Project Description

This project aims to develop a machine learning-based system for early detection and classification of Alzheimer's disease using the Open Access Series of Imaging Studies (OASIS) dataset. The system will analyze tabular clinical and demographic data to predict cognitive decline and classify patients into different stages of dementia. By leveraging advanced machine learning algorithms, we will create a predictive model that can assist healthcare professionals in early diagnosis and intervention planning for Alzheimer's disease.

The project will focus on preprocessing clinical data, feature engineering from neuroimaging measurements and cognitive assessments, and building robust classification models to distinguish between non-demented, very mild dementia, mild dementia, and moderate dementia cases.

Group Members & Roles

Member 1: Shady Kishk

Role: *Team Leader / Data Scientist*

Responsibilities:

- Coordinates project workflow and ensures team collaboration.
- Leads model selection, feature engineering, and training.
- Oversees model evaluation and final integration.

Member 2 Mostafa Ali

Role: *Data Analyst*

Responsibilities:

- Handles data collection, cleaning, and preprocessing.
- Conducts exploratory data analysis (EDA).
- Identifies trends and correlations within the dataset.

Member 3: Kirolos Emad

Role: *Machine Learning Engineer*

Responsibilities:

- Builds and tunes predictive models (Random Forest, XGBoost, etc.).
- Optimizes hyperparameters and evaluates performance.
- Ensures reproducibility of ML experiments.

Member 4: Mohamed Sherif

Role: *AI Model Developer*

Responsibilities:

- Implements, trains, and optimizes deep learning models.
- Experiments with different architectures (ANN, CNN, etc.)
- Compares traditional ML models with neural network approaches.

Member 5: Amr Ahmed

Role: *Model Evaluation & Deployment Engineer*

Responsibilities:

- Evaluates model performance using accuracy, F1-score, and confusion matrices.
- Deploys the final predictive model for demonstration.
- Contributes to report writing and literature review.

Team Leader

Name: Shady Samy Kishk (shadykishk77@gmail.com)

Role: Responsible for task coordination, timeline management, and reviewing final submissions to ensure project quality.

Objectives

1. **Data Preparation**: Clean and preprocess the OASIS tabular dataset, handling missing values and ensuring data quality for reliable model training.
2. **Feature Engineering**: Extract and engineer relevant features from clinical measurements including age, education level, socioeconomic status, MMSE scores, CDR ratings, eTIV, nWBV, and ASF.
3. **Model Development**: Build and train multiple machine learning classification models (Random Forest, XGBoost, Neural Networks) to predict Alzheimer's disease stages.
4. **Performance Optimization**: Achieve model accuracy of at least 85% with balanced precision and recall across all dementia stages.
5. **Deployment**: Create a RESTful API that can accept patient data and return predictions in real-time for clinical decision support.
6. **Clinical Validation**: Ensure the model's predictions are interpretable and clinically relevant for healthcare professionals.

Tools & Technologies

Programming Languages & Frameworks

- **Python 3.9+**: Primary programming language
- **scikit-learn**: Machine learning model development
- **Pandas & NumPy**: Data manipulation and numerical computing

Data Processing & Visualization

- **Pandas**: Data preprocessing and analysis
- **Matplotlib & Seaborn**: Data visualization and EDA
- **SMOTE (imbalanced-learn)**: Handling class imbalance

Deployment & API

- **Flask/FastAPI**: RESTful API development
- **Docker**: Containerization for deployment
- **AWS EC2/Azure ML**: Cloud deployment platform

- **Postman**: API testing

Development Tools

- **Jupyter Notebook**: Exploratory data analysis and prototyping
- **Git/GitHub**: Version control and collaboration
- **VS Code**: Development environment

Milestones, Key Deliverables & Deadlines

Milestone 1: Data Collection, Exploration & Preprocessing
Key Deliverables:

- EDA Report
- Interactive Visualizations
- Cleaned & Structured Dataset
  Deadline: October 20, 2025

Milestone 2: Data Analysis, Visualization & Feature Engineering
Key Deliverables:

- Data Analysis Report
- Enhanced Visualizations
- Feature Engineering Summary
  Deadline: October 30, 2025

Milestone 3: Model Development & Optimization
Key Deliverables:

- Trained Machine Learning / Deep Learning Models
- Hyperparameter Optimization Results

- Model Evaluation Report
  Deadline: November 10, 2025

Milestone 4: MLOps, Deployment & Monitoring
Key Deliverables:

- Deployed Predictive Model
- MLOps Pipeline Setup
- Monitoring Dashboard & Performance Logs
  Deadline: November 20, 2025

Milestone 5: Final Documentation & Presentation
Key Deliverables:

- Final Project Report
- Presentation Slides
- Demonstration Video (if applicable)
  Deadline: November 25, 2025

KPIs (Key Performance Indicators)

1. Data Quality

- **Percentage of missing values handled**: **≥ 98%**
  - Target: Successfully impute or handle at least 98% of missing data points using appropriate statistical methods (mean/median imputation for continuous variables, mode for categorical variables)
- **Data accuracy after preprocessing**: **≥ 99.5%**
  - Target: Ensure data integrity through validation checks, outlier detection, and consistency verification across all features

- **Dataset diversity (representation of different categories)**: **≥ 85%** ○ Target: Maintain balanced representation across age groups (60-70, 70-80, 80+), gender (male/female), education levels, and dementia stages (CDR 0, 0.5, 1, 2)

2. Model Performance

- **Model accuracy (Accuracy/F1-Score)**: **≥ 95%**
  - ○ Target Accuracy: ≥ **87%** overall classification accuracy
  - ○ Target F1-Score ≥ **85%** weighted F1-score across all dementia stages
- **Model prediction speed (Latency)**: **≤ 3.8 milliseconds**
  - ○ Target: Single prediction inference time under 50ms for real-time clinical use
  - ○ Batch prediction (100 patients): ≈ **0.38 seconds**
- **Error rate (False Positive/False Negative Rate)**: **6% / 10%**
  - ○ False Positive Rate: ≤ **8%** (healthy patients misclassified as having dementia)
  - ○ False Negative Rate: ≤**10%** (dementia cases missed - more critical, but acceptable for early-stage screening)
  - ○ Emphasis on minimizing false negatives for clinical safety

4. Business Impact & Practical Use

- **Reduction in manual effort**: **≥ 60%**
  - Target: Automate preliminary screening process, reducing clinician time spent on initial assessment
  - Estimated time savings: 15-20 minutes per patient evaluation
- **Expected cost savings**: **≥ 40%**
  - Target: Reduce costs associated with unnecessary advanced imaging (MRI/PET scans) through accurate preliminary screening ○ Early detection leading to reduced long-term care costs
- **User satisfaction**: **≥ 85%**
  - Target: Achieve high satisfaction scores from healthcare professionals using the system

- Measured through: ease of use (≥90%), prediction reliability (≥85%), clinical utility (≥80%), and documentation quality (≥85%)
- Collected via post-deployment surveys and feedback forms from pilot users