
miR-6761-3p	AD and NC, VaD and NC	miR-6831-5p	AD and NC, VaD and NC
miR-4314	AD and NC, All and NC	miR-6836-3p	AD and NC, All and NC
miR-6836-3p	AD and NC, All and NC	miR-920	AD and NC, All and NC
miR-6829-3p	AD and NC, All and NC	miR-1471	AD and NC, All and NC
miR-4749-3p	AD and NC, All and NC		
miR-4713-5p	AD and NC, All and NC		
miR-4486	AD and NC, All and NC		

The table shows features that are shared in two or more subsets. The miRNAs selected by RF and RFE and DL are shown in the feature column and corresponding subsets are shown in the subset column.

Table 2.

RF and RFE		Deep Learning	
Feature	**Subset**	**Feature**	**Subset**
miR-3184-3p	VaD and NC	miR-125b-1-3p	DLB and NC
miR-371b-3p	All and NC	miR-150-3p	VaD and NC
miR-4290	DLB and NC	miR-1470	AD and NC
miR-1539	AD and NC	miR-6760-5p	All and NC

The table shows the miRNAs which have the highest importance value (calculated through the feature selection model) in each subset. These miRNAs only appeared in on subset, and are not shared in two or more subsets. The miRNAs selected by RF and RFE and DL are shown in the feature column and corresponding subsets are shown in the subset column.

The results of LR done on the selected features identified in Tables 2 and 3 are recorded in Table 3.

We can find from Table 3 that miR-6777-3p is a protective factor in the subset DLB and NC, while all other features are risk factors. The p-value of miR-1470 is greater than 0.05 in the subset AD and NC, meaning its finding is not significant. Based on the odds ratio of each feature, we find that miR-6806-5p in subset VaD and NC has the highest odds ratio among all the features.

We added the clinical features that are not selected by the RF and RFE and DL in the experiment to the 20 selected biometric features and performed logistic regression. The results are shown in Table 4. We can find from the table that age, APOE4, and sex are all risk factors, as expected.

Table 3. The odds ratio and p-value of the selected features

RF and RFE				Deep Learning			
Feature	**Subset**	**Odds Ratio**	**P-value**	**Feature**	**Subset**	**Odds Ratio**	**P-value**
miR-6777-3p	AD and NC	1.224	3.70E-19	miR-4419a	DLB and NC	1.381	6.30E-06
	DLB and NC	0.924	1.86E-15		VaD and NC	1.650	3.40E-05
	VaD and NC	1.725	8.34E-18		All and NC	1.619	2.97E-05
	All and NC	1.361	2.32E-23				
miR-1471	AD and NC	1.324	3.61E-66	miR-208a-5p	AD and NC	1.433	4.09E-18
	VaD and NC	1.429	9.07E-12		DLB and NC	2.193	6.39E-15
	All and NC	1.357	4.73E-24		All and NC	1.674	1.92E-21
miR-6806-5p	AD and NC	1.054	2.38E-14				
	VaD and NC	3.381	4.22E-11				
	All and NC	1.118	1.48E-15				
miR-1539	AD and NC	1.111	6.21E-05	miR-1470	AD and NC	1.281	1.85E-01
miR-4290	DLB and NC	1.182	8.68E-11	miR-125b-1-3p	DLB and NC	1.327	4.11E-02
miR-3184-3p	VaD and NC	1.428	3.51E-14	miR-150-3p	VaD and NC	1.435	4.02E-05
miR-371b-3p	All and NC	1.295	6.58E-24	miR-6760-5p	All and NC	0.843	1.04E-07

Table 4. The odds ratio and p-value of the clinical features

Feature	Subset	Odds Ratio (RF&RFE)	Odds Ratio (DL)	P-value
age	AD and NC	3.359	3.871	3.61E-66
	DLB and NC	4.526	4.346	3.46E-32
	VaD and NC	3.019	3.589	7.54E-20
	All and NC	3.264	3.570	5.43E-69
APOE4	AD and NC	2.520	2.483	1.44E-15
	DLB and NC	1.473	1.564	2.35E-03
	VaD and NC	1.298	1.256	2.12E-01
	All and NC	2.302	2.294	3.10E-13
sex	AD and NC	1.363	1.223	1.36E-12
	DLB and NC	1.241	1.222	1.12E-02
	VaD and NC	0.807	0.907	2.50E-01
	All and NC	1.324	1.267	6.80E-10

Pathway analysis results (Figure 2-4)

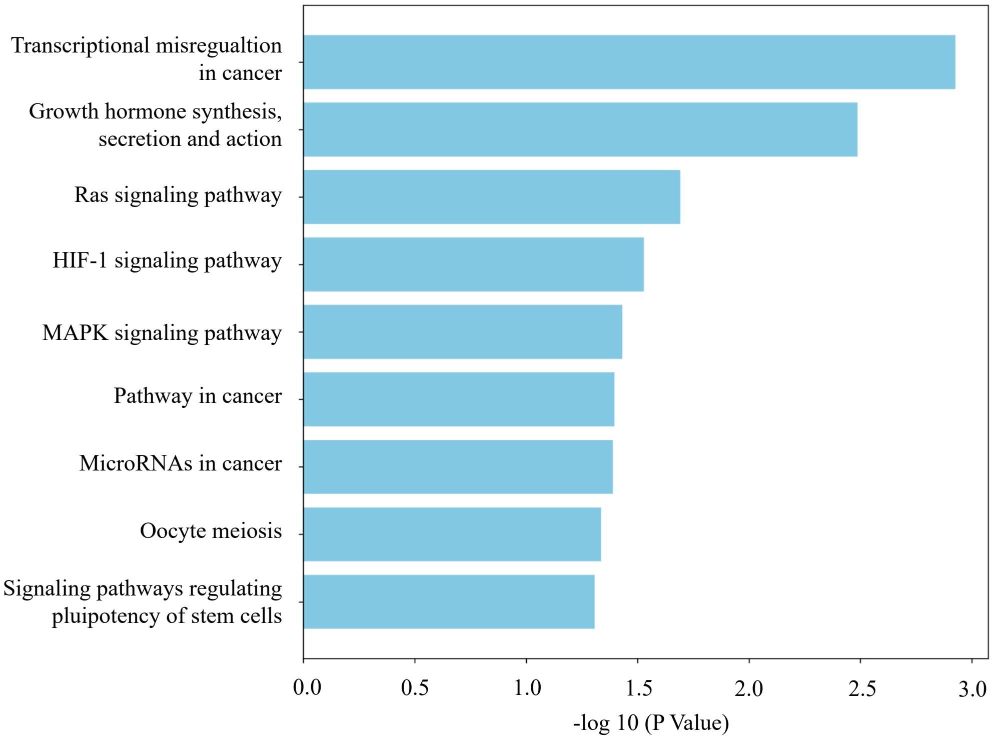

Figure 2. miR-371b-3p pathway analysis. A list of 44 target genes of miR-371b-3p were identified through miRTarBase. The list was then used to conduct pathway analysis using DAVID. The results of the pathways which have a p-value smaller than 0.05 are shown above. The y-axis shows the list of related pathways while the x-axis shows the value of -log10 (p-value).

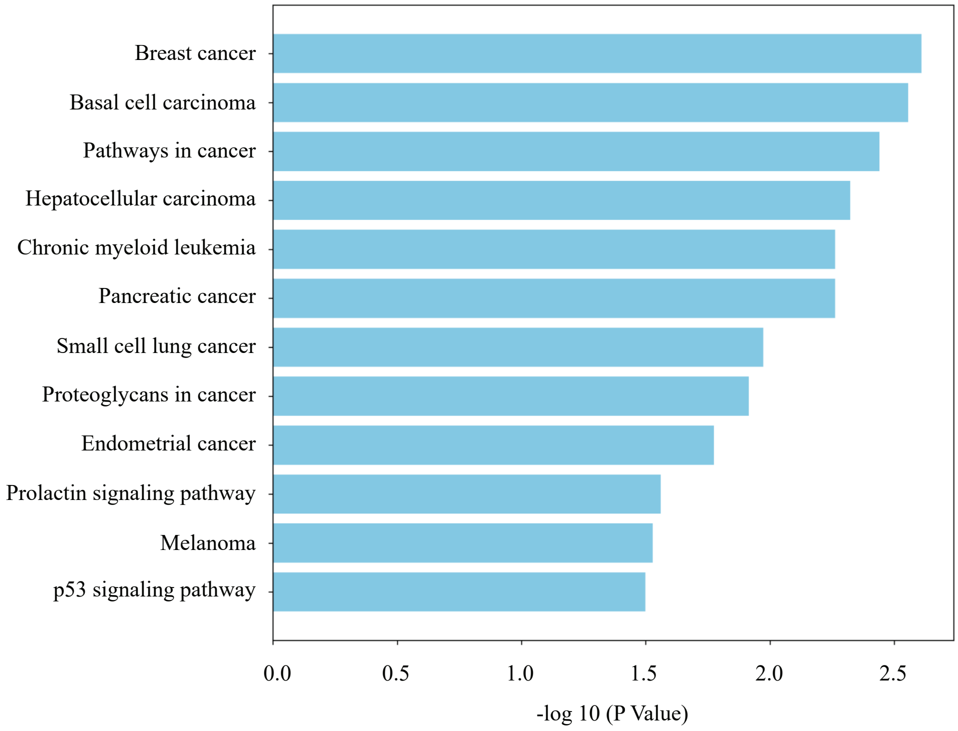

Figure 3. miR-4290 pathway analysis. A list of 89 target genes of miR-4290 were identified through miRTarBase. The list was then used to conduct pathway analysis using DAVID. The results of the most important pathways are shown above. The y-axis shows the list of related pathways while the x-axis shows the value of -log10 (p-value).

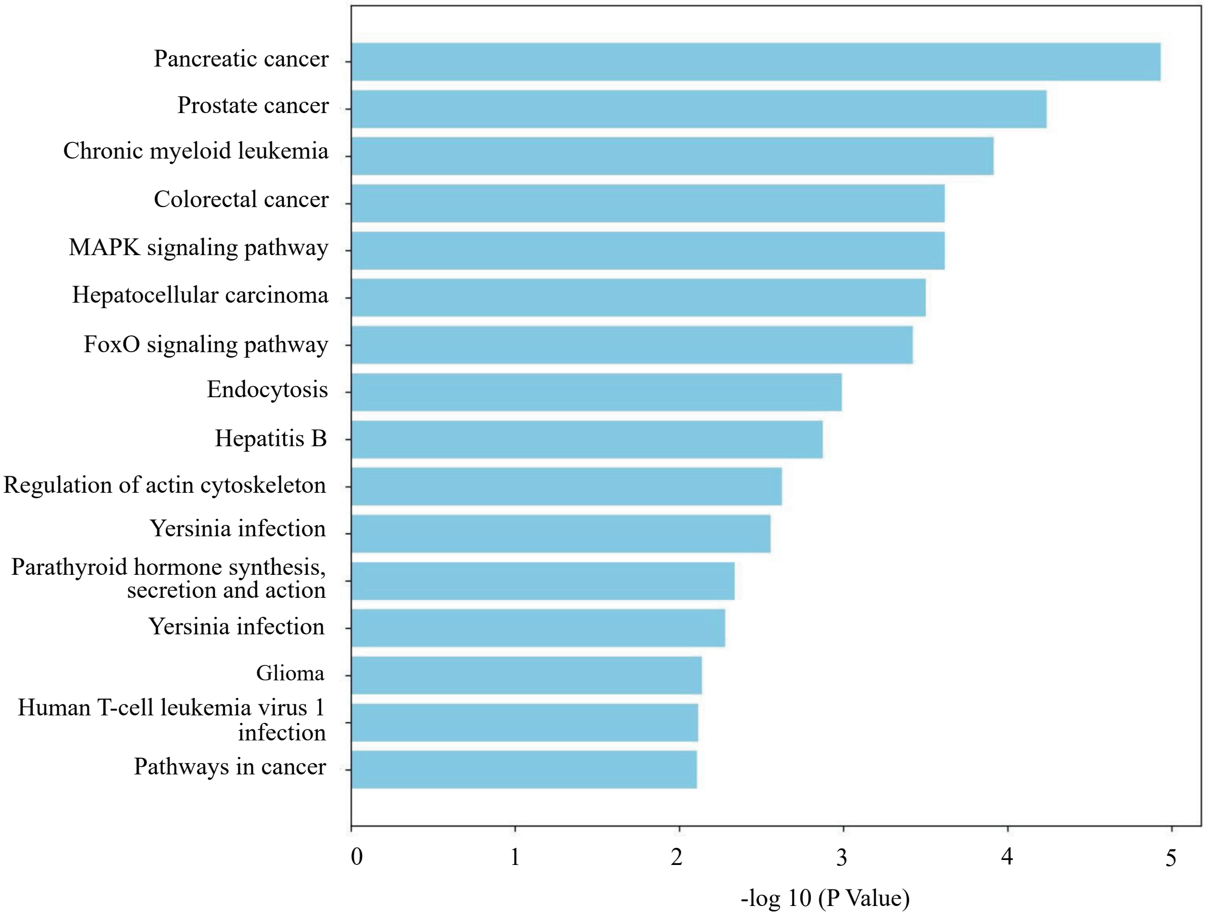

Figure 4. Pathway analysis of the biomarkers identified. A list of 208 shared target genes were identified through miRTarBase. The list was then used to conduct pathway analysis using DAVID. The relevant pathways which have a p-value smaller than 0.01 are shown above. The y-axis shows the list of related pathways while the x-axis shows the value of -log10(p-value).

DISCUSSION

Deep learning

From Figure 1, we can find that Deep Learning Model achieved the best accuracy scores overall, with slightly worse results for the combined (All) and Normal Cognitive (NC) subsets. These results demonstrate that deep learning is better suited to handle complex biological data than traditional machine learning techniques. As one of the best performing methods, deep learning has the potential to become an important tool in upcoming research projects, especially for constructing models where accuracy is the top priority. Deep Learning might augur a sea change in the standard approaches for machine learning in biological data analysis, especially in the area of dementia research.

Diagnosis of dementia

By conducting a systematic literature review, we found that miR-6777-3p has obvious lower expression levels in vascular dementia compared to non-dementia controls, and it might be a key biomarker of VaD (33).

Research certified that the expression level of miR-1471 was remarkably reduced in the plasma of intracranial aneurysms (IA) patients compared to control group, which suggests that miR-1471 could take part in the pathogenesis of IA (34). Studies have shown the connection of IA and increase in risk of dementia, due to the reduction of cerebral blood flow and neuronal damage caused by IA (35). As miR-1471 is associated with IA, it may also be connected with the development of dementia.

MiR-6806-5p has also been found to be differentially altered in patients with unruptured or ruptured intracranial aneurysms, with reduced expression level suggesting its role in the pathogenesis of IA (36). Therefore, it can serve as a potential biomarker for dementia.

A study in the field of AD has discovered miR-4419a as a regulator of STAT3 signaling pathway, associated with synaptic dysfunction and neuronal injury in AD, emphasizing its involvement in the transition from MCI to AD (37).

At the same time, miR-208a-5p might induce synaptic dysfunction and neuroinflammation due to its potential influence on the PI3K-AKT signaling pathway (37). Other studies have also given similar results, supporting miR-208a-5p as an early diagnostic biomarker for AD (38).

The interaction of miR-6760-5p on the 3'UTR region of the SNCA gene, and its significant alteration in expression levels in patients with Parkinson disease (39), demonstrates the role of miR-6760-5p in neurodegenerative diseases and connection with dementia.

There is no direct literature evidence that miR-371b-3p is associated with neurodegenerative diseases from previous studies. We implemented pathway analysis on its target genes to validate its potential as a newly discovered miRNA biomarker for dementia. The results are shown in Figure 2.

Subtype differentiation of dementia

MiR-3184-3p was found to be a potential target of TFDM, which may be associated with multiple signaling pathways in the pathology of VaD, making it a potential biomarker for the diagnosis of VaD.

Studies have found more remarkable changes in miR-125b-1-3p exosomal serum levels between AD patients and healthy controls, supporting the finding that miR-125b-1-3p participates in the regulation of AD and could serve as a potential biomarker for AD diagnosis (40).

MiR-150-3p has been linked to neuroprotective pathways in neurodegenerative diseases. By packaging it into exosomes secreted by neural stem cells (NSCs), this miRNA inhibits Caspase-2 signaling pathway and reduces neuron apoptosis (41). It has also been found that miR-150-5p has expression level upregulation in AD patients (42).

Because miR-1539 is specifically present in AD and NC subsets, and not in other subsets, it is likely to have a general contribution to the development of AD. It is also the most important unique biomarker in the All and NC subset. No previous studies have mentioned the connection between miR-1539 and dementia. The high level of miR-1539 expression in exosomes and cancer tissues of CRC patients suggests its potential function as an oncogene involved in tumor stem cell-like characteristic and cancer progression, and offers potential explanations for its wide biological relevance (43, 44). To validate it as a potential AD biomarker, we further performed pathway analysis on its target genes. The findings are presented in the pathway analysis section below.

MiR-4290 is only detected in the DLB and NC subset, providing a potential biomarker for DLB. Although miR-4290 is found to be associated with cancer developments, such as larynx cancer (45), there is no direct literature evidence that this miRNA is associated with neurodegenerative diseases. However, it has the potential to be an miRNA biomarker for dementia which was not identified before. In order to verify this assumption, we performed pathway analysis on the target genes of miR-4290. The results are shown in Figure 3.

Pathway Analysis

From Figure 2, we can see that miR-371b-3p is closely connected to growth hormone synthesis, secretion and action and Ras signaling pathway, which are identified to be related to the development of dementia in other studies (46, 47). Therefore, miR-371b-3p has the potential to be a novel miRNA biomarker for dementia.

MiR-1539 only has 15 target genes and its gene enrichment results showed limited insights of miR-1539 because of its relatively small gene numbers. However, we found that its genes are associated with negative regulation of transcription from RNA polymerase II promoter, having a p-value < 0.05. This procedure was found to be linked with the development of neurodegenerative disorders such as AD (48). Therefore, miR-1539 may contribute to the development of dementia through negative regulation of transcription from RNA polymerase II promoter.

We can find from Figure 3 that miR-4290 is related with p53 signaling pathway. P53 was found to play an important role in the development of neurodegenerative diseases (49). Additionally, it was found to be associated with AD (50). Therefore, miR-4290's connection with p53 signaling pathway can lead to the development of dementia.

From Figure 4 we found that the identified miRNA biomarkers for dementia were also closely related to cancer, endocytosis and efferocytosis, hormones and infections. The relationship between cancer and dementia has been extensively researched (51). Endocytosis and efferocytosis and hormones were also found to be connected with dementia development (52). Although associations between infectious disease and dementia have been noticed (53), there are currently no studies doing research on this topic on a molecular level. The results from pathway analysis provide evidence for the connection between dementia and infections on a molecular level, such as Hepatitis B and Yersinia infection. Infections may increase the risk of developing dementia, and people with dementia may have a higher risk of being infected. This finding can provide support for the higher death rates of patients diagnosed with dementia due to their increased likelihood of infection. Additionally, as age is the main factor in the development of dementia, it may also contribute to the risk of infection. Since females are more likely to be diagnosed with dementia, they may also have higher risks of being infected. In conjunction with the analysis above, this study considers the pathways listed to have connections with the development of dementia and can serve as potential diagnostic and therapeutic targets for dementia.

CONCLUSION

Through a comprehensive machine learning process and a holistic literature review, we are able to identify 7 miRNAs, miR-6777-3p, miR-1471, miR-6806-5p, miR-4419a, miR-208a-5p, miR-371b-3p, and miR-6760-5p, as the most potential biomarkers for dementia diagnosis. Additionally, 5 miRNAs, miR-1539, miR-4290, miR-3184-3p, miR-125b-1-3p, and miR-150-3p, are identified as the best potential biomarkers for dementia subtype differentiation. Among them, miR-371b-3p, miR-1539, and miR-4290 are newly discovered biomarkers that have not been mentioned in any studies before. One notable finding is that deep learning can be used to find new biomarkers and understand dementia's connections with other diseases, demonstratinges the power of integrating deep learning with traditional machine learning techniques. This comprehensive approach can be applied to other complex diseases, paving the way for better results and new findings. The results of pathway analysis on the identified miRNA biomarkers also show the potential association of infectious disease and dementia from a molecular level, which has never been conducted in previous studies. Future studies should focus on the broader biological insights and the implementation of miR-371b-3p, miR-1539 and miR-4290 in real life. They should also focus on the practice of deep learning algorithms in future biological research on the area of dementia and the relationship between infectious disease and dementia.

ACKNOWLEDGEMENTS

This paper was supported by Ahmet Ay, professor of biology and mathematics in Colgate University.

FUNDING

The study was not funded by any schools or institutions.

CONFLICT OF INTEREST

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCE

1. Scheltens P, Blennow K, Breteler MMB, de Strooper B, Frisoni GB, Salloway S & Van der Flier WM. Alzheimer's disease. The Lancet. 2016; 388 (10043): 505–517.
2. Robinson L, Tang E & Taylor JP. Dementia: timely diagnosis and early intervention. BMJ. 2015; 350: h3029.
3. Ngandu T, Lehtisalo J, Solomon A, Levälahti E, Ahtiluoto S, Antikainen R, ... Kivipelto M. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. The Lancet. 2015; 385 (9984): 2255-2263.
4. Jack CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, ... & Weiner MW. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2010; 7 (3): 257-262.
5. Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, ... & Trojanowski JQ. Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. Annals of Neurology. 2009; 65 (4): 403-413.
6. Weiner MW, et al. The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. Alzheimer's & Dementia. 2017; 13 (6): 1-18.
7. O'Bryant SE, et al. Blood-based biomarkers in Alzheimer's disease: Current state of the science and a novel collaborative paradigm for advancing from discovery to clinic. Alzheimer's & Dementia. 2019; 15 (3): 431-434.
8. Schneider R, McKeever P, Kim T, Graff C, van Swieten JC, Karydas A, Boxer A, Rosen H, Miller BL, Laforce R, Jr., Galimberti D, Masellis M, Borroni B, Zhang Z, Zinman L, Rohrer JD, Tartaglia MC, Robertson J & Genetic FTD Initiative (GENFI). Downregulation of exosomal miR-204-5p and miR-632 as a biomarker for FTD: A GENFI study. Journal of Neurology, Neurosurgery, and Psychiatry. 2018; 89 (8): 851-858.
9. Gámez-Valero A, Campdelacreu J, Vilas D, Ispierto L, Reñé R, Álvarez R, Armengol MP, Borràs FE & Beyer K. Exploratory study on microRNA profiles from

plasma-derived extracellular vesicles in Alzheimer's disease and dementia with Lewy bodies. Translational Neurodegeneration. 2019; 8: 31.

10. Bzdok D, Altman N & Krzywinski M. Points of significance: Statistics versus machine learning. Nature Methods. 2018; 15: 233-234.
11. Hastie T, Tibshirani R & Friedman J. The elements of statistical learning: Data mining, inference, and prediction. Springer. 2009.
12. Libbrecht MW & Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015; 16 (6): 321-332.
13. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, ... & Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596 (7873): 583-589.
14. Shigemizu D, Akiyama S, Asanomi Y, Boroevich KA, et al. Risk prediction models for dementia constructed by supervised principal component analysis using miRNA expression data. Communications Biology. 2019; 2: 77.
15. Li ZD, Guo W, Ding SJ, et al. Identifying key microRNA signatures for neurodegenerative diseases with machine learning methods[J]. Frontiers in Genetics. 2022; 13: 880997.
16. Goodfellow I, Bengio Y & Courville A. Deep Learning. MIT Press. 2016.
17. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, ... & DePristo MA. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology. 2018; 36 (10): 983-987.
18. Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, & Ferri CP. The global prevalence of dementia: A systematic review and meta-analysis. Alzheimer's & Dementia. 2013; 9 (1): 63-75.e2.
19. Strittmatter WJ, Saunders AM, Georgiou CD, Yin KJ, Schlossmacher MG, Perkin JA, ... & Roses AD. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proceedings of the National Academy of Sciences of the United States of America. 1993; 90 (5): 1977-1981.
20. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, ... & van Duijn CM. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA. 1997; 278 (16): 1349-1356.
21. Vegeto E, Villa A, Della Torre S, Crippa V, Rusmini P, Cristofani R, Galbiati M, Maggi A, & Poletti A. The Role of Sex and Sex Hormones in Neurodegenerative Diseases. Endocrine reviews. 2020; 41 (2): 273–319.
22. Castro-Aldre L, Haller S, Poloni C, Marechal O, & Ibanez A. Sex and gender considerations in Alzheimer's disease: The Women's Brain Project contribution. Frontiers in Aging Neuroscience. 2023; 15: 1105620.
23. Azad NA, Al Bugami M & Loy-English I. Gender differences in dementia risk factors. Gender Medicine. 2007; 4 (2): 120-129.
24. Cheadle C, Vawter MP, Freed WJ & Becker KG. Analysis of microarray data using Z score transformation. The Journal of molecular diagnostics. 2003; 5 (2): 73-81.
25. Guyon I & Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research. 2003; 3 (Mar): 1157-1182.
26. Cristianini N & Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press. 2000.
27. Guyon I, Weston J, Barnhill S & Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002; 46 (1-3): 389-422.
28. Breiman L. Random forests. Machine Learning. 2001; 45 (1): 5-32.
29. Rakotomamonjy A. Variable selection using SVM-based criteria. Journal of Machine Learning Research. 2003; 3: 1357-1370.
30. Hosmer DW, Lemeshow S & Sturdivant RX. Applied Logistic Regression. John Wiley & Sons. 2013.
31. Agresti A & Franklin C. Statistics: The Art and Science of Learning from Data. Pearson. 2013.
32. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC & Lempicki RA. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biology. 2003; 4 (5): R60.
33. Liang H, Xu X, Wang L, Wang W, Li D, Wu H, ... & Li X. Identification of potential diagnostic and therapeutic targets for cerebral ischemia/reperfusion injury using integrated analysis of circRNA-miRNA-mRNA network. Frontiers in Pharmacology. 2021; 12: 796628.
34. Jin H, Li C, Ge H, Jiang Y & Li Y. Circulating microRNA: A novel potential biomarker for early diagnosis of intracranial aneurysm rupture: A case control study. Journal of Translational Medicine. 2013; 11 (1): 296.
35. Rincon F & Wright CB. Vascular cognitive impairment. Current opinion in neurology. 2013; 26 (1): 29–36.
36. Liao B, Zhou MX, Zhou FK, Luo XM, Zhong SX, Zhou YF, Qin YS, Li PP & Qin C. Exosome-Derived MiRNAs as Biomarkers of the Development and Progression of Intracranial Aneurysms. Journal of atherosclerosis and thrombosis. 2020; 27 (6): 545–610.
37. Han YH, Xiang HY, Lee DH, Feng L, Sun HN, Jin MH & Kwon T. Identification and diagnostic potential of serum microRNAs as biomarkers for early detection of Alzheimer's disease. Aging. 2023; 15 (21): 12085–12103.
38. Su L, Zhang Y, Wang Y & Wei H. Identification of a lncRNA/circRNA-miRNA-mRNA ceRNA network in Alzheimer's disease. Journal of Integrative Neuroscience.

2023; 22 (6): 136.

39. Toffoli M, Dreussi E, Cecchin E, Valente M, Sanvilli N, Montico M, Gagno S, Garziera M, Polano M, Savarese M, Calandra-Buonaura G, Placidi F, Terzaghi M, Toffoli G & Gigli GL. SNCA 3'UTR genetic variants in patients with Parkinson's disease and REM sleep behavior disorder. Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology. 2017; 38 (7): 1233–1240.
40. Duan X, Zheng Q, Liang L & Zhou L. Serum Exosomal miRNA-125b and miRNA-451a are Potential Diagnostic Biomarker for Alzheimer’s Diseases. Degenerative Neurological and Neuromuscular Disease. 2024; 14: 21–31.
41. Yang R, Li Z, Xu J, Luo J, Qu Z, Chen X, Yu S & Shu H. Role of hypoxic exosomes and the mechanisms of exosome release in the CNS under hypoxic conditions. Frontiers in Neurology. 2022; 13.
42. Chia SY, Vipin A, Ng KP, Tu H, Bommakanti A, Wang BZ, Tan YJ, Zailan FZ, Ng ASL, Ling SC, Okamura K, Tan EK, Kandiah N & Zeng L. Upregulated Blood miR-150-5p in Alzheimer's Disease Dementia Is Associated with Cognition, Cerebrospinal Fluid Amyloid-β, and Cerebral Atrophy. Journal of Alzheimer's disease: JAD. 2022; 88 (4): 1567–1584.
43. Cui X, Lv Z, Ding H, Xing C & Yuan Y. MiR-1539 and Its Potential Role as a Novel Biomarker for Colorectal Cancer. Frontiers in oncology. 2021; 10: 531244.
44. Liu J, Ma L, Wang Z, Wang L, Liu C, Chen R & Zhang J. MicroRNA expression profile of gastric cancer stem cells in the MKN-45 cancer cell line. Acta Biochimica et Biophysica Sinica. 2014; 46 (2): 92–99.
45. Karatas OF, Suer I, Yuceturk B & Creighton CJ. . Identification of microRNA profile specific to cancer stem-like cells directly isolated from human larynx cancer specimens. BMC Cancer. 2016; 16 (1): 853.
46. Rollero A, Murialdo G, Fonzi S, Garrone S, Gianelli MV, Gazzerro E, ... & Polleri A. Relationship between cognitive function, growth hormone and insulin-like growth factor I plasma levels in aged subjects. Neuropsychobiology. 1998; 38 (2): 73-79.
47. Kang K, Bai J, Zhong S, Zhang R, Zhang X, Xu Y, Zhao M, Zhao C & Zhou Z. Down-Regulation of Insulin Like Growth Factor 1 Involved in Alzheimer’s Disease via MAPK, Ras, and FoxO Signaling Pathways. Oxidative Medicine and Cellular Longevity. 2022; 2022: 8169981.
48. Bakulski KM, Dolinoy DC, Sartor MA, Paulson HL & Konen JR. Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex. Journal of Alzheimer’s Disease. 2012; 29 (3): 571-588.
49. Chang JR, Ghafouri M, Mukerjee R, Bagashev A, Chabrashvili T & Sawaya BE. Role of p53 in neurodegenerative diseases. Neuro-degenerative diseases. 2012; 9 (2): 68–80.
50. Wolfrum P, Fietz A, Schnichels S & Hurst J. The function of p53 and its role in Alzheimer’s and Parkinson’s disease compared to age-related macular degeneration. Frontiers in Neuroscience. 2022; 16: 1029473.
51. Sun M, Wang Y, Sundquist J, Sundquist K & Ji J. The association between cancer and dementia: A national cohort study in Sweden. Frontiers in Oncology. 2020; 10: 73.
52. Taheri F, Taghizadeh E, Navashenaq JG & Sahebkar A. The role of efferocytosis in neuro-degenerative diseases. Neurological Sciences. 2022; 43 (5): 1593–1603.
53. Dunn NDM, Mullee M, Perry VH & Holmes C. Association between dementia and infectious disease: Evidence from a case-control study. Alzheimer Disease & Associated Disorders. 2005; 19 (2): 91-94.
54. Cox DR. The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological). 1958; 20 (2): 215-232.