ITCS444 Data Mining

Faculty of Information and Communication Technology

Project 2: Alzheimer's Disease

Submitted by

6688072	Kemjira	Nugboon
6688099	Nuttanan	Reamprasert
6688129	Patsatraporn	Thongdeesakul
6688168	Pornthita	Sukprapaipat

Presented to

Assoc. Prof. Dr. Suppawong Tuarob

Introduction

Alzheimer’s disease (AD) is a common neurodegenerative disorder affecting millions worldwide, and early risk identification is crucial for effective prevention. Traditional diagnostics like clinical tests and brain imaging are often costly and time-consuming, leading researchers to adopt data mining and machine learning techniques to uncover hidden patterns in patient health and lifestyle data.

This project explores clinical, behavioral, and lifestyle factors related to Alzheimer’s by analyzing a dataset containing demographic information, cognitive scores, and health metrics. Using dimensionality reduction methods such as PCA and t-SNE, the study examines whether patient groups can be visually separated, helping reveal patterns and early indicators associated with Alzheimer’s risk.

Objective

1. Apply PCA and t-SNE to reduce and visualize high-dimensional patient data, enabling clearer interpretation of cognitive, behavioral, and health-related features.
2. Explore hidden structures and patterns within the dataset by examining how key variables such as MMSE, Memory Complaints, Behavioral Problems, and Age affect patient clustering.
3. Identify natural groupings of patients that may reflect Alzheimer’s diagnosis status by using non-linear (t-SNE) and linear (PCA) dimensionality reduction.
4. Analyze key relationships between cognitive scores, demographic factors, and clinical indicators, using correlation analysis and visualization to highlight the most influential features in distinguishing Alzheimer’s from non-Alzheimer’s groups.

Methodology

Dimensionality Reduction Approach

- PCA (Principal Component Analysis)

A linear dimensionality reduction technique used to capture overall variance in the dataset. It provides insight into major variance components and shows broad clustering trends among patients.

- t-SNE (t-Distributed Stochastic Neighbor Embedding)

A non-linear technique designed to preserve local relationships in high-dimensional data. It is especially effective for revealing small, complex clusters related to cognitive and behavioral patterns.

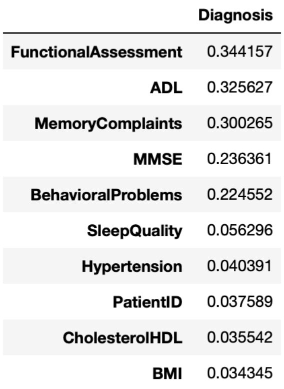
Key Methodological Steps

1. Data Loading and Initial Inspection

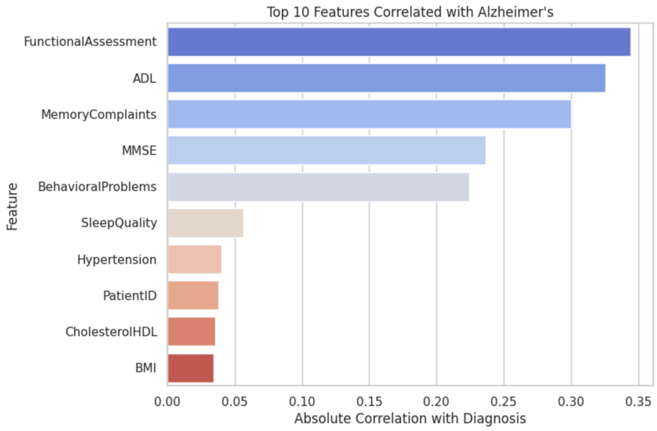
The dataset was imported, explored through summary statistics, and reviewed for missing values, outliers, and structural consistency.

2. Correlation-Based Feature Selection

A correlation matrix identified features most strongly associated with Alzheimer's diagnosis. The top features included MMSE, ADL, FunctionalAssessment, MemoryComplaints, and BehavioralProblems.



	Diagnosis
FunctionalAssessment	0.344157
ADL	0.325627
MemoryComplaints	0.300265
MMSE	0.236361
BehavioralProblems	0.224552
SleepQuality	0.056296
Hypertension	0.040391
PatientID	0.037589
CholesterolHDL	0.035542
BMI	0.034345



3. Handling Missing Values

Missing entries in selected features were replaced using mean imputation to maintain dataset completeness while minimizing bias.

4. Feature and Target Preparation

The feature matrix (X) was constructed from the top correlated variables, while the target vector (y) represented diagnostic categories (Normal vs. Alzheimer's).

5. Standardization

Features were standardized using StandardScaler to ensure equal scaling across variables prior to PCA and t-SNE.

6. Dimensionality Reduction Setup
   - PCA: Applied to evaluate major variance components and produce 2D visualizations of overall structure.
   - t-SNE: Applied to the top five PCA components (perplexity = 30; 2000 iterations) to reveal nonlinear clusters not captured by PCA.
7. Visualization and Interpretation

Both PCA and t-SNE embeddings were color-coded by Diagnosis, Age, MMSE, Memory Complaints, and Behavioral Problems to explore cluster formation and variable influence.

Data Preparation

The dataset underwent several preprocessing steps to ensure analytic validity:

- Feature Selection: Correlation analysis identified MMSE, ADL, FunctionalAssessment, MemoryComplaints, BehavioralProblems, and Age as the most diagnostic-relevant features.
- Missing Values: Mean imputation was applied to maintain consistency.
- Feature Target Setup: X included the selected features, and y consisted of the binary diagnosis label.
- Standardization: All selected features were standardized for uniform variance contribution.
- Dimensionality Reduction: PCA quantified variance distributions, and t-SNE created 2D nonlinear embeddings for deeper structural analysis.
- Visualization: PCA and t-SNE embeddings were visualized and color-coded based on variables such as Diagnosis, Age, MMSE, Memory Complaints, and Behavioral Problems. These visualizations provided insight into how clinical and cognitive features influence patient separation and cluster formation.

Results

1. Feature Preparation and Correlation Analysis

Correlation heatmaps revealed strong relationships among cognitive-functional measures such as MMSE, ADL, and FunctionalAssessment, indicating a shared cognitive decline dimension. MemoryComplaints and BehavioralProblems showed moderate correlations with these cognitive features, suggesting distinct but partially related symptom domains.

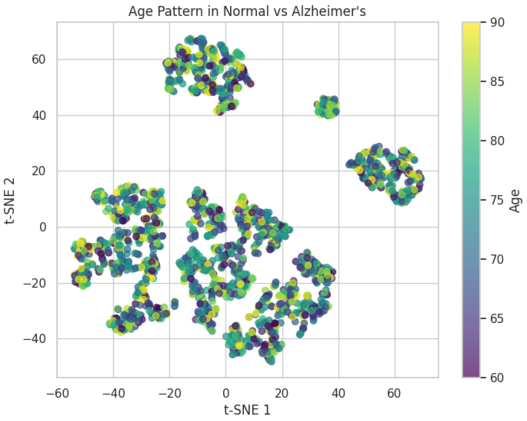
2. PCA Results
   - Explained Variance:

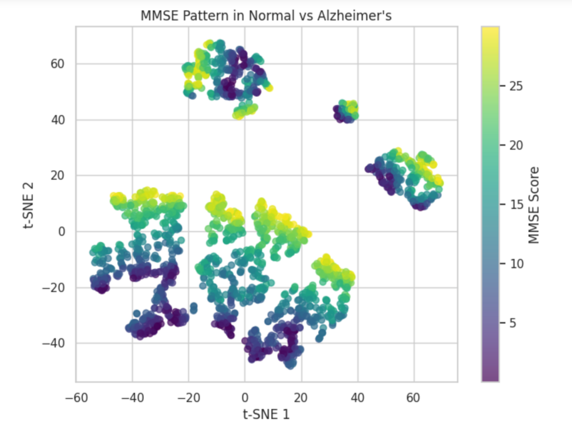
PC1 accounted for the majority of dataset variance, primarily driven by cognitive and functional metrics. PC2 explained a smaller but meaningful portion, while subsequent components showed diminishing returns. Approximately 90% of variance was captured by a relatively small number of components.

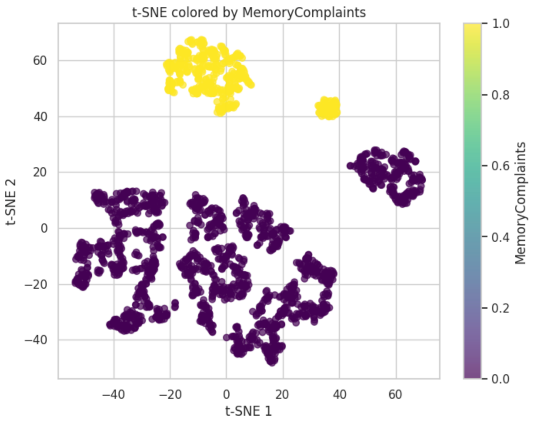
- PCA 2-D Projection:

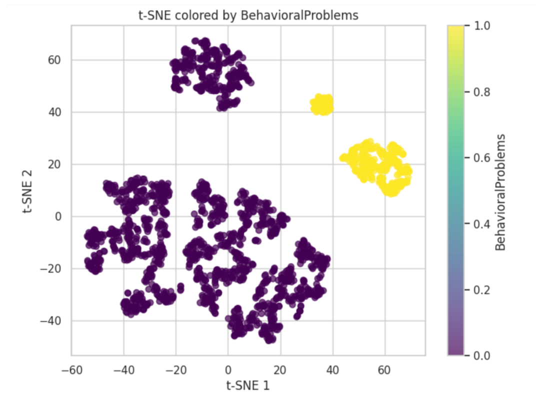
When projected onto PC1 and PC2, Alzheimer's and Normal groups displayed substantial overlap. This indicates that linear variance alone does not align strongly with diagnostic differences, resulting in diffuse and non-separable clusters.

3. t-SNE Results (after PCA reduction to 5 components)
    - The t-SNE embedding produced clearer and more compact patterns than PCA. Alzheimer's cases formed distinct, dense clusters, whereas Normal individuals appeared more dispersed.
    - Variable Overlays:


Age Pattern in Normal vs Alzheimer's

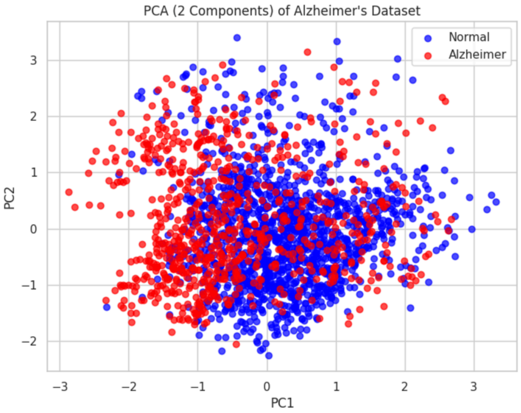

MMSE Pattern in Normal vs Alzheimer's

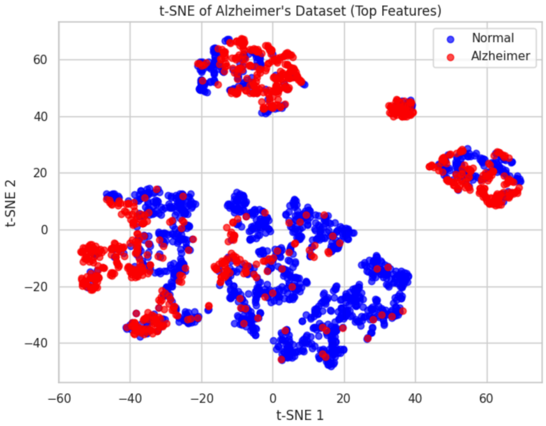

t-SNE colored by MemoryComplaints


t-SNE colored by BehavioralProblems

        - Age: Distributed broadly, showing minimal influence on cluster boundaries.
        - MMSE: Strongly aligned with Alzheimer's clusters; low scores concentrated in distinct regions.
        - Memory Complaints: Formed visibly separated clusters corresponding to Alzheimer's diagnosis.
        - Behavioral Problems: Created well-defined cluster boundaries that closely matched Alzheimer's group patterns.

Interpretation





PCA revealed that Alzheimer’s and Normal groups share substantial linear variability, which limits PCA’s ability to distinguish them. PC1 primarily reflected cognitive-functional decline, strongly influenced by MMSE, ADL, and FunctionalAssessment scores, while PC2 captured secondary behavioral and memory-related variability. The overlap in PCA space indicates that diagnosis-relevant differences are not linearly distributed.

t-SNE, however, uncovered meaningful nonlinear structures that PCA could not reveal. Alzheimer’s patients clustered tightly, suggesting consistent nonlinear interactions among cognitive decline, memory issues, and behavioral symptoms. In contrast, the Normal group exhibited greater variability and dispersion. MMSE, memory complaints, and behavioral problems emerged as the strongest drivers of cluster formation, while age did not significantly influence structural patterns. Overall, t-SNE demonstrated superior capability in capturing the complexity of Alzheimer’s-related symptom interactions.

Conclusion

The analysis shows that PCA captures the overall variance in cognitive and functional features but cannot clearly separate Alzheimer’s from Normal groups due to the nonlinear nature of symptom interactions. Meanwhile, t-SNE produces clearer and more clinically meaningful clusters, driven mainly by low MMSE scores, memory complaints, and behavioral symptoms revealing that Alzheimer’s patterns emerge from complex nonlinear relationships rather than age alone. This suggests that nonlinear dimensionality reduction is more effective for identifying early indicators of cognitive decline.

From a business perspective, these insights support the development of smarter healthcare tools. Organizations such as hospitals, clinics, and digital health companies can use nonlinear analytics to build early-screening systems, risk-monitoring dashboards, and more accurate decision-support tools. This demonstrates how machine-learning–driven analysis can translate into practical solutions that improve early detection and enhance care management.