Alzheimer's Disease Prediction Using Machine Learning

Stanford CS229 Project

Anusheh Chaudry
Department of Computer Science
Stanford University
anusheh@stanford.edu

Annie Fan
ICME
Stanford University
anniefan@stanford.edu

Amelia Farinas
Neurosciences IDP
Stanford University
afarinas@stanford.edu

1 Introduction

Alzheimer's disease(AD) is the most common form of dementia, affecting over 6 million people in the US alone. Despite its prevalence, it is difficult to diagnose, often relying on subjective clinical criteria or invasive procedures to sample CSF. We applied machine learning models to predict AD status from plasma proteomic measurements, which are significantly less invasive than a lumbar puncture procedure as they only require a blood draw. The inputs to our algorithms were demographic variables (age, sex, race) and blood protein measurements from individual patients, and the output was a prediction of whether the individual was a control subject or an AD patient. We tested a variety of methods including logistic regression, support vector machines, neural networks, and XGBoost.

2 Related Work

Early AD diagnosis has seen significant advancements with the integration of machine learning (ML) and deep learning (DL) models, as well as the strategic utilization of multimodal data. Traditional ML methods, including Decision Trees, Random Forest, and Support Vector Machines, offer a foundational approach to AD prediction. Kavitha and et al. (2022) demonstrated the effectiveness of these techniques in early-stage AD detection, highlighting their strength in handling structured data. However, their reliance on manual feature engineering and difficulty in processing high-dimensional data represent significant limitations. Advancements in DL, as discussed by Zhang and et al. (2022), have addressed some of these challenges by efficiently analyzing complex datasets such as blood-based proteomic data. The superior feature extraction capabilities of DL architectures, particularly convolutional neural networks (CNNs), have set new benchmarks in medical imaging analysis. Despite these advances, DL models often suffer from issues related to interpretability and require extensive computational resources. Dara et al. (2023) have emphasized the importance of integrating various machine learning approaches to enhance diagnostic precision further .

The integration of multimodal data sources represents a promising avenue for enhancing diagnostic accuracy. Diogo et al. (2022) and You et al. (2023) have demonstrated the potential of combining diverse datasets, including MRI data and plasma proteins, to develop more comprehensive diagnostic models. This approach aligns with the trend towards personalized medicine but introduces challenges in data harmonization and integration. In our opinion, the multimodal data integration strategy stands out for its holistic approach to understanding AD, leveraging the strengths of both ML and DL while mitigating their individual weaknesses. The state-of-the-art in AD diagnosis is increasingly moving away from manual, subjective assessments towards automated, data-driven models that offer higher precision and objectivity. This shift has been facilitated by the advancements in computational power and the availability of large-scale biomedical datasets.

3 Dataset and Features

Our dataset, from the Washington University Knight Alzheimer's Disease Research Center, contains clinical, demographic, and proteomic data from 3,076 elderly patients. The dataset is private; we have access due to a collaboration between the Cruchaga lab at Washington University and the

Wyss-Coray lab at Stanford. As preprocessing, we removed subjects who were missing labels (classification of AD or healthy control by an expert physician), as well as removing subjects with blood collected before 2007, whose proteomic profiles formed a distinct cluster on a PCA plot, a batch effect that could reflect changes in sample collection or storage over time. This left us with 2091 subjects with data for all demographic and proteomic variables. We used a 70/15/15 split for our training, validation, and test sets, with 1463 subjects in the training group and 314 each in the validation and test sets.

Proteomics data was obtained through the SomaScan assay, which provides measurements of 7,289 protein levels in the blood, without missing values. Measurements with values below the assay's limit of detection were retained as 'best guess' estimates for the respective protein concentration. Log10 transformation was applied to the raw protein measurements as they are distributed log-normally, and z-score normalization was applied to all quantitative measures before modeling. Other sets of features that we utilized included genetic information (copy numbers of APOE2 and APOE4 alleles, which represent, respectively, the strongest known protective and detrimental genetic variants affecting Alzheimer's risk), and clinical features (25 clinical variables such as BMI, blood pressure, comorbidities such as diabetes, and lifestyle risk factors such as tobacco use). 681 people people had at least one missing value for the clinical variables (although only 22 were missing APOE genotyping); we removed subjects missing values when utilizing these variables in our models, leaving us with 2069 subjects for the models with APOE genotype and 1410 subjects for the models including clinical variables.

4 Methods

We first used logistic regression as our baseline model in this binary classification task. A logistic regression model predicts the probability of an instance belonging to a class by applying a sigmoid function to transform the input features into a probability score between 0 and 1. Using logistic regression, we could also study the linear association between our features and AD and perform feature selection using L1 regularization. We used the LogisticRegression implementation from sklearn.linear_model, using the 'liblinear' solver which implements coordinate descent.

Next, we applied Support Vector Machines (SVMs) for distinguishing between control subjects and AD patients, utilizing demographic and plasma proteomic data. SVM, a supervised learning model, operates by constructing an optimal hyperplane in a high-dimensional space, which serves the purpose of classification or regression. The essence of SVM lies in maximizing the margin between the nearest points of different classes (functional margin), which is crucial for minimizing the classifier's generalization error. Our implementation leveraged LinearSVC from sklearn.svm, optimized for linear SVMs, adhering to the minimization problem: $\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$, where w denotes the weight vector, b the bias, and C the regularization parameter balancing classification accuracy and margin maximization. This mathematical formulation allows for a nuanced control over the trade-off between a larger margin and the classifier's performance on training data, with C acting as a key regulator in this delicate balance.

Next, we implemented XGboost using the XGBClassifier from the package XGBoost. XGBoost is a decision tree based ensemble algorithm employing a gradient boosting framework that incrementally builds decision trees to correct the residuals of previous trees, thereby improving the model's accuracy with each iteration. The core of XGBoost lies in its objective function: $Obj(\Theta) = \sum_i l(y_i, \hat{y}_i) + \sum \Omega(f_k)$. Here, l represents the loss function measuring prediction errors, $\hat{y}_i$ are the predicted values, y_i the actual values, f_k the individual trees, and Ω a regularization term penalizing model complexity to prevent overfitting. Its robustness in classification and its scalability make it well-suited for our objective and our dataset.

In our research, we employed a feedforward neural network with Rectified Linear Unit (ReLU) activation functions to classify AD patients from proteomic data. The architecture of our model includes multiple layers where each neuron connects forward without loops, allowing complex pattern recognition through the network's depth. The ReLU function, defined as $f(x) = \max(0, x)$, introduces necessary non-linearity, enabling the model to learn intricate data relationships beyond linear separability. The output layer uses a sigmoid function, $f(x) = \frac{1}{1+e^{-x}}$, to provide a probability score between 0 and 1, indicative of disease presence. Our training process involved minimizing the binary cross-entropy loss, a measure of prediction error, and iteratively updating model weights via

backpropagation to enhance diagnostic accuracy. This setup is particularly effective for biomedical datasets like ours, where the ability to discern subtle patterns can significantly impact the model's predictive performance.

5 Experiments / Results / Discussion

5.1 Logistic Regression

We first trained logistic regression models with L2 regularization to predict whether a patient has AD - one using only demographic data (age, sex, race) and another one on demographic and proteomics data. Our evaluation metrics included accuracy, precision, recall, and the Area Under the Receiver Operating Characteristic (AUC-ROC) curve. These metrics provided a comprehensive understanding of our model's performance, with AUC-ROC being particularly valuable due to its insensitivity to class imbalance, a common issue in medical datasets. We subsequently applied logistic regression with L1 regularization to perform feature selection by filtering out features that have a coefficient of 0. We chose the optimal regularization strength (c = 0.0375) based on the highest AUC achieved in the validation set. Including proteomic data significantly improved model performance (test AUC = 0.84 vs AUC = 0.64).

We further explored the use of the genetic and clinical variables described above. For these models, we again used L1 regularization selecting optimal regularization strength for each set of features using the validation set. Although, as expected, APOE status and clinical variables had improved predictive power in terms of test AUC compared to demographic variables alone, they did not provide any improvement over proteomic models, and including all variables (demographic, proteomic, clinical, genetic) also did not improve on our proteomic models, indicating that this information is already represented by the proteomic data. Therefore, we decided to focus on models containing demographic and proteomic data alone, which have the added advantage of serving as an easy and independent test for AD that only requires a blood sample, with no need for medical history or genetic testing. For our proteomics model, feature selection reduced the dimension of our input features from 7296 to 182. The features with the largest coefficients in the L1 regularized model are represented below, with red bars indicating proteins with positive coefficients (levels increased in AD patients) and blue bars indicating proteins with negative coefficients (levels decreased in AD patients). Top proteins included acetylcholinesterase, SPC25, and neuronal pentraxin receptor, indicating that these proteins may be important in AD pathogenesis. We performed gene ontology analysis using gProfiler on the 181 selected proteins to identify their shared biological functions, indicating pathways that may become dysfunctional in AD patients.

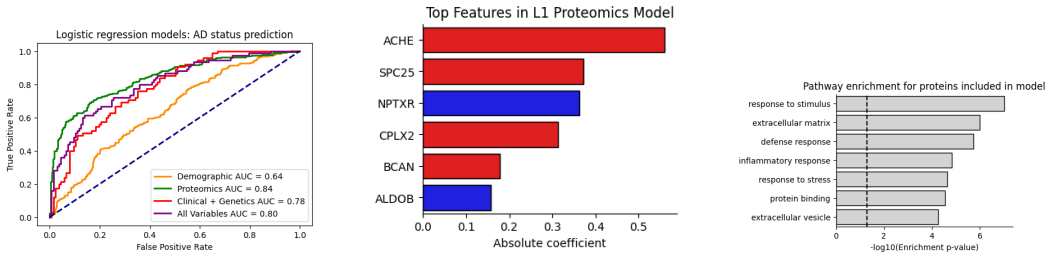


Figure 1: Logistic Regression Results: ROC, Top Feature, Pathway Enrichment

5.2 SVM

We employed SVM for its proficiency in high-dimensional and binary classification contexts, applying a 70/30 data split for training and testing. It was necessary to add feature scaling due to the high dimensionality of proteomic data, which we used StandardScaler for. The choice of LinearSVC from sklearn.svm was due to its capability to efficiently manage large, sparse datasets, and we also employed CalibratedClassifierCV for probability estimation critical in clinical diagnostics. We evaluated our model using accuracy, precision, recall, and AUC-ROC. AUC-ROC is particularly valuable for the way it deals with class imbalance, which is a frequent issue in medical datasets. The SVM model demonstrated modest performance with demographic data alone but showed significant

improvement when proteomics data was included, highlighting the potential of biological markers in AD diagnosis over demographic information. Despite this improvement, a moderate recall rate indicates a tendency toward false negatives, suggesting a need for further model refinement or exploring alternative algorithms. The performance improvement with proteomics data, depicted in ROC curves and a confusion matrix, confirms the value of integrating biological markers for a more nuanced understanding of AD, though the false negative rate calls for further model enhancement.

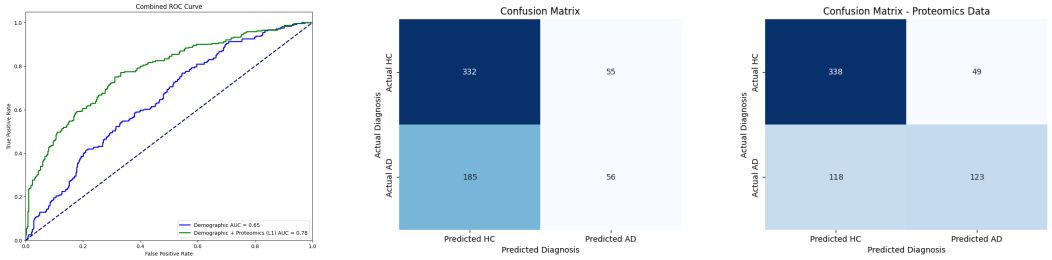


Figure 2: SVM Results: ROC, Confusion Demographic, Confusion Proteomics

5.3 XGBoost

We fitted an XGBoost classifier on demographic variables plus either the entire proteomic feature set, or a set of 100 top proteins whose levels were most significantly different between controls and AD patients in the training set. We applied grid search using 5-fold cross-validation to find a hyperparameter combination that optimized the ROC-AUC scores for both training sets. For XGBoost classifiers, one for each feature set, the learning rate is 0.1 and the objective is 'binary:logistic'. XGBoost trained on demographic data and top 100 proteins had better performances in the evaluation metrics than the model trained using demographic and the entire proteomics data. Specifically, the former showed a 10% increase in both precision and recall, which suggests that proteins whose levels were most significantly different between controls and AD patients have higher predictive power in our task. Though XGBoost yielded promising results, it did not outperform logistic regression using L1 regularization.

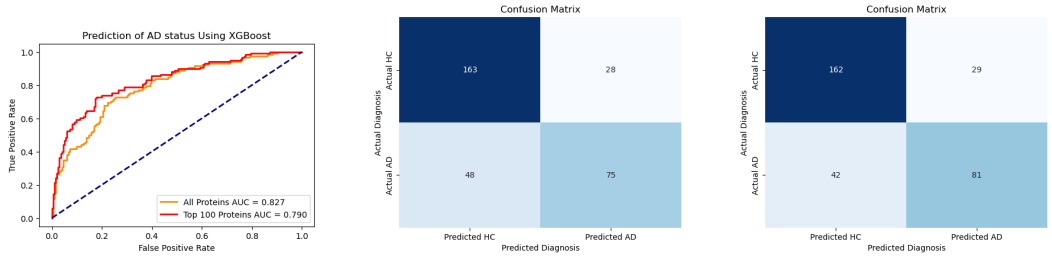


Figure 3: XGBoost Results: ROC, Confusion Baseline, Confusion Selected Features

5.4 Neural Network

To explore whether adding more complexity to the neural network would increase the predictive power, we used grid search with 5-fold cross validation to experiment with different combinations of numbers of layers (ranging from 1 to 4) and the number of units in each layer (values of 32, 64, and 128). To prevent overfitting, we added early stopping in our training process where we terminated the training if there was no decrease in the validation loss after 3 epochs. We used Relu as an activation function in the hidden layers and applied sigmoid on the output layer to a probability score between 0 and 1 of the input being in a positive class. We used a batch size of 64 and we trained for 20 epochs. The result from our grid search suggested that a simple neural network of 1 hidden layer with 64 units had the highest AUC score. The performances of the two neural networks indicated that adding complexity or regularization to the network did not offer much improvement. The performance of both neural networks was very comparable to the XGBoost classifier fitted on demographic data and top 10 proteins. Still, these more complex models did not outperform logistic regression.

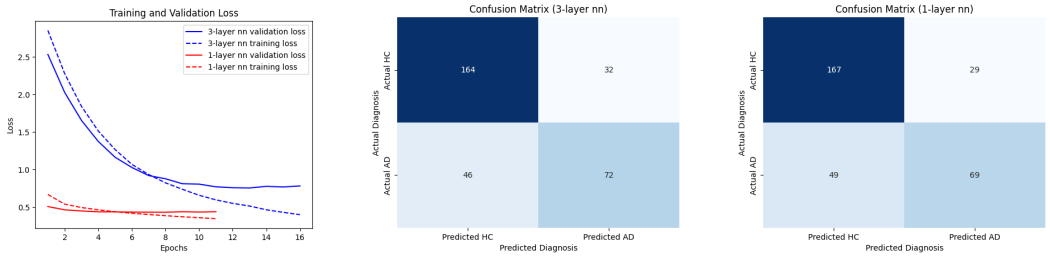


Figure 4: Neural Net Results: Loss Graph, Confusion 3-layer, Confusion 1-layer

5.5 Model Results

Model	Features	AUC	Accuracy	Precision	Recall
Logistic Regression (L2 Regularization)	Age, Sex, Race	0.644	0.619	0.508	0.270
Logistic Regression (L1 Regularization)	Age, Sex, Race + proteomics data	0.837	0.779	0.728	0.676
SVM (L2 Regularization)	Age, Sex, Race	0.645	0.618	0.505	0.232
SVM (L1 Regularization)	Age, Sex, Race + proteomics data	0.778	0.734	0.715	0.510
XGBoost	Age, Sex, Race + full proteomics data	0.790	0.713	0.663	0.483
XGBoost	Age, Sex, Race + top 100 proteomics data	0.827	0.777	0.761	0.593
1 hidden layer of 64 units	Age, Sex, Race + top 100 proteins	0.826	0.752	0.704	0.585
• Hidden layer with 64 units • Hidden layer with 128 units • Dropout layer of rate 0.2 • Hidden layer with 64 units • L2 regularization	Age, Sex, Race + top 100 proteomics data	0.814	0.752	0.692	0.610

Table 1: Combined Results

6 Conclusion / Future Work

In conclusion, among the models we tested, logistic regression with proteomic data emerged as the most effective, achieving the highest AUC score. This shows the significant role plasma proteomics have as a powerful, non-invasive biomarker for AD, providing insights that surpass traditional demographic data in predictive capability. The high performance of logistic regression, in comparison to more complex models like XGBoost and neural networks, might be attributed to its ability to effectively capture the linear relationships between the proteomic features and the disease status. The simplicity and interpretability of logistic regression also make it an appealing choice for clinical applications. Given that logistic regression, XGBoost, and a simple neural network model all showed promising results, as our future exploration we can apply a boosting and ensemble framework to dynamically combine the predictions of the three algorithms. Through this approach, we aim to incorporate the diverse strengths of each model to potentially enhance the prediction accuracy and robustness.

7 Contributions

Amelia performed dataset cleaning and worked on logistic regression, neural network models, and introduction, dataset, methods, and experiments/results/discussion parts of report. Annie worked on the

baseline of logistic regression, XGboost, neural network, and methods, experiments/results/discussion, and future works parts of report. Anusheh worked on svm model, neural network, and related works, methods, experiments/results/discussion, and conclusion parts of report.

References

OA Dara, JM Lopez-Guede, HI Raheem, J Rahebi, E Zulueta, and U Fernandez-Gamiz. 2023. Alzheimer's disease diagnosis using machine learning: A survey. *Applied Sciences*, 13(14):8298.

VS Diogo, HA Ferreira, D Prata, and et al. 2022. Early diagnosis of alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alz Res Therapy*, 14:107.

C Kavitha and et al. 2022. Early-stage alzheimer's disease prediction using machine learning models. *Frontiers in public health*, 10:853294.

J. You, Y. Guo, Y. Zhang, and et al. 2023. Plasma proteomic profiles predict individual future health risk. *Nat Commun*, 14:7817.

Yuting Zhang and et al. 2022. Predicting at(n) pathologies in alzheimer's disease from blood-based proteomic data using neural networks. *Frontiers in aging neuroscience*, 14:1040001.