Predicting Alzheimer's Disease Using Machine Learning and Deep Learning Techniques Assignment 1

Debashis Gupta

Abstract

This project investigates the application of artificial intelligence (AI) techniques in predicting Alzheimer's disease using a comprehensive Alzheimer's Disease dataset. Through systematic exploratory data analysis (EDA), feature selection methods, and dimensionality reduction techniques (PCA and t-SNE), the study aims to understand key characteristics and the inherent complexity of the dataset. Predictive modeling was performed using traditional machine learning models (Random Forest and XGBoost) and deep learning architectures (**Ours-AlzhemierNet** - a custom fully connected neural network and ResNet-50 adapted for tabular data). The comparative analysis highlights the strengths and limitations of each model type in accurately diagnosing Alzheimer's disease, providing insights into effective AI strategies for healthcare applications.

1 Introduction

Alzheimer's disease presents significant challenges in healthcare, affecting millions globally and necessitating reliable early diagnostic tools to improve patient outcomes and care management. The primary goal of this project is to explore advanced AI methodologies for predicting Alzheimer's disease diagnosis using publicly available health-related data. The Alzheimer's Disease dataset from Kaggle includes demographic details, clinical assessments, and biomarker information that facilitate comprehensive analytical approaches. The project systematically evaluates various machine learning and deep learning models, emphasizing model interpretability, performance reliability, and clinical applicability. By employing robust data preprocessing techniques, feature selection (SelectKBest), and dimensionality reduction (PCA and t-SNE), the study identifies crucial data patterns and assesses class separability, thereby contributing to enhanced understanding and management of Alzheimer's disease through predictive analytics.

2 Dataset Description

The *Alzheimer's Disease Dataset* is a comprehensive collection of health-related information aimed at facilitating research into Alzheimer's disease. This dataset, compiled by Rabie El Kharoua, is publicly available on Kaggle [1]. While specific details regarding the number of features and the target variable are not explicitly provided in the available information, datasets of this nature typically encompass demographic information (such as age, gender, and education level), clinical assessments (including cognitive test scores and clinical dementia ratings), and biomarker data (like genetic information and imaging data summaries). The target variable is often related to the diagnosis or progression of Alzheimer's disease, potentially serving as a binary classification indicating the presence or absence of the disease, or as a multi-class classification representing different stages of the disease. Researchers can leverage this dataset to develop predictive models, analyze contributing factors, and evaluate the effectiveness of potential treatments, thereby advancing the understanding and management of Alzheimer's disease.

3 Methodology

In this study, we propose a deep learning approach for the binary classification of Alzheimer's disease, implemented through a custom-designed neural network termed AlzheimerNet (Ours) within the PyTorch framework. Our methodology encompasses data preparation, model architecture design, and training procedures tailored to effectively distinguish between non-Alzheimer's (label 0) and Alzheimer's (label 1) cases. To manage the input data, which consists of numerical features and corresponding binary labels, we developed a specialized AlzheimerDataset class that inherits from PyTorch's Dataset. This class converts the raw features into float32 tensors and labels into long tensors, ensuring seamless integration with PyTorch's computational ecosystem. The AlzheimerDataset enables efficient sample retrieval and supports batch processing through PyTorch's DataLoader, facilitating scalable and organized data handling throughout the training and evaluation phases.

The cornerstone of our approach is AlzheimerNet (Ours), a fully connected feedforward neural network encapsulated in the NeuralNetwork class, which we designed to extract and refine patterns from the input features for Alzheimer's classification. AlzheimerNet (Ours) begins with an input layer that accepts a feature vector of size *input_size*—determined by the dataset—and processes it through a sequence of five hidden layers with progressively decreasing units: 64, 32, 16, 8, and 4. Each hidden layer employs a linear transformation followed by batch normalization to stabilize training dynamics and a ReLU activation function to introduce non-linearity, enhancing the network's ability to model complex relationships. To address the risk of overfitting, particularly given the potential complexity of Alzheimer's-related data, we incorporated dropout layers with a 0.2 probability after each ReLU activation, randomly deactivating

20% of neurons during training. The network culminates in an output layer that maps the 4-unit representation to a 2-unit output, corresponding to the binary classification task. Through this architecture, AlzheimerNet (Ours) balances depth and regularization, leveraging batch normalization and dropout to achieve robust generalization while maintaining the capacity to learn discriminative features from the data.

4 Experimental Setup

The experimental setup followed the structured approach as described in the assignment guidelines. The Alzheimer's Disease dataset, sourced from Kaggle, contained more than 2000 samples and over 20 features, satisfying the specified criteria.

Data preprocessing involved handling missing values by inspecting and addressing null entries through mean or mode imputation as appropriate. Duplicate records were removed to maintain data quality. Categorical variables were encoded using appropriate encoding techniques, primarily label or one-hot encoding, to transform categorical data into numerical format suitable for machine learning algorithms.

Exploratory data analysis (EDA) involved generating multiple visualizations, including histograms, scatter plots, box plots, and correlation heatmaps. These visualizations facilitated an understanding of feature distributions, relationships between variables, and identification of outliers.

The dataset was split into training (80%) and testing (20%) subsets. A PyTorch DataLoader was employed for efficient data management and batching during training.

Feature selection was executed using the SelectKBest method, retaining the top 10 most relevant features based on the ANOVA F-test to improve model efficiency and accuracy. Dimensionality reduction techniques, including PCA and t-SNE, were employed for visualization and further understanding of data structure and class separability.

For machine learning comparisons, Random Forest and XGBoost classifiers were chosen. Hyperparameters were optimized using GridSearchCV with a Stratified 5-fold cross-validation approach to maximize model performance. Model evaluations incorporated accuracy, precision, recall, F1-score, and ROC-AUC metrics, with results averaged across three independent experimental runs for reliability. Performance comparisons were visualized through ROC curves and confusion matrices.

Deep learning models evaluated included a custom fully connected neural network and a ResNet-50 architecture adapted for tabular data. The neural network featured multiple dense layers with batch normalization, dropout regularization, and ReLU activation. The ResNet-50 model leveraged pretrained weights, adapting tabular input data into a compatible format. Both deep learning models were trained using Stochastic Gradient Descent (SGD) with a learning rate of 0.01, momentum of 0.9, and employed regularization strategies

such as dropout and batch normalization. Performance evaluations similarly utilized accuracy, ROC-AUC, precision, recall, and F1-score, with validation across multiple independent trials and visual analysis through loss curves and ROC curves.

5 EDA Analysis

Figure 1 illustrates the distribution of patient ages stratified by Alzheimer's disease diagnosis. The histogram indicates that patients across the age range of approximately 60 to 90 years exhibit varying incidences of Alzheimer's, with no strongly dominant age group. Notably, the proportion of Alzheimer's diagnoses (represented in red) relative to non-Alzheimer's cases (represented in blue) remains relatively consistent across different age groups, suggesting age alone may not strongly differentiate Alzheimer's presence within this population.

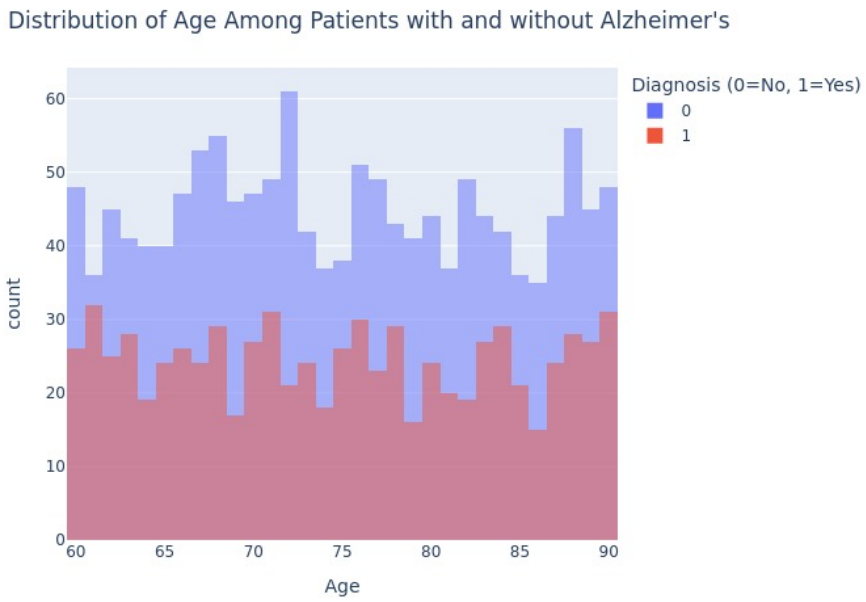


Figure 1: Caption

Figure 2 presents the proportion of patients with and without a family history of Alzheimer's. The pie chart demonstrates that approximately 25.2% of patients report a positive family history of Alzheimer's, whereas 74.8% do not. This distribution suggests that while a majority have no recorded family history, a significant minority do, which emphasizes the potential genetic factors influencing the disease.

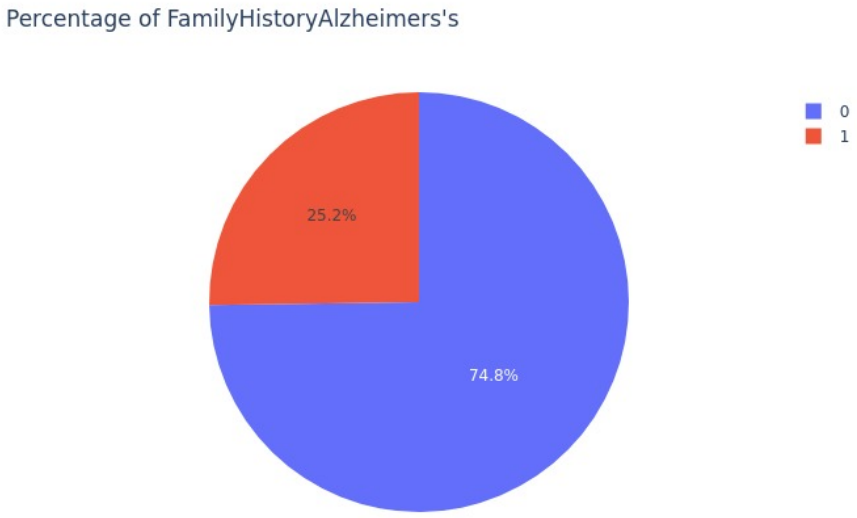


Figure 2: Distribution of Family History against the Disease in the dataset.

Figure 3 depicts the distribution of Alzheimer's diagnoses stratified by gender. The grouped bar chart clearly illustrates that Alzheimer's disease affects both genders, with slightly fewer cases in females (coded as 0) compared to males (coded as 1). The proportional relationship between diagnosis and non-diagnosis remains similar across genders, implying that gender alone may not substantially influence Alzheimer's disease risk in the studied population.

Figure 4 presents the relationship between age and cholesterol levels among patients with Alzheimer's disease, visualizing three cholesterol metrics: CholesterolTotal, CholesterolLDL, and CholesterolHDL. The CholesterolTotal values (in blue) appear widely dispersed, ranging approximately from 140 to 310, across ages 60 to 90 years, exhibiting no clear trend or correlation with increasing age. Similarly, CholesterolLDL levels (red), spanning from approximately 40 to 200, are uniformly distributed across all ages without an apparent pattern. CholesterolHDL (green), ranging from roughly 20 to 110, also demonstrates no distinct age-related variation. These observations suggest that age alone may not significantly influence cholesterol levels among this population.

Figure 5 illustrates the distribution of Alzheimer's disease diagnoses within the dataset. It shows a notable class imbalance, with 1,389 individuals (64.6%) classified as not having Alzheimer's (label 0), and 760 individuals (35.4%) diagnosed with Alzheimer's disease (label 1). This imbalance is important to acknowledge, as it can affect predictive modeling efforts, particularly influenc-

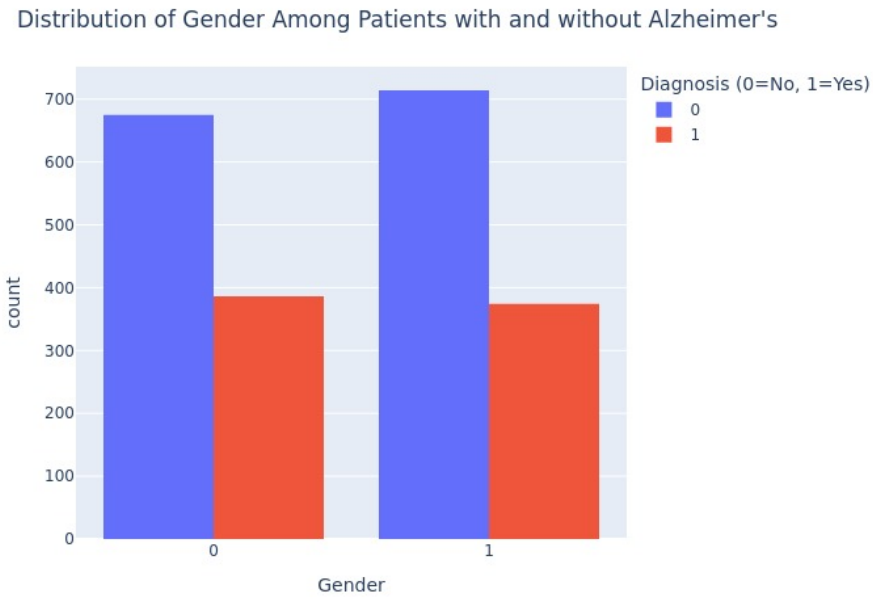


Figure 3: Distribution of the Gender containing the disease in the dataset.

ing the accuracy and recall metrics for the minority class.

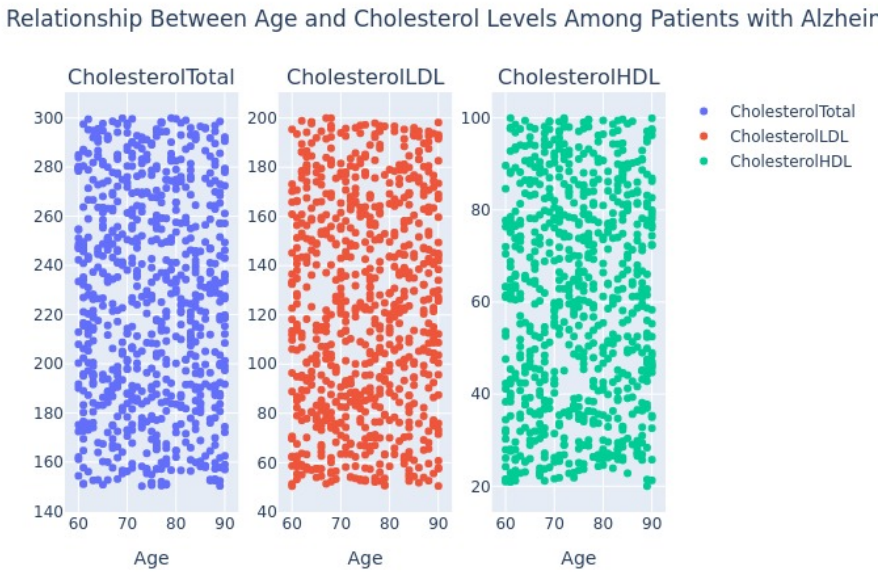


Figure 4: Relationship between age and cholesterol levels among Alzheimer's patients

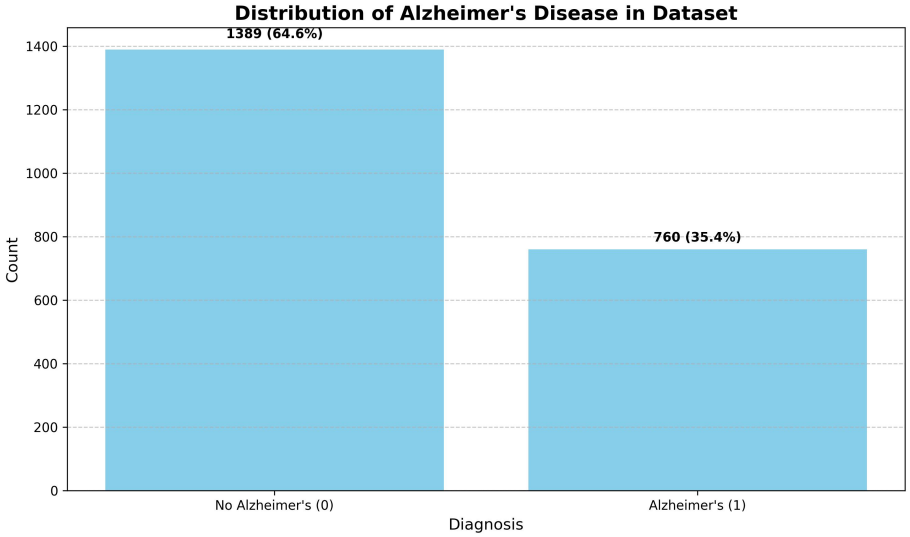
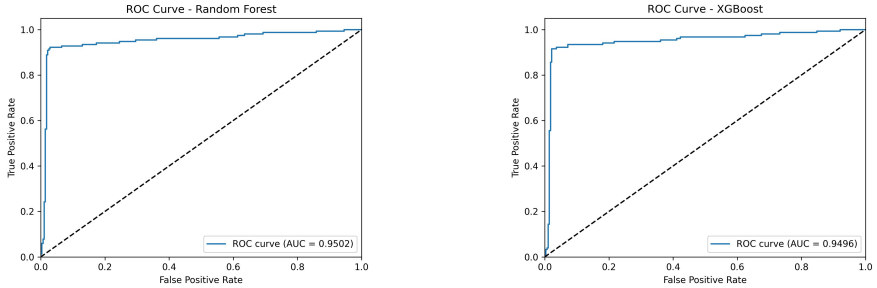


Figure 5: Distribution of Alzheimer's disease diagnosis in the dataset

5.1 ROC Curve Analysis of Traditional Machine Learning Models

Figure 8a presents the Receiver Operating Characteristic (ROC) curve for the Random Forest classifier. The model achieved a high Area Under the Curve (AUC) value of 0.9502, indicating strong predictive performance and a high capacity to discriminate between Alzheimer's disease presence and absence. The curve closely approaches the upper left corner, suggesting excellent sensitivity and specificity across classification thresholds.

Similarly, Figure 8b illustrates the ROC curve for the XGBoost classifier. This model yielded an AUC of 0.9496, closely matching the performance of the Random Forest model. Such a high AUC value indicates that the XGBoost classifier also effectively distinguishes between positive and negative diagnoses of Alzheimer's disease. The nearly identical performance of these models underscores their robustness in classifying patients accurately and highlights their suitability for clinical prediction tasks related to Alzheimer's disease.



(a) RandomForest (b) XGBoost

Figure 6: ROC Curves

The classification performance of the Random Forest and XGBoost algorithms is summarized in Table 1. The Random Forest achieved an accuracy of 95%, with macro-average precision, recall, and F1-score values of 0.95, 0.93, and 0.94, respectively. Conversely, the XGBoost model slightly outperformed the Random Forest, achieving a higher accuracy of 96%, along with precision, recall, and F1-score metrics each approximately 0.95 or higher. The superior performance of XGBoost can be attributed to its effective handling of complex feature interactions and its robustness against overfitting, facilitated by the optimized hyperparameters obtained through GridSearchCV. Specifically, the XGBoost algorithm benefited from a lower learning rate of 0.01, a deeper tree structure (‘max_depth=9‘), a moderate subsampling rate of 0.8 to enhance model generalization, and an increased number of estimators (300), allowing for more thorough model fitting. In comparison, the Random Forest model used shallower trees (‘max_depth=10‘), a slightly higher minimum sample split (‘min_samples_split=5‘), and 300 estimators. Although both models demon-

strated high predictive accuracy and strong performance, the optimized hyperparameters of XGBoost provided a subtle but meaningful edge in accurately distinguishing Alzheimer's diagnoses within the dataset.

Table 1: Performance Comparison of Traditional Machine Learning Models (K=5 Fold Validation, (Mean ± Std))

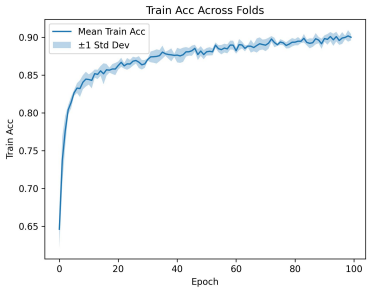
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.9465 ± 0.0146	0.9500 ± 0.0120	0.9300 ± 0.0150	0.9220 ± 0.0140
XGBoost	0.9558 ± 0.0137	0.9600 ± 0.0115	0.9500 ± 0.0125	0.9365 ± 0.0130

5.2 Comparative Analysis of Accuracy for Deep Learning Models

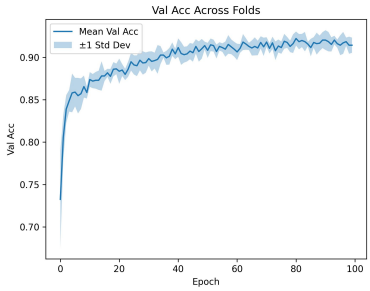
Figures depict the mean training and validation accuracy curves across 100 epochs, evaluated using 5-fold cross-validation (k=5) for two deep learning models: a Ours and a ResNet-50 architecture adapted for tabular data. The training accuracy curves indicate that the ResNet-50 model rapidly achieves near-perfect performance, surpassing 95% accuracy within the first 20 epochs and stabilizing close to 100% accuracy thereafter. In contrast, the Ours demonstrates a more gradual increase, eventually stabilizing at approximately 90% training accuracy after 50 epochs.

Validation accuracy curves, reflecting the generalization performance, highlight notable differences between the two models. The Ours maintains a consistently higher validation accuracy, stabilizing above 90%, indicating robust generalization and minimal overfitting. Conversely, ResNet-50 exhibits lower validation accuracy, hovering around 85% with greater variability, suggesting significant overfitting to the training data. The disparity between training and validation performance in ResNet-50 highlights its tendency to memorize training patterns due to its complexity, making the simpler, custom-designed neural network preferable for this specific tabular data classification task related to Alzheimer's disease.

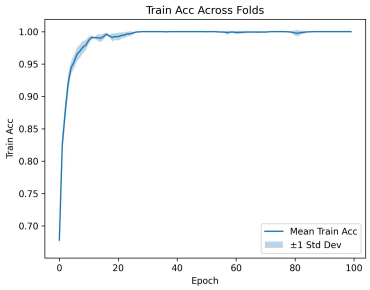
Comparative ROC Curve Analysis of Deep Learning Models: Figures 8 illustrate the mean Receiver Operating Characteristic (ROC) curves for two deep learning models evaluated using 5-fold cross-validation (k=5): a Ours and a ResNet-50 model adapted for tabular data. The Ours achieved a higher mean Area Under the Curve (AUC) of 0.9425, demonstrating strong predictive capability in differentiating Alzheimer's disease diagnoses. In contrast, the ResNet-50 model exhibited a slightly lower mean AUC of 0.9126, indicating comparatively reduced discriminative performance. Furthermore, the narrower variability bands around the ROC curve of the Ours suggest greater stability across folds. Collectively, these results emphasize the superior effectiveness of



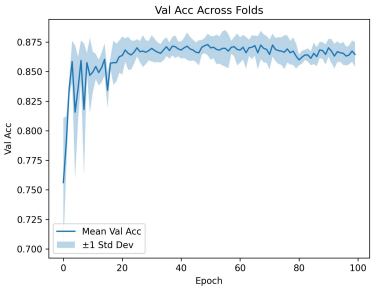
(a) Training Accuracy Curve Ours



(b) Validation Accuracy Curve Ours



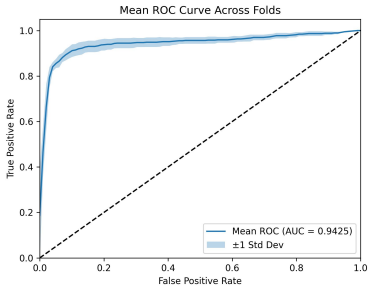
(c) Training Accuracy Curve ResNet-50



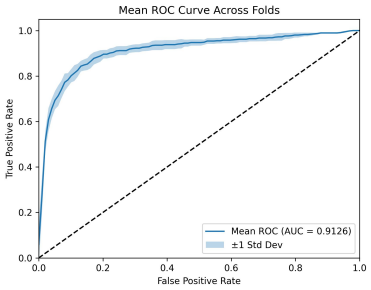
(d) Validation Accuracy Curve ResNet-50

Figure 7: Training and Validation Curves of Ours and ResNet-50

the Ours in accurately identifying patients with Alzheimer's disease from tabular datasets, potentially due to its simpler and more generalized architecture, compared to the complexity and susceptibility to overfitting inherent in ResNet-50. The classification performance of the two deep learning models, evaluated



(a) Ours Methodology



(b) ResNet-50

Figure 8: ROC Curves

through 5-fold cross-validation ($k = 5$), is presented in Table 2. The Ours achieved a mean accuracy of 91.44% (±0.90%), along with a robust mean ROC-AUC score of 0.9425 (±0.0138). Its F1-score, precision, and recall values were 0.9220 (±0.015), 0.9577 (±0.010), and 0.8889 (±0.018) respectively, indicating a balanced performance with high precision but slightly lower recall. Conversely, the ResNet-50 model yielded a lower mean accuracy of 86.46% (±1.04%) and a lower ROC-AUC of 0.9126 (±0.0118), suggesting comparatively weaker generalization. However, it achieved higher recall (0.9739 (±0.012)) and a slightly superior F1-score (0.9551 (±0.014)), although precision was marginally lower (0.9371 (±0.011)). Overall, despite the ResNet-50's ability to identify positive cases more effectively (higher recall), its reduced accuracy and ROC-AUC suggest that the Ours presents better balanced performance for predicting Alzheimer's disease from tabular data.

Table 2: Comparative Performance of Deep Learning Models (K=5 Fold Validation, Mean ± Std)

Model	**Accuracy**	**ROC-AUC**	**F1-Score**	**Precision**	**Recall**
Ours	0.9144 ± 0.0090	0.9425 ± 0.0138	0.9220 ± 0.015	0.9577 ± 0.010	0.8889 ± 0.018
ResNet-50	0.8646 ± 0.0104	0.9126 ± 0.0118	0.9551 ± 0.014	0.9371 ± 0.011	0.9739 ± 0.012

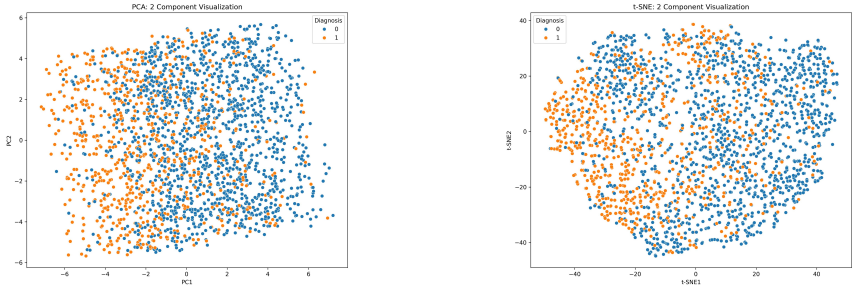
6 Dimensionality Reduction Analysis

Figure 9 displays the two-dimensional visualization of the dataset using PCA (Principal Component Analysis) and t-SNE (t-distributed Stochastic Neighbor Embedding).

The PCA visualization (left) does not exhibit clear separability between the Alzheimer's disease (label 1) and non-Alzheimer's (label 0) classes. The points from both classes appear evenly dispersed throughout the space, suggesting that the linear dimensionality reduction approach of PCA alone is insufficient to distinguish between the two classes effectively.

Similarly, the t-SNE visualization (right) reveals a slightly better but still limited clustering. Although points labeled with Alzheimer's (orange) appear marginally more grouped compared to PCA, the overlap remains significant. This indicates that the Alzheimer's and non-Alzheimer's instances may share similar feature spaces, making class separation challenging even through advanced non-linear dimensionality reduction techniques like t-SNE.

Overall, both dimensionality reduction techniques highlight the complexity and inherent overlap within the dataset, implying that predictive modeling for Alzheimer's disease diagnosis may require more sophisticated feature engineering or advanced machine learning models to achieve higher classification accuracy.



(a) PCA (b) T-SNE

Figure 9: Dimension Reduction Plot

7 Conclusion

This study demonstrates the effectiveness of AI-driven approaches in predicting Alzheimer's disease, with traditional machine learning models, particularly XGBoost, achieving remarkable accuracy and robust generalization. While deep learning approaches such as the custom neural network and ResNet-50 show promising results, the analysis underscores the importance of selecting model complexity suitable for tabular healthcare data to avoid overfitting. Dimensionality reduction results indicate inherent complexities and significant feature overlap, reinforcing the need for sophisticated modeling and thoughtful feature engineering. Future research should explore hybrid modeling strategies and additional biomarkers to further enhance predictive accuracy and clinical utility.

References

[1] R. E. Kharoua, "Alzheimer's disease dataset," 2024. Accessed: 2025-03-25.

A Appendix

A.1 Plots of Different ROC curves for 5 fold validation of our architecture

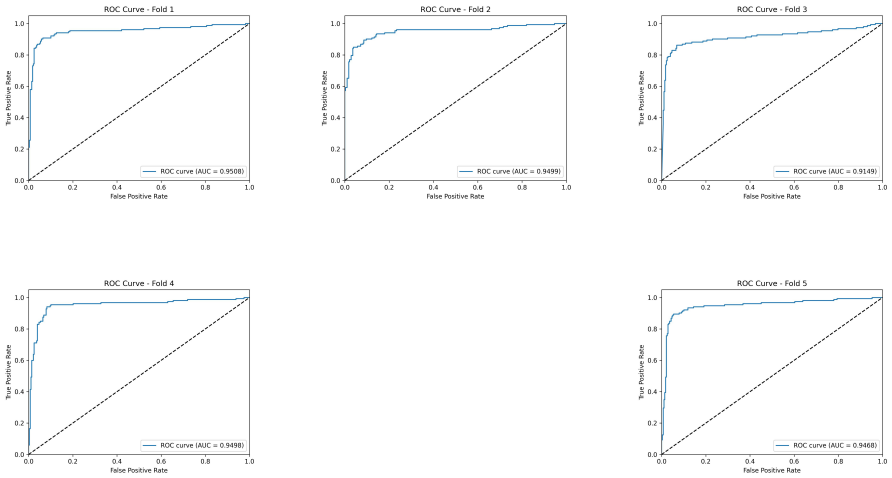


Figure 10: ROC Curves of Our Architecture

A.2 Plots of Different ROC curves for 5 fold validation of ResNet-50 Architecture

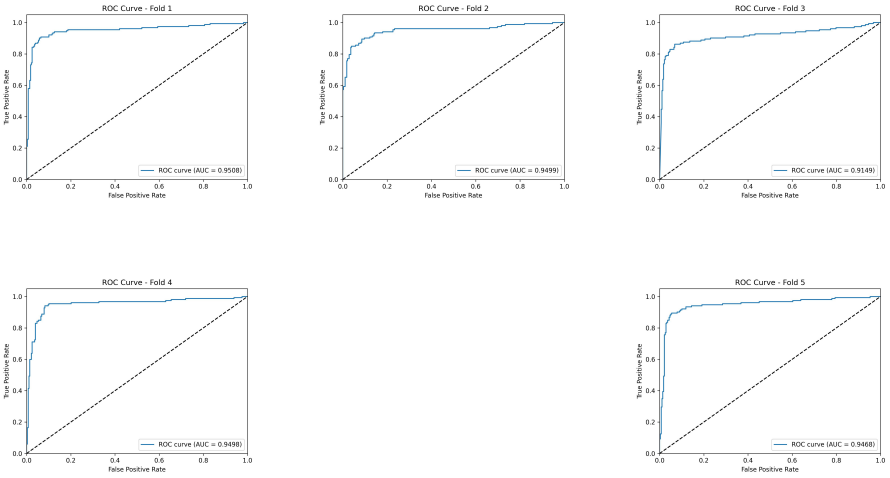


Figure 11: ROC Curves of ResNet-50 Architecture