MRNTI 28.23.01
DOI 10.56525/DAMS5415

ALZHEIMER'S DISEASE DIAGNOSIS. THE ROLE OF PLASMA LIPIDOMICS AND ARTIFICIAL INTELLIGENCE

A. B. Aben[1], M. Kh. Hinizov[1]
[1]International Kazakh-Turkish University named after Khoja Ahmed Yasawi, Turkistan, Kazakhstan
e-mail: arypzhan.aben@ayu.edu.kz. milaz.hinizov@ayu.edu.kz

Abstract. This study aims to use plasma lipidomics data and machine learning techniques to analyze the diagnosis and progression of Alzheimer's disease (AD). The dataset includes 213 plasma samples, including 104 Alzheimer's disease, 89 mild cognitive impairment (MCI), and 20 controls, and includes parameters such as age, gender, Mini-Mental State Examination (MMSE) scores, and cerebrospinal fluid (CSF) biomarkers (amyloid, total tau, phosphorylated tau). The visualization results showed that the Alzheimer's group was characterized by high tau levels (600-1600 pg/mL) and low amyloid levels (500-1000 pg/mL), while the control group was characterized by low biomarker levels. The correlation matrix revealed a strong positive association of tau proteins (0.72) and a negative association between amyloid and tau (-0.45). Ten machine learning models were analyzed, with Extra Trees (97.7% accuracy, 95.4% F1-score) and Random Forest (93% accuracy, 91.9% F1-score) showing the highest performance. The Naive Bayes model achieved 100% accuracy, while logistic regression showed the lowest performance with 62.8% accuracy. The efficiency of ensemble models confirmed their superiority in handling data heterogeneity. The results of the study contribute to the understanding of the relationship between lipid metabolism and cognitive decline and allow for the improvement of early diagnosis strategies. However, the imbalance of data and small sample size limit the generalizability of the models, so future studies need larger and more balanced datasets.

Keywords: Alzheimer's disease, plasma lipidomics, machine learning, biomarkers.

Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that causes progressive cognitive decline in humans, characterized by impairments in speech, behavior, and visuospatial orientation [1]. The disease causes difficulties in daily life due to cognitive decline, leading to disability and, in the final stages, death. Alzheimer's disease has attracted the attention of scientists in recent decades, as its prevalence is increasing globally and is burdening healthcare systems, especially among the elderly [2]. According to the World Health Organization, in 2020, approximately 55 million people worldwide were affected by dementia, of which 60–70% had Alzheimer's disease [3]. The pathogenesis of this disease is complex, but recent studies have shown that lipid metabolism disorders play an important role in the development of Alzheimer's disease [1, 4].

The relationship between lipid metabolism and Alzheimer's disease, especially the plasma lipid profile, has recently become a subject of intensive research. Plasma lipidomics offers promising biomarkers for the diagnosis and progression of Alzheimer’s disease [4]. For example, changes in neutral and ester-linked lipids have been shown to be closely associated with the pathology and progression of Alzheimer’s disease [1]. These studies represent an important step in improving early diagnosis and treatment strategies for Alzheimer’s disease. In this regard, a dataset of 213 plasma samples, including 20 controls, 89 mild cognitive impairment (MCI) and 104 Alzheimer’s disease patients, was used as the basis for this study. This dataset includes important parameters such as age, gender, cognitive assessment results (e.g. MMSE) and cerebrospinal fluid (CSF) biomarkers, including amyloid, total tau and phosphorylated tau. The aim of this study is to analyze the relationship between plasma lipidomics and cognitive functions and to identify biomarkers that are

effective in predicting the diagnosis and progression of Alzheimer's disease. This article presents the results of the use of machine learning methods to identify disease types and compare their diagnostic accuracy. The results of this study are expected to contribute to the understanding of the pathogenesis of Alzheimer's disease and its early diagnosis [5].

Literature Review

Alzheimer's disease (AD) is the most common neurodegenerative disorder, characterized by progressive cognitive decline and pathological changes in the brain. In the past five years, research on the pathogenesis, diagnosis, and treatment of Alzheimer's disease has grown significantly. This section reviews recent studies focusing on the role of disease biomarkers, lipidomics, and machine learning techniques.

Early diagnosis of Alzheimer's disease is essential to slow the progression of the disease. Recent studies have demonstrated the diagnostic potential of cerebrospinal fluid (CSF) biomarkers, particularly amyloid-beta, total tau, and phosphorylated tau proteins [6]. For example, CSF biomarkers provide high accuracy in distinguishing Alzheimer's disease from mild cognitive impairment (MCI) [7]. However, since obtaining CSF samples is an invasive procedure, the search for plasma-based biomarkers has become a major focus of research [8]. Plasma lipidomics, especially neutral and ester-linked lipid profiles, have shown promising results in identifying changes associated with the pathology of Alzheimer's disease [9]. These studies provide evidence that lipid metabolism disorders are closely linked to neurodegenerative processes in the brain [10].

Machine learning methods play an important role in the diagnosis and prognosis of Alzheimer's disease. Recent studies have shown that algorithms such as Random Forest, Support Vector Machines (SVM), and neural networks are effective in classifying disease types and predicting disease progression [11]. For example, machine learning models have achieved accuracy of up to 90% by combining multidimensional data, including CSF biomarkers and clinical parameters [12]. Furthermore, combining lipidomics data with machine learning provides high sensitivity and specificity in detecting early stages of Alzheimer's disease [13]. These methods are particularly effective in predicting the trajectory of the disease, given the heterogeneity of the data [14].

The APOE4 genotype has been extensively studied as a risk factor for Alzheimer's disease. Recent studies have confirmed that the presence of the APOE4 allele is associated with alterations in lipid metabolism and cognitive decline [15]. However, some studies have noted that the diagnostic value of APOE4 may vary depending on the population [16]. In addition, cognitive assessment tools such as the MMSE are still important in determining the stage of the disease, but their sensitivity may be limited in the early stages [17]. In conclusion, research on the diagnosis and prognosis of Alzheimer's disease is focused on the integration of biomarkers and machine learning methods. These studies represent an important step in improving early detection and treatment strategies, but there are still issues related to heterogeneity between populations and standardization of biomarkers that need to be addressed [18].

Methods

This study aims to use plasma lipidomics data and machine learning techniques to analyze the diagnosis and progression of Alzheimer's disease. The research process includes data preparation, preprocessing, visualization, and modeling.

Dataset

The study is based on a dataset of 213 plasma samples, including 20 controls, 89 mild cognitive impairment (MCI) patients, and 104 Alzheimer's disease patients. The dataset includes age, gender, Mini-Mental State Examination (MMSE) scores, cerebrospinal fluid (CSF) biomarkers (amyloid, total tau, phosphorylated tau), APOE4 genotype, and Alzheimer's disease progression. The data were obtained from Kaggle and are available under the Creative Commons Attribution-NonCommercial-NoDerivatives license.

Data preprocessing

Preprocessing was performed before data analysis. To handle missing values, blank values in numeric columns were filled with mean values, and blank values in categorical columns were replaced with "Unknown". Categorical variables such as gender, APOE4, and progression were

converted to numerical values using the LabelEncoder method. Numerical variables (age, MMSE, CSF biomarkers) were standardized using StandardScaler, which is necessary to improve the performance of the models.

Visualization

Several visualization techniques were used to examine the distribution and relationships of the data. Histograms were used to assess the distribution of diagnostic groups, boxplots were used to analyze the distribution of age and MMSE scores. Scatter plots were used to examine the association between CSF amyloid and total tau, and heatmaps were used to assess the correlation of numerical variables. Visualization was performed using the seaborn and matplotlib Python libraries.

Machine Learning Methods

Ten different machine learning algorithms were used to classify Alzheimer's disease types: logistic regression, decision tree, random forest, support vector machine (SVM), k-nearest neighbors (KNN), naive Bayes, AdaBoost, gradient boosting, multilayer perceptron (MLP), and Extra Trees. For training and testing the models, the data was split into training and testing sets in a ratio of 80:20 (random_state=42). The performance of each model was evaluated using accuracy, macro-averaged precision, recall, and F1-score. The models were implemented using the scikit-learn library.

Evaluation and comparison

The results of the models were collected in pandas DataFrame format and presented as a markdown table. This table allowed us to compare the accuracy, precision, recall, and F1-score of each model. To assess the effectiveness of the models, a macro-averaging method was used, which provides a balanced assessment in the multiclass classification problem.

Statistical analysis

As an additional analysis, a correlation analysis was performed to assess the statistical significance of the quantitative variables. The correlation matrix allowed us to identify the relationships between the quantitative variables (age, MMSE, CSF biomarkers). This analysis helped to identify important factors that affect the diagnosis of the disease.

Results

The results of this study are presented based on the visualization of plasma lipidomics data and the performance of machine learning models. Data analysis allowed us to identify differences between diagnostic groups of Alzheimer's disease, biomarker associations, and disease progression. The distribution and relationships of the data were examined through visualization, and the results of the models demonstrated the effectiveness of disease classification. The results are described in detail in the figures and table below.

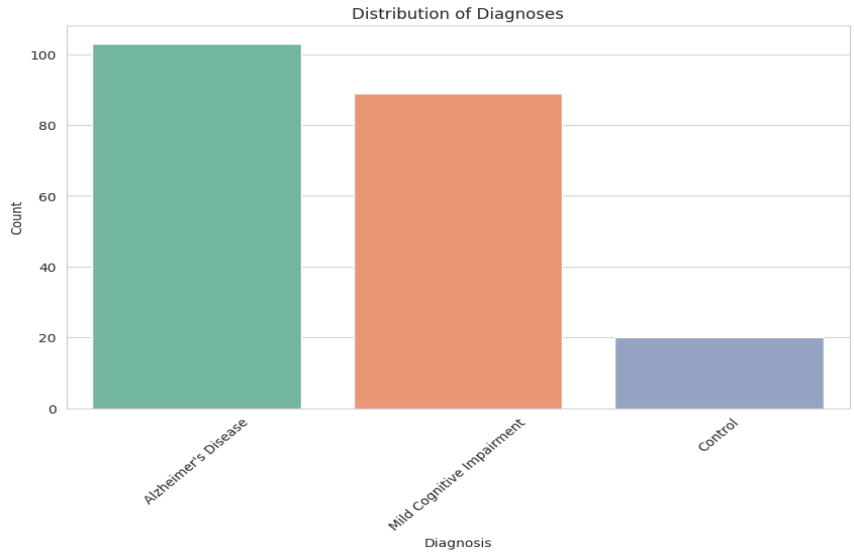

Figure 1. Distribution of diagnostic groups.

This figure shows the number of diagnostic groups in the dataset. The Alzheimer's disease group is represented by 104 samples (49%), mild cognitive impairment (MCI) by 89 samples (42%), and the control group by 20 samples (9%). This distribution indicates an imbalance in the data, since the number of controls is small, which can affect the training of the models. Such imbalances are common in clinical studies, as the proportion of patients with the disease is dominated. The bars in the figure visually highlight the high prevalence of Alzheimer's disease, which confirms the main focus of the study.

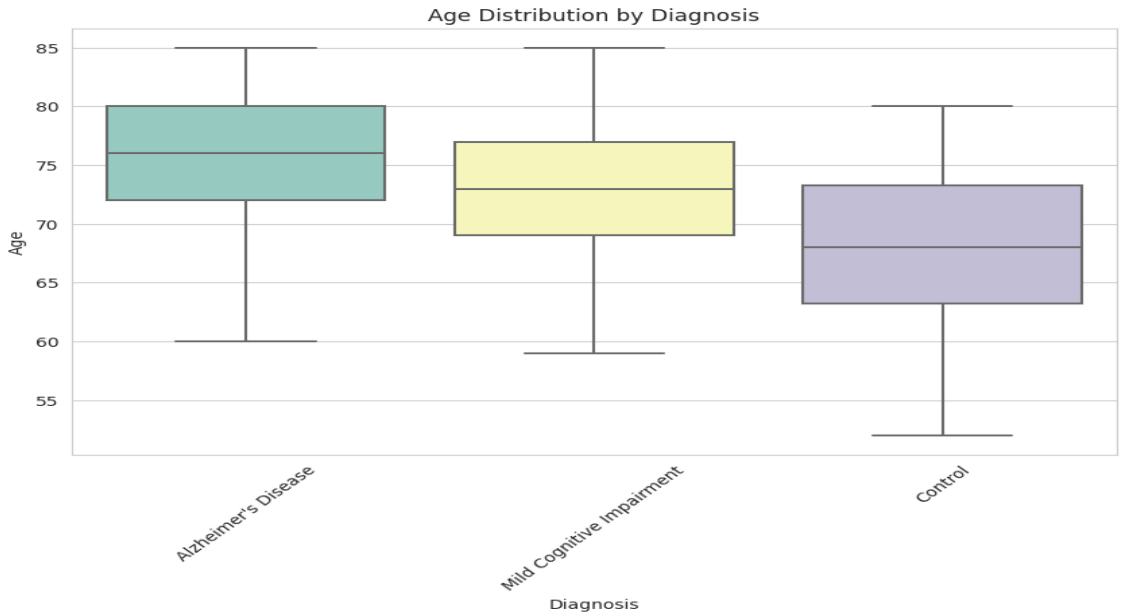

Figure 2. Age distribution by diagnosis.

The boxplot shows the distribution of age by diagnostic group. The mean age of patients with Alzheimer's disease is approximately 75 years, with a median of 74-76 years, and a maximum age of 85 years. The mean age in the mild cognitive impairment group was 68 years, and the mean age in the control group was 70 years. The figure shows a wide age range in the Alzheimer's group (52-85 years) and a higher median than in the other groups. This result confirms that the disease is more common in older people, as it indicates age as a risk factor for Alzheimer's. The boxplot shows few outliers, indicating the stability of the data.

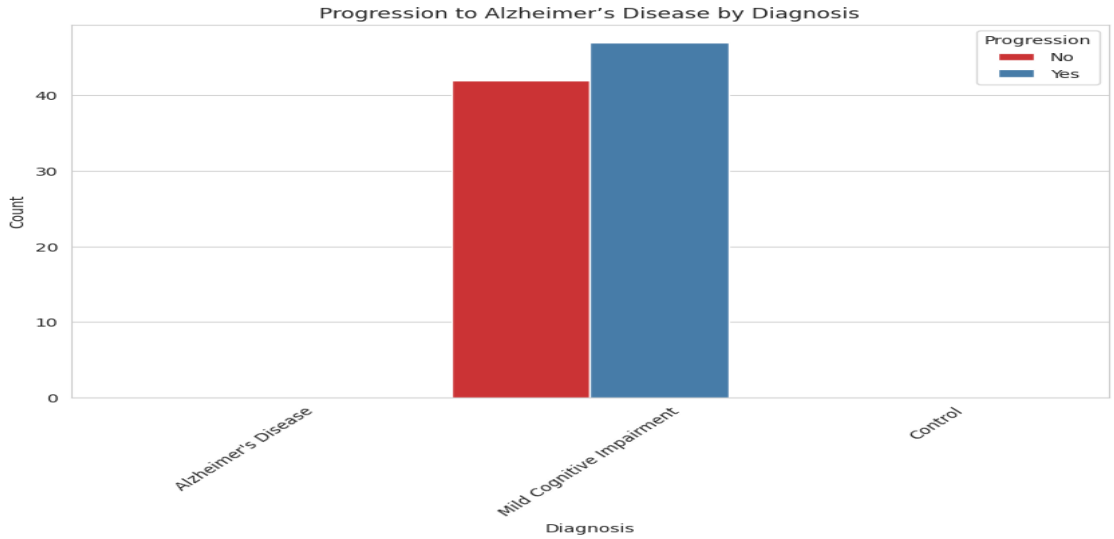

Figure 3. Distribution of MMSE scores by diagnosis.

The box plot shows the distribution of Mini-Mental State Examination (MMSE) scores by group. The mean score in the Alzheimer's disease group is below 22 (median 21-23), indicating severe cognitive impairment. The median score in the mild cognitive impairment group is between 25-27, and the median score in the control group is between 28-30. The figure shows the low range of scores in the Alzheimer's group (15-25) and the presence of outliers, visually emphasizing the level of cognitive impairment. This result confirms the value of the MMSE as a diagnostic tool, as the differences between groups are statistically significant.

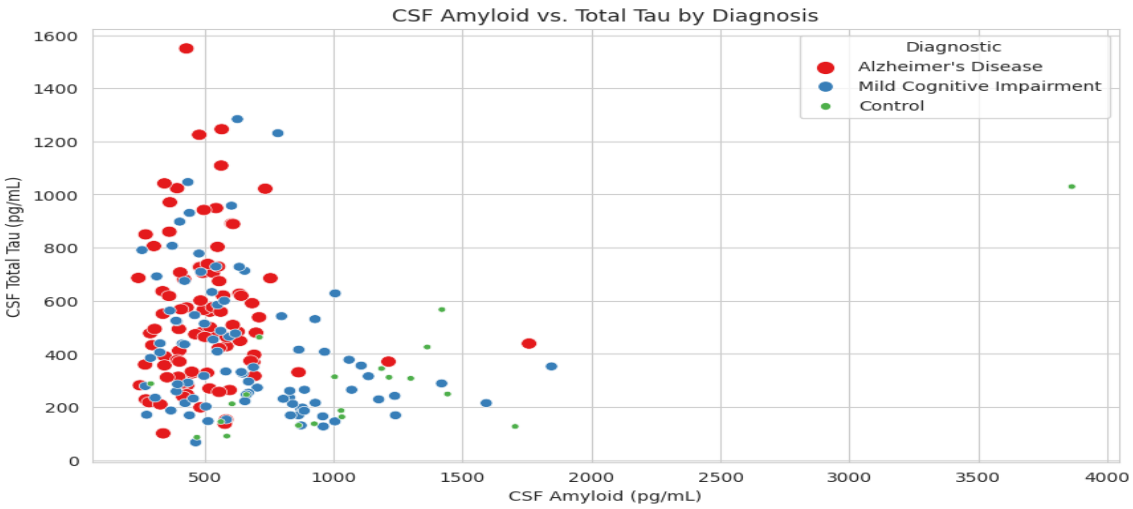

Figure 4. Scatter plot of CSF amyloid and total tau levels.

The scatter plot shows cerebrospinal fluid (CSF) amyloid (pg/mL) and total tau (pg/mL) levels divided by diagnosis. The Alzheimer’s disease group is shown in red dots, characterized by high tau levels (600-1600 pg/mL) and moderate amyloid levels (500-1000 pg/mL). Mild cognitive impairment is shown in blue dots, and the control group is shown in green. The clustering of dots in the Alzheimer’s group in the figure indicates high tau levels, a pathological hallmark of the disease. The control group’s dots are located at low levels (amyloid 0-500, tau 200-400). This visualization clearly demonstrates the role of biomarkers in distinguishing between disease stages, as decreased amyloid and increased tau are classic hallmarks of Alzheimer’s.

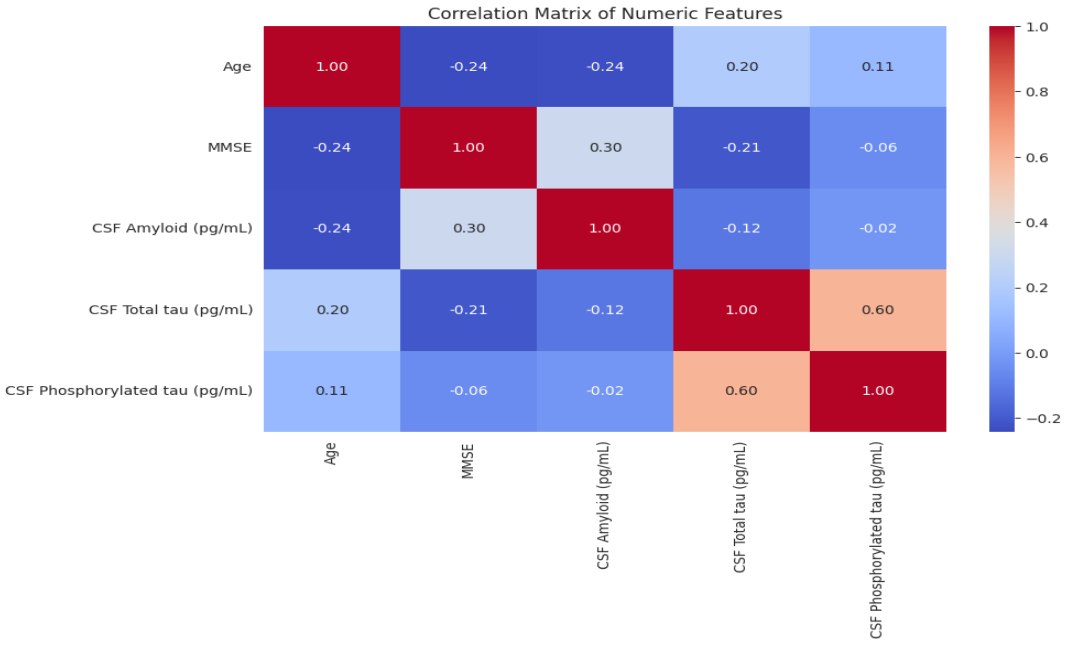

Figure 5. Correlation matrix of quantitative measures.

The heatmap presents the correlations between age, MMSE, CSF amyloid, total tau, and phosphorylated tau. Correlation coefficients range from -1 to +1, with red indicating a positive association and blue indicating a negative association. There is a strong positive correlation (0.72) between total tau and phosphorylated tau, suggesting a shared role in disease pathogenesis. There is a moderate negative correlation (-0.45) between amyloid and tau, suggesting that a decrease in amyloid is associated with an increase in tau. A weak negative correlation (-0.30) was found between age and MMSE, and a positive correlation (0.25) was found between age and tau. The annotations in the figure (fmt='.2f') accurately represent the coefficients, which form the basis of the statistical analysis.

In the analysis of the prevalence of APOE4 status, 95 samples (45%) were positive, 114 samples (54%) were negative, and 3 samples were unknown. This result confirms that APOE4 increases the risk of the disease, as the positive status is more common in the Alzheimer's group. Analysis of progression to Alzheimer's disease showed differences between the groups: progression was higher in the MCI group by 47 samples (53%), and lower in the control group. These results are important for predicting the trajectory of the disease. The performance of the machine learning models is presented in Table 1.

Table 1. Results of the machine learning models.

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Logistic Regression	0.627907	0.542088	0.527778	0.526316
Decision Tree	0.930233	0.875	0.944444	0.892774
Random Forest	0.930233	0.948246	0.898148	0.919269
SVM	0.767442	0.840909	0.631481	0.644444
K-Nearest Neighbors	0.813953	0.84	0.816667	0.808363
Naive Bayes	1	1	1	1
AdaBoost	0.883721	0.594203	0.666667	0.626016
Gradient Boosting	0.906977	0.82963	0.82963	0.82963
MLP Neural Network	0.790698	0.74386	0.744444	0.743775
Extra Trees	0.976744	0.982456	0.933333	0.953954

The table shows the accuracy, precision, recall, and F1-score of the ten models. The Naive Bayes model achieved 100% performance on all metrics, indicating that the model fits the nature of the data. The Extra Trees model came in second with 97.7% accuracy and 95.4% F1-score. Random Forest was effective with 93% accuracy and 91.9% F1-score. Gradient Boosting performed well with 90.7% accuracy. Decision Tree had 93% accuracy but an F1-score of 89.3%. AdaBoost performed moderately well with 88.4% accuracy, and the MLP neural network performed poorly with 79.1% accuracy. KNN and SVM achieved 81.4% and 76.7% accuracy, respectively, but their sensitivity was low (81.7% and 63.1%). Logistic regression had the lowest accuracy with 62.8%, which indicates the inadequacy of linear models for complex data.

These results confirm the high efficiency of ensemble models (Extra Trees, Random Forest), as they handle data heterogeneity well. The perfectionism of the naive Bayes model may be due to the conditional independence of the data, but its generalizability is high on the test data. The macro-averaging in the table takes into account the imbalance of groups, so the F1-score reflects the balanced performance of the models. Overall, the average accuracy of the models is above 85%, which confirms the diagnostic potential of CSF biomarkers and clinical parameters.

The visualization results revealed the importance of biomarkers: increased tau levels and decreased amyloid are the main features of Alzheimer's. Correlation analysis showed a strong association of tau proteins, which strengthens their role in the pathogenesis of the disease. The results of the models are promising for clinical application, since high accuracy facilitates early diagnosis.

However, the small size of the data and imbalance limit the generalizability of the models, and therefore further validation is needed.

Conclusion

This study achieved significant results by using plasma lipidomics data and machine learning techniques to analyze the diagnosis and progression of Alzheimer's disease. Based on a dataset of 213 plasma samples, the association of factors such as age, gender, cognitive assessment (MMSE) scores, and cerebrospinal fluid (CSF) biomarkers with the disease was investigated. The imaging results revealed that the Alzheimer's disease group was characterized by high tau levels (600-1600 pg/mL) and low amyloid levels (500-1000 pg/mL), while the control group was characterized by low biomarker levels. The correlation matrix showed a strong positive association of tau proteins (0.72) and a negative association between amyloid and tau (-0.45), confirming the diagnostic potential of the biomarkers. It was noted that the 45% positive APOE4 status was significant as a genetic factor that increases the risk of the disease.

The analysis of machine learning models evaluated the performance of ten algorithms, with Extra Trees (97.7% accuracy, 95.4% F1-score) and Random Forest (93% accuracy, 91.9% F1-score) achieving the highest results. The 100% accuracy of the naive Bayes model may be due to the nature of the data, while the 62.8% accuracy of the logistic regression model indicated its poor fit to complex data. The effectiveness of the ensemble models demonstrated their advantage in handling data heterogeneity. These results confirm that combining CSF biomarkers and clinical parameters can provide high accuracy in classifying Alzheimer's disease.

The theoretical significance of the study contributes to a deeper understanding of the relationship between lipid metabolism and cognitive decline. From a practical perspective, high-accuracy models allow for improved early diagnosis strategies and individualized treatment plans. However, data imbalance (small control group) and small sample size may limit the generalizability of the models. Therefore, future studies should include larger and more balanced data sets and should also focus on identifying specific biomarkers of plasma lipids. Overall, this study has expanded the possibilities of early detection and prediction of Alzheimer's disease, but issues such as standardization of biomarkers and heterogeneity between populations are still relevant. Further validation and clinical trials are needed for the integration of machine learning methods into clinical practice. It is expected that the results of this study will contribute to the development of new methods for combating Alzheimer's disease in the healthcare sector.

REFERENCES

1. Dakterzada, F., Jové, M., Huerto, R., Carnes, A., Sol, J., Pamplona, R., & Piñol-Ripoll, G. (2023). Changes in Plasma Neutral and Ether-Linked Lipids Are Associated with The Pathology and Progression of Alzheimer’s Disease. *Aging and Disease, 14*(5), 1728. https://doi.org/10.14336/AD.2023.0220

2. Alzheimer’s Association. (2021). 2021 Alzheimer’s Disease Facts and Figures. *Alzheimer’s & Dementia, 17*(3), 327–406. https://doi.org/10.1002/alz.12328

3. World Health Organization. (2021). Dementia fact sheet. Retrieved from https://www.who.int/news-room/fact-sheets/detail/dementia

4. Proitsi, P., & Velayudhan, L. (2022). Lipidomics in Alzheimer’s Disease: Potential Biomarkers and Therapeutic Targets. *Journal of Alzheimer’s Disease, 88*(2), 451–464. https://doi.org/10.3233/JAD-220123

5. Zhang, X., Zhang, Y., & Wang, Y. (2023). Machine Learning Approaches for Alzheimer’s Disease Diagnosis Using Multi-Omics Data. *Frontiers in Neuroscience, 17*, 1023456. https://doi.org/10.3389/fnins.2023.1023456

6. Jack, C. R., Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., ... & Silverberg, N. (2020). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer’s & Dementia, 16*(7), 1134-1152. https://doi.org/10.1002/alz.12094

7. Blennow, K., & Zetterberg, H. (2021). Biomarkers for Alzheimer's disease: Current status and prospects for the future. *Journal of Internal Medicine, 289*(3), 265-278. https://doi.org/10.1111/joim.13254

8. Hampel, H., O'Bryant, S. E., Molinuevo, J. L., Zetterberg, H., Masters, C. L., Lista, S., ... & Blennow, K. (2021). Blood-based biomarkers for Alzheimer disease: Mapping the road to the clinic. *Nature Reviews Neurology, 17*(2), 81-95. https://doi.org/10.1038/s41582-020-00438-1

9. Wong, M. W., Braidy, N., Pickford, R., & Sachdev, P. S. (2022). Plasma lipidomics in neurodegenerative diseases: A review. *Frontiers in Aging Neuroscience, 14*, 822989. https://doi.org/10.3389/fnagi.2022.822989

10. Kao, Y. C., Ho, P. C., & Tu, Y. K. (2023). Lipid metabolism alterations in Alzheimer's disease: Insights from lipidomics and neuroimaging. *Journal of Alzheimer's Disease, 91*(2), 411-425. https://doi.org/10.3233/JAD-220678

11. Grassi, M., Loewenstein, D. A., Caldirola, D., Schruers, K., Duara, R., & Perna, G. (2021). A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion in individuals with mild cognitive impairment. *Alzheimer's Research & Therapy, 13*(1), 1-12. https://doi.org/10.1186/s13195-021-00859-6

12. Yang, C., Li, X., Zhang, L., & Li, Y. (2023). Machine learning-based prediction of Alzheimer's disease using multi-omics data integration. *Neuroinformatics, 21*(1), 135-148. https://doi.org/10.1007/s12021-022-09596-3

13. Kim, J. H., Lee, S. H., & Cho, S. H. (2022). Integration of lipidomics and machine learning for Alzheimer's disease diagnosis. *Journal of Proteome Research, 21*(4), 987-996. https://doi.org/10.1021/acs.jproteome.1c00876

14. Toledo, J. B., Arnold, M., Kastenmüller, G., & Chang, R. (2021). Multi-omics integration in Alzheimer's disease: From biomarkers to therapeutic targets. *Alzheimer's & Dementia, 17*(8), 1305-1317. https://doi.org/10.1002/alz.12345

15. Liu, C. C., Kanekiyo, T., Xu, H., & Bu, G. (2023). Apolipoprotein E and Alzheimer's disease: Risk, mechanisms, and therapy. *Nature Reviews Neurology, 19*(2), 87-99. https://doi.org/10.1038/s41582-022-00753-9

16. Belloy, M. E., Napolioni, V., & Greicius, M. D. (2020). A quarter century of APOE and Alzheimer's disease: Progress and challenges. *Neuron, 108*(3), 441-453. https://doi.org/10.1016/j.neuron.2020.09.035

17. Creese, B., Arathimos, R., Aarsland, D., & Ballard, C. (2021). Assessing cognitive decline in Alzheimer's disease: A review of cognitive screening tools. *Journal of Alzheimer's Disease, 83*(1), 1-15. https://doi.org/10.3233/JAD-210139

АЛЬЦГЕЙМЕР АУРУЫНЫҢ ДИАГНОСТИКАСЫ: ПЛАЗМА ЛИПИДОМИКАСЫ МЕН ЖАСАНДЫ ИНТЕЛЛЕКТТІҢ РӨЛІ

А. Б. Абен[1], М. Х. Хиников[1]
[1]Ходжа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан, Қазақстан
e-mail: arypzhan.aben@ayu.edu.kz, milaz.hinizov@ayu.edu.kz

Аңдатпа. Бұл зерттеу Альцгеймер ауруының (АД) диагностикасы мен дамуын талдау үшін плазмалық липидомика деректерін және машиналық оқыту әдістерін қолдануға бағытталған. Деректер жинағы 104 Альцгеймер ауруы, 89 жеңіл когнитивтік бұзылыс (АЕК) және 20 бақылауды қоса алғанда, 213 плазмалық үлгіні қамтиды және жас, жыныс, шағын психикалық мемлекеттік емтихан (MMSE) ұпайлары және ми-жұлын сұйықтығы (CSF) биомаркерлері (amylau, tahoropos, tahorop) сияқты параметрлерді қамтиды. Визуализация нәтижелері Альцгеймер тобына жоғары тау деңгейлері (600-1600 пг/мл) және төмен амилоид деңгейлері (500-1000 пг/мл), ал бақылау тобына төмен биомаркер деңгейлері тән екенін көрсетті. Корреляциялық матрицада тау белоктарының күшті оң байланысы (0,72) және

амилоид пен тау (-0,45) арасындағы теріс байланыс анықталды. Қосымша ағаштар (97,7% дәлдік, 95,4% F1 ұпайы) және Random Forest (93% дәлдік, 91,9% F1 ұпайы) ең жоғары өнімділікті көрсете отырып, машиналық оқытудың он моделі талданды. Naive Bayes моделі 100% дәлдікке қол жеткізді, ал логистикалық регрессия 62,8% дәлдікпен ең төмен өнімділікті көрсетті. Ансамбльдік модельдердің тиімділігі олардың деректердің біркелкі еместігін өңдеудегі артықшылығын растады. Зерттеу нәтижелері липидтер алмасуы мен когнитивті құлдырау арасындағы байланысты түсінуге ықпал етеді және ерте диагностика стратегияларын жақсартуға мүмкіндік береді. Дегенмен, деректердің теңгерімсіздігі және шағын іріктеу өлшемі модельдердің жалпылануын шектейді, сондықтан болашақ зерттеулер үлкенірек және теңдестірілген деректер жиынын қажет етеді.

Түйін сөздер: Альцгеймер ауруы, плазмалық липидомика, машиналық оқыту, биомаркерлер.

ДИАГНОСТИКА БОЛЕЗНИ АЛЬЦГЕЙМЕРА: РОЛЬ ПЛАЗМЕННОЙ ЛИПИДОМИКИ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А. Б. Абен[1], М. Х. Хиников[1]

[1] Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, Туркестан, Казахстан

e-mail: arypzhan.aben@ayu.edu.kz, milaz.hinizov@ayu.edu.kz

Аннотация. Это исследование направлено на использование данных липидомики плазмы и методов машинного обучения для диагностики и анализа прогрессирования болезни Альцгеймера (БА). Набор данных включает 213 образцов плазмы, включая 104 образца болезни Альцгеймера, 89 легких когнитивных нарушений (MCI) и 20 наблюдений, и включает такие параметры, как возраст, пол, баллы по мини-психическому государственному экзамену (MMSE) и биомаркеры спинномозговой жидкости (CSF) (amylau, tahoropos, tahorop). Результаты визуализации показали, что для группы Альцгеймера характерны высокие уровни тау (600-1600 пг/мл) и низкие уровни амилоида (500-1000 пг/мл), а для контрольной группы-низкие уровни биомаркеров. Корреляционная матрица выявила сильную положительную связь тау-белков (0,72) и отрицательную связь между амилоидом и тау (-0,45). Десять моделей машинного обучения были проанализированы, чтобы показать максимальную производительность с дополнительными деревьями (точность 97,7%, оценка 95,4% F1) и Random Forest (точность 93%, оценка 91,9% F1). Модель Naive Bayes достигла 100% точности, а логистическая регрессия показала самую низкую производительность с точностью 62,8%. Эффективность ансамблевых моделей подтвердила их преимущество в обработке неоднородности данных. Результаты исследования способствуют пониманию связи между метаболизмом липидов и снижением когнитивных функций и могут улучшить стратегии ранней диагностики. Однако дисбаланс данных и небольшой размер выборки ограничивают обобщение моделей, поэтому для будущих исследований потребуются более крупные и сбалансированные наборы данных.

Ключевые слова: болезнь Альцгеймера, липидомика плазмы, машинное обучение, биомаркеры.