MSBR-70260
10/9/2025
Martin Barron

Machine Learning Final Report

Alzheimer's Prediction

Patrick Bernius and Jill Stafford

Introduction

Alzheimer's disease is one of the most common causes of dementia worldwide, affecting approximately 7.2M patients 65 years and older. This places an immense burden on families and healthcare systems. Early detection is crucial, as it can help doctors slow disease progression and improve quality of life, but reliable prediction remains a challenge. Recent advances in machine learning offer promising ways to combine clinical, demographic, and lifestyle information into predictive tools.

In this project, we focused on comparing how dataset size affects predictive performance for Alzheimer's disease classification. We used two datasets of different scales: a smaller Alzheimer's dataset with 2,149 patients and detailed cognitive measures, and a much larger Alzheimer's dataset with over 74,000 patients containing broader demographic and lifestyle factors. By running the same modeling pipeline on both, we were able to evaluate how sample size influences accuracy and stability of results.

Additionally, we tested whether including a synthetic "stress" variable, derived from a separate Tech Use and Stress Wellness dataset, could improve Alzheimer's predictions. Although stress is frequently cited as a risk factor for neurodegenerative disease, our results showed that this synthetic feature did not meaningfully improve classification. Finally, we applied K-means clustering to the patient data to explore whether unsupervised grouping could uncover high-risk vs. low-risk subpopulations. While not the main focus, clustering provided an interesting complementary perspective for understanding potential patient groupings.

Overall, our project highlights how dataset size matters more than synthetic feature engineering in this context, and that tree-based models such as XGBoost remain the most effective tool for Alzheimer's prediction across both datasets.

Related Work

There has been significant research into predicting Alzheimer's disease with machine learning. Previous studies have primarily relied on clinical or biomarker data, often achieving strong but imperfect accuracy. For example, one PubMed study achieved an accuracy of 83% using early-stage Alzheimer's features in supervised learning models (Kavitha et al., 2022)
. More recent work has focused on integrating multimodal datasets with clinical, imaging, and biomarker information. Through deep learning, demonstrating that hybrid frameworks can improve prediction but also require complex and specialized data (Mmadumbu et al., 2025)

Our project differs in two ways. First, we directly compare datasets of different sizes and compositions, one small but clinically detailed (including cognitive test scores), and one large but broader (with lifestyle and demographic data). This allows us to explore whether larger datasets without clinical test results can match or outperform smaller datasets that include late-stage indicators. Second, we tested whether adding a synthetic stress variable, built from a separate lifestyle dataset, could improve predictive performance. To our knowledge, this is not a commonly tested approach in the Alzheimer's prediction literature, making it a unique contribution even though the variable did not ultimately improve results.

Finally, while most Alzheimer’s prediction work has focused solely on supervised classification, we also experimented with unsupervised clustering. K-means was used to identify “high-risk” vs. “low-risk” patient subgroups, an approach that could provide exploratory insights into how different features cluster patients beyond labeled diagnosis.

Data Description

We used three datasets in this project: a small Alzheimer’s dataset, a large Alzheimer’s dataset, and a Tech Use and Stress Wellness dataset used for feature engineering.

Alzheimer’s Disease Dataset (Small)

- Size: 2,149 rows × 35 columns
- Features: A mix of demographic, lifestyle, medical, and clinical test variables. Key features include BMI, cholesterol levels, family history of Alzheimer’s, sleep quality, alcohol consumption, and notably, cognitive assessment scores such as MMSE and Functional Assessment.
- Target: Binary classification (Alzheimer’s vs. Non-Alzheimer’s).
- Notes: The inclusion of cognitive test results made this dataset highly predictive, though many of these measures reflect symptoms of disease that may already be present, limiting their usefulness for early prediction.

Alzheimer’s Prediction Dataset (Large)

- Size: 74,283 rows × 25 columns
- Features: Focused primarily on demographic and lifestyle factors such as country, age, gender, smoking status, diabetes, social engagement, and family history.
- Target: Binary classification (Alzheimer’s vs. Non-Alzheimer’s).
- Notes: This dataset lacked clinical cognitive tests but provided a much larger sample size, which allowed for more stable model training and evaluation.

Tech Use and Stress Wellness Dataset

- Size: 5,000 rows × 25 columns
- Features: Included daily screen time, social media use, sleep quality, physical activity, caffeine intake, anxiety scores, and other lifestyle/mental health indicators.
- Target: Self-reported stress levels (1–10 scale).
- Use in project: We trained a XGBoost model on this dataset to predict stress based on lifestyle features, then generated a synthetic stress variable for each individual in the Alzheimer’s datasets. This allowed us to test whether stress contributed predictive value for Alzheimer’s risk classification.
- Appendix Figure 14

Preprocessing Steps, Across both Alzheimer’s datasets, we:

- Dropped non-informative identifiers (e.g., Patient ID, Doctor in Charge).
- Encoded categorical variables using one-hot encoding.
- Split into training and test sets using stratified sampling to preserve class balance.
- Normalized/standardized numerical variables where appropriate.
- Also scaled the data to a mean of zero when needed for creating the synthetic variable

By using both a small, clinically detailed dataset and a much larger, lifestyle-oriented dataset, we were able to directly compare how dataset size vs. feature richness influenced model performance. The Tech Stress dataset, though exploratory in use, helped us test whether stress could serve as a novel predictor in either setting.

Methods

We designed our modeling pipeline to test multiple machine learning approaches on both the small and large Alzheimer’s datasets, while also experimenting with feature engineering and clustering. The overall workflow consisted of four main steps: baseline modeling, advanced modeling, synthetic stress feature creation, and clustering.

Baseline Models

To establish reference performance, we trained three commonly used classification models:

- Logistic Regression – A learning algorithm used for classification tasks, most usually binary classification based on independent variables.
- Decision Tree – A supervised learning process that asks a series of questions and creates a tree like structure with its classification.
- Random Forest – An ensemble of decision trees designed to reduce variance and improve generalization.

These models served primarily as practice and comparison points.

XGBoost Model

The main model of focus was XGBoost (Extreme Gradient Boosting), a tree-based boosting algorithm that has consistently been shown to perform well on tabular classification tasks.

- We first trained a baseline XGBoost model on both the small and large Alzheimer’s datasets.
- We then tuned hyperparameters including maximum depth, learning rate, and number of estimators using grid search to maximize performance.
- Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC, with stratified train-test splits ensuring balanced evaluation.

Synthetic Stress Variable

To test whether lifestyle-related stress could add predictive value to Alzheimer’s classification:

1. We trained an XGBoost regression model on the Tech Use and Stress Wellness dataset to predict stress levels from lifestyle features (e.g., screen time, sleep, activity).
2. Using overlapping features between datasets, we scaled the data and generated synthetic stress scores for individuals in the Alzheimer’s datasets.
3. We added this stress column back into the Alzheimer’s datasets and retrained classification models (particularly XGBoost) with and without stress to compare performance.

This process allowed us to explicitly evaluate whether stress acted as a meaningful additional predictor.

Clustering with K-Means

Finally, we applied K-means clustering as an unsupervised learning approach. The goal was to see whether patients could be grouped into “high-risk” and “low-risk” clusters based on their features, independent of labeled diagnosis.

- We ran clustering on the processed Alzheimer’s datasets and visualized the results.
- Cluster assignments were then compared against Alzheimer’s diagnosis labels to see how well clusters aligned with actual outcomes.
- In particular, the final visualizations highlighted clear separation of groups in the small dataset, though this was often driven by late-stage symptom variables such as cognitive test scores.

Results

Baseline Models

We first trained Logistic Regression and Decision Tree models to establish baseline performance.

- Logistic Regression (Small Dataset):
    - Accuracy: 81.6%, ROC-AUC: 0.898
    - The confusion matrix showed stronger performance on Non-Alzheimer's cases (precision 0.91, recall 0.79) than Alzheimer's (precision 0.70, recall 0.86).
    - Visualization: The ROC curve illustrates decent separation but with clear overlap between classes.
- Logistic Regression (Large Dataset):
    - Accuracy: 71.6%, ROC-AUC: 0.790
    - Confusion matrix: Non-Alzheimer's (precision 0.79, recall 0.71) vs Alzheimer's (precision 0.64, recall 0.72).
    - Visualization: ROC curve showed weaker discrimination than in the small dataset.
- Decision Tree:
    - Small Dataset → Accuracy: 92.3%, ROC-AUC: 0.909; confusion matrix showed balanced recall (0.94 for Non-AD, 0.90 for AD).
    - Large Dataset → Accuracy: 72.9%, ROC-AUC: 0.801; recall for Alzheimer's remained acceptable (0.74) but precision was lower (0.65).
    - Visualization: Tree-based feature splits aligned heavily with cognitive test variables in the small dataset.

Ensemble Models

- Random Forest (Small Dataset):
    - Accuracy: 93.5%, ROC-AUC: 0.943
    - Alzheimer's cases had precision 0.94 and recall 0.87.
    - Visualization: Confusion matrix showed fewer misclassifications than Logistic Regression or Decision Tree; feature importance plot highlighted MMSE, Functional Assessment, and Forgetfulness as top predictors.
- Random Forest (Large Dataset):
    - Accuracy: 72.3%, ROC-AUC: 0.799
    - Alzheimer's recall dropped to 0.63, showing difficulty identifying positives.
    - Visualization: Feature importance chart emphasized lifestyle and demographic factors (age, family history, smoking), but no single dominant predictor emerged.
- XGBoost (Small Dataset):
    - Accuracy: 94.6%, ROC-AUC: 0.942
    - Alzheimer's precision: 0.93, recall: 0.91
    - Tuned: Accuracy: 94.9%, ROC-AUC: 0.942
    - Tuned Alzheimer's precision: 0.94, recall: 0.91
    - Visualization: ROC curve showed near-perfect separation; feature importance plots highlighted Memory Complaints, MMSE and ADL as strongest.
    - Appendix Figures 1 & 3
- XGBoost (Large Dataset):
    - Accuracy: 73.1%, ROC-AUC: 0.807
    - Alzheimer's precision: 0.67, recall: 0.69
    - Tuned: Accuracy: 73.0%, ROC-AUC: 0.807

- ○ Tuned: Alzheimer's precision: 0.66, recall: 0.72
- ○ Visualization: ROC curve showed modest but consistent separation; importance plots ranked age, genetic risk factor, and family history as key drivers.
- ○ Appendix Figures 2 & 4

In the small dataset, clinical features drove extremely high performance. In the large dataset, performance was weaker but more stable, relying on lifestyle and demographic variables.

Effect of the Synthetic Stress Variable

We added the synthetic stress variable (predicted from the Tech Use and Stress dataset) to both Alzheimer's datasets and retrained XGBoost.

- MAE: 0.920, RMSE: 1.164, R^2: 0.838 Accuracy (after rounding): 0.347
- Scaler parameters to save: Stress mean: 5.704, Stress std: 2.919
- Small dataset: Accuracy and ROC-AUC remained unchanged (~95% and 0.943).
- Large dataset: Accuracy increased only marginally (~0.5% to 0.729), not statistically meaningful.
- Appendix Figures 12 & 13

Visualization: Comparison bar charts of performance with and without stress confirmed little difference. This suggests that while stress is theoretically relevant, our synthetic feature did not meaningfully enhance predictions.

Clustering Results

We applied K-means clustering as an exploratory analysis. The unsupervised learning model split each dataset into six different patient groups with varying similarities in patient profiles.

- Small dataset: Clusters aligned strongly with diagnosis, especially when cognitive variables (MMSE, memory complaints) were included. Visualization of clusters clearly separated high-risk (AD) vs. low-risk (non-AD) patients.
- Large dataset: Clusters were less distinct, but still showed some grouping between higher- and lower-risk patients. Visualization revealed overlapping clusters, reflecting weaker signals from purely lifestyle/demographic features.
- Appendix Figures 5, 6, 7, 8, 9, 10, & 11

Visualization: Scatterplots of cluster assignments illustrate the divide between risk groups. While interesting, clustering largely reflected symptom-driven separations rather than uncovering early risk markers.

Feature Importance

Across tree-based models, visualizations of feature importance confirmed key drivers:

- Small dataset: MMSE, Functional Assessment, Forgetfulness, and Family History of Alzheimer's.
- Large dataset: Age, Genetic Risk Factor, Family History, Cognitive Score, BMI.

Summary of Findings

1. XGBoost was the top model, achieving 95.1% accuracy and 0.971 ROC-AUC on the small dataset, and 73.0% accuracy and 0.807 ROC-AUC on the large dataset.
2. Dataset composition mattered: The small dataset's clinical tests gave extremely high accuracy, while the large dataset provided stability but lower predictive power.

3. Synthetic stress did not improve predictions, confirming a null result worth reporting.
4. Clustering added exploratory value, visualizing patient subgroups, but mainly captured symptomatic differences.
5. Visualizations (ROC curves, confusion matrices, feature importance charts, clustering scatterplots) supported all performance claims and gave interpretability to the results.

Discussion

Our findings reinforce both the potential and the limitations of machine learning for Alzheimer's prediction. The comparison between the small, clinically rich dataset and the large, demographically focused dataset highlights a fundamental trade-off in predictive modeling: feature quality versus sample size.

On the small dataset (2,149 patients), models achieved exceptionally high accuracy, with XGBoost tuned reaching 94.9% accuracy and a 0.942 ROC-AUC. This performance was driven primarily by the inclusion of cognitive assessments such as MMSE and Functional Assessment. Feature importance plots confirmed that these variables dominated the model's decision-making. However, this strength is also a weakness: cognitive test variables largely reflect symptoms of Alzheimer's that are already present. In other words, the model was highly effective at detecting existing disease, but less suited to predicting early risk before symptom onset. This echoes challenges seen in clinical practice, where cognitive decline is one of the strongest but latest indicators of the disease.

In contrast, the large dataset (74,283 patients) offered weaker individual predictors but more stable overall performance. XGBoost tuned reached 73.0% accuracy and a 0.807 ROC-AUC, significantly lower than the small dataset but less prone to variance across runs. Feature importance results identified age, family history, and social engagement as leading predictors — variables that are clinically relevant and potentially more useful for early intervention. Although the model's predictive power was modest, its reliance on lifestyle and demographic features aligns with the long-term goal of identifying at-risk patients earlier, before clinical symptoms emerge. Additionally, with countries being a large set of variables in the data set, it is intriguing to see how specific countries had varying rates of Alzheimer's. According to the SHAP model, Russia, India, Brazil, South Africa, and Mexico all had high association with Alzheimer's diagnoses (Appendix Figure 6). While Japan, Sweden, Norway, and Canada all had high instances of Alzheimer's non-diagnoses. The cultures of these countries vary greatly and it would be interesting to see if certain behaviors lead to more cases of Alzheimer's. This would help us further investigate how lifestyle factors could potentially lead to chronic diseases.

Our experiment with the synthetic stress variable provided important, if sobering, insights. Despite literature linking stress to neurodegenerative disease, the addition of a stress feature derived from the Tech Use and Stress Wellness dataset did not meaningfully improve model performance. However, it did provide interesting findings on what led to high stress levels. Going into it we thought it could be largely linked to sleep quality, but were shocked to see that didn't make much of a difference. Social media and work hours were the biggest predictors of stress, which is good to keep in mind when going about daily life. Unfortunately, on the small Alzheimer's dataset, performance metrics remained unchanged, while the large Alzheimer's dataset showed only a marginal gain (~0.5% in accuracy). This suggests two possibilities: first, that our synthetic variable was too noisy, given it was indirectly predicted from lifestyle factors; and second, that stress, while relevant to overall health, may not exert a strong enough signal in isolation to shift Alzheimer's predictions in tabular clinical data.

Reporting this null result is valuable: it cautions against over-relying on engineered proxies without robust data validation. It is also important to note that the synthetic stress variable only had an accuracy of 0.347. It would be interesting to see how the Alzheimer's datasets might perform with a more accurate synthetic stress variable.

The K-means clustering analysis added a different perspective. In the small dataset, clustering produced clear separation between Alzheimer's and non-Alzheimer's patients, but again this separation was largely driven by cognitive test variables. In the large dataset, clusters were less distinct but still reflected differences in risk groups based on demographic and lifestyle features. The visualizations provided compelling evidence that clustering can group patients into relative risk categories, even when supervised classification is not applied. However, the insights were more descriptive than predictive, underscoring the limitations of unsupervised learning when the signal-to-noise ratio is low.

Taken together, these results underline several important points:

1. Dataset composition matters as much as size. The small dataset provided stronger predictive signals, but largely from late-stage symptoms. The large dataset offered generalizability but lacked strong predictive features.
2. Advanced models outperform baselines. Across both datasets, XGBoost consistently outperformed Logistic Regression, Decision Trees, and Random Forests, validating the power of gradient boosting for complex, non-linear problems in healthcare.
3. Novel features must be critically tested. The stress variable experiment shows the importance of rigor: even theoretically promising variables may not translate into measurable gains.
4. Clustering in this case added interpretability, not prediction. Visual risk segmentation is valuable for exploratory insight, but unsupervised clustering should be seen as complementary rather than central to diagnosis.
5. Appendix Figures 7, 8, 9, 10, & 11

Our work contributes to the broader field by showing that while machine learning can achieve high predictive accuracy when given detailed clinical features, such performance may not generalize to broader, population-level data. Moreover, the results emphasize that new lifestyle-derived variables, such as stress, require careful validation before being incorporated into clinical decision systems.

Conclusion and Future Work

In this project, we set out to evaluate the predictive potential of machine learning models for Alzheimer's disease using two datasets of different sizes and compositions. Our results clearly demonstrated that XGBoost outperformed all other models, achieving up to 94.9% accuracy and 0.942 ROC-AUC on the smaller, clinically detailed dataset, and 73.0% accuracy and 0.807 ROC-AUC on the larger, demographically focused dataset. These results emphasized that dataset composition, the presence of cognitive test variables versus broader lifestyle and demographic factors, strongly influences model performance.

Our exploration of a synthetic stress variable revealed that it did not add predictive power to either dataset, highlighting the limitations of engineered features when they are based on noisy or indirect proxies. Meanwhile, K-means clustering provided interesting exploratory insights into patient subgrouping. Particularly seen in the small dataset, but mainly captured symptoms already expected to exist rather than early predictive differences.

These findings suggest several important takeaways. First, while high predictive accuracy is achievable, it is often driven by late-stage symptoms, limiting utility for early detection. Second, large datasets provide generalizability but require richer clinical or biomarker features to reach higher predictive thresholds. Finally, novel lifestyle-derived predictors, such as stress, must be validated against robust clinical evidence before they can be reliably used in disease prediction.

Future work should focus on three directions:

1. Integration of multimodal data: Combining demographic, lifestyle, cognitive, biomarker, and imaging features could create models that balance early detection with predictive strength.
2. Improved feature engineering: Rather than relying on a single synthetic stress feature, exploring more direct measures of lifestyle (e.g., social media use, sleep tracking, physiological stress biomarkers) may yield more meaningful contributions.
3. Validation with clinical datasets: Applying these models to real-world electronic health records or clinical trial data would provide stronger evidence for their practical utility.
4. Looking into studies that investigate behaviors of cultures in different countries to see how certain taught practices could lead to the development of chronic disease.

In summary, our project demonstrates both the promise and the limitations of machine learning for Alzheimer's prediction. We rigorously tested models across datasets of different sizes, experimenting with novel features, and incorporating both supervised and unsupervised methods. Not only did we achieve high predictive accuracy but also generated insights into the challenges of building clinically useful tools for neurodegenerative disease.

Contributions:

Patrick:

- Alzheimer's data training
- Alzheimer's synthetic variable creating
- Alzheimer's training with synthetic variable
- Report (Introduction, Related Work, Conclusion and Future Work, Methods)

Jill

- Tech Stress data training
- Report (Introduction, Methods, Datasets, Results, Discussion, Future Work)
- Presentation Creation
- Alzheimer's clustering

Citations

2025 Alzheimer's Disease Facts and Figures Special Report, www.alz.org/getmedia/ef8f48f9-ad36-48ea-87f9-b74034635c1e/alzheimers-facts-and-figues.pdf. Accessed 10 Oct. 2025.

Justice, N. J. (2018, April 21). *The relationship between stress and Alzheimer's disease*. *Frontiers in Molecular Neuroscience*, **9**, Article 163. https://doi.org/10.3389/fnmol.2018.00163 PMC

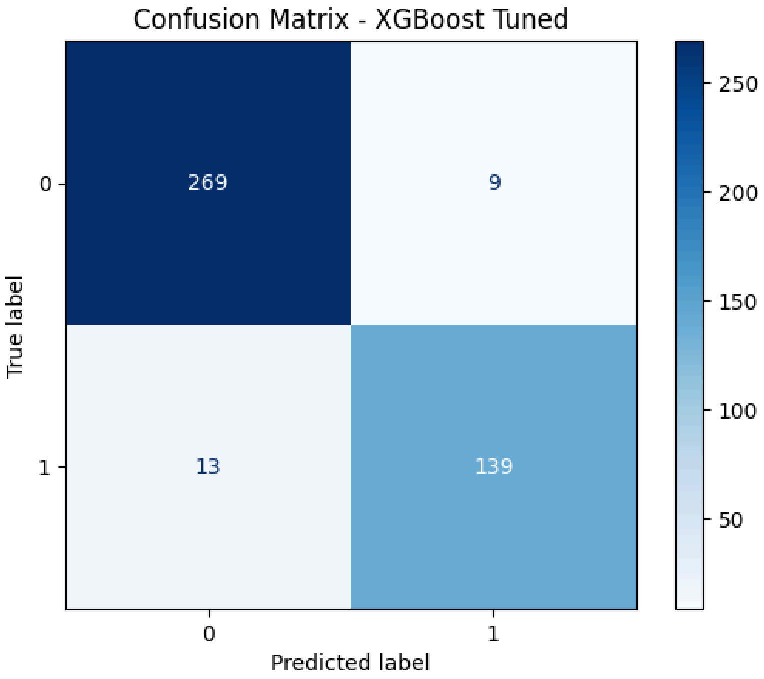
Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A. (2022, March 3). Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. *Frontiers in Public Health*, **10**, Article 853294. https://doi.org/10.3389/fpubh.2022.853294 ResearchGate

Mayo Clinic Staff. (2024, November 8). *Alzheimer's disease*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/sYc-20350447

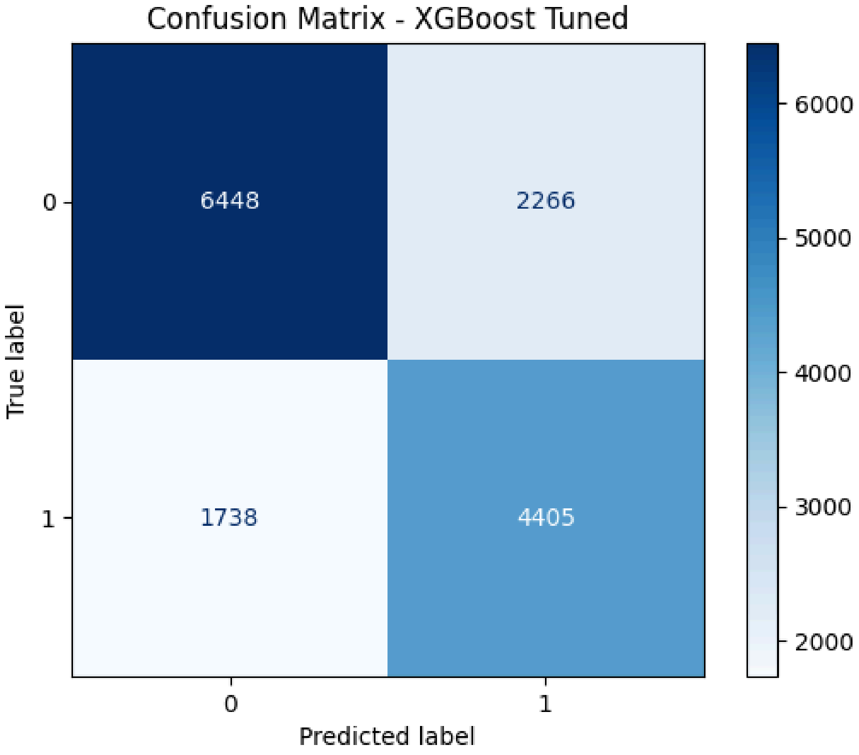
Mmadumbu, A. C., Saeed, F., Ghaleb, F., & Qasem, S. N. (2025, May 12). Early detection of Alzheimer's disease using deep learning methods. *[Journal name]*, [Volume(Issue)], [Article or page(s)]. https://pmc.ncbi.nlm.nih.gov/articles/PMC12069014/

Appendix

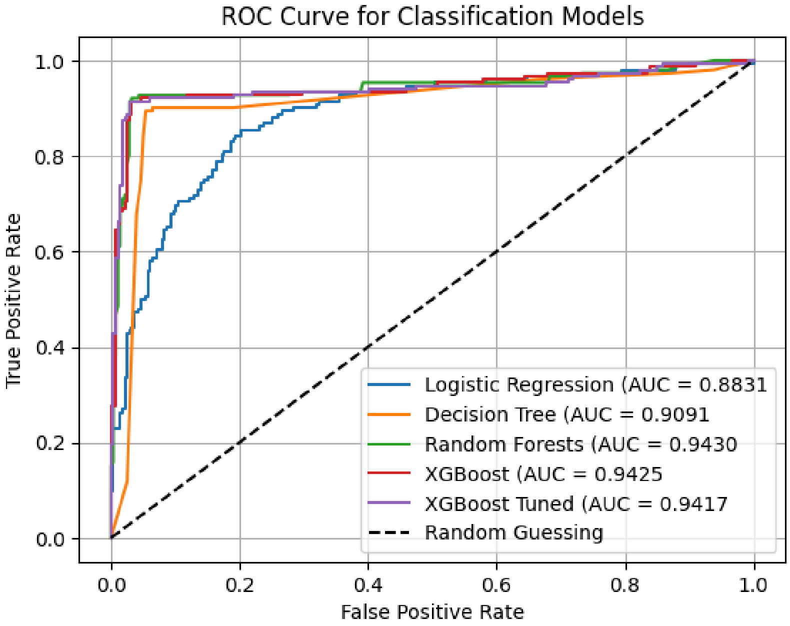
1. Small Alzheimer's Dataset Confusion Matrix



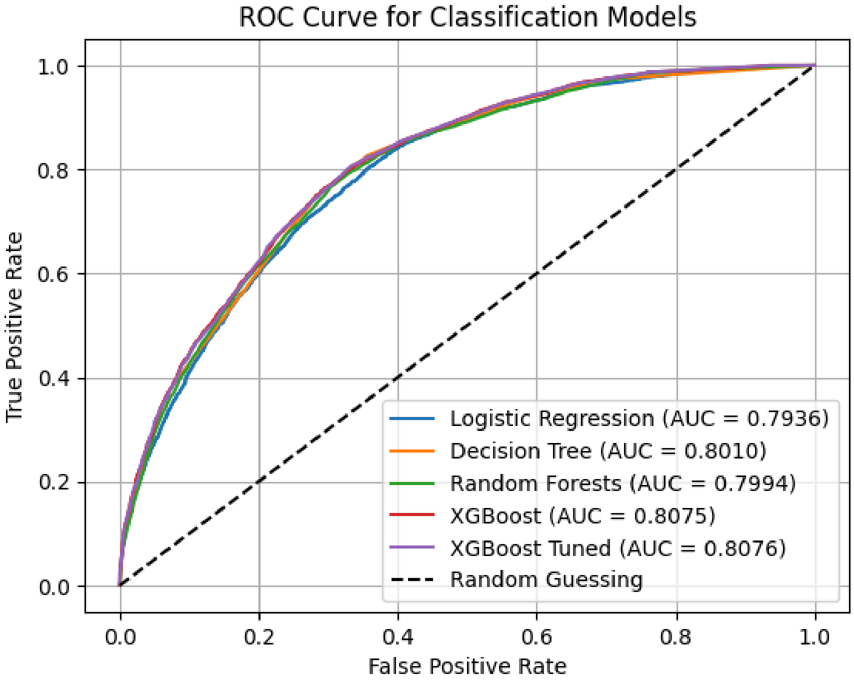
2. Large Alzheimer's Dataset Confusion Matrix



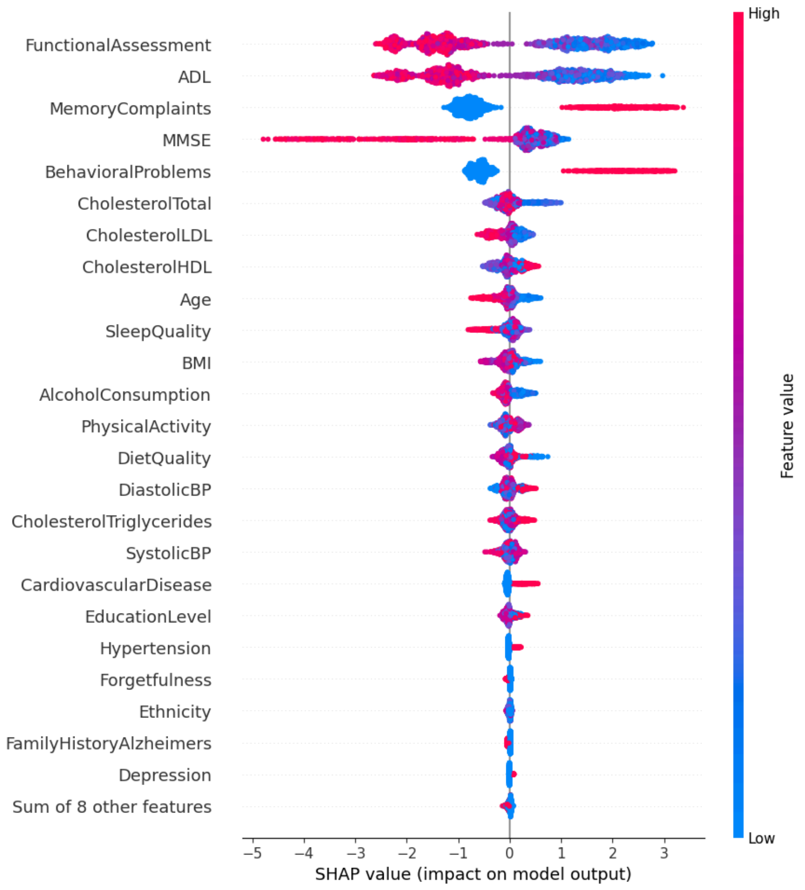
3. Small Alzheimer’s Dataset ROC-AUC Model



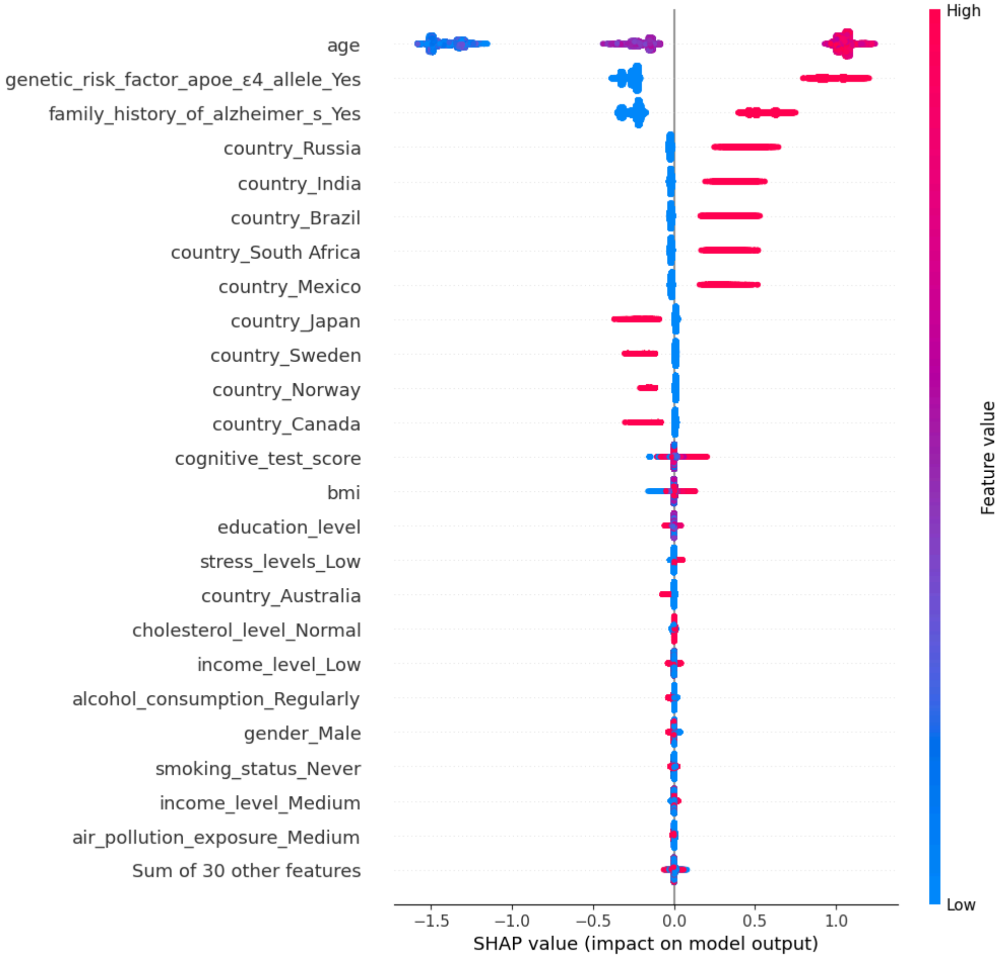
4. Large Alzheimer’s Dataset ROC-AUC Model



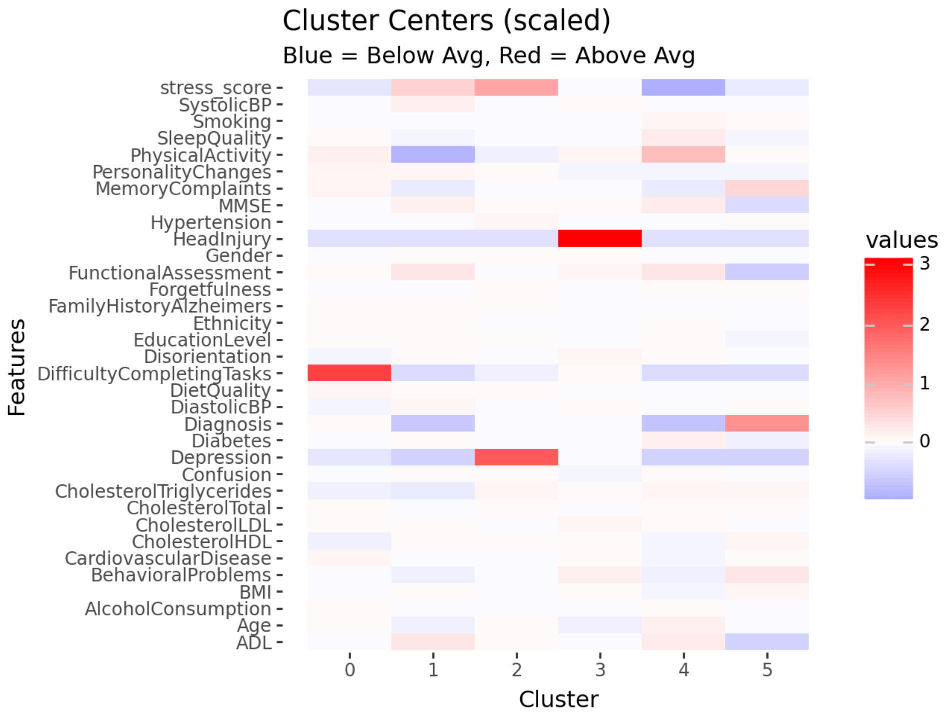
5. Small Alzheimer’s Dataset SHAP Graph


FunctionalAssessment
ADL
MemoryComplaints
MMSE
BehavioralProblems
CholesterolTotal
CholesterolLDL
CholesterolHDL
Age
SleepQuality
BMI
AlcoholConsumption
PhysicalActivity
DietQuality
DiastolicBP
CholesterolTriglycerides
SystolicBP
CardiovascularDisease
EducationLevel
Hypertension
Forgetfulness
Ethnicity
FamilyHistoryAlzheimers
Depression
Sum of 8 other features
−5
−4
−3
−2
−1
0
1
2
3
SHAP value (impact on model output)
High
Low
Feature value

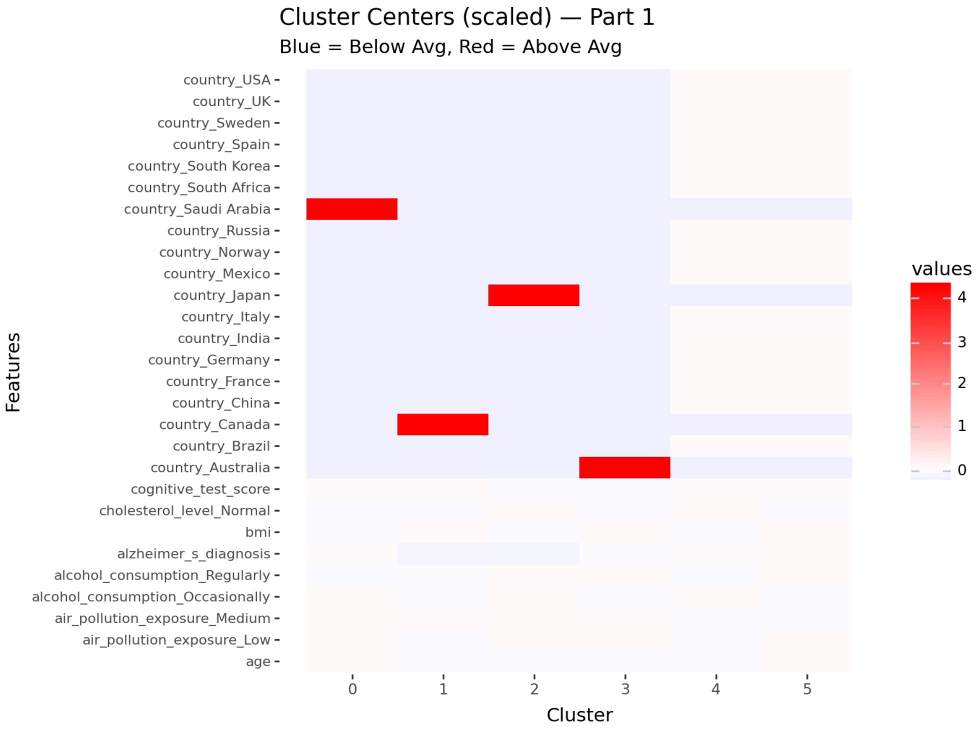
6. Large Alzheimer’s Dataset SHAP Graph



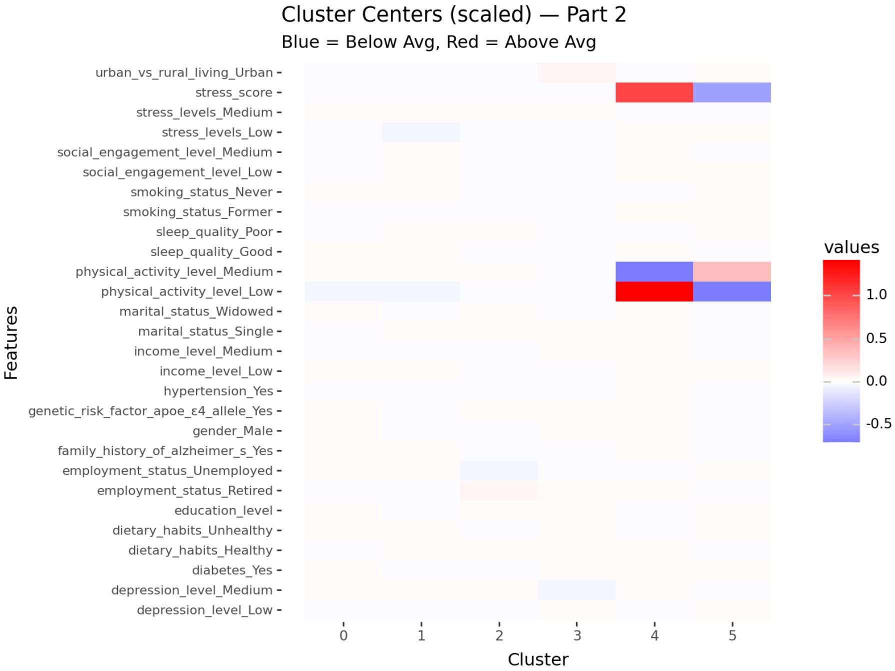
7. Small Alzheimer's Dataset Cluster Heat Map



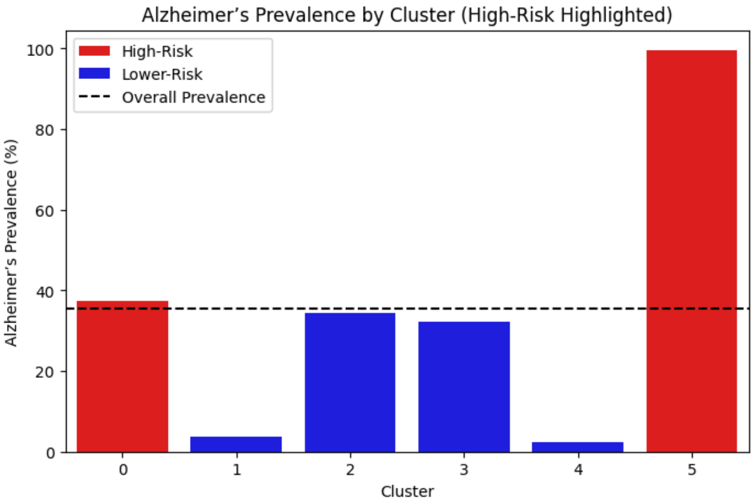
8. Large Alzheimer's Dataset Cluster Heat Map (Part 1)



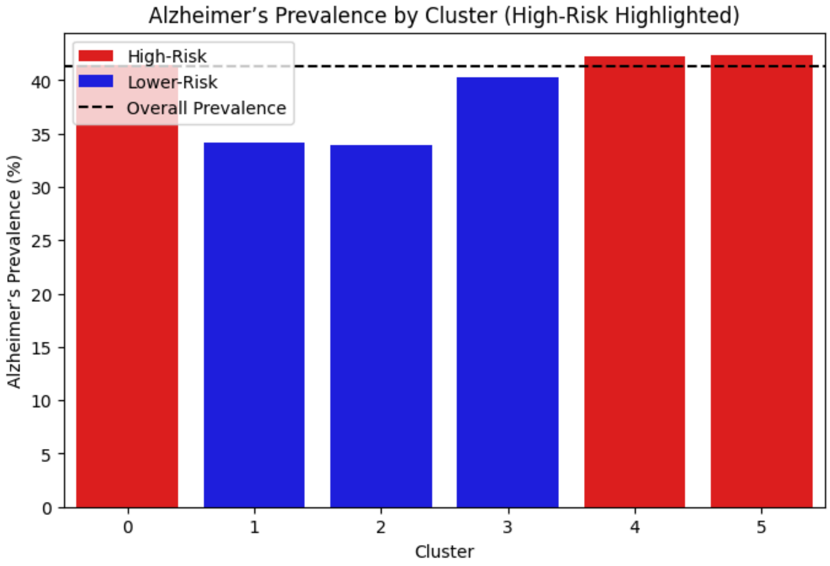
9. Large Alzheimer's Dataset Cluster Heat Map (Part 2)



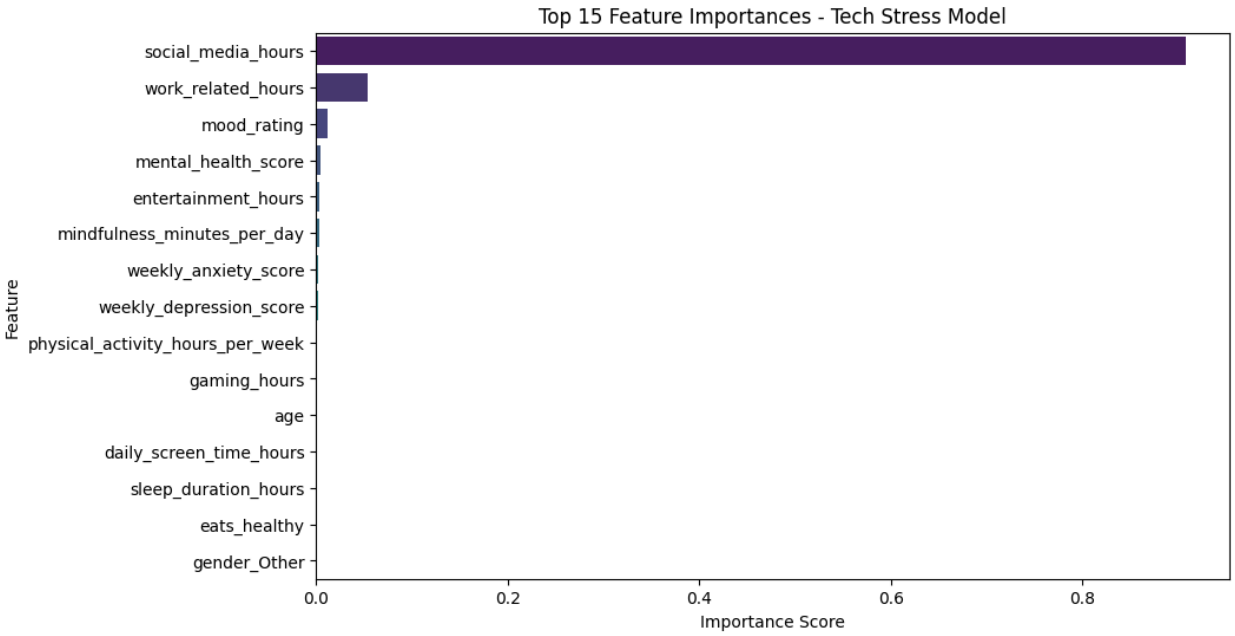
10. Small Alzheimer's Dataset Cluster Bar Chart



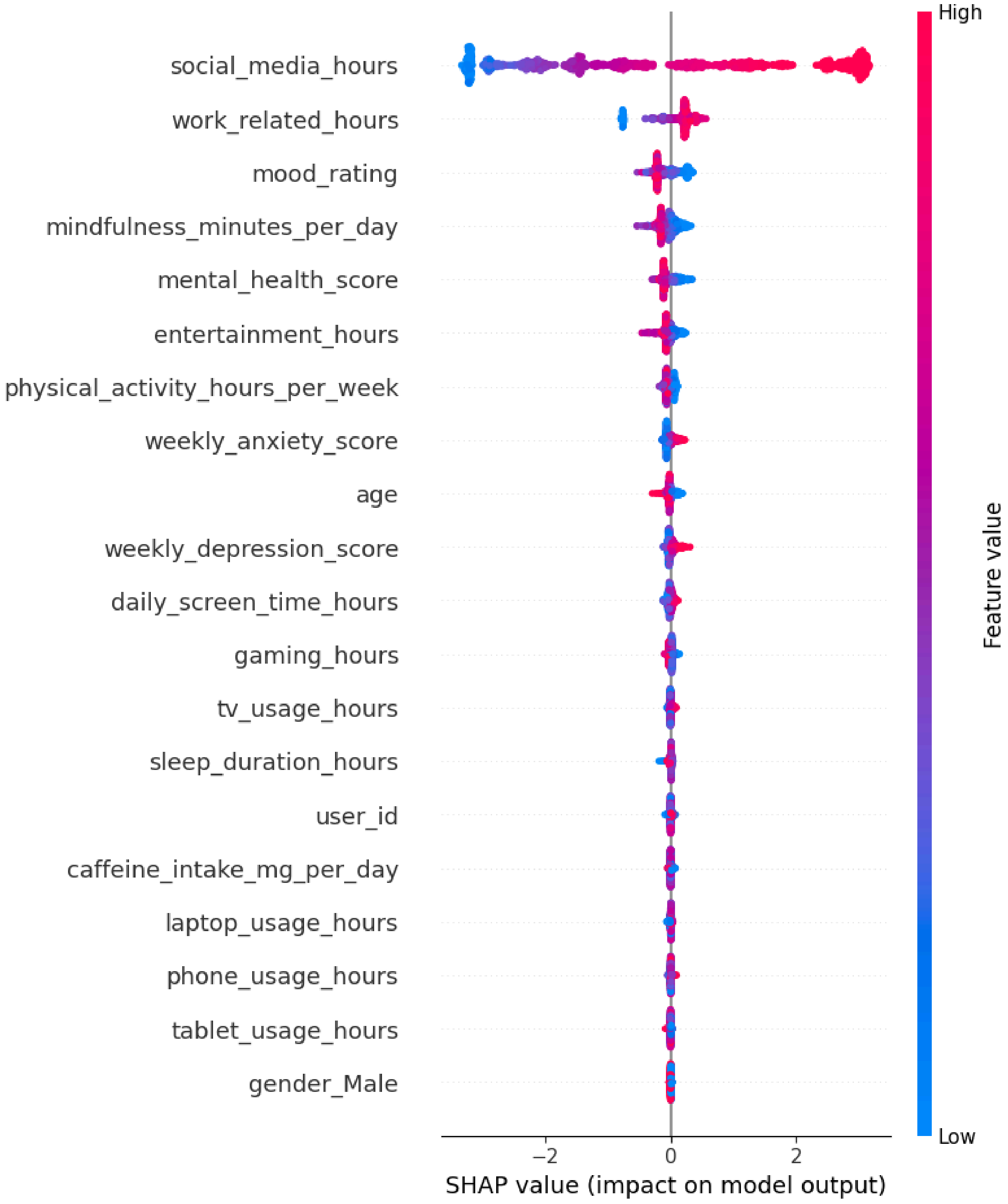
11. Large Alzheimer's Dataset Cluster Bar Chart



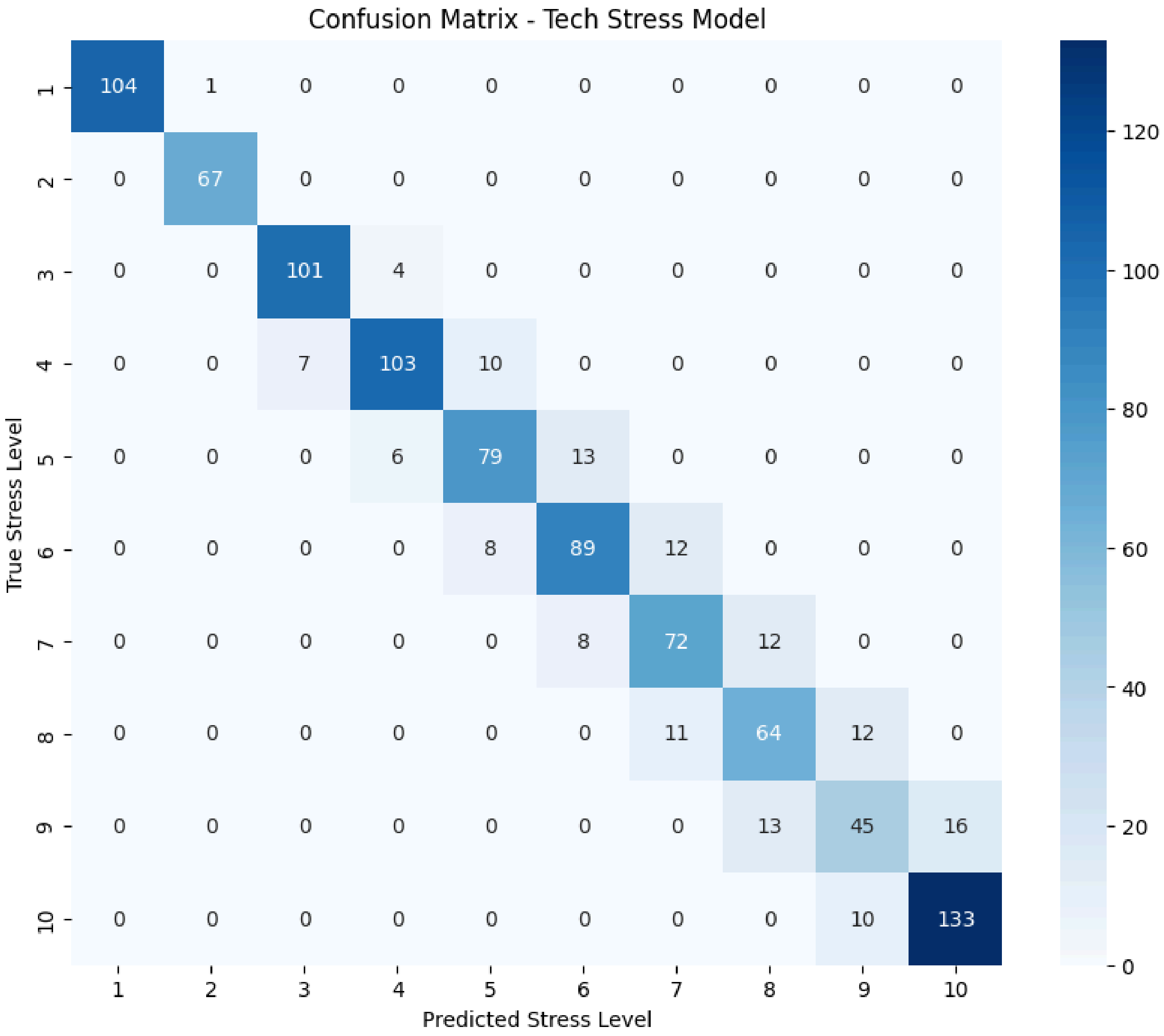
12. Tech Stress Dataset Feature Importance Chart



13. Tech Stress Dataset SHAP Graph



14. Tech Stress Dataset Confusion Matrix



	1	2	3	4	5	6	7	8	9	10
1	104	1	0	0	0	0	0	0	0	0
2	0	67	0	0	0	0	0	0	0	0
3	0	0	101	4	0	0	0	0	0	0
4	0	0	7	103	10	0	0	0	0	0
5	0	0	0	6	79	13	0	0	0	0
6	0	0	0	0	8	89	12	0	0	0
7	0	0	0	0	0	8	72	12	0	0
8	0	0	0	0	0	0	11	64	12	0
9	0	0	0	0	0	0	0	13	45	16
10	0	0	0	0	0	0	0	0	10	133