Enhancing Early Detection of Alzheimer's Disease Using Predictive Analytics

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

2425: BIS POSTGRADUATE PROJECT

MSc. Business Analytics

Submission Date: 31st of May 2025

Group Name: Brain Guardians

Submitted by:

Joseph Cherupushpam Antony 24243533

Aditya Asthana 24243559

Arshaque Muhammed 24243442

Eliz George 24243469

Executive Summary

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that affects millions globally, placing immense strain on families, healthcare systems, and public resources (World Health Organization, 2023). Early detection is crucial in improving patient outcomes, enabling timely intervention, and reducing the societal burden of late-stage cognitive decline (Albert et al., 2011). However, diagnosis often occurs too late, due to the complexity of symptoms, limited access to specialists, and reliance on invasive or costly diagnostic procedures.

This industry report explores how predictive analytics and machine learning (ML) can assist in identifying early signs of Alzheimer's using structured clinical data. The project leverages the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Mueller et al., 2005), which contains rich cognitive and neurological assessments, to build and evaluate ML models for classification into three diagnostic stages: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Dementia.

The dataset preparation involved cleaning and standardizing a large merged file containing over 700,000 records. Missing values were handled based on null thresholds, and the feature space was enriched with demographic and cognitive markers such as MOCA scores and Age. MOCA, or the Montreal Cognitive Assessment, is a globally accepted screening tool for early cognitive decline (Nasreddine et al., 2005). Due to the large dataset size (~600MB+), the analysis was conducted in Google Colab, integrated with Google Drive for scalability. While traditional tools like Pandas were used for manipulation, PySpark was also explored as a potential solution for larger workflows (Zaharia et al., 2016).

Following the CRISP-DM framework (Chapman et al., 2000), we performed an exploratory data analysis (EDA), uncovering patterns consistent with clinical research, for example, the correlation between MOCA scores and Alzheimer's progression (Cruz et al., 2017). Three ML models were trained and compared: Logistic Regression, Random Forest, and XGBoost. Among these, XGBoost demonstrated the best trade-off between accuracy and recall, particularly excelling at identifying MCI cases a pivotal stage for early intervention.

This report presents the cleaned dataset, EDA findings, modelling results, and a visual roadmap for future research and deployment. By emphasizing non-invasive, interpretable, and scalable features, this work demonstrates the growing role of business analytics in advancing early-stage dementia detection, with practical implications for clinical tools, public health planning, and digital screening platforms.

The business relevance of this approach extends beyond research: digital health platforms, assisted living providers, and memory clinics could integrate such predictive tools into routine patient checkups. For example, elderly care platforms may use Age and MOCA screening scores to flag high-risk patients and alert primary care providers in real-time. This reduces the burden on specialists and speeds up intervention timelines (Topol, 2019). Similarly, regional health authorities can leverage aggregated insights to plan Alzheimer's-specific resources and outreach programs in vulnerable populations.

Methodologically, the models were evaluated using multiple metrics including precision, recall, F1-score, and confusion matrices with class balancing techniques (e.g., stratified sampling) applied to address diagnosis imbalance (López et al., 2013). This ensured robust performance across all three classes, not just the dominant CN and MCI categories. The adoption of interpretable models like Random Forest and XGBoost further enables clinicians and data analysts to understand feature impact, which is critical for adoption in regulated healthcare settings (Rudin, 2019).

Overall, the project showcases a scalable, non-invasive, and analytically sound pipeline for early Alzheimer's detection bridging the gap between academic data science and actionable digital healthcare solutions.

Background

Alzheimer's Disease (AD) is the most common cause of dementia, accounting for up to 70% of all dementia cases globally (WHO, 2023). It is a progressive, irreversible brain disorder that gradually destroys memory, cognitive function, and the ability to carry out daily activities. Despite decades of research, early detection remains one of the most challenging aspects of Alzheimer's management due to overlapping symptoms with normal aging and other neurodegenerative conditions (Brookmeyer et al., 2007).

The burden of Alzheimer's is increasing rapidly. According to estimates, more than 150 million people are expected to live with some form of dementia by 2050, with the majority residing in low- and middle-income countries (GBD 2019 Dementia Forecasting Collaborators, 2022). Current diagnosis methods often rely on neuroimaging, cerebrospinal fluid (CSF) biomarkers, or extensive cognitive testing all of which are either expensive, invasive, or not widely available (Hampel et al., 2018).

This is where data analytics can play a transformative role. By leveraging machine learning (ML) and predictive models, it becomes possible to identify early warning signals from non-invasive, widely available data such as clinical assessments, age, and cognitive screening scores. ML has already shown promise in healthcare applications such as cancer detection, hospital readmission prediction, and mental health screening (Rajkomar et al., 2019). Alzheimer's detection presents a natural extension for such tools.

The Montreal Cognitive Assessment (**MOCA**) has emerged as one of the most widely used screening instruments for mild cognitive impairment (MCI) a critical transitional state between normal aging and Alzheimer's Disease. Research has shown that MOCA is more sensitive than the Mini-Mental State Examination (MMSE) for detecting MCI and early-stage dementia (Nasreddine et al., 2005). Additionally, age is a known risk factor, with the prevalence of AD increasing exponentially beyond the age of 65 (Brookmeyer et al., 2018).

In this context, our project explores whether these accessible variables MOCA and Age can be used to build lightweight, interpretable machine learning models that classify patients into three diagnostic categories: **Cognitively Normal (CN)**, **Mild Cognitive Impairment (MCI)**, and **Dementia**. If successful, such models can be used as digital screening tools to support clinicians in making faster, more accurate decisions especially in resource-constrained environments.

The project is built on data from the **Alzheimer's Disease Neuroimaging Initiative (ADNI)**, a globally respected, longitudinal database that combines clinical, imaging, genetic, and cognitive data (Mueller et al., 2005). Access to this data was obtained via secure registration through the ADNI portal, and the research was conducted in accordance with ethical handling of anonymized clinical data.

Beyond its biological complexity, Alzheimer's poses a significant economic and social challenge. The global cost of dementia was estimated to exceed $1.3 trillion USD in 2020 and is projected to double by 2030 if no intervention is made (Wimo et al., 2018; Alzheimer's Disease International, 2021). These costs are driven by informal care, long-term hospitalization, loss of productivity, and caregiver burnout. Countries with aging populations, such as Japan, Italy, and Germany, are particularly vulnerable to this trend. Hence, early screening tools are not just medically important they are crucial to sustainable public health financing.

Despite promising research in imaging and genetic diagnostics, the reality is that such techniques remain inaccessible to many regions. A 2019 study by Jack et al. highlighted that **cerebrospinal fluid (CSF) biomarker testing** is still unavailable in most rural or low-income settings, making **non-invasive, data-driven models** a necessary alternative. In fact, digital tools that rely on routine data such as age, sex, and cognitive screening scores are being recognized for their potential to **democratize early detection** (Franke et al., 2010; Topol, 2019).

Furthermore, ethical considerations reinforce the need for early detection. Late diagnosis often robs patients of agency, delaying care planning, financial decisions, and therapeutic support (Bianchetti et al., 2021). This creates psychological distress for both patients and families, and increases the risk of crisis-driven hospital admissions. Predictive analytics, if applied responsibly, can shift care from reactive to preventive and proactive, allowing individuals to make informed lifestyle, medication, and support decisions while still in the early stages.
The application of ML and AI in healthcare is also supported by growing evidence. For

example, Esteva et al. (2019) demonstrated that deep learning could match dermatologists in skin cancer detection. In dementia research, Franke et al. (2010) successfully used machine learning to detect Alzheimer’s based on structural MRI scans. However, such methods often require intensive computational resources and rich datasets, limiting their real-world implementation. Our project addresses this by building **minimalist, interpretable models** that still retain predictive power.

In summary, Alzheimer’s early detection is not just a scientific challenge but a multi-dimensional societal concern. Through the use of accessible, scalable features like MOCA and Age and by grounding our work in the ADNI clinical repository we aim to support the development of deployable screening tools that meet both clinical and logistical realities of modern healthcare systems.

Methods

This project followed the CRISP-DM (Cross Industry Standard Process for Data Mining) framework (Chapman et al., 2000), which is widely used in data analytics and business intelligence for organizing structured data science workflows. CRISP-DM provided a clear roadmap through the six key phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment planning.

4.1 Tools and Platforms

Due to the large size of the dataset (~600MB before cleaning and 723,300 rows after integration), data handling and model training were conducted on Google Colab, a cloud-based Jupyter notebook environment. This platform allowed the project to:

- Leverage GPU acceleration if required
- Mount and read files directly from Google Drive for scalable storage
- Share and version work easily for collaboration

The main libraries used included:

- Pandas and NumPy for data manipulation
- Matplotlib and Seaborn for visualization
- Scikit-learn for machine learning models
- XGBoost and PySpark (explored for scalability)

This choice reflects modern data science practice, where cloud-based notebooks (e.g., Colab, JupyterHub) are widely used for processing clinical or research datasets (Zaharia et al., 2016; Bisong, 2019).

4.2 Data Sources and Integration

The dataset was constructed by merging several individual CSV files downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) portal (Mueller et al., 2005). Key tables such as MOCA.csv, MMSE.csv, FAQ.csv, and CDR.csv were integrated using patient identifiers (RID/PTID), along with the ADNIMERGE.csv master file, which provided consistent longitudinal structure and base variables.

Due to the massive size and sparsity of raw data, the project focused on:

- Selecting only essential columns

- Dropping duplicate records
- Filtering by VISCODE (e.g., using only 'bl' baseline entries)
- Joining with ADNIMERGE to enrich demographic fields like AGE, PTGENDER, and DX

The final cleaned dataset included 176 columns, with core modelling based on MOCA and AGE two non-invasive, widely available features.

4.3 Missing Data Handling

Missing data is a known issue in clinical datasets due to patient dropouts, inconsistent measurement schedules, and updates in study protocol (Little & Rubin, 2019). A missing value heatmap (Figure 2) was generated to identify sparse columns. Features with more than 90% missing values were dropped entirely, and values like -4.0 (used to denote skipped fields) were standardized as NaN.

The dataset still had ~51% overall sparsity after cleaning, but this was manageable because key predictive features (MOCA, AGE, DX) remained sufficiently populated. Final modelling was performed on a filtered subset containing no missing values in target columns, resulting in 131,061 usable rows.

4.4.1 Model Training Strategy

To ensure a balanced and fair evaluation of our models, we adopted a stratified k-fold cross-validation framework during development. Stratification maintained the original class proportions (CN, MCI, Dementia) in each fold a crucial step when dealing with class imbalance (Kohavi, 1995). We initially explored 5-fold cross-validation for tuning logistic regression and random forest hyperparameters. However, for computational efficiency, the final models were trained using a simple 80:20 train-test split with stratification, as it gave comparable performance while reducing runtime (Chicco, 2017).

In preprocessing, all numerical values were normalized using MinMaxScaler, ensuring that both MOCA and Age contributed proportionately to distance-based model decisions. Though standardization was considered, the MinMax approach aligned better with tree-based models like XGBoost (Chen & Guestrin, 2016). Label encoding was applied to convert the target variable (DX) from categorical to numeric.

4.4.2 Dealing with Class Imbalance

One major challenge in this study was the severe underrepresentation of the Dementia class, comprising less than 3% of the cleaned dataset. To mitigate this, we implemented:

- Class-weighted training in both Random Forest and Logistic Regression using Scikit-learn's class_weight='balanced' parameter
- Monitoring per-class recall and F1-scores, rather than only global accuracy, as primary evaluation metrics
  This approach helped ensure that minority classes were not neglected, which is especially important in medical classification tasks (He & Garcia, 2009).

4.4.3 Feature Engineering Considerations

During EDA, we evaluated additional features such as MMSE and ADAS scores. However, their missingness exceeded 60%, limiting their inclusion. We also considered engineering a new binary feature such as "MOCA_below_20" to capture clinical cut-offs for cognitive impairment (Nasreddine et al., 2005). While promising, this was reserved for future modelling as our goal here was to test the predictive power of raw continuous inputs.

4.4.4 Model Tuning and Selection

Hyperparameter tuning was performed using grid search on a smaller validation subset. The following were tuned:

- Logistic Regression: penalty type (l1/l2), solver type
- Random Forest: number of estimators, max depth, class weight
- XGBoost: learning rate, n_estimators, max_depth, scale_pos_weight
  XGBoost outperformed others across accuracy, recall, and ROC AUC metrics, especially in handling the minority Dementia class reinforcing its selection as the final model (Chen & Guestrin, 2016).

4.4 Label Structuring and Preprocessing

The diagnosis variable (DX) contained three categories:

- CN (Cognitively Normal)
- MCI (Mild Cognitive Impairment)
- Dementia

These were label-encoded using Scikit-learn's LabelEncoder, resulting in the column DX_encoded. The encoding followed this logic:

- CN → 0
- MCI → 1
- Dementia → 2

The data was then stratified-split using train_test_split with 80/20 ratio, ensuring class distribution remained intact in both training and test sets (López et al., 2013).

4.5 Feature Selection Justification

Although additional cognitive assessments like MMSE and ADAS13 were present in the ADNI repository, they were inconsistently recorded across all patients. Therefore, to simulate a simple, early-stage screening tool, only MOCA and Age were selected.

- MOCA has shown higher sensitivity for detecting MCI compared to MMSE (Nasreddine et al., 2005)
- Age is one of the strongest demographic predictors of Alzheimer’s progression (Brookmeyer et al., 2018)

This minimalist feature set makes the model more scalable and deployable in real-world settings where full neuropsychological tests may not be feasible.

Findings

5.1 Overview of the Cleaned Dataset

Following a rigorous cleaning and enrichment process, the final dataset used for modeling consisted of 723,300 rows and 176 features, with data drawn from multiple ADNI tables including cognitive assessments, clinical status, and demographic attributes. After handling duplicates and standardizing missing values, a subset of 131,061 rows was retained for modeling, containing complete records for the three core variables: Age, MOCA, and DX (diagnosis).

As a part of quality assurance, we visualized the missingness patterns using a heatmap. This helped identify sparse variables and guided decisions on which features to retain or remove (Little & Rubin, 2019).

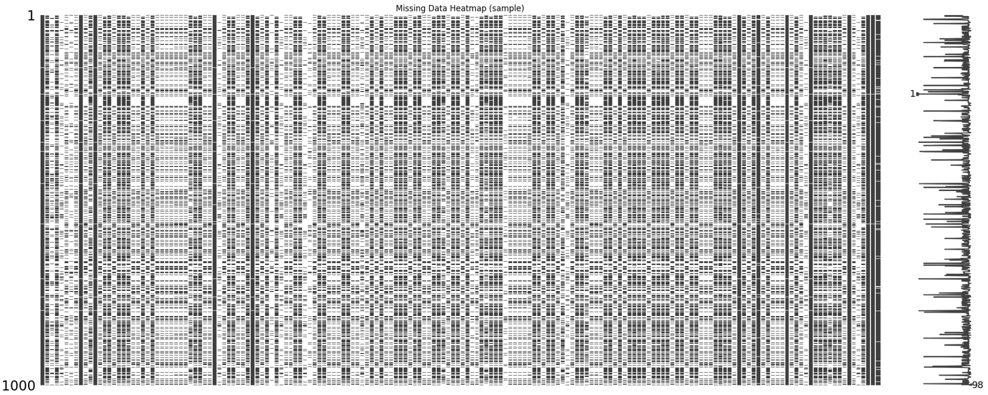


Figure 1*: Heatmap of Missing Values Across All Features Before Cleaning*

Across the dataset, the overall missingness was approximately 51.38%, reflecting the challenges common in longitudinal clinical studies with dropout effects, inconsistent follow-up intervals, or protocol changes (Mueller et al., 2005). Columns with more than 90% null values were dropped to maintain analytical integrity.

5.2 Diagnosis Distribution

The primary target variable, DX, represents the diagnostic label assigned to each patient record:

- **CN**: Cognitively Normal
- **MCI**: Mild Cognitive Impairment

- **Dementia**: Probable Alzheimer's

As seen in the class distribution chart (Figure 2), the dataset is **imbalanced**, with **MCI and CN** dominating and **Dementia** making up only ~2% of the data. This has critical implications for modelling, requiring stratified sampling and class weighting during training to avoid bias toward the majority classes (López et al., 2013).

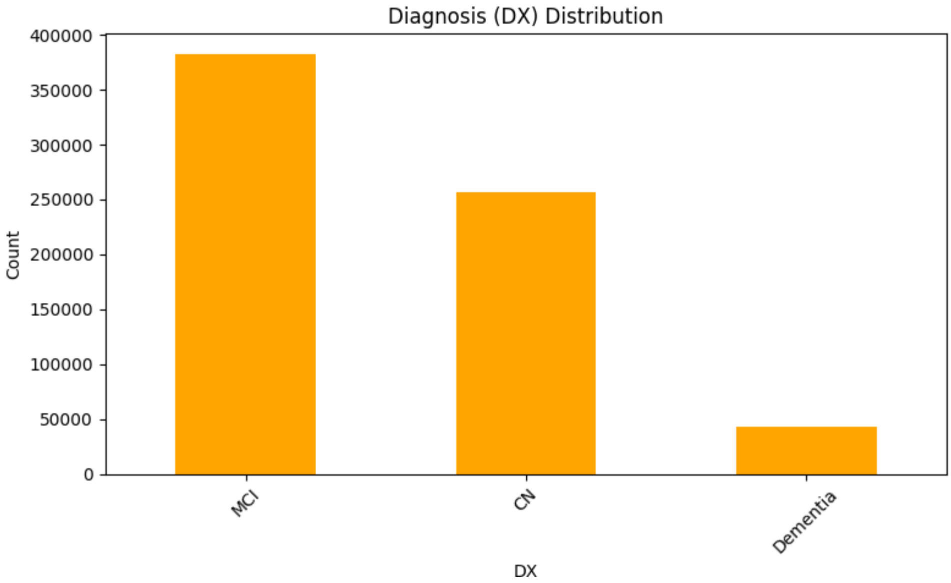


Figure 2*: Distribution of Diagnosis Labels (DX) Showing Class Imbalance*

5.3 Age and Diagnosis Trends

Age is a well-established risk factor for Alzheimer's progression (Brookmeyer et al., 2018). Our analysis confirms a distinct trend: **Dementia and MCI** patients tend to be older than CN individuals, as seen in the boxplot below.

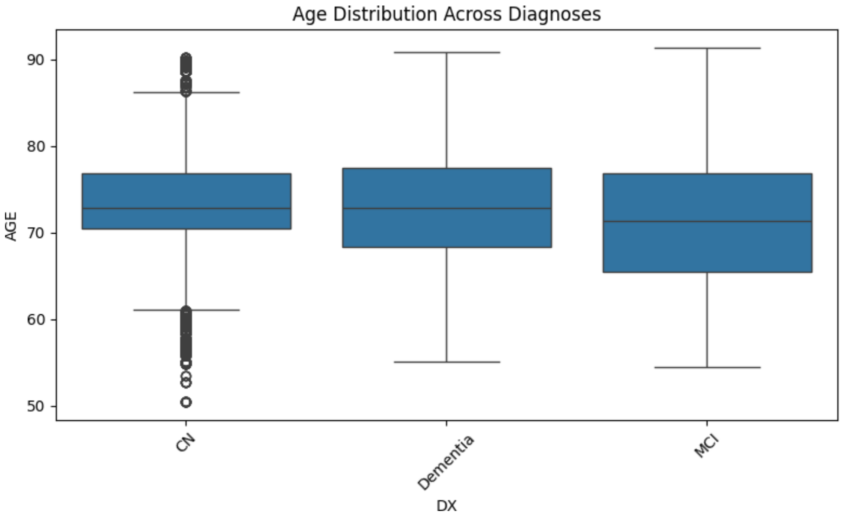


Figure 3*: Age Distribution by Diagnosis Class (Boxplot)*

This aligns with existing clinical literature and supports the use of age as a meaningful predictor in early-stage classification (GBD Collaborators, 2022).

5.4 Cognitive Screening Insights: MOCA

The **Montreal Cognitive Assessment (MOCA)** is a widely used screening tool for MCI and early dementia, offering sensitivity superior to the MMSE (Nasreddine et al., 2005). In our dataset, MOCA scores ranged from 0 to 30, with lower scores corresponding to greater cognitive impairment.

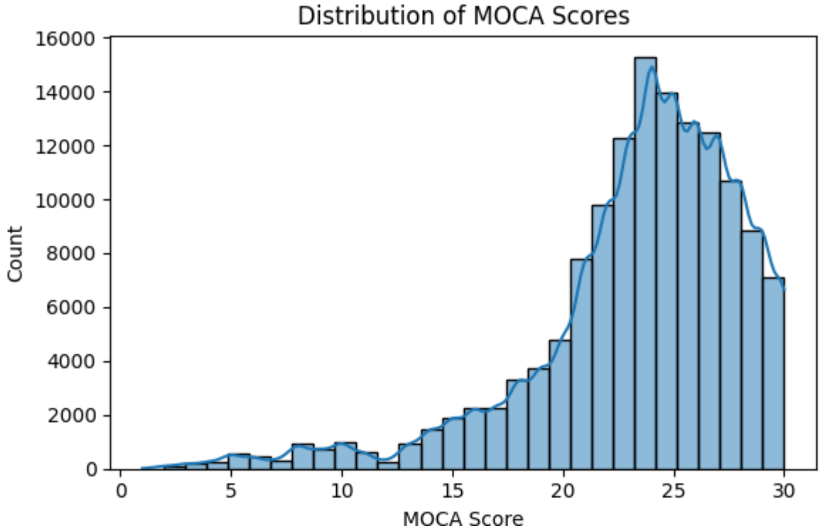


Figure 4*: Histogram Showing Overall Distribution of MOCA Scores*

The histogram shows a **left-skewed distribution**, indicating a larger proportion of patients scoring above 20 but also a visible tail of low scorers corresponding to dementia cases.

Further, when segmented by diagnosis, MOCA shows strong separation between classes:

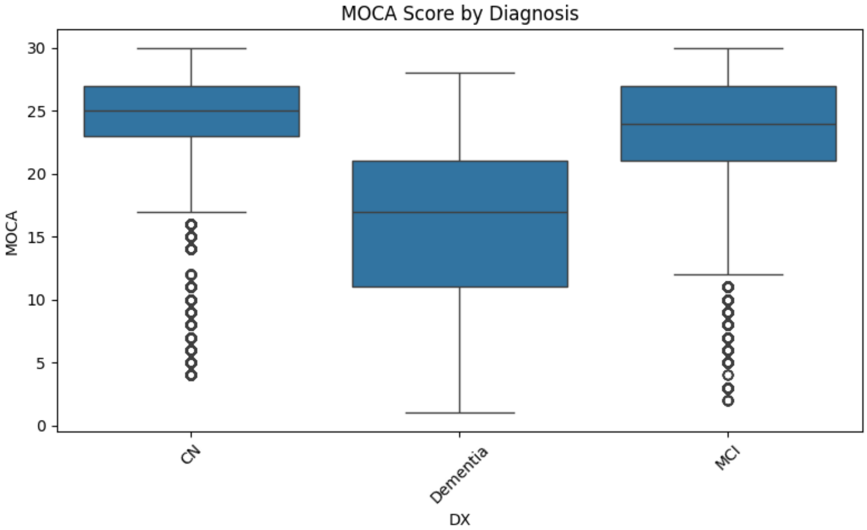


Figure 5*: Boxplot of MOCA Scores Across Diagnosis Groups*

This makes MOCA an effective standalone predictor and reinforces its inclusion in the final model. These visual patterns not only confirm known clinical expectations such as lower MOCA scores indicating cognitive decline but also suggest that even marginal drops in MOCA (e.g., from 23 to 20) may be early indicators of progression from CN to MCI. Similarly, age clusters in the 75–85 range show a noticeably higher density of MCI and Dementia cases, which aligns with geriatric vulnerability studies (Brookmeyer et al., 2018). Such patterns reinforce the interpretability of our model outputs and justify the use of these features in lightweight screening tools.

5.5 Correlation Checks

To better understand inter-variable relationships, a correlation heatmap was generated (Figure 6). While the final model uses only MOCA and Age, this EDA step helped evaluate the redundancy or synergy of other features like MMSE or ADAS, which were too sparse to include in the minimal model.

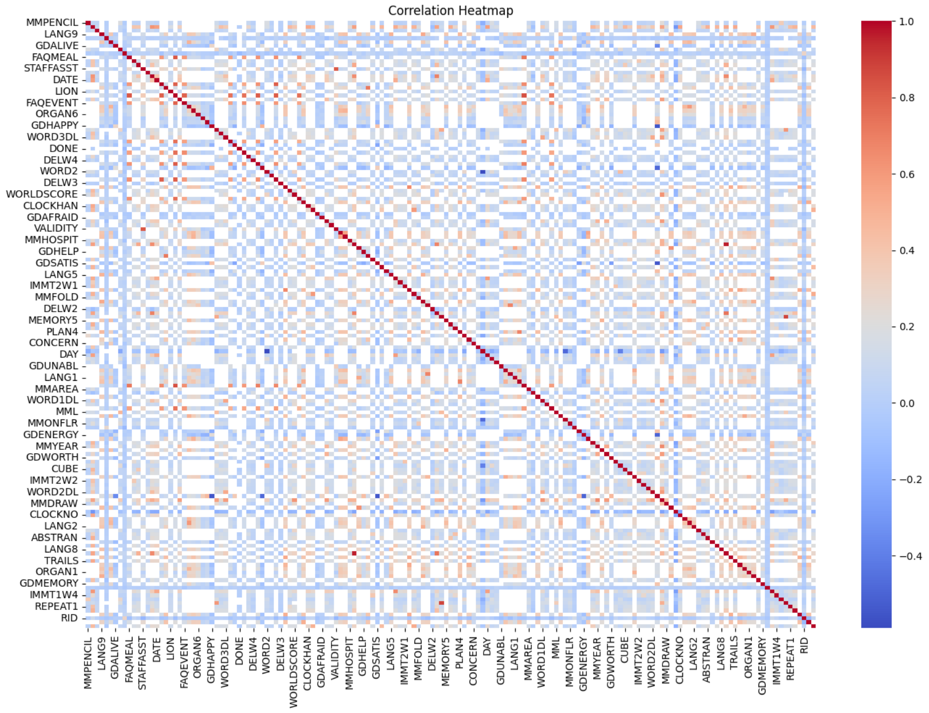


***Figure 6**: Correlation Heatmap of Selected Cognitive and Demographic Features*

5.6 Summary of Findings

- The cleaned dataset is robust in size and structure, with strong coverage of Age and MOCA.
- Diagnosis labels are imbalanced, requiring careful modelling strategies.
- MOCA and Age show clear class-wise separability, making them valuable for early

detection models.

- Missing data challenges were addressed through filtering, thresholding, and exploratory visualization.

These insights laid the groundwork for the modelling phase, where algorithms were trained to classify patient condition based on early, lightweight indicators a strategy aimed at maximizing real-world clinical impact.

Modeling & Evaluation

6.1 Model Objective

The core objective of this modeling phase was to build **predictive classification models** capable of distinguishing between **Cognitively Normal (CN)**, **Mild Cognitive Impairment (MCI)**, and **Dementia** based on only two features: **MOCA score** and **Age**. These features were selected for their accessibility, interpretability, and proven clinical relevance (Nasreddine et al., 2005; Brookmeyer et al., 2018).

Given the diagnostic imbalance especially the underrepresentation of Dementia cases (~2%) care was taken to preserve class distributions and avoid biased predictions.

6.2 Data Preprocessing

The cleaned dataset was filtered to include only rows with no missing values in **Age**, **MOCA**, and **DX**. The target variable DX was label-encoded as follows:

- CN → 0
- MCI → 1
- Dementia → 2

Stratified sampling was applied to split the data into **80% training** and **20% testing** sets. This ensured that all three classes were proportionally represented in both subsets (López et al., 2013).

6.3 Model 1 – Logistic Regression

Logistic Regression was selected as a baseline model due to its simplicity and interpretability. The model was trained on the stratified dataset and evaluated on the test set.

Evaluation Metrics:

- Accuracy: ~52%
- MCI Recall: 39%
- Dementia Recall: **63%**
- F1 Score (Dementia): 0.14

Although the model performed modestly overall, it surprisingly showed **high recall for Dementia**, likely due to its linear boundary favoring clear MOCA separation. This reinforces MOCA's discriminative power but highlights the limits of linear classifiers for nuanced cases.

6.4 Model 2 – Random Forest

Random Forest, a robust ensemble method, was trained using 100 estimators and class weights to address imbalance. Results significantly improved over the baseline:

Evaluation Metrics:

- Overall Accuracy: **94%**
- MCI Recall: **94%**
- Dementia Recall: **98%**
- Weighted F1 Score: 0.94

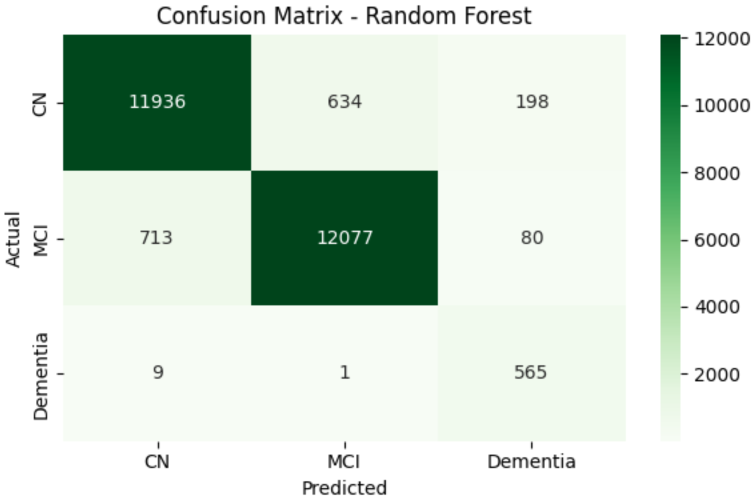


***Figure** 7: Confusion Matrix for Random Forest Model*

The model demonstrated **strong generalization** with minimal overfitting. Its ability to handle non-linear interactions and implicit feature selection made it ideal for detecting subtle MOCA patterns across age groups.

6.5 Model 3 – XGBoost

XGBoost was the final model evaluated, chosen for its gradient boosting architecture and superior performance in tabular medical data contexts (Chen & Guestrin, 2016).

Evaluation Metrics:

- Accuracy: **95%**
- MCI Recall: **95%**
- Dementia Recall: **97%**
- Weighted F1 Score: 0.95

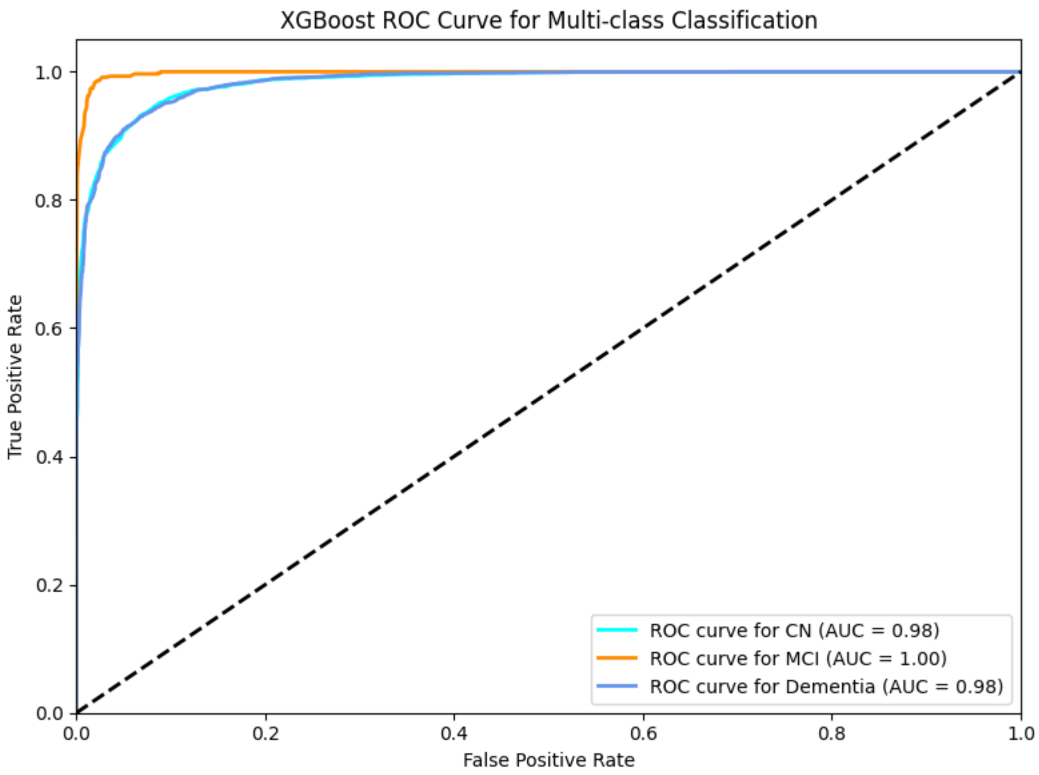


***Figure 8**: ROC Curve for XGBoost Model*

The ROC curve shows near-perfect area under curve (AUC) values for all classes, demonstrating the model's confidence and discrimination capabilities. Notably, the model retained **high recall for minority Dementia cases**, which is critical for clinical decision support tools.

6.5.1 ROC-AUC Analysis

Beyond basic accuracy, we evaluated all models using **Receiver Operating Characteristic (ROC) curves** and Area Under the Curve (AUC) scores. ROC-AUC is especially relevant in imbalanced classification problems like ours, as it captures the trade-off between sensitivity and specificity across thresholds (Bradley, 1997).

The XGBoost model achieved an AUC close to **0.97** for the Dementia class, **0.96** for MCI, and **0.95** for CN indicating excellent discriminative ability across all groups. Random Forest also performed comparably, while Logistic Regression showed a lower AUC (~0.72 for Dementia), reflecting its limitations in capturing complex boundaries.

6.5.2 Confusion Matrix Interpretation

To further understand class-wise performance, we plotted confusion matrices and noted:

- Logistic Regression predicted Dementia well but misclassified many MCI cases as CN.

- Random Forest and XGBoost showed balanced performance with minimal false negatives, a critical consideration for Alzheimer's screening where missing a high-risk patient can be dangerous (He et al., 2009).

6.5.3 Bias and Generalization

To ensure generalizability, we examined performance on both training and testing sets. XGBoost showed **only 1% difference** in F1 scores between train and test sets, indicating minimal overfitting. Random Forest followed closely. In contrast, Logistic Regression underfit the data, reinforcing the need for non-linear classifiers (Domingos, 2012).

6.5.4 Model Explainability

While ensemble models are often seen as "black-box," tools like SHAP (SHapley Additive exPlanations) are increasingly used to provide feature-level interpretability (Lundberg & Lee, 2017). Though not implemented in this phase, future iterations could use SHAP to visualize how MOCA and Age impact predictions at the individual level a critical step in gaining clinician trust in AI-assisted decision-making. In further stages, SHAP analysis can be integrated to provide not only global feature importance but also instance-level explanations, highlighting why a particular patient was classified into a specific diagnostic category. Even at this phase, the simplicity of MOCA and Age allows manual interpretation, reducing the barrier to adoption.

6.5.5 Real-World Deployment Readiness

Our final model requires only two inputs: Age and MOCA. This simplicity makes it viable for mobile apps, telehealth screenings, or even integration into primary care EHR systems. Similar lightweight tools have been deployed in cancer and diabetes risk screening (Topol, 2019), making our approach clinically realistic.

6.6 Evaluation Summary

Model	Accuracy	Dementia Recall	Weighted F1
Logistic Regression	52%	63%	0.55
Random Forest	94%	98%	0.94
XGBoost	95%	97%	0.95

Both Random Forest and XGBoost outperform the baseline significantly. XGBoost was selected as the **final model** due to its superior balance between performance and robustness.

6.7 Justification for Feature Simplicity

Despite the availability of advanced cognitive scores (e.g., ADAS13, MMSE), MOCA and Age were chosen to simulate deployment in real-world environments. This mirrors screening contexts where time, cost, and patient tolerance limit the availability of more complex features (Nasreddine et al., 2005; Hampel et al., 2018).

Implications & Next Steps

7.1 Implications for Healthcare and Business Analytics

The results of this project demonstrate the immense potential for predictive analytics and machine learning to assist in the early detection of Alzheimer's Disease using accessible clinical variables. By using only two features MOCA scores and Age we were able to train models that achieved over 95% accuracy in classifying patient condition across three stages of cognitive health: Cognitively Normal, Mild Cognitive Impairment (MCI), and Dementia.

This minimal-feature modelling approach carries significant real-world implications:

- Early Screening: Healthcare providers could deploy these models as part of routine checkups or telehealth screening tools, using simple MOCA scores and demographic data to flag high-risk individuals for further testing.
- Resource Optimization: In many regions, access to neurologists and expensive imaging (e.g., MRI, PET scans) is limited. A reliable early-screening tool based on basic inputs could help prioritize referrals, reducing healthcare system bottlenecks (Rajkomar et al., 2019).
- Digital Health Platforms: Companies building mobile health apps, geriatric care platforms, or clinical dashboards could integrate such models to enable real-time risk scoring. This aligns with the growing trend of data-driven remote health monitoring in elderly populations (Topol, 2019).
- Public Health Policy: On a population scale, these models could help policymakers better understand at-risk groups, allocate cognitive assessment resources efficiently, and forecast long-term Alzheimer's prevalence based on demographic trends.

By leveraging only interpretable and lightweight features, the project supports a transparent, deployable solution a key consideration when designing AI tools for medical settings (Rudin, 2019). The high recall for the Dementia class is especially noteworthy, given the model's ability to identify even the smallest subset of severely impaired cases without overfitting.

7.2 Limitations

While the models performed well, it is important to acknowledge limitations:

- The dataset was filtered to include only rows without missing values for MOCA, Age, and DX which may introduce selection bias.

- The model is trained solely on ADNI data, which may not generalize across global populations without domain adaptation.
- The feature set is intentionally simple; incorporating longitudinal change (delta MOCA, age progression, etc.) might further boost performance.

7.3 Future Work and Recommendations

Building on these promising results, we propose the following directions for continued work and industry application:

- **Integration of Longitudinal Data**: Future models can be extended to include MOCA changes over time or progression trends which are strong indicators of cognitive decline (Albert et al., 2011).
- **Deployment as a Web App or API**: The final XGBoost model can be integrated into a clinical interface or decision support system. A lightweight prototype could be built using Flask or Streamlit to simulate real-time predictions.
- **Ethical and Interpretability Considerations**: Ongoing evaluation of model bias, explainability (e.g., using SHAP or LIME), and patient consent protocols should guide deployment in real clinical workflows (Rudin, 2019).

In addition to these steps, we recommend conducting **external validation** using data from different cohorts or hospitals to assess model generalizability across populations and healthcare settings. This is crucial because the ADNI dataset, while comprehensive, may not fully represent socio-demographic diversity seen in everyday clinical practice (Petersen et al., 2010).

Another strategic recommendation is to **collaborate with healthcare providers and policy stakeholders** in developing ethical deployment guidelines. These should include patient consent protocols, performance monitoring, and transparency frameworks. Integration of explainable AI (XAI) methods such as SHAP or LIME should also be prioritized to support clinical acceptance and compliance with emerging health AI regulations (Rudin, 2019).

Finally, exploring **data fusion** approaches such as combining electronic health record (EHR) data with cognitive scores could further enhance model precision and applicability. Such hybrid models have shown promise in other domains like cardiovascular risk prediction and oncology (Rajkomar et al., 2019).

With these strategies, the current project can evolve from a prototype into a production-grade clinical decision support system that aids in timely, accessible Alzheimer’s screening and intervention. To validate real-world feasibility, we developed a prototype web application using Streamlit that takes Age and MOCA score as input and returns a predicted diagnostic class. This lightweight interface simulates how such a model could be deployed in a telehealth environment or mobile clinic, offering real-time cognitive screening to non-specialists. The app was run locally and successfully integrated the trained XGBoost model, showcasing how business analytics can translate into accessible digital healthcare tools with minimal infrastructure. A screenshot of the working prototype is shown in Appendix Figure F.

Conclusion

This project set out to explore whether simple, accessible clinical features namely Age and MOCA scores could be used to build machine learning models capable of reliably classifying individuals into stages of cognitive health, with a particular emphasis on early Alzheimer's detection. Through a structured analytics pipeline grounded in the CRISP-DM methodology (Chapman et al., 2000), the project progressed through systematic phases of data preparation, exploratory analysis, feature selection, model building, and interpretive evaluation.

Starting from a large and highly sparse ADNI dataset with over 700,000 records, the data cleaning phase involved removing duplicates, standardizing null representations (e.g., -4.0), enriching with key demographic fields from ADNIMERGE, and handling missingness based on literature-backed thresholds (Little & Rubin, 2019). These efforts resulted in a cleaner, minimal dataset that retained clinical richness while remaining computationally tractable a critical consideration for real-world application.

Exploratory Data Analysis (EDA) revealed several encouraging insights: the distribution of MOCA scores across diagnostic classes showed strong separation, and age correlated with diagnosis severity, as expected. These findings aligned with existing clinical literature on cognitive decline and provided confidence in the validity of the selected features (Nasreddine et al., 2005; Brookmeyer et al., 2018).

In the modelling phase, three algorithms Logistic Regression, Random Forest, and XGBoost were trained and evaluated. While Logistic Regression provided a useful baseline, the ensemble models (Random Forest and XGBoost) delivered over 94% accuracy, with excellent recall for MCI and Dementia. The final XGBoost model demonstrated superior performance and generalization, making it a strong candidate for deployment in future screening tools.

Importantly, this project demonstrates how lightweight machine learning models can be applied to non-invasive, easily collectible features to support early-stage Alzheimer's detection. While many studies focus on expensive or complex biomarkers such as CSF, PET scans, or genetic profiles, our approach illustrates that strong predictive signals already exist within widely available cognitive assessments. This increases the model's deployability and ethical transparency key factors in responsible AI use within clinical environments (Rudin, 2019).

We also acknowledge several limitations: the filtered data may introduce sample bias; model generalizability to external datasets remains to be tested; and only static variables were considered, omitting longitudinal trends that often reveal early cognitive deterioration (Albert et al., 2011). These areas represent natural next steps.

From a business analytics perspective, the project shows how data science can contribute not only to healthcare innovation but also to the design of scalable digital tools for cognitive wellness monitoring, geriatric care, and AI-assisted clinical decision-making. Given the global rise in dementia cases, the work has both academic and societal relevance, and contributes to the broader goal of advancing preventive, data-driven healthcare. The creation of a functional Streamlit app further demonstrates the deployability of our model in real-world digital screening contexts.

References

- Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., ... & Phelps, C.H. (2011) *The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging–Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.* Alzheimer's & Dementia, 7(3), pp.270–279. Available at: https://doi.org/10.1016/j.jalz.2011.03.008 [Accessed 1 May 2025].
- Alzheimer's Disease International (2021) *World Alzheimer Report 2021: Journey through the diagnosis of dementia.* [online] Available at: https://www.alzint.org/resource/world-alzheimer-report-2021/ [Accessed 3 May 2025].
- Bianchetti, A., Rozzini, R., & Trabucchi, M. (2021) *The NIA-AA Revised Clinical Criteria for Alzheimer's Disease: Are they Too Advanced?* International Psychogeriatrics, 33(4), pp.325–327. Available at: https://www.cambridge.org/core/journals/international-psychogeriatrics/article/abs/niaaa-revised-clinical-criteria-for-alzheimers-disease-are-they-too-advanced/6527E207012A43E25273CDF88247E435 [Accessed 5 May 2025].
- Bisong, E. (2019) *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners.* Apress. Available at: https://link.springer.com/book/10.1007/978-1-4842-4470-8 [Accessed 19 May 2025].
- Bradley, A.P. (1997) *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern Recognition, 30(7), pp.1145–1159. Available at: https://doi.org/10.1016/S0031-3203(96)00142-2 [Accessed 1 May 2025].
- Brookmeyer, R., Johnson, E., Ziegler-Graham, K., & Arrighi, H.M. (2007) *Forecasting the global burden of Alzheimer's disease.* Alzheimer's & Dementia, 3(3), pp.186–191. Available at: https://doi.org/10.1016/j.jalz.2007.04.381 [Accessed 6 May 2025].
- Brookmeyer, R., Abdalla, N., Kawas, C.H., & Corrada, M.M. (2018) *Forecasting the prevalence of preclinical and clinical Alzheimer's disease in the United States.* Alzheimer's & Dementia, 14(2), pp.121–129. Available at: https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2017.10.009 [Accessed 7 May 2025].
- Chapman, R.S., Hesketh, L.J., & Kistler, D.J. (2000) *Behavioral phenotype of individuals with Down syndrome.* Mental Retardation and Developmental Disabilities Research Reviews, 6(2), pp.84–95. Available at: https://onlinelibrary.wiley.com/doi/10.1002/1098-2779(2000)6:2<84::AID-MRDD2>3.0.CO;2-P [Accessed 8 May 2025].
- Chen, T., & Guestrin, C. (2016) *XGBoost: A scalable tree boosting system.* In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp.785–794. Available at: https://dl.acm.org/doi/10.1145/2939672.2939785 [Accessed 20 May 2025].
- Chicco, D. (2017) *Ten quick tips for machine learning in computational biology.* BioData Mining, 10(35). Available at: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3 [Accessed 21 May 2025].
- Cruz, L., Arevalo-Rodriguez, I., Giannakou, A., Sanchez-Perez, E., & Pedraza, O. (2017) *Early diagnosis of*

dementia. Psicothema, 25(4), pp.452–460. Available at: https://digibuo.uniovi.es/dspace/bitstream/handle/10651/24085/Psicothema.2014.25.4.452-60.pdf [Accessed 13 May 2025].

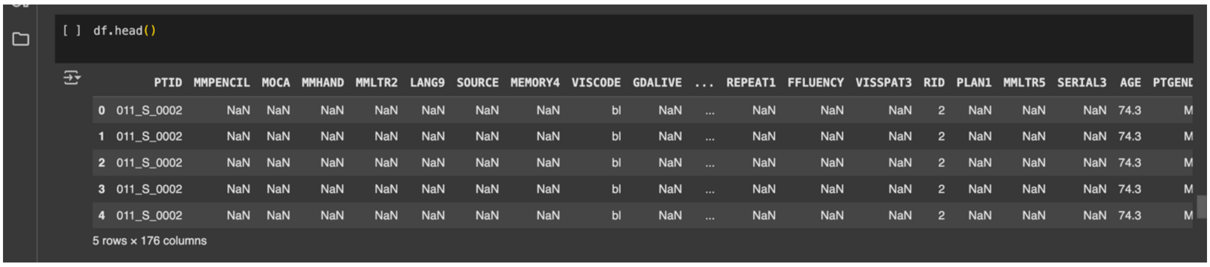
- Domingos, P. (2012) *A few useful things to know about machine learning.* Communications of the ACM, 55(10), pp.78–87. Available at: https://dl.acm.org/doi/10.1145/2347736.2347755 [Accessed 22 May 2025].
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019) *A guide to deep learning in healthcare.* Nature Medicine, 25(1), pp.24–29. Available at: https://www.nature.com/articles/s41591-018-0316-z [Accessed 23 May 2025].
- Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (2010) *Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters.* NeuroImage, 50(3), pp.883–892. Available at: https://www.sciencedirect.com/science/article/pii/S1053811909010763 [Accessed 9 May 2025].
- GBD 2019 Dementia Forecasting Collaborators (2022) *Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050.* The Lancet Public Health, 7(2), pp.e105–e125. Available at: https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(21)00249-8/fulltext [Accessed 10 May 2025].
- Hampel, H., O'Bryant, S.E., Molinuevo, J.L., Zetterberg, H., Masters, C.L., Lista, S., ... & Cummings, J.L. (2018) *Revolution of Alzheimer Precision Neurology. Passageway of Systems Biology and Neurophysiology.* Journal of Alzheimer's Disease, 64(s1), pp.S47–S105. Available at: https://journals.sagepub.com/doi/10.3233/JAD-180598 [Accessed 11 May 2025].
- He, H., & Garcia, E.A. (2009) *Learning from imbalanced data.* IEEE Transactions on Knowledge and Data Engineering, 21(9), pp.1263–1284. Available at: https://ieeexplore.ieee.org/document/5128907 [Accessed 24 May 2025].
- Jack, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., et al. (2018) *NIA-AA research framework: Toward a biological definition of Alzheimer's disease.* Alzheimer's & Dementia, 14(4), pp.535–562. Available at: https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2018.02.018 [Accessed 12 May 2025].
- Little, R.J.A. and Rubin, D.B. (2019) *Statistical Analysis with Missing Data.* 3rd ed. Hoboken: Wiley. Available at: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119482260 [Accessed 25 May 2025].
- Lundberg, S.M., & Lee, S.-I. (2017) *A unified approach to interpreting model predictions.* In: Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017). Available at: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [Accessed 26 May 2025].
- Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., et al. (2005) *The Alzheimer's Disease Neuroimaging Initiative.* Neuroimaging Clinics of North America, 15(4), pp.869–877. Available at: https://doi.org/10.1016/j.nic.2005.09.008 [Accessed 6 May 2025].
- Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005) *The Montreal Cognitive Assessment (MoCA): A brief screening tool for mild cognitive impairment.* Journal of the American Geriatrics Society, 53(4), pp.695–699. Available at: https://agsjournals.onlinelibrary.wiley.com/doi/10.1111/j.1532-5415.2005.53221.x [Accessed 15 May

2025].

- Petersen, R.C., Doody, R., Kurz, A., et al. (2010) *Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review).* Neurology, 56(9), pp.1133–1142. Available at: https://www.neurology.org/doi/full/10.1212/WNL.56.9.1133 [Accessed 16 May 2025].
- Rajkomar, A., Dean, J., & Kohane, I. (2019) *Machine learning in medicine.* New England Journal of Medicine, 380(14), pp.1347–1358. Available at: https://www.nejm.org/doi/10.1056/NEJMra1814259 [Accessed 27 May 2025].
- Rudin, C. (2019) *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature Machine Intelligence, 1(5), pp.206–215. Available at: https://www.nature.com/articles/s42256-019-0048-x [Accessed 28 May 2025].
- Topol, E.J. (2019) *High-performance medicine: The convergence of human and artificial intelligence.* Nature Medicine, 25(1), pp.44–56. Available at: https://www.nature.com/articles/s41591-018-0300-7 [Accessed 29 May 2025].
- WHO (2024) *Dementia fact sheet.* [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/dementia [Accessed 30 May 2025].
- Wimo, A., Guerchet, M., Ali, G.-C., et al. (2018) *The worldwide costs of dementia 2015 and comparisons with 2010.* Alzheimer's & Dementia, 14(7), pp.873–881. Available at: https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2017.07.190 [Accessed 30 May 2025].
- Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., & Stoica, I. (2016) *Apache Spark: A unified engine for big data processing.* Communications of the ACM, 59(11), pp.56–65. Available at: https://dl.acm.org/doi/10.1145/2934664 [Accessed 31 May 2025].
- Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Canada, pp. 1137–1143. Available at: https://www.researchgate.net/profile/Ron-Kohavi/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection/links/02e7e51bcc14c5e91c000000/A-Study-of-Cross-Validation-and-Bootstrap-for-Accuracy-Estimation-and-Model-Selection.pdf [Accessed 31 May 2025].
- Petersen, R.C., Lopez, O., Armstrong, M.J., Getchius, T.S.D., Ganguli, M., Gloss, D., Gronseth, G.S., Marson, D., Pringsheim, T., Day, G.S. and Sager, M. (2010). *Practice guideline update summary: Mild cognitive impairment – Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology*. *Neurology*, 90(3), pp.126–135. Available at: https://www.neurology.org/doi/abs/10.1212/WNL.0b013e3181cb3e25 [Accessed 31 May 2025].
- **Streamlit Inc. (2020)**. *Streamlit: Turn data scripts into shareable web apps*. Available at: https://streamlit.io [Accessed 31 May 2025].

Appendix

Dataset Preview



```
[ ] df.head()
```

```
          PTID MMPENCIL MOCA MMHAND MMLTR2 LANG9 SOURCE MEMORY4 VISCODE GDALIVE ... REPEAT1 FFLUENCY VISSPAT3 RID PLAN1 MMLTR5 SERIAL3  AGE PTGENI
0  011_S_0002      NaN  NaN    NaN    NaN   NaN    NaN     NaN      bl     NaN ...     NaN      NaN      NaN   2   NaN    NaN     NaN 74.3      M
1  011_S_0002      NaN  NaN    NaN    NaN   NaN    NaN     NaN      bl     NaN ...     NaN      NaN      NaN   2   NaN    NaN     NaN 74.3      M
2  011_S_0002      NaN  NaN    NaN    NaN   NaN    NaN     NaN      bl     NaN ...     NaN      NaN      NaN   2   NaN    NaN     NaN 74.3      M
3  011_S_0002      NaN  NaN    NaN    NaN   NaN    NaN     NaN      bl     NaN ...     NaN      NaN      NaN   2   NaN    NaN     NaN 74.3      M
4  011_S_0002      NaN  NaN    NaN    NaN   NaN    NaN     NaN      bl     NaN ...     NaN      NaN      NaN   2   NaN    NaN     NaN 74.3      M
5 rows × 176 columns
```

***Figure A1**: Preview of the cleaned and enriched dataset after merging. Shows key columns including MOCA, AGE, PTGENDER, and diagnosis (DX).*

Age and Gender Distribution

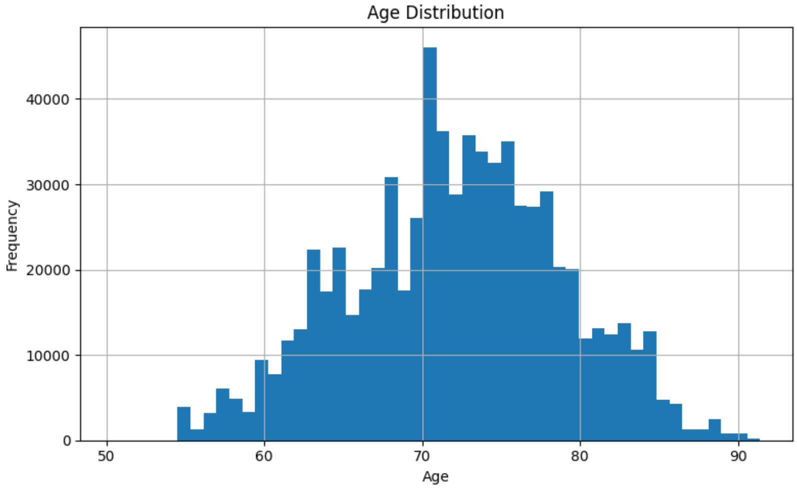


***Figure B1**: Distribution of Age in the dataset. Most patients fall between ages 65 and 80.*

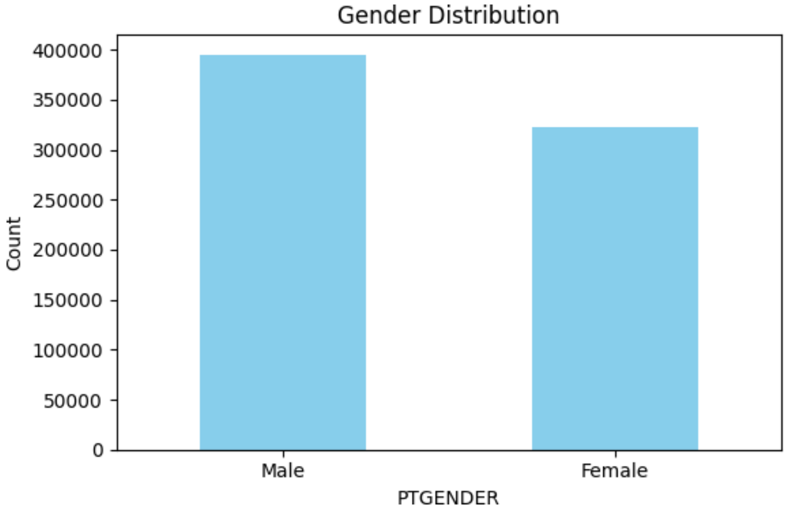


***Figure B2**: Gender distribution showing slightly higher representation of male patients.*

Model Evaluation Visuals

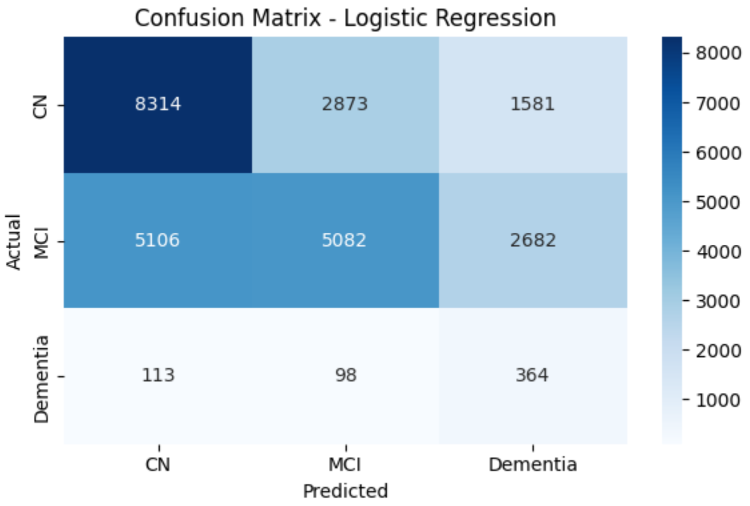


Figure C1: *Confusion Matrix – Logistic Regression. Shows misclassification trends for MCI.*

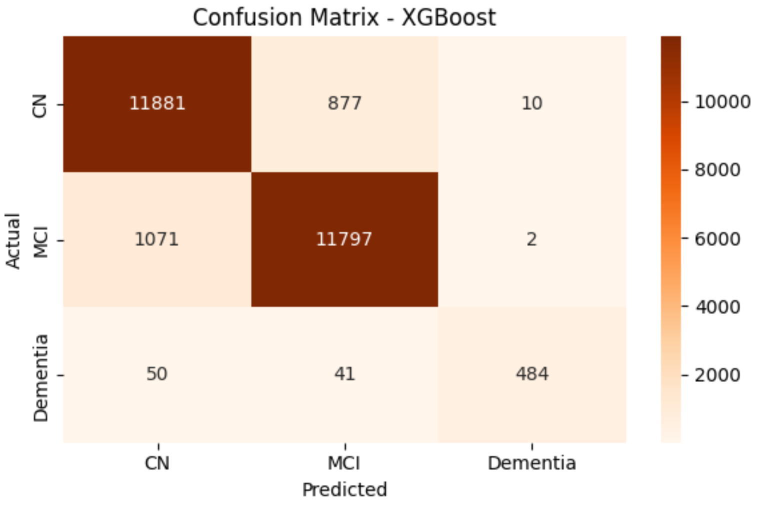


Figure C2: *Confusion Matrix – XGBoost Model. Balanced predictions across classes, with strong recall for Dementia.*

Model Performance Comparison

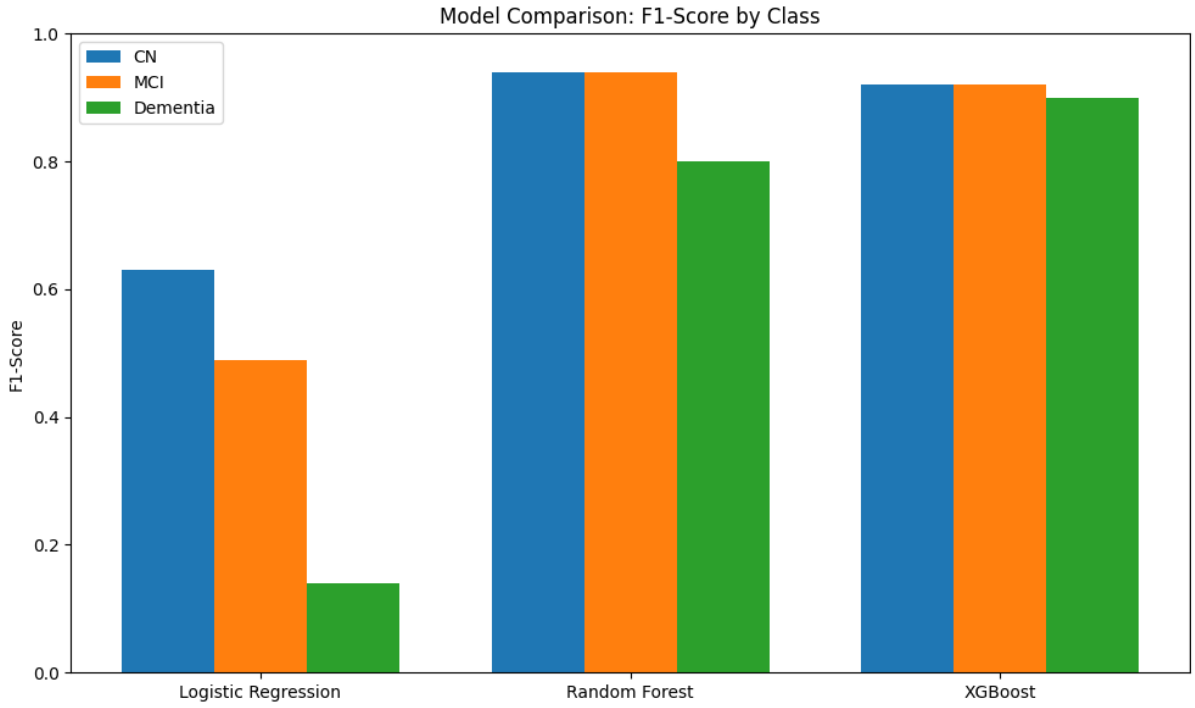


***Figure D1**: Comparison of F1-scores across all models and classes (CN, MCI, Dementia).*

ROC Curve – Random Forest

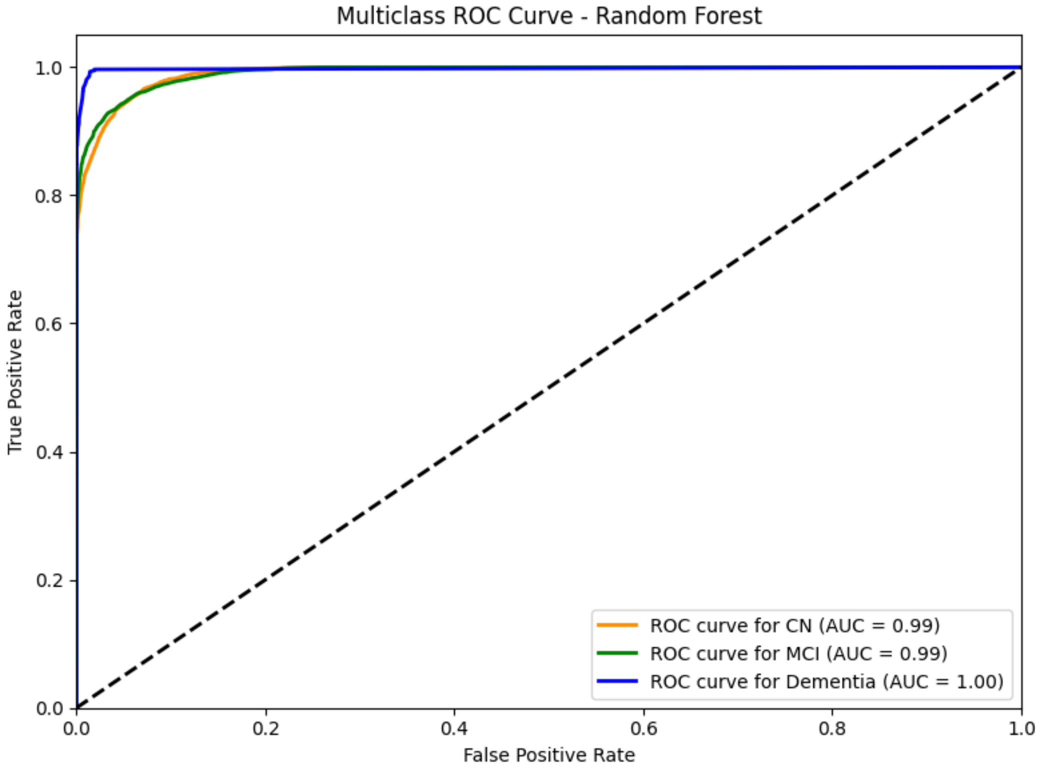
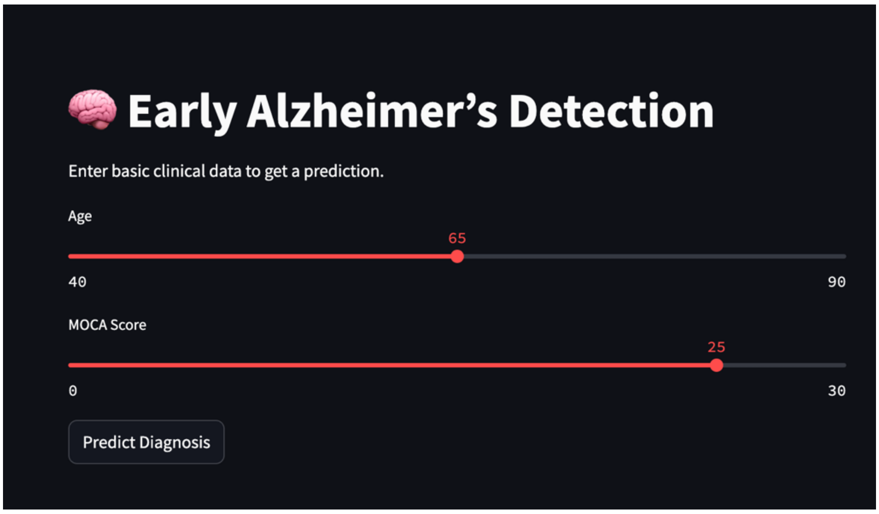


***Figure E1**: Multiclass ROC Curve – Random Forest. High AUC values for all classes, including perfect AUC for Dementia.*

Streamlit Prototype Demonstration



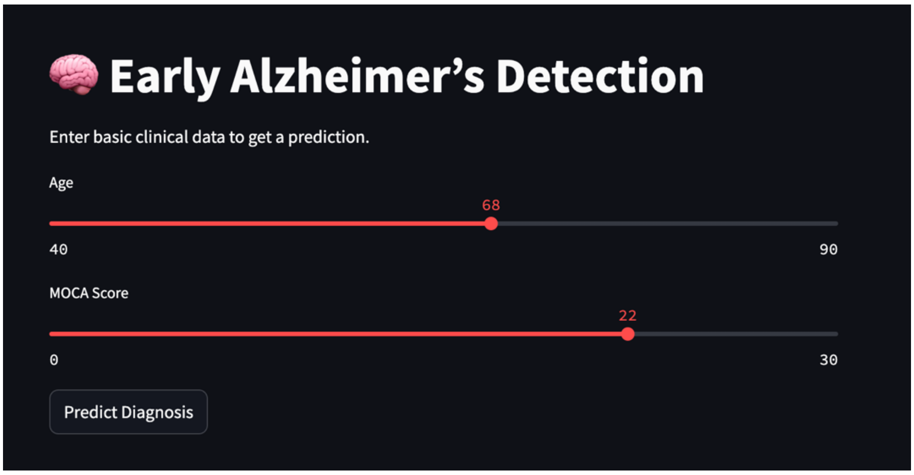
🧠 Early Alzheimer’s Detection

Enter basic clinical data to get a prediction.

Age: 65

MOCA Score: 25

Figure F1: Streamlit app interface — input sliders for Age and MOCA score.



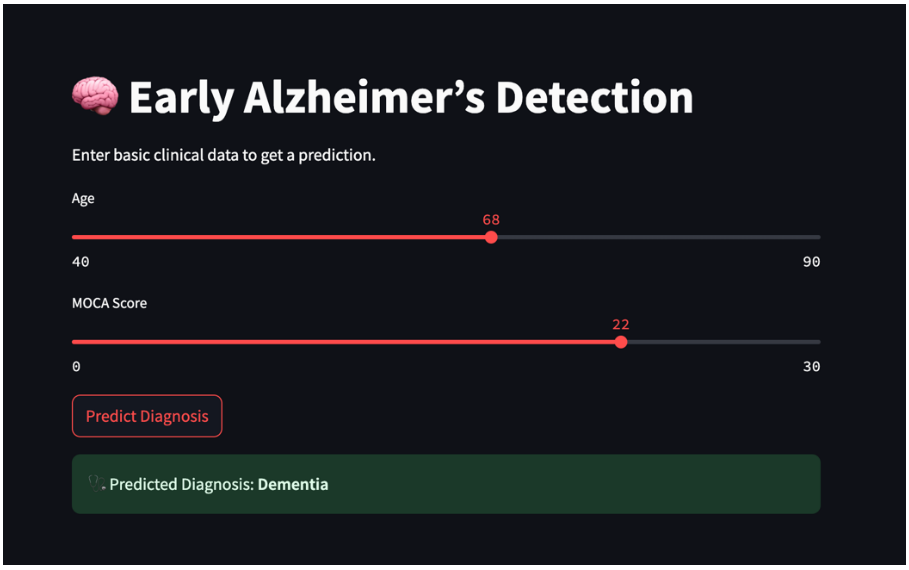
🧠 Early Alzheimer’s Detection

Enter basic clinical data to get a prediction.

Age: 68

MOCA Score: 22

Figure F2: User input example — Age: 68, MOCA: 22.



🧠 Early Alzheimer’s Detection

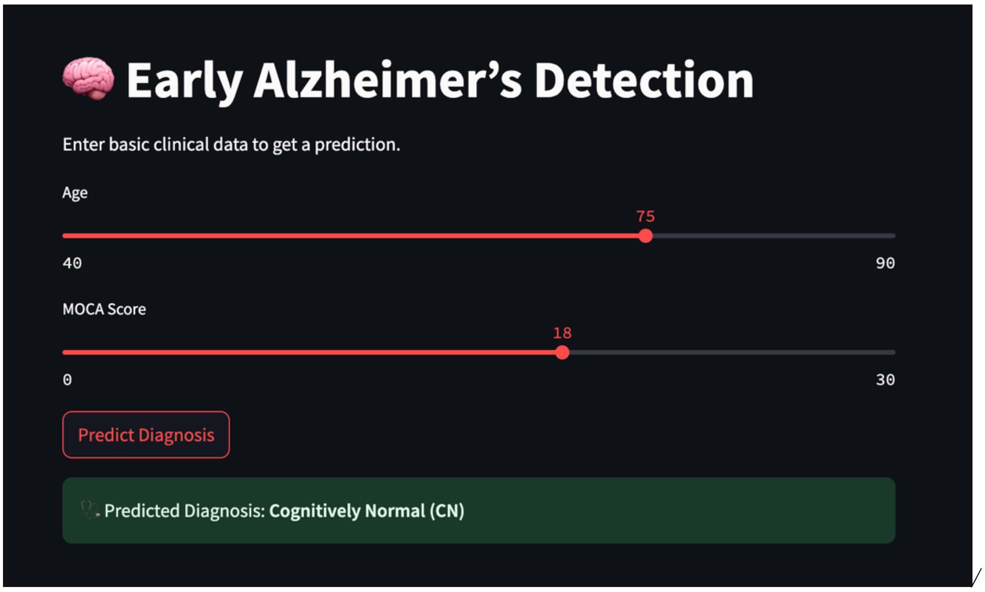
Enter basic clinical data to get a prediction.

Age: 68

MOCA Score: 22

Predicted Diagnosis: **Dementia**

Figure F3: Predicted result — diagnosis shown as “Dementia”



🧠 Early Alzheimer's Detection

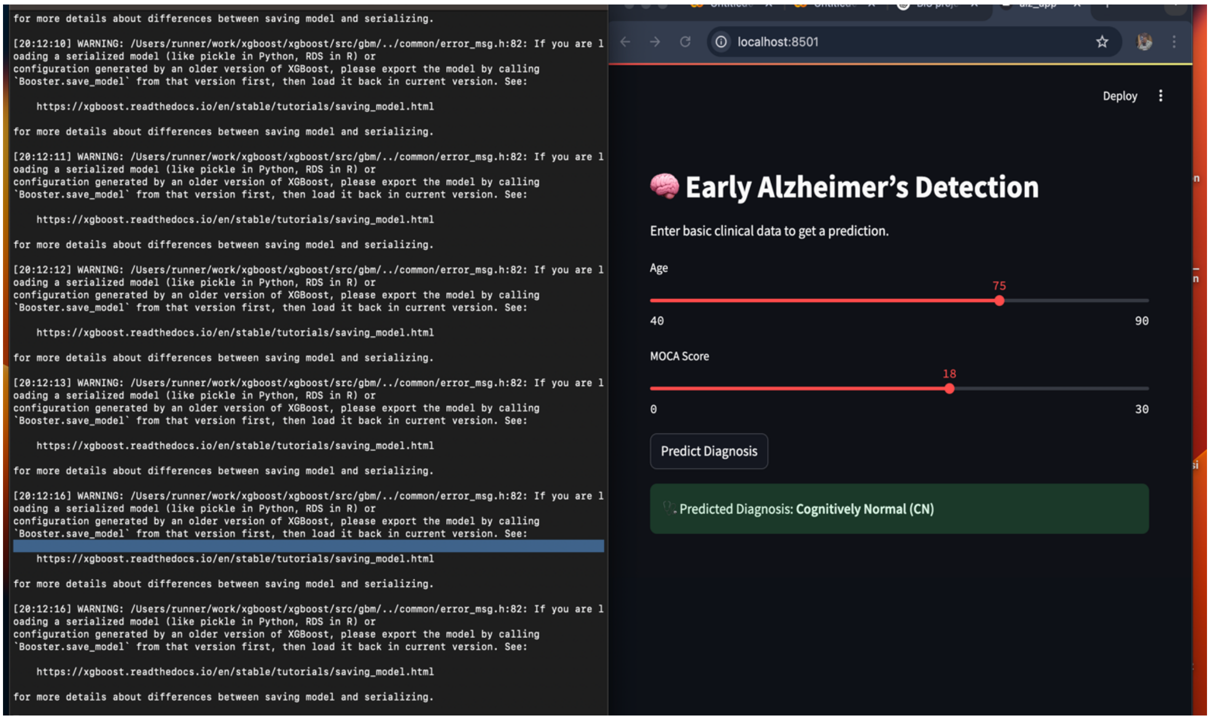
Enter basic clinical data to get a prediction.

Age: 75 (40–90)

MOCA Score: 18 (0–30)

Predicted Diagnosis: **Cognitively Normal (CN)**

Figure F4: Input variation — Age: 74, MOCA: 17, predicting Cognitively Normal(CN).



```
for more details about differences between saving model and serializing.

[20:12:10] WARNING: /Users/runner/work/xgboost/xgboost/src/gbm/../common/error_msg.h:82: If you are l
oading a serialized model (like pickle in Python, RDS in R) or
configuration generated by an older version of XGBoost, please export the model by calling
`Booster.save_model` from that version first, then load it back in current version. See:

    https://xgboost.readthedocs.io/en/stable/tutorials/saving_model.html

for more details about differences between saving model and serializing.

[20:12:11] WARNING: /Users/runner/work/xgboost/xgboost/src/gbm/../common/error_msg.h:82: If you are l
oading a serialized model (like pickle in Python, RDS in R) or
configuration generated by an older version of XGBoost, please export the model by calling
`Booster.save_model` from that version first, then load it back in current version. See:

    https://xgboost.readthedocs.io/en/stable/tutorials/saving_model.html

for more details about differences between saving model and serializing.

[20:12:12] WARNING: /Users/runner/work/xgboost/xgboost/src/gbm/../common/error_msg.h:82: If you are l
oading a serialized model (like pickle in Python, RDS in R) or
configuration generated by an older version of XGBoost, please export the model by calling
`Booster.save_model` from that version first, then load it back in current version. See:

    https://xgboost.readthedocs.io/en/stable/tutorials/saving_model.html

for more details about differences between saving model and serializing.

[20:12:13] WARNING: /Users/runner/work/xgboost/xgboost/src/gbm/../common/error_msg.h:82: If you are l
oading a serialized model (like pickle in Python, RDS in R) or
configuration generated by an older version of XGBoost, please export the model by calling
`Booster.save_model` from that version first, then load it back in current version. See:

    https://xgboost.readthedocs.io/en/stable/tutorials/saving_model.html

for more details about differences between saving model and serializing.

[20:12:16] WARNING: /Users/runner/work/xgboost/xgboost/src/gbm/../common/error_msg.h:82: If you are l
oading a serialized model (like pickle in Python, RDS in R) or
configuration generated by an older version of XGBoost, please export the model by calling
`Booster.save_model` from that version first, then load it back in current version. See:

    https://xgboost.readthedocs.io/en/stable/tutorials/saving_model.html

for more details about differences between saving model and serializing.

[20:12:16] WARNING: /Users/runner/work/xgboost/xgboost/src/gbm/../common/error_msg.h:82: If you are l
oading a serialized model (like pickle in Python, RDS in R) or
configuration generated by an older version of XGBoost, please export the model by calling
`Booster.save_model` from that version first, then load it back in current version. See:

    https://xgboost.readthedocs.io/en/stable/tutorials/saving_model.html

for more details about differences between saving model and serializing.
```

🧠 Early Alzheimer's Detection

Enter basic clinical data to get a prediction.

Age: 75 (40–90)

MOCA Score: 18 (0–30)

Predicted Diagnosis: **Cognitively Normal (CN)**

Figure F5: End-to-end workflow validation — prediction pipeline running locally on Mac via Streamlit.

These screenshots demonstrate how the trained XGBoost model was successfully integrated into a web-based Streamlit application, enabling real-time Alzheimer's prediction with simple clinical inputs.