Proceedings of the 6th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/2026.32825

Predicting Alzheimer's Disease: A Machine Learning Approach Using Protein Biomarkers and Demographic Features

Huilin Lou

Business School, Renmin University of China, Beijing, China
mavislou1111@outlook.com

Abstract. Alzheimer's disease (AD) is a dementia that primarily affects the elderly, characterized by a decline in cognitive function and daily living abilities. With the increasing global aging population, low-cost and convenient early screening methods for Alzheimer's disease can effectively promote early and regular screening among the elderly, thereby driving early intervention and treatment. This study uses clinical assessments, protein biomarkers, and demographic data from the ADNI dataset, employing random forest, XGBoost, and LightGBM models for analysis. The results show that the random forest model, optimized with GridSearchCV, performed best in both accuracy and recall. Based on this, the importance of features output by the random forest model is analyzed, and the impact of each feature on the prediction results is further explored, using the `predict_proba` function of `RandomForestClassifier`, while misclassification analysis is also performed. The results show that ABETA42, PTAU, and age are key features in early Alzheimer's disease screening, while ABETA42, PTAU, and sex play important roles in misclassification analysis. This study provides an effective machine learning method for early screening of Alzheimer's disease, which reduces screening costs while ensuring accuracy and has broad clinical application potential.

Keywords: Alzheimer's disease, early screening, demographic characteristics, protein biomarkers, random forest

1. Introduction

Alzheimer's disease (AD) is a degenerative central nervous system disease affecting the elderly population, mainly manifested as memory impairment, language impairment, apraxia, agnosia, and visuospatial impairment. With the increasing global aging process, the prevalence of Alzheimer's disease is constantly rising [1]. Statistics show that approximately 10 million new cases of Alzheimer's disease (AD) are diagnosed globally each year, with the highest-risk group concentrated in the elderly population aged 65 and above. With changing demographics, AD has become a significant challenge to global public health and socioeconomic development, particularly in aging countries where it places increasing pressure on healthcare resources and social support systems.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.32825

Therefore, early diagnosis of Alzheimer's disease is of great significance for delaying disease progression, improving patients' quality of life, and the timing of early intervention [2].

This study proposes a machine learning model that combines representative protein biomarkers of Alzheimer's disease with demographic characteristics. The goal is to reduce the time and economic cost of screening while ensuring diagnostic accuracy through multi-feature fusion, laying the foundation for future research on multi-feature prediction models of Alzheimer's disease.

2. Literature review

Biomarkers in cerebrospinal fluid (CSF) and blood have been shown to play an important role in the early diagnosis of Alzheimer's disease [3-5]. For example, total tau protein, phosphorylated tau protein, and ABETA42, three CSF biomarkers, are considered to have the highest diagnostic accuracy in early AD diagnosis and can predict the progression of AD in patients with mild cognitive impairment (MCI) [4]. In addition, plasma biomarkers can also support early diagnosis without invasive testing, such as plasma pTau-217 was found to have a high consistency with the ratio of pTau-181 and ABETA42 in cerebrospinal fluid [6,7]. The large amount of data generated by non-invasive technologies such as blood component monitoring not only provides a platform for developing more accurate and reliable biomarkers, but also significantly reduces patient suffering, psychological impact, risk of complications, and medical costs [8].

With the improvement of computing power, machine learning and deep learning methods have been widely used in the early diagnosis of Alzheimer's disease [9,10]. Deep learning-based frameworks utilize convolutional neural networks for AD staging and classification, achieving high accuracy and good robustness [2,11]. Structural magnetic resonance imaging (MRI) is also widely used to identify brain region features associated with AD. These methods can automatically identify key brain region features and effectively distinguish between MCI and mild AD, demonstrating outstanding generalization ability across datasets and accuracy in early diagnosis [12,13].

Despite progress in the early diagnosis of Alzheimer's disease, machine learning models that combine biomarkers with demographic characteristics are still relatively few. Studies have shown that the heterogeneity of AD may be related to demographic factors (such as age, sex, education level, and socioeconomic status) [14]. This study can reduce diagnostic costs while ensuring accuracy, providing a practical initial screening solution for Alzheimer's disease.

3. Methodology

3.1. Experimental design

First, relevant data were selected from three ADNI datasets. The first dataset, Clinical Dementia Rating, includes basic information about the participants and various clinical indicators used to assess dementia severity. The second dataset, UPENN CSF Biomarkers, contains protein biomarker data of participants. The third dataset, Subject Demographics, covers the demographic characteristics of participants, including age, sex, education level, and other relevant factors [15].

Next, these datasets were merged using an inner join based on five common features (PHASE, PTID, RID, VISCODE/VISCODE2, and VISDATE) to ensure that samples recorded in all datasets were retained. The merged dataset contained 2658 data points. Based on these, the participants' ages were calculated using their birth and test dates, and an "Age" column was added to the dataset.

Then, the columns were filtered, retaining the following variables: CDGLOBAL, reflecting the severity of dementia in patients; PTGENDER, representing the patient's gender; PTEDUCAT,

DOI: 10.54254/2753-8818/2026.32825

indicating the patient's education level; PTETHCAT and PTRACCAT, reflecting the patient's ethnic and racial classification; Age, representing the patient's age; ABETA40, ABETA42, TAU and PTAU, which denote the concentrations of ABETA40, ABETA42, TAU and PTAU proteins.

Since the ABETA40 missing rate was as high as 75%, this variable was directly deleted, retaining other biomarkers. Based on this, multicollinearity among features was assessed by calculating the correlation matrix and variance inflation factor (VIF). The results showed that the correlation between TAU and PTAU was 0.977, and their VIF values were 26.47 and 27.17, respectively, which may lead to model instability. Because PTAU is more representative of Alzheimer's disease than TAU, and because clarifying the clinical significance of each biomarker is crucial for understanding disease mechanisms in medical research, this study decided against PCA and instead opted to directly delete TAU while retaining PTAU. This reduces multicollinearity while preserving information closely related to the disease.

Next, missing values were handled and data standardized. Skewness was calculated, and imputation methods were selected based on its magnitude. For variables with a skewness greater than 0.5 or less than -0.5 (PTAU and ABETA42), the median was used for imputation; for variables with a skewness close to 0 (Age), the mean was used for imputation.

Finally, StandardScaler was used to standardize the numerical features (Age, ABETA42, PTAU) so that their mean is 0 and their variance is 1, ensuring their effective application in machine learning algorithms.

For categorical variables (PTGENDER, PTEDUCAT, PTETHCAT, PTRACCAT), One-Hot Encoding was used to convert them into numerical variables to avoid the model misinterpreting the order relationship between categories.

In this experiment, if the dependent variable CDGLOBAL is divided into three categories, the third category has only 41 samples. Even with methods such as SMOTE to address sample imbalance, the model performance is still unsatisfactory. To ensure model performance, a binary classification task was decided upon. A CDGLOBAL value of 0 indicates normal, and values of 0.5, 1, and 2 are grouped as AD-related cognitive impairment (CDGLOBAL 0.5 represents very early AD).

3.2. Experimental procedure

This study established three random forest models [16]. In Model 1, ABETA42 and PTAU were used as independent variables. The dataset was divided into training and test sets in an 80/20 ratio. RandomForestClassifier was used to train the training set, with a random seed of 42 for consistency of experimental results. After training, a classification report is output, including metrics like accuracy, precision, recall, and F1-Score. The prediction performance of each category is analyzed using a confusion matrix. To evaluate model stability, 5-fold cross-validation is used to calculate the average cross-validation score. Meanwhile, the model's feature_importances_ attribute is used to rank features by importance, analyzing which features have a dominant effect on the model.

Model 2, based on Model 1, adds the Age feature and retrains the model to observe any improvement in performance. Subsequently, based on Model 2, gender, education level, and race/ethnicity features are included as independent variables to train Model 3, and the performance of Model 3 is compared with the first two models.

To further optimize model performance, based on Model 3, the hyperparameters of the random forest model are tuned using GridSearchCV, mainly adjusting the number of trees, maximum depth, and minimum sample splits. Furthermore, based on the independent variable combinations of Model 3, XGBoost and LightGBM models are tried and compared [16]. Finally, to further analyze the

DOI: 10.54254/2753-8818/2026.32825

model's predictive behavior, the `predict_proba` method of `RandomForestClassifier` was used to obtain the predicted probability that each test sample belongs to the dementia class. These predicted probabilities were stratified into different levels using a threshold from 0.0 to 1.0 (with a step size of 0.1). Each level corresponds to a different prediction confidence level, and samples were classified according to confidence level from low to high. This allowed for a deeper understanding of the model's performance at different confidence levels and the impact of various features on the model's prediction results at different prediction confidence levels. Based on this, further attention was paid to misclassified samples. By analyzing their mean and distribution characteristics, key factors that might lead to model misclassification were identified.

4. Results

Using ABETA42 and PTAU as independent variables, Model 1 achieved an accuracy of 0.64. However, the recall for the normal class (Category 0) was low at only 0.53, indicating a significant error in predicting normal individuals. For the dementia category (Category 1), the recall was 0.73, demonstrating that the model could effectively identify dementia patients. Feature importance data obtained using `feature_importances_` showed that ABETA42 had an importance of 0.53, while PTAU had an importance of 0.47, indicating both biomarkers contributed similarly to the model.

Model 2, by adding the Age feature to Model 1, showed a significant improvement in performance. The classification report showed that Model 2 achieved an accuracy of 0.71, and the recall for categories 0 and 1 increased from 0.53 and 0.73 to 0.61 and 0.79, respectively. Building upon this foundation, the introduction of four additional demographic features—gender, education level, race, and ethnicity—improved the accuracy of Model 3 to 0.81, with both classes showing increased recall, reaching 0.74 and 0.86 respectively. The cross-validation score of the three models also improved from 0.641 to 0.685 to 0.688.

Figure 1 reveals that demographic features other than age have relatively low importance in the model. However, despite the relatively small individual contribution of demographic features, their combination with biomarkers helps the model identify potential Alzheimer's disease-related patterns and provides more accurate predictions. Therefore, their role in multivariate models is indispensable, especially in complex clinical data, where they help to capture more potential influencing factors.

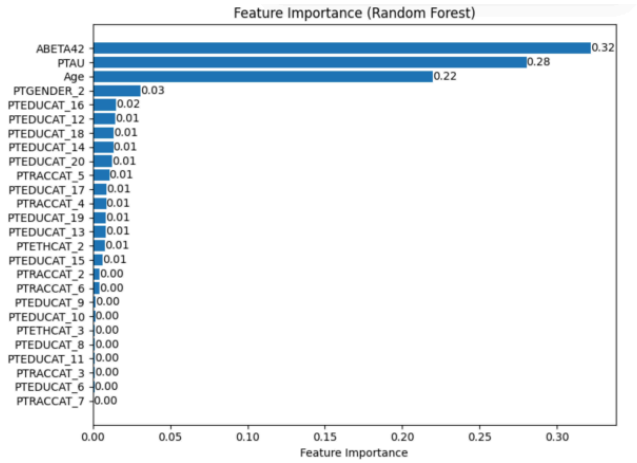

Figure 1. Feature importance of key variables in Alzheimer's disease prediction model

DOI: 10.54254/2753-8818/2026.32825

To further improve model performance, hyperparameter tuning of Model 3 was performed using GridSearchCV. The results showed limited performance improvement, with only the recall rate for category 0 increasing from 0.74 to 0.75, indicating that the random forest model was approaching its performance ceiling. However, the mean cross-validation score improved from 0.688 to 0.705, suggesting improved stability after using GridSearchCV.

In addition to Random Forest, this study evaluated XGBoost and LightGBM models; however, both demonstrated lower accuracy and recall compared to the Random Forest model. Overall, the best performing model was Random Forest Model 3 with hyperparameter tuning using GridSearchCV; therefore, this model will continue to be used in subsequent analyses.

Subsequently, the predicted probability of each sample in Model 3 was obtained using the `predict_proba` method of `RandomForestClassifier`, and stratified analysis was performed based on these probabilities. As shown in Figure 2, in the high-confidence levels (Layer 9 and Layer 10) where the model predicts dementia, the mean values of Age and ABETA42 are negative, indicating that the increase of these two variables is negatively correlated with the probability of predicting dementia. However, this negative age effect reflects ADNI's sampling bias: controls were older (75.8 y) than cases (72.4 y). PTAU is positive in the high levels and has the highest absolute value among all features, indicating a significant positive correlation between PTAU and the model's prediction of dementia. These findings provide foundation for model optimization and feature adjustment, while the high values of PTETHCAT_2 and PTETHCAT_5 are due to data skewness and do not provide any further information.

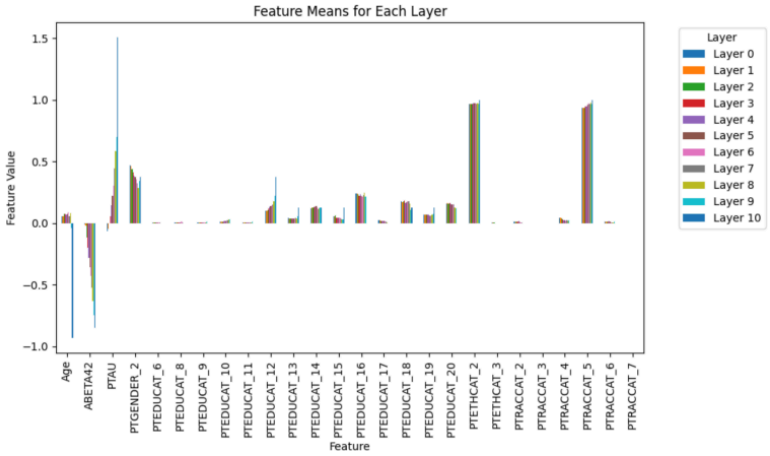

Figure 2. Feature means across layers for Alzheimer's disease prediction

Building upon this foundation, a misclassification analysis was conducted to identify key features influencing model misclassification. Analysis of the means and distributions of false positive (FP) and false negative (FN) samples revealed significant deviations from 0 for three features: ABETA42, PTAU, and PTGENDER_2. Among these, only the ABETA42 mean showed a difference in positive and negative signs between FP and FN: the standardized mean was -0.18 in false positive samples and 0.33 in false negative samples. This may be because ABETA42 levels are typically significantly reduced in Alzheimer's patients; therefore, lower ABETA42 levels in healthy individuals or higher levels in patients can lead to misclassification, further illustrating the crucial role of ABETA42 in early screening.

The mean values of PTAU and PTGENDER_2 showed consistent differences in direction across both types of misclassified samples. The standardized mean of PTAU was -0.24 in false positives and -0.44 in false negatives, indicating that PTAU values were low in misclassified samples. The

DOI: 10.54254/2753-8818/2026.32825

standardized mean for sex in false positives was 0.40, and in false negatives it was 0.42, indicating a higher proportion of females in misclassified samples. This suggests that the roles of PTAU and sex in false positives and false negatives may be complex. Especially regarding sex, given the relatively balanced sex distribution in this study, females may have more complex physiological mechanisms leading to model misclassification.

5. Conclusion

The results show that the random forest model 3, optimized with hyperparameters using GridSearchCV and incorporating all independent variables (protein biomarkers and demographic characteristics), performed best in predicting Alzheimer's disease. ABETA42, PTAU, and age contributed the most to the model and significantly influenced dementia prediction at high confidence levels. ABETA42 was negatively correlated with the prediction of dementia, while PTAU exhibited a significant positive correlation with the prediction of dementia. Further misclassification analysis revealed that the mean values of ABETA42, PTAU, and PTGENDER_2 in the misclassified samples deviated significantly from the overall sample mean. In particular, women may have more complex physiological mechanisms that significantly influence misclassification in dementia prediction.

Although Model 3 achieved an accuracy of 81%, which is good for early screening, there is still room for improvement compared to some existing studies with accuracy rates exceeding 95%. Therefore, the current model should primarily be used as a preliminary screening tool rather than a sole diagnostic method until further optimization. Additionally, two methodological limitations should be noted. First, our use of inner join to merge datasets retained only participants with complete data, which may introduce selection bias. Second, while CDGLOBAL 0.5 was grouped as early AD-related cognitive impairment, it may also arise from non-AD etiologies (e.g., vascular cognitive impairment). Furthermore, the data came from protein biomarkers in cerebrospinal fluid. Future research could investigate blood protein biomarkers to potentially lower costs, particularly in populations with complex physiological conditions.

References

[1] Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A. (2022). Early-stage Alzheimer's disease prediction using machine learning models. Frontiers in public health, 10, 853294.

[2] Arafa, D. A., Moustafa, H. E. D., Ali, H. A., Ali-Eldin, A. M., & Saraya, S. F. (2024). A deep learning framework for early diagnosis of Alzheimer's disease on MRI images. Multimedia Tools and Applications, 83(2), 3767-3799.

[3] Rani, S., Dhar, S. B., Khajuria, A., Gupta, D., Jaiswal, P. K., Singla, N., ... & Barnwal, R. P. (2023). Advanced overview of biomarkers and techniques for early diagnosis of alzheimer's disease. Cellular and Molecular Neurobiology, 43(6), 2491-2523.

[4] Papaliagkas, V., Kalinderi, K., Vareltzis, P., Moraitou, D., Papamitsou, T., & Chatzidimitriou, M. (2023). CSF biomarkers in the early diagnosis of mild cognitive impairment and Alzheimer's disease. International Journal of Molecular Sciences, 24(10), 8976.

[5] Troike, K., Zilka, S., Bekris, L. M., Leverenz, J., Khrestian, M., DiFilippo, F. P., ... & Colón-Franco, J. M. (2024). A-120 Analytical and Clinical Verification of the Elecsys® Alzheimer's Disease Cerebrospinal Fluid Biomarkers pTau/Abeta42 and Total-Tau/Abeta42. Clinical Chemistry, 70(Supplement_1), hvae106-119.

[6] Lewczuk, P., Łukaszewicz-Zając, M., Kornhuber, J., & Mroczko, B. (2024). Clinical significance of plasma candidate biomarkers of Alzheimer's Disease. Neurologia i Neurochirurgia Polska, 58(4), 363-379.

[7] Cecchetti, G., Agosta, F., Rugarli, G., Spinelli, E. G., Ghirelli, A., Zavarella, M., ... & Filippi, M. (2024). Diagnostic accuracy of automated Lumipulse plasma pTau-217 in Alzheimer's disease: a real-world study. Journal of Neurology, 271(10), 6739-6749.

DOI: 10.54254/2753-8818/2026.32825

[8] Vrahatis, A. G., Skolariki, K., Krokidis, M. G., Lazaros, K., Exarchos, T. P., & Vlamos, P. (2023). Revolutionizing the early detection of Alzheimer's disease through non-invasive biomarkers: the role of artificial intelligence and deep learning. Sensors, 23(9), 4184.
[9] Alamro, H., Thafar, M. A., Albaradei, S., Gojobori, T., Essack, M., & Gao, X. (2023). Exploiting machine learning models to identify novel Alzheimer's disease biomarkers and potential targets. Scientific reports, 13(1), 4979.
[10] Helaly, H. A., Badawy, M., & Haikal, A. Y. (2022). Deep learning approach for early detection of Alzheimer's disease. Cognitive computation, 14(5), 1711-1727.
[11] Marwa, E. G., Moustafa, H. E. D., Khalifa, F., Khater, H., & AbdElhalim, E. (2023). An MRI-based deep learning approach for accurate detection of Alzheimer's disease. Alexandria Engineering Journal, 63, 211-221.
[12] Diogo, V. S., Ferreira, H. A., Prata, D., & Alzheimer's Disease Neuroimaging Initiative. (2022). Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. Alzheimer's Research & Therapy, 14(1), 107.
[13] Liu, S., Masurkar, A. V., Rusinek, H., Chen, J., Zhang, B., Zhu, W., ... & Razavian, N. (2022). Generalizable deep learning model for early Alzheimer's disease detection from structural MRIs. Scientific reports, 12(1), 17106.
[14] Duara, R., & Barker, W. (2022). Heterogeneity in Alzheimer's disease diagnosis and progression rates: implications for therapeutic trials. Neurotherapeutics, 19(1), 8-25.
[15] Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., DeCarli, C., Green, R. C., ... & Alzheimer's Disease Neuroimaging Initiative. (2022). Using the Alzheimer's Disease Neuroimaging Initiative to improve early detection, diagnosis, and treatment of Alzheimer's disease. Alzheimer's & Dementia, 18(4), 824-857.
[16] Uddin, K. M. M., Alam, M. J., Uddin, M. A., & Aryal, S. (2023). A novel approach utilizing machine learning for the early diagnosis of Alzheimer's disease. Biomedical Materials & Devices, 1(2), 882-898.