Science, Engineering and Technology Vol. 5, No. 1, pp. 192-202
www.setjournal.com *https://doi.org/10.54327/set2025/v5.i1.189*

Identification of Plasma Proteins Associated with Alzheimer's Disease Using Feature Selection Techniques and Machine Learning Algorithms

Zakaria Mokadem[1], Mohamed Djerioui[1], Bilal Attallah[1], Youcef Brik[1]

[1]*LASS Laboratory, Faculty of Technology, University of M'sila, University Pole Road Bordj Bou Arreridj, M'sila 28000, Algeria.*

Abstract

Alzheimer's disease (AD) is a chronic, progressive neurodegenerative disorder that typically affects elderly individuals. Detecting Alzheimer's using plasma proteins is a critical step toward improving treatment results for this disease. This study aims to use computational algorithms to explore the relationship between plasma proteins and AD progression by identifying a panel of plasma proteins that can serve as biomarkers for tracking and diagnosing AD. We applied two feature selection methods, Sequential Backward Feature Selection (SBFS) and Analysis of Variance (ANOVA) to extract significant proteins from a dataset of 146 proteins. The data was collected from the plasma of 566 individuals, comprising both Alzheimer's patients and healthy controls. The SBFS technique generated all possible combinations of protein groups from the 146 proteins, which were then trained and tested using five machine learning models: Decision Tree, Random Forest, Extremely Randomized Trees, Extreme Gradient Boosting, and Adaptive Boosting. Subsequently, ANOVA was applied to refine and reduce the selected panel size. Finally, we used XGBoost and AdaBoost models to validate the final panel. The findings introduce a plasma protein panel consisting of A2Macro, BNP, BTC, PPP, and PYY proteins for diagnosing AD. This panel achieved a sensitivity of 88.88%, a specificity of 66.66%, and an AUC of 0.85. These results demonstrate that plasma protein biomarkers can facilitate timely interventions, potentially slowing disease progression and improving patient outcomes. This non-invasive and affordable diagnostic method has the potential to make Alzheimer's screening accessible to a broader population.

Keywords: *Alzheimer's disease, ANOVA, Blood biomarker, Feature selection, Machine learning, Plasma proteins*.

1. Introduction

Alzheimer's disease (AD) is a chronic, progressive neurodegenerative disorder that typically affects individuals aged 60 years and older [1]. Alzheimer's is gradient degeneration of the essential cognitive activities such as memory, thinking, and cognition [2]. Generally, AD is characterized by the gradient death of nerve cells, which is caused by the accumulation of extracellular amyloid-β (Aβ) plaques and interneuronal neurofibrillary (NFL) tangles composed of forms of the Tau-protein [3]. Diagnosing AD is a complicated process that involves a combination of neuropsychological assessments, blood tests, cerebrospinal fluid (CSF) analysis, and neuroimaging. These biomarkers require careful evaluation to exclude the other neurodegenerative disorders that share similar symptoms with AD. However, despite the effectiveness of PET and CSF biomarkers in the clinical diagnostic process for AD, the high cost of PET scans and CSF tests restricts their accessibility and generalizability as diagnostic tools [4]. These limitations could be countered by the use of blood protein biomarkers in AD diagnosis [5]. Blood testing is a well-established part of clinical practice, as it is easy to perform, safe, and does not require additional training for healthcare professionals [6], [7]. Blood is a complex liquid tissue

Corresponding author: *Zakaria Mokadem (zakaria.mokadem@univ-msila.dz)*
Received: *28 August 2024*; Revised: *3 January 2025*; Accepted: *16 January 2025*; Published: *5 February 2025*
© 2025 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License

composed of cells and extracellular fluid. It contains a diverse array of molecules, including proteins, nucleic acids, lipids, and other metabolic products, which can be observed in plasma, serum, and cellular compartments [8]. Brain-derived biomarkers are typically present at relatively low concentrations in the blood, due to the blood-brain barrier (BBB), which restricts the free passage of molecules between the central nervous system (CNS) and blood [9]. However, the progressive damage to the BBB in AD patients may allow some proteomics molecules to pass into the blood, thereby enabling the possibility of diagnosing patients with AD and the progression of the disease based on plasma proteins.

Several studies have attempted to identify blood biomarkers associated with AD by profiling a wide range of proteins in the blood and investigating their correlation with AD progression. For example, Ray et al. [10] proposed 18 proteins in blood plasma that can be used to classify blinded samples from Alzheimer's and control subjects with close to 90% accuracy using an unsupervised clustering algorithm called Shrunken centroid algorithm. Thambisetty et al. [11] discovered a panel of 15 plasma proteins that were differentially expressed in AD using the correlation analyses method. O'Bryant et al. [12] identified a panel with 30 serum proteins with sensitivity (SN) of 88%, specificity (SP) of 82%, and area under receiver operating curve (AUC) of 0.91. Laske et al [13] used a support vector machine (SVM) classifier to identify a panel of 3 serum markers of AD with 93.8% sensitivity, and 80.0% specificity. Llano et al. [14] developed a classifier model and identified 4 different proteomic signatures with 86.5% SN, 84.2% SP, and AUC of 0.85. Guo et al. [15] applied Six algorithms, Linear Discriminant Analysis (LDA), Logistic Regression (LR), Partial Least Squares (PLS), Random Forests (RF), Nearest Shrunken Centroids (NSC), and Support Vector Machine (SVM) to identify the 5 plasma panels of AD using SPSS and logistic regression analysis software and achieved 89.36% of sensitivity and 79.17% of specificity. Morgan et al. [16] developed a method based on LR for identifying a panel with 5 inflammatory markers in plasma with 84% SN, 70% SP, and AUC of 0.79. Stamate et al [17] proposed a study that evaluated the use of Deep Learning (DL), RF, and extreme gradient boosting (XGBoost) algorithms for AD classification and achieved an AUC of 0.88 with XGBoost and 0.85 with DL and RF. Zhao et al. [18] identified a panel with 12 serum markers using multivariate RF machine learning with 90% SN and 66.7% specificity. More recently, Eke et al. [19] developed a method based on SVM to identify plasma-based biomarkers for early AD detection, by using feature selection and evaluation modalities, 5 panels were identified that achieved sensitivity (SN) > 80%, specificity (SP) > 70%, and AUC of at least 0.80.

However, most of these studies used no more than one feature selection technique, making the panels generated by these techniques unstable and unreliable as biomarkers for AD, thereby highlighting the need for further investigation and scrutiny.

The main objective of this study is to use computational algorithms to identify a panel of plasma proteins that can serve as biomarkers for AD detection. This study aims to introduce a novel plasma protein panel for the accurate diagnosis of AD using Sequential Backward Feature Selection (SBFS) techniques utilized with five ML models: Decision Tree (DT), Random Forest (RF), Extremely randomized trees (Extra Trees), eXtreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost).

Proteins often play a role in biological activities within the body that change during the progression of AD, making them valuable indicators and biomarkers. By computing protein concentrations in the plasma of AD patients, we can identify proteins associated with the disease and develop a practical, cost-effective algorithm for its diagnosis. This study has significant implications: diagnosing AD using the selected protein panel enables timely interventions that can slow disease progression and improve patient outcomes. Furthermore, by focusing on plasma-based biomarkers, the study emphasizes a non-invasive, scalable, and affordable diagnostic method, making AD screening accessible to a broader population.

The rest of this paper is organized as follows: Section 2 presents the materials and methods, including the dataset description, an introduction to the machine learning (ML) algorithms, panel validation process, and feature selection techniques used. Section 3 presents the results, including the outcomes of the Sequential Backward Feature Selection (SBFS) technique, one-way Analysis of Variance (ANOVA) outcomes, and the results of the classification models that were used for the validation of the selected plasma protein panel. The discussion of the results is in Section 4. Section 5 draws

important conclusions, and outlines potential directions for future research.

2. Materials and Methods

2.1. Data description

This study used samples collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI) to qualify multiplex panels in plasma proteomics. ADNI provides unlimited data access and encourages researchers to develop potential methods for analyzing the progression of AD [20].

The data lists 566 subjects, which represent baseline data on the concentration of 146 blood plasma proteins derived from a cohort of 112 Alzheimer's disease (AD) patients, 396 mild cognitive impairment (MCI) patients, and 58 healthy controls (HC). The available neuropsychological assessments in ADNI, such as the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR), were used to categorize the clinical groups. A list of the 146 proteins is provided in the supplementary materials. Table 1 presents the demographics of the baseline subjects.

Table 1. Demographics of the baseline subjects.

Group	HC	MCI	AD
Nbr baseline	58	396	112
Age	75.3 (62-90)	74.9 (55-90)	75.4 (55-89)
Gender M/F	30/28	256/140	65/47
MMSE	28.9 (25-30)	27.0 (23-30)	23.6 (20-27)

2.2. Machine learning algorithms

2.2.1. Decision Tree

DT is a hierarchical supervised machine learning algorithm that employs decision rules to divide the feature space of a dataset into subsets belonging to a single class. It achieves this by recursively partitioning the feature space of the training data to determine an optimal set of decision rules [21].

Feature selection methods are employed to identify the most effective partitions that separate different classes. Entropy (E) and Information Gain (IG) are used to evaluate each node within the decision tree. The feature with the highest Information Gain or the lowest Entropy is chosen as the optimal candidate for splitting the data at a given node. The formula for entropy is as follows:

$$E = -\sum_{i=1}^{n} p_i \log(p_i) \quad (1)$$

Where the P_i are the ratios of elements of each class. The formula for Information Gain (IG) is as follows:

$$IG = E_{parent} - Average(E_{children}) \quad (2)$$

2.2.2. Random Forest

RF is an ensemble learning algorithm that enhances prediction accuracy by combining multiple decision trees. Each decision tree in the Random Forest is built using a bootstrapped sample of the training data, ensuring that every sample has an equal probability of being chosen. The algorithm selects a random subset of the training data with the replacement for constructing each tree. In the ensemble, every tree operates as an independent classifier, predicting the class label of an unlabeled instance. The final classification decision is made using a majority voting technique, where the most frequent prediction among the trees determines the output [22], [23].

2.2.3. Extremely Randomized Trees

The Extra Trees algorithm is an ensemble learning method that leverages decision tree principles for both classification and regression tasks. During the tree-building process, Extra Trees randomly selects split points at each node to construct multiple decision trees. Each tree is built using a randomly sampled subset of the data without replacement, ensuring that each tree has unique samples. A random subset of features is selected for each tree from the total feature set. The entire dataset is used to build the trees, enhancing efficiency. By combining random splits and aggregating the results from multiple trees, Extra Trees reduces computational costs, improves processing speed, and performs effectively on large, high-dimensional datasets.

2.2.4. Extreme Gradient Boosting

XGBoost is a flexible supervised learning algorithm known for its fast execution and support for parallel computing. XGBoost introduces a regularized model formulation to control overfitting, achieving better performance results. During the boosting process, each new model learns from the errors of previous iterations, creating trees sequentially. Each decision tree is adjusted

to address the mistakes of the prior model. This tree ensemble approach enhances the predictive power compared to a single decision tree, enabling boosting to overcome the limitations of poor predictive performance. Additionally, random sampling in the XGBoost model reduces the variance of the final model. It improves prediction accuracy by using only a random subset of the data to fit each new tree [24].

2.2.5. Adaptive Boosting

AdaBoost algorithm is an ensemble learning method designed to improve the predictive performance of weak classifiers by combining them into a strong classifier. It works iteratively, adapting to the data at each step. Initially, all training samples are assigned equal weights. A weak classifier is then trained to minimize the classification error on the weighted dataset. After training, the algorithm calculates the error of the classification and assigns it a weight based on its performance. The weights of the training samples are then updated to emphasize those that were misclassified, making them more influential in the next iteration. This process is repeated for a predetermined number of iterations or until the desired accuracy is achieved [25].

2.3. Feature selection techniques

2.3.1. Sequential backward feature selection

SBFS is a feature selection technique that uses a top-down search approach to exclude features iteratively. It begins with the full feature set, involving all the features, and applies the basic Sequential Backward Selection (SBS) method. Features are progressively excluded, and at each iteration, the most significant feature from the remaining set is conditionally included if it improves the performance of the previous subsets. This process continues until the optimal feature subset is identified [26]. In the first step, the SBFS technique was applied iteratively to train five machine learning classifiers: Decision Tree (DT), Random Forest (RF), Extremely Randomized Trees (Extra Trees), Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost). The technique evaluated all possible subset combinations of the 146 proteins for each classifier. For each ML model, the accuracy was recorded to identify the subsets of proteins that achieved the best classification performance. By the end, five combinations of significant features, each corresponding to the results of one of the ML algorithms, were improved as the most relevant for Alzheimer's disease classification. Figure 1 presents the SBFS diagram.

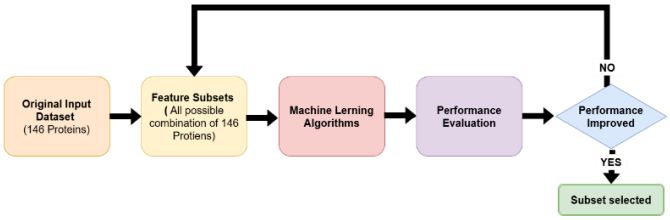

Figure 1. Diagram of the sequential backward feature selection technique.

2.3.2. One-way analysis of variance

ANOVA is a statistical technique used to analyze the differences between group means and their associated variances. It helps determine whether at least one group's mean is significantly different from others. ANOVA can help reduce the dimensionality of a dataset by identifying features that have a significant relationship with the target variable.

The second step of the framework involved using one-way ANOVA to reduce the size of the panels obtained from the first step of the feature selection by comparing the means and the variances of each protein in two clinical groups, HC and AD. This particular application of ANOVA enabled the exclusion of irrelevant proteins or those not associated with AD from the panels, specifically those with similar concentration values in the two groups. ANOVA provides an F-statistic and a p-value as results. A threshold (p-value=0.05) was applied to determine whether to reject or retain features. Features with non-significant p-values (high p-values) were excluded, as they did not contribute to distinguishing between the classes. While features with significant p-values (low p-values) were retained on the protein panel. The null hypothesis (H_0) states that the means of the two groups (HC and AD) are equal, indicating no statistically significant difference in the tested feature (protein). In such cases, the feature was eliminated from the panel. The alternative hypothesis (H_1) states that the means of the two groups are different, indicating a statistically significant difference. These features were kept in the panel. This statistical analysis was applied to all features selected in

the first step, enabling the identification of a plasma protein panel for AD prediction. We performed this statistical analysis with all the features that had been selected in the first step.

2.4. Panel validation process

The panel validation process evaluates the robustness, accuracy, and generalizability of the selected plasma protein panel using two ML algorithms: XGBoost and AdaBoost. The dataset was split into training and testing subsets, with 80% allocated for training the models and 20% for testing. To reduce the risk of overfitting and enhance predictive performance, five-fold cross-validation was employed. The effectiveness of the models was assessed using key performance metrics, including sensitivity (the ability to correctly identify AD patients), specificity (the ability to correctly identify healthy controls), accuracy (the overall prediction correctness), and the Area Under the Curve (AUC), which measures the ability of ML models to distinguish between AD and healthy controls. These metrics provided a comprehensive evaluation of the predictive performance achieved with the selected protein panel. The primary goal of the panel validation process is to confirm that the selected plasma protein panel serves as a reliable and accurate tool for early-stage AD diagnosis, highlighting its potential for clinical application. Figure 2 illustrates an overall diagram of the framework, which includes the three steps.

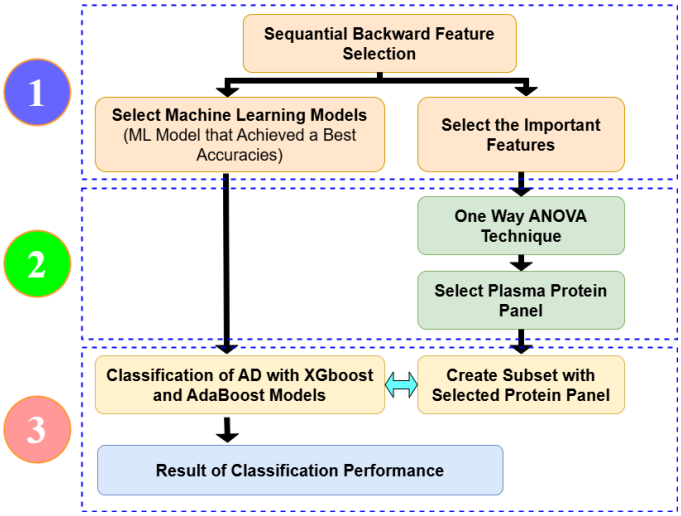

Figure 2. Proposed framework diagram.

3. Results

3.1. SBFS technique outcomes

In the first step, the SBFS technique generated subsets (groups of proteins) from the 164 proteins available in the dataset. Each subset was trained and tested using the ML models (DT, RF, Extra Trees, XGBoost, and AdaBoost). Table 2 presents the preselected proteins for each ML model along with their achieved accuracies, while Figure 3 and Figure 4 illustrate the overall results of the SBFS technique, showing the evaluation accuracies as a function of the subset size. The SBFS technique selects the group of proteins that achieves the highest accuracy for each group size. Proteins yielding accuracies greater than 90% are accepted, while those producing results below 90% are eliminated. From the results, the DT and RF classifiers achieved accuracies below 90%. However, the Extra Trees, XGBoost, and AdaBoost models achieved accuracies of 90.58%, 93.52%, and 95.88%, respectively. The selected proteins for the second step constitute the combination of proteins that achieved more than 90% accuracy in the classification process. This includes all protein groups identified by the Extra Trees, XGBoost, and AdaBoost classifiers. The selected proteins are A1Micro, Apo A-II, BTC, CD5L, Cystatin-C, IL-16, PLGF, Proinsulin-Intact, Proinsulin-Total, PYY, TECK, TN-C, ACE, A-IV, Apo B, Apo E, BMP-6, BNP, IgM, IL-8, MIP-1 alpha, SGOT, Sortilin, AGRP, CA-19-9, FRTN, HGF, IFN-gamma, PPP, RAGE, Transferrin, TTR, and VKDPS.

Table 2. Preselected proteins from SBFS for ML models with their best-achieved accuracies.

Machine learning	Preselected proteins*	Accuracy
Decision tree	Apo A-II, AXL, I-309, IL-16, MDC, PLGF, PPP, PRL, and Sortilin	86.47%
Random forest	Apo A-II, Apo A-IV, Apo B, Apo E, Apo H, Cystatin-C, IL-16, MIGI, MIP-1 alpha, NGL, PLGF, PYY, TIMP-1, and TTR	89.41%
Extra trees	A1Micro, Apo A-II, BTC, CD5L, Cystatin-C, IL-16, PLGF, Proinsulin-Intact, Proinsulin-Total, PYY, TECK, TN-C	90.58%
XGBoost	A1Micro, ACE, Apo A-II, Apo A-IV, Apo B, Apo E, BMP-6, BNP, BTC, CD5L, IGM, IL-16, IL-8, MIP-1 alpha, PYY, SGOT, Sortilin, TN-C.	93.52%
AdBoost	AGRP, Apo A-II, Apo B, Apo E, BMP-6, BNP, BTC, CA-19-9, FRTN, HGF, IL-16, IFN-gamma, PPP, Proinsulin-Total, PYY, RAGE, Transferrin, TTR, VKDP.	95.88%

(*) **Abbreviation of the preselected proteins:**

Apo A-II: Apolipoprotein A-II.
AXL: AXL Receptor Tyrosine Kinase.
I-309: T Lymphocyte-Secreted Protein I-309.
IL-16: Interleukin-16.
MDC: Macrophage-Derived Chemokine.
PLGF: Placenta Growth Factor.
PPP: Pancreatic Polypeptide.
PRL: Prolactin.
Apo A-IV: Apolipoprotein A-IV.
Apo B: Apolipoprotein B.
Apo E: Apolipoprotein E.
Apo H: Apolipoprotein H.
MIGI Monokine Induced by Gamma Interferon.
MIP-1 alpha: Macrophage Inflammatory. Protein-1 alpha.
NGL: Neutrophil Gelatinase-Associated Lipocal.
PYY: Peptide YY.
BTC: Betacellulin.
TECK: Thymus-Expressed Chemokine.
TN-C: Tenascin-C.
ACE: Angiotensin-onverting Enzyme.
BMP-6: Bone Morphogenetic Protein 6.
BNP: Brain Natriuretic Peptide.
IGM: Immunoglobulin; IL-8: Interleukin-8.
SGOT: Serum Glutamic Oxaloacetic Transaminase.
AGRP: Agouti-Related Protein.
CA-19-9: Cancer Antigen 19-9.
FRTN: Ferritin; HGF: Hepatocyte Growth Factor.
IFN-gamma: Interferon gamma Induced Protein 10.
PPP: Pancreatic Polypeptide; RAGE: Receptor for advanced glycosylation end.
VKDPS: Vitamin K-Dependent Protein S.
TIMP-1: Tissue Inhibitor of Metalloproteinases 1.
TTR: Transthyretin; 1Micro: Alpha-1-Microglobulin.

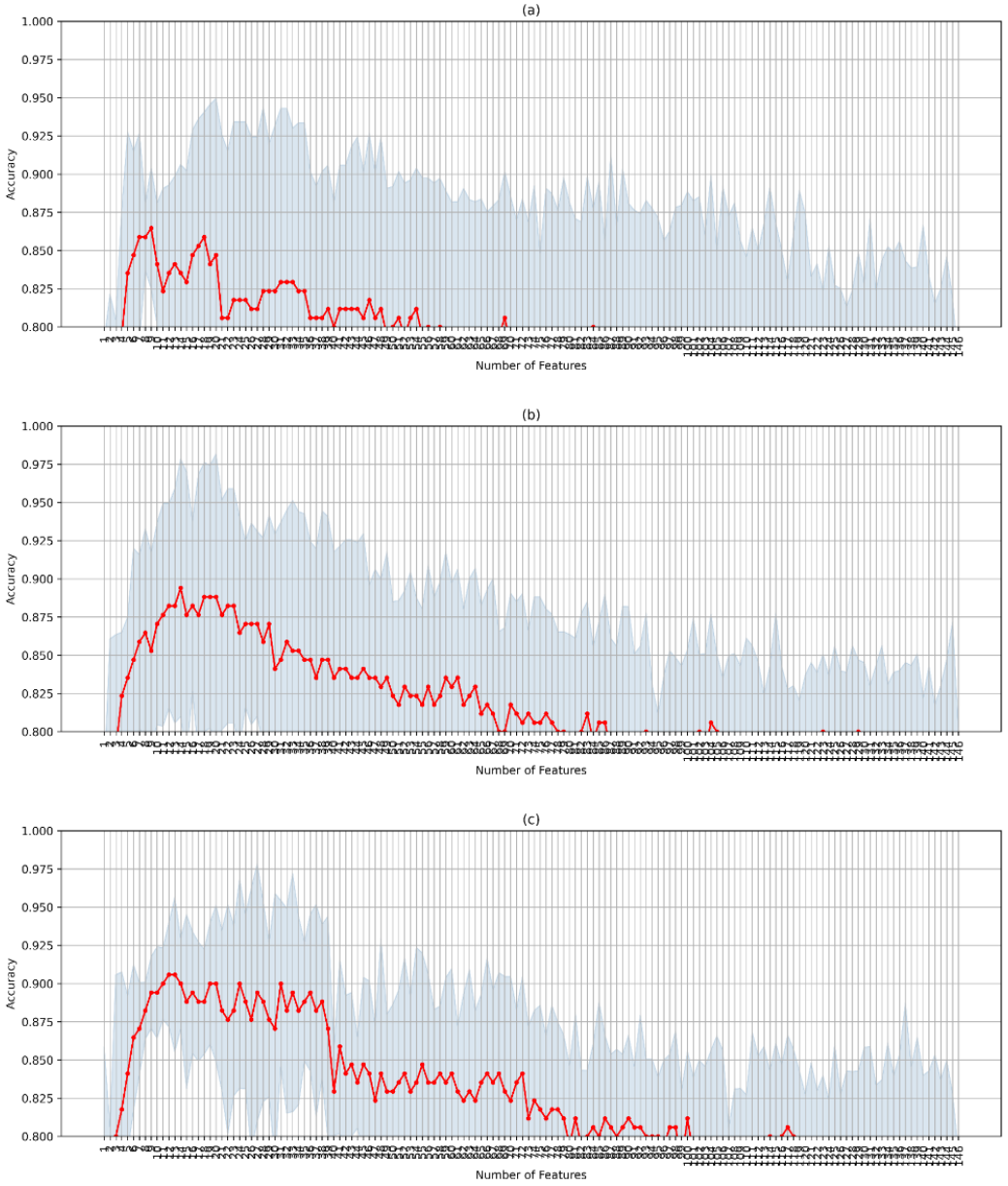

Figure 3. Results of SBFS for AD vs. HC. Performance accuracy as a function of the subset size for: (a) decision tree model, (b) random forest model, and (c) extra trees model.

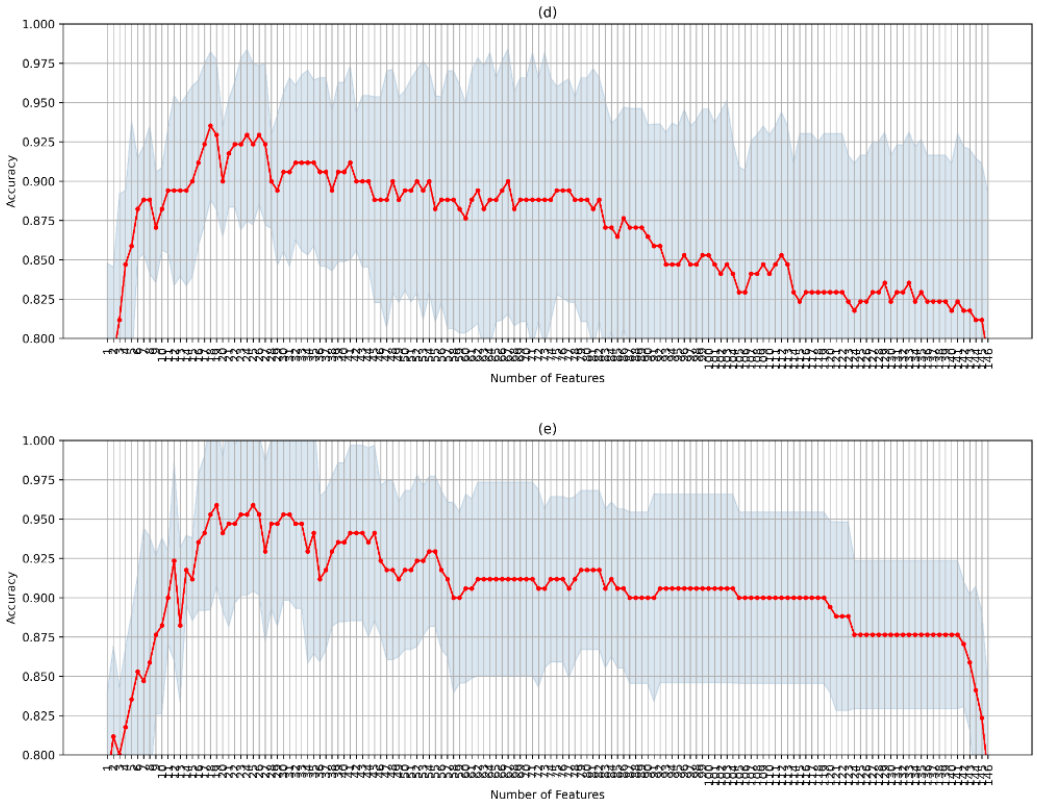

Figure 4. Results of SBFS for AD vs. HC. Performance accuracy as a function of the subset size for: (d) gradient boost model and (e) AdaBoost model.

3.2. ANOVA outcomes

The purpose of the ANOVA test was to determine whether to retain or exclude each preselected protein. For this, the two clinical groups, AD and HC, were separated, and a one-way ANOVA test was performed for each preselected protein. Proteins with high p-values between the classes were excluded, while those with low p-values were retained based on a threshold of (p-value =0.05). This process enabled the exclusion of all proteins with non-significant p-values and the creation of a final plasma protein panel containing only the significant proteins. Table 3 presents the result of the one-way ANOVA test. The final plasma protein panel selected consists of five proteins, namely A2Macro, BNP, BTC, PPP, and PYY.

Table 3. Panel selected by one-way ANOVA test for proteins that achieved a P-value<0.05.

	Protein	P-value
A2Macro	Alpha-2-Macroglobulin	0.016
BNP	Brain Natriuretic Peptide	0.014
BTC	Betacellulin	0.013
PPP	Pancreatic Polypeptide	0.025
PYY	Peptide YY	0.003

3.3. Outcomes of the classification models

We chose two machine learning algorithms, namely XGBoost and AdaBoost, to test and validate the selected protein panel. The performance parameters were measured using 5-fold Cross-validation including accuracy, sensitivity, specificity, and AUC. Both models achieved the same test accuracy of 76.47%, with slightly different values in other performance metrics. Additionally, the XGBoost model yielded an AUC of 0.85, while the AdaBoost model achieved only 0.78. Table 4 presents the performance metrics of the two machine learning by using a selected panel. Figure 5 shows the receiver operator characteristic (ROC) curves of the XGBoost and AdaBoost models.

Table 4. Performance metrics of the XGBoost and AdaBoost models using plasma protein panel selected (A2Macro, BNP, BTC, PPP, and PYY).

Machine learning	Accuracy	Sensitivity	Specificity	AUC
XGBoost	76.47%	81.81%	66.66%	0.85
AdaBoost	76.47%	88.88%	62.50%	0.78

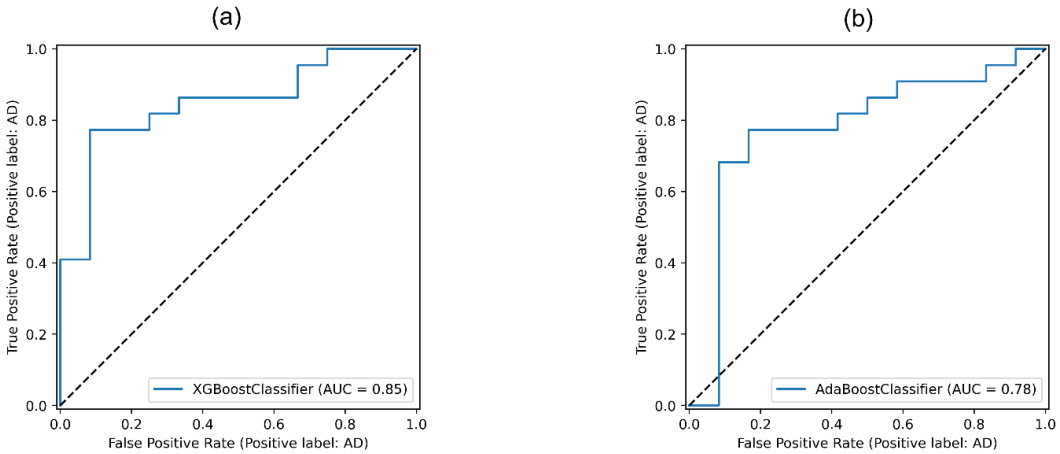

Figure 5. Receiver Operator Characteristic (ROC) curves of the classification models using plasma protein panel selected (A2Macro, BNP, BTC, PPP, and PYY). (a) XGBoost model and (b) AdaBoost model.

4. Discussion

There is a significant need for a fast and cost-effective method to diagnose AD patients. Therefore, our study aimed to identify an optimal panel of blood protein biomarkers that could serve as a first-line diagnostic tool. We proposed a novel non-amyloid plasma panel consisting of five proteins: A2Macro, BNP, BTC, PPP, and PYY, for AD diagnosis. We used the SBFS technique to train five ML algorithms, followed by one-way ANOVA analysis, to extract five significant proteins for our panel from a dataset containing 146 proteins. SBFS is a reliable process as it incorporates all possible protein combinations during the training of ML models. However, it requires complex computations and consumes substantial time. Our findings are consistent with several recent studies that aimed to identify blood proteins associated with AD. For instance, A2Macro, BNP, and PYY proteins were identified in the panel proposed by Liano et al. and Eke et al. [14], [19], while PPP and BTC proteins were selected in the study conducted by Liano et al. [14]. However, the current study proposes a smaller panel size compared to many panels suggested in previous studies. For example, Eke et al. proposed a panel with six proteins [19], and O'Bryant et al. identified 30 proteins for AD detection [12]. The smaller size of the panel makes it more interpretable and less costly to implement in practical applications, such as machine learning applications. We tested and validated the proposed panel using XGBoost and AdaBoost classifiers. The XGBoost model achieved an AUC of 0.85, indicating its ability to distinguish AD patients from the control group. While, the AdaBoost classifier achieved a specificity of 88.88%, suggesting that the model is more accurate in the classification of AD group than the HC group. Table 5 provides a comparison between our results and the recent related works.

Table 5. Comparison of our findings with recent related works.

Author	Algorithm	Panel	Results		
			SN	SP	AUC
Eke C.S. et al. [19]	SVM	Apo E, PYY, SGOT, A2M, BNP, EoT3 and RAGE	88.9	73.8	0.89
Zhao X. et al. [18]	RF	12 miRNA	90.0	67.0	0.77
Stamate D. et al. [17]	XGBoost	20 ranked metabolites	-	-	0.88
	RF		-	-	0.85
	DNN		-	-	0.85
Morgan et al. [16]	LR	-	84.0	70.0	0.79
Habbiburr R et al. [27]	LR/LASSO*	Apo E, AMBP, C3, IL-16, IGFBP2 and Apo D	-	-	0.85
Tianchi Z et al. [28]	LASSO	Apo E, CgA, CRP, CCL26, CCL20, Nr-CAM, and PYY	-	-	0.77
Current study	XGBoost	A2Macro, BNP, BTC, PPP, and PYY	81.81	66.66	0.85
	AdaBoost		88.88	62.5	0.78

*Lasso: Least Absolute Shrinkage and Selection Operator

It is crucial to interpret our findings while considering certain limitations. Our study focused on AD patients and healthy controls, excluding other AD stages such as early-MCI, MCI, and late-MCI. Therefore, more extensive studies are required, particularly with a larger dataset that includes all these categories of subjects. Additionally, the small size of the available dataset led to higher variance as it did not fully represent the diversity of individuals, potentially limiting the ability to generalize the model to new data.

The main findings of the current study confirm that plasma proteins can be used as a biomarker signature for diagnosing AD patients. We identified a panel of five proteins that can be effectively used with machine learning algorithms to predict AD. These findings facilitate the diagnostic process, and we highly recommend using this panel as a first-line diagnostic tool.

5. Conclusion

This study focused on exploiting the ML algorithms to identify a robust panel of plasma proteins associated with AD detection. Through the application of two feature selection techniques SBFS followed by one-way ANOVA, we successfully extracted a significant protein panel from a dataset of 146 non-Amyloid-β and non-Tau proteins. The identified panel, consisting of five proteins (A2Macro, BNP, BTC, PPP, and PYY), demonstrated strong potential as reliable biomarkers for AD detection. The validation process, using XGBoost and AdaBoost models, achieved a sensitivity of 88.88%, a specificity of 66.66%, and an AUC of 0.85, further emphasizing the effectiveness of the panel. These results confirm that the proposed plasma protein panel could serve as a non-invasive and cost-effective diagnostic tool for AD, significantly accelerating diagnosis and enabling timely interventions. This study has certain limitations. The dataset used in this research was relatively small, which may limit the generalizability of the findings. In future work, we will focus on the validation of the plasma protein panel using larger and more diverse datasets. Additionally, we will compare our findings with panels extracted from CSF to confirm the association between the identified proteins and AD. Furthermore, we will expand the evaluation techniques to include multiclass classification tasks and to determine the protein concentration levels for each class. These efforts will help refine the diagnostic framework and enhance its applicability for broader clinical use.

Acknowledgments

We acknowledge the Alzheimer's Disease Neuroimaging Initiative (ADNI) for granting us access to their database.

Competing Interest Statement

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability

The data used in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu) and is available with permission to all researchers.

Supplementary Data

List of the proteins.

References

[1] Tan, C.-C., Yu, J.-T., Tan, L., "Biomarkers for preclinical Alzheimer's disease", *Journal of Alzheimer's Disease*, vol. 4, no. 42, pp. 1051–1069, 2014.

[2] Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack Jr, R., Kaye, J., Montine, T.J., et al., "Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on aging-alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease", *Alzheimer's & dementia*, vol. 3, no. 7, pp. 280–292, 2011.

[3] Kumar, A., Sidhu, J., Goyal, A., & Tsao, J. W., "Alzheimer's disease", 2018.

[4] PLANCHE, Vincent, BOUTELOUP, Vincent, PELLEGRIN, Isabelle, et al., "Validity and performance of blood biomarkers for Alzheimer disease to predict dementia risk in a large clinic-based cohort", *Neurology,* vol. 100, no 5, pp. e473-e484, 2023.

[5] Lista, S., O'Bryant, S.E., Blennow, K., Dubois, B., Hugon, J., Zetterberg, H., Hampel, H., "Biomarkers in sporadic and familial Alzheimer's disease", *Journal of*

Alzheimer's Disease, vol. 2, no. 47, pp. 291–317, 2015.

[6] Henriksen, K., O'Bryant, S.E., Hampel, H., Trojanowski, J.Q., Montine, T.J., Jeromin, A., Blennow, K., L¨onneborg, A., Wyss-Coray, T., Soares, H., et al., "The future of blood-based biomarkers for Alzheimer's disease", *Alzheimer's & Dementia*, vol.1, no. 10, pp. 115–131, 2014.

[7] Schneider, P., Hampel, H., Buerger, K., "Biological marker candidates of Alzheimer's disease in blood, plasma, and serum", *CNS neuroscience & therapeutics*, vol. 4, no. 15, pp. 358–374, 2009.

[8] Lista, S., Faltraco, F., Prvulovic, D., Hampel, H., "Blood and plasma-based proteomic biomarker research in Alzheimer's disease", *Progress in Neurobiology*, vol.101, pp. 1–17, 2013.

[9] Thambisetty, M., Lovestone, S., "Blood-based biomarkers of Alzheimer's disease: challenging but feasible", *Biomarkers in medicine*, vol. 1, no. 4, pp. 65–79, 2010.

[10] Ray, S., Britschgi, M., Herbert, C., Takeda-Uchimura, Y., Boxer, A., Blennow, K., Friedman, L.F., Galasko, D.R., Jutel, M., Karydas, A., et al., "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins", *Nature Medicine*, vol. 11, no. 13, pp. 1359–1362, 2007.

[11] Thambisetty, M., Hye, A., Foy, C., Daly, E., Glover, A., Cooper, A., Simmons, A., Murphy, D., Lovestone, S., "Proteome-based identification of plasma proteins associated with hippocampal metabolism in early Alzheimer's disease", *Journal of Neurology* vol. 255, pp. 1712–1720, 2008.

[12] O'Bryant, S.E., Xiao, G., Barber, R., Reisch, J., Hall, J., Cullum, C.M., Doody, R., Fairchild, T., Adams, P., Wilhelmsen, K., et al., "A blood-based algorithm for the detection of Alzheimer's disease", *Dementia and geriatric cognitive disorders*, vol. 1, no. 32, pp. 55–62, 2011.

[13] Laske, C., Leyhe, T., Stransky, E., Hoffmann, N., Fallgatter, A.J., Dietzsch, J., "Identification of a blood-based biomarker panel for classification of Alzheimer's disease", *International Journal of Neuropsychopharmacology,* vol. 9, no. 14, pp. 1147–1155, 2011.

[14] Llano, D.A., Devanarayan, V., Simon, A.J., (ADNI, A.D.N.I., et al., "Evaluation of plasma proteomic data for Alzheimer's disease state classification and for the prediction of progression from mild cognitive impairment to Alzheimer's disease", *Alzheimer Disease & Associated Disorders,* vol. 3, no. 27, pp. 233–243, 2013.

[15] Guo, L.-H., Alexopoulos, P., Wagenpfeil, S., Kurz, A., Perneczky, R., "Initiative, A.D.N., et al. Plasma proteomics for the identification of Alzheimer's disease", *Alzheimer's disease and associated disorders*, vol. 4, no. 27, 2013.

[16] Morgan, A.R., Touchard, S., Leckey, C., O'Hagan, C., Nevado-Holgado, A.J., Barkhof, F., Bertram, L., Blin, O., Bos, I., Dobricic, V., et al.," Inflammatory biomarkers in Alzheimer's disease plasma", *Alzheimer's & dementia*, vol. 6, no. 15, pp. 776–787, 2019.

[17] Stamate, D., Kim, M., Proitsi, P., Westwood, S., Baird, A., Nevado-Holgado, A., Hye, A., Bos, I., Vos, S.J., Vandenberghe, R., et al., "A metabolite-based machine learning approach to diagnosing Alzheimer-type dementia in blood: Results from the European medical information framework for Alzheimer disease biomarker discovery cohort Alzheimer's & Dementia", *Translational Research & Clinical Interventions*, vol. 1, no. 5, pp. 933–938, 2019.

[18] Zhao, X., Kang, J., Svetnik, V., Warden, D., Wilcock, G., David Smith, A., Savage, M.J., Laterza, O.F., "A machine learning approach to identify a circulating microRNA signature for Alzheimer's disease", *The journal of applied laboratory medicine*, vol. 1, no. 5, pp. 15–28, 2020.

[19] Eke, C.S., Jammeh, E., Li, X., Carroll, C., Pearson, S., Ifeachor, E., "Early detection of Alzheimer's disease with blood plasma proteins using support vector machines", *IEEE Journal of Biomedical and Health Informatics*, vol. 1, no. 25, pp. 218–226, 2020

[20] Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI, www.loni.ucla.edu/ADNI, Accessed 4th August 2024.

[21] Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., "An introduction to decision tree modeling", *Journal of Chemometrics*, vol. 6, no. 18, pp. 275–285, 2004.

[22] Breiman, L. "Random forests", *Machine learning,* vol. 1, no. 45, pp. 5–32, 2001.

[23] Ali, J., Khan, R., Ahmad, N. Maqsood, I., "Random forests and decision trees", *International Journal of Computer Science Issues (IJCSI)*, vol. 5, no. 9, pp. 272, 2012.

[24] Ramón, A., Torres, A. M., Milara, J., Cascón, J., Blasco, P., & Mateo, J.," eXtreme Gradient Boosting-based method to classify patients with COVID-19", *Journal of Investigative Medicine,* vol. 7, no. 70, pp. 1472-1480, 2022.

[25] Feng, D. C., Liu, Z. T., Wang, X. D., Chen, Y., Chang, J. Q., Wei, D. F., & Jiang, Z. M., "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach", *Construction and Building Materials*, vol. 2020 pp. 230, 2020.

[26] Pudil, P., Novovičová, J., & Kittler, J.," Floating search methods in feature selection", *Pattern recognition letters*, vol.11, no. 15, pp. 1119-1125, 1994.

[27] REHMAN, Habbiburr, ANG, Ting Fang Alvin, TAO, Qiushan, et al. "Comparison of commonly measured plasma and cerebrospinal fluid proteins and their significance for the characterization of cognitive impairment status", *Journal of Alzheimer's Disease*, vol. 97, no 2, pp. 621-633, 2024.

[28] ZHUANG, Tianchi, YANG, Yingqi, REN, Haili, et al. "Novel plasma protein biomarkers: A time-dependent predictive model for Alzheimer's disease", *Archives of Gerontology and Geriatrics*, vol. 129, pp. 105650, 2025.