Machine Learning Report

Course: Introduction to Machine Learning (instructor: Dr. Eng. Krzysztof Smółka)

Comparative analysis of machine learning methods for the classification of neurodegenerative diseases

Date: 2025-01-21

Abstract

The goal of this project is to develop a classification model to predict patients' cognitive status based on demographic data, neuropsychological test results, and MRI volumetric measures (OASIS-2). Preprocessing includes imputation of missing data (median for SES, removal of MMSE when missing), one-hot encoding of categorical variables, and class distribution balancing with SMOTE. Random Forest (200 trees, class_weight='balanced') was selected as the classifier, as it handles heterogeneous features and nonlinearities well without the need for scaling. The model was evaluated with 5-fold cross-validation and a held-out test set, achieving macro-F1=0.93 and AUC-ROC=0.99. We additionally compared performance before and after feature selection, showing that four key attributes (MR Delay, SES, MMSE, CDR) ensure high interpretability with only a minimal drop in performance.

Keywords: dementia, random forest, SMOTE, OASIS-2, feature selection, neurodegenerative diseases

1 Introduction

Neurodegenerative diseases such as Parkinson's disease, Alzheimer's disease, and dementia are among the most pressing challenges in modern medicine and social care. Progressive loss of cognitive and motor functions negatively affects the quality of life of patients and their families, and increasing incidence in aging populations intensifies pressure on healthcare systems. Early detection of pathological changes can substantially extend the period during which supportive treatment and rehabilitation interventions remain effective. In recent years, the development of machine learning algorithms has enabled automated pattern recognition in large, high-dimensional datasets from neuroimaging, clinical tests, and phonation measurements.

This work presents a comparative analysis of selected classification methods — including tree-based ensemble models, support vector machines (SVM), gradient boosting algorithms, and

stacking techniques — on three publicly available datasets related to neurodegenerative diseases: UCI Parkinson's Telemonitoring, Alzheimer's Clinical & Demographic (Kaggle), and OASIS-2 Dementia Prediction. We applied unified preprocessing procedures across experiments (imputation, categorical encoding, and SMOTE oversampling) and evaluated the models via cross-validation and independent test sets.

Our objective is not only to compare the predictive performance of each algorithm (metrics: accuracy, precision, recall, F1-score, AUC-ROC) but also to identify trade-offs between predictive performance and interpretability — a key consideration for diagnostic deployment.

2 Materials and Methods

We used three open datasets: UCI Parkinson's Telemonitoring, Alzheimer's Clinical & Demographic (Kaggle), and OASIS-2 Dementia Prediction. After loading, we applied standardized preprocessing: (i) missing-value imputation — median for SES; removal of records without MMSE; (ii) one-hot encoding for all nominal variables; and (iii) standardization for models that are scale-sensitive (e.g., SVM, gradient boosting). Data were split into training (80%) and test (20%) sets with stratification, and in the training set we balanced class distribution using SMOTE.

We compared the following classifiers: (1) SVM, (2) Random Forest, (3) Gradient Boosting, and (4) Stacking ensembles. Models were trained with 5-fold cross-validation and evaluated on the held-out test set using accuracy, precision, recall, F1-score, and AUC-ROC. For OASIS-2, we additionally conducted feature selection using Random Forest feature importances to reduce predictors to the four most salient.

2.1 Datasets

Dataset	Source / link	Samples	Features	Label type	License
Alzheimer's disease dataset (Kaggle)	Kaggle	2,149	Clinical & demographic features incl. MMSE, BP, cholesterol, ADL, etc.	Binary: 0 – no diagnosis, 1 – Alzheimer's	CC BY 4.0
DARWIN (Diagnosis Alzheimer WIth Handwriting)	UCI	174	Handwriting dynamics features (air/paper time, pressure stats, etc.)	Binary: H – healthy, P – patient	CC BY 4.0
Augmented Alzheimer MRI	Kaggle	~40,000	MRI brain images (JPEG/PNG),	4-class: Non, Very Mild,	LGPL 3.0

Dataset		images	processed as pixel matrices	Mild, Moderate Demented	
Parkinson's (UCI, disease)	UCI	197	Voice features incl. jitter, shimmer, RPDE, DFA, spread1/2, PPE	Binary: 0 – healthy, 1 – Parkinson's	CC BY 4.0
Parkinson's (UCI, telemonitoring)	UCI	5,875	Demographics/time + extensive voice features	Regression: motor_UPDRS, total_UPDRS	CC BY 4.0
Parkinson's Disease Classification	UCI	756	Audio-derived features (acoustic, spectral, statistical)	Binary: 1 – Parkinson's, 0 – healthy	CC BY 4.0
Dementia Prediction Dataset (OASIS-2)	Mendeley	373 sessions (150 persons)	Demographics, MRI volumetry, cognitive tests (MMSE, CDR)	3-class: Nondemented, Demented, Converted	CC BY-NC 3.0

2.2 Machine Learning Methods

For the Parkinson's data we applied SMOTE (k_neighbors=2, random_state=42) to balance the dataset and used H2O AutoML (max_runtime_secs=600, balance_classes=False, seed=42), which selected StackedEnsemble_BestOfFamily_4 as the best model. For the Alzheimer's data we first applied SMOTE (random_state=42) and then trained a Gradient Boosting classifier (scikit-learn; StandardScaler + GradientBoostingClassifier(random_state=42)) on 34 clinical and demographic features; in the feature-selection variant, we used a MinMaxScaler → SelectKBest(chi², k=13) → GradientBoostingClassifier pipeline. For OASIS-2 we imputed SES (median) and removed rare missing MMSE values, one-hot encoded 'M/F' and 'Hand', performed an 80/20 stratified split (random_state=42), balanced the training set with SMOTE (k_neighbors=5, random_state=42), and chose RandomForestClassifier (200 trees, class_weight='balanced', random_state=42). We also tested feature selection via SelectFromModel(threshold='mean').

3 Experiments and Results

3.1 Parkinson's Telemonitoring (UCI)

After SMOTE balancing (k_neighbors=2, random_state=42), H2O AutoML (max_runtime_secs=600, balance_classes=False, seed=42) selected

StackedEnsemble_BestOfFamily_4, achieving AUC ≈ 0.999, accuracy 96.7%, recall 100%, and precision 92.9%. For comparison, literature reports include Little et al. (QDA) accuracy 91.8–95.4%, Dutta et al. (ANN) 95.89% accuracy and 93.75% precision, and Kumar et al. (Random Forest) 94.92% accuracy, F1 ≈ 95%, and AUC = 1.00.

3.2 Alzheimer's Clinical & Demographic (Kaggle)

We addressed class imbalance with SMOTE (random_state=42), then trained Gradient Boosting (scikit-learn; StandardScaler + GradientBoostingClassifier(random_state=42)) on the full set of 34 features. In the feature-selection variant, MinMaxScaler → SelectKBest(chi², k=13) → GradientBoostingClassifier achieved AUC ≈ 0.96, accuracy 91%, precision 95%, and recall 88%. In the literature, a CNN (INFEB Journal 2024) reported accuracy 88.65% (precision 88.84%, recall 88.65%, F1 88.62%), and Mahamud et al. (Voting LGBM+RF with SMOTE) reported accuracy 96.35% (precision 92%, recall 97%, F1 95%).

3.3 Dementia Prediction (OASIS-2)

Preprocessing involved SES imputation (median) and removal of missing MMSE, one-hot encoding of 'M/F' and 'Hand', an 80/20 stratified split (random_state=42), SMOTE balancing (k_neighbors=5, random_state=42), and RandomForestClassifier (200 trees, class_weight='balanced', random_state=42). Feature selection via SelectFromModel(threshold='mean') was also tested. On the full feature set, the model achieved: AUC-ROC = 0.99, Accuracy = 0.93, Precision = 0.93, Recall = 0.93, F1-score = 0.93. For comparison, Battineni et al. (SVM RBF) report accuracy 68.75% and precision 64.18%; Rawat et al. (stacking GBM+ANN) achieved accuracy 0.89; Vinayak et al. (XGBoost) reported accuracy 97.87%.

Table 1. Experimental results

Dataset	Task	Reference (best)	Our model (key params)	Our result
Parkinson's (UCI)	Disease classification from voice	Little 2007 (QDA) Acc 91.8–95.4%; Dutta 2018 (ANN) Acc 95.89%, Prec 93.75%; Kumar 2020 (RF) Acc 94.92%, F1≈95%, AUC=1.0	H2O AutoML StackedEnsemble; SMOTE(k=2), seed=42, 600s	AUC 0.999; Acc 96.72%; Prec 92.86%; Recall 100%; F1 96.30%
Alzheimer's (Kaggle)	Diagnosis from clinical & demographic features	CNN (INFEB 2024): Acc 88.65%; Mahamud 2025 (Voting LGBM+RF+SMOTE):	GradientBoosting; SelectKBest(chi²,k=13); SMOTE; random_state=42	AUC 0.96; Acc 0.91; Prec 0.95; Rec 0.88; F1 0.91

		Acc 96.35%, Prec 92%, Rec 97%, F1 95%		
Dementia Prediction (OASIS-2)	Dementia classification (3 classes)	Battineni 2019 (SVM RBF): Acc 68.75%; Vinayak 2020 (XGB): Acc 97.87%	RandomForest (200 trees, class_weight='balanced'); SMOTE(k=5)	AUC 0.99; Acc 0.93; Prec 0.93; Rec 0.93; F1 0.93

4 Discussion

Our experiments show that ensemble and boosting techniques consistently outperform baseline models across all three neurodegenerative disease classification tasks. For the UCI Parkinson's dataset, the H2O AutoML Stacked Ensemble (AUC ≈ 0.999, recall = 1.00) surpasses literature baselines (best SVM/RF with AUC=1.00 and accuracy up to 95%), benefiting from automatic multi-model ensembling and prior SMOTE oversampling. On the Alzheimer's task (Kaggle), Gradient Boosting with standardization and 13 selected features achieved AUC = 0.96 and F1 = 0.91, comparable to Mahamud et al. (LGBM + RF ensemble) and exceeding simpler CNN approaches. For OASIS-2, Random Forest with full preprocessing and SMOTE achieved AUC = 0.99 and F1 = 0.93, substantially improving over Battineni et al. (SVM RBF: accuracy ≈ 0.69) and close to XGBoost (accuracy ≈ 0.98).

All models benefited from unified preprocessing — median imputation, one-hot encoding, and SMOTE — which mitigated small-sample and class-imbalance issues. Feature selection (OASIS-2) reduced input dimensionality to four key attributes with only a minor metric drop (AUC from 0.99 to 0.97), suggesting the possibility of simpler, more interpretable models without major performance loss. The main limitations are dataset size and heterogeneity; to confirm generalizability, future work should include external cohort testing and long-term stability analyses under longitudinal patient monitoring.

Conclusions

Ensemble methods (stacking, boosting) and Random Forest — supported by SMOTE oversampling and careful preprocessing (imputation, one-hot encoding) — substantially outperform traditional SVMs in neurodegenerative disease classification tasks. The Stacked Ensemble achieved near-perfect AUC for Parkinson's, Gradient Boosting delivered high precision and recall for Alzheimer's detection, and Random Forest on OASIS-2 yielded AUC=0.99 and F1=0.93. Feature selection based on feature importance allowed reduction to four key predictors with a minimal performance drop, improving interpretability. Before clinical deployment, further validation on independent cohorts and stability analysis in long-term patient monitoring settings are required.

5 References

Little, M. A., et al. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE TBME 56(4), 1015–1022.

Dutta, D., & Banerjee, S. (2018). Classification of Parkinson's disease using neural networks with Levenberg–Marquardt algorithm. IJCA 181(35), 1–5.

Kumar, S., & Sharma, P. (2020). Random forest based classification of Parkinson's disease. IJMET 11(4), 145–155.

Battineni, G., Chintalapudi, N., & Amenta, F. (2019). ML in medicine: classification and prediction of dementia by SVM. Informatics in Medicine Unlocked 16, 100200.

Battineni, G., Amenta, F., & Chintalapudi, N. (2019). Data for: ML in medicine... Mendeley Data, V1. doi:10.17632/tsy6rbc5d4.1.

Rawat, R. M., et al. (2020). Dementia detection using ML by stacking models. Proc. ICCES 2020, 434–439.

Vinayak, S. S., Shahina, A., & Khan, A. N. (2020). Dementia prediction on OASIS using supervised & ensemble learning. IJEAT 10(1), 244–253.

Kaggle (2024). Alzheimer disease dataset. Accessed: 2025-06-20.

UCI (2024). DARWIN dataset. Accessed: 2025-06-20.

Kaggle (2024). Augmented Alzheimer MRI Dataset. Accessed: 2025-06-20.

UCI (2024). Parkinson's Telemonitoring Dataset. Accessed: 2025-06-20.

6 Attachments

Archive structure (/.zip):
data/Alzheimer/ {data.csv, alzheimers_disease_data.csv, Augmented Alzheimer_MRI_Dataset/...}
data/Dementia/ {dementia_dataset.csv, participants.tsv, TIHM_datasets/...}
data/Parkinsons/ {pd_speech_features.csv, ParkinsonsUCI/...}
notebooks/ {anhelina_mendohralo_dementia.ipynb, miraslau_alkhovik_alzheimer.ipynb, tymur_huselnykov_parkinsons.ipynb}

7 Annexes

Annex A. Dementia Prediction Dataset (OASIS-2) — attribute descriptions

Attribute	Description
Subject ID	Unique patient identifier
MRI ID	Unique MRI exam identifier
Group	Target class: Nondemented (healthy), Converted (MCI → dementia), Demented
Visit	Clinical visit number (temporal order)
MR Delay	Delay between MRI date and clinical assessment (days)
M/F	Sex (M = male, F = female)
Hand	Dominant hand (Left/Right)
Age	Age in years
EDUC	Years of education
SES	Socioeconomic status (1–5)
MMSE	Mini-Mental State Examination (0–30)
CDR	Clinical Dementia Rating (0 = none, 0.5 = MCI, ≥1 = dementia)
eTIV	Estimated Total Intracranial Volume (ml)
nWBV	Normalized Whole Brain Volume (% of cranial volume)
ASF	Atlas Scaling Factor

Annex B. Parkinson's Telemonitoring Dataset (UCI) — selected attributes

Attribute	Description
name	Patient name (ASCII) and recording number
MDVP:Fo(Hz)	Mean fundamental frequency of voice

MDVP:Fhi(Hz)	Max fundamental frequency
MDVP:Flo(Hz)	Min fundamental frequency
MDVP:Jitter(%)	Percent jitter of fundamental frequency
MDVP:Jitter(Abs)	Absolute jitter
MDVP:RAP	Relative Average Perturbation
MDVP:PPQ	Period Perturbation Quotient
Jitter:DDP	Derivative of Difference of Periods
MDVP:Shimmer	Amplitude variability
MDVP:Shimmer(dB)	Amplitude variability in dB
Shimmer:APQ3	Amplitude Perturbation Quotient (3 periods)
Shimmer:APQ5	Amplitude Perturbation Quotient (5 periods)
MDVP:APQ	Average amplitude irregularity
Shimmer:DDA	Derivative of Difference of Amplitude
NHR	Noise-to-Harmonics Ratio
HNR	Harmonics-to-Noise Ratio
status	Health status: 1 – Parkinson's, 0 – healthy
RPDE	Recurrence Period Density Entropy
D2	Correlation Dimension
DFA	Detrended Fluctuation Analysis
spread1	Nonlinear variability of fundamental frequency
spread2	Another nonlinear variability measure
PPE	Pitch Period Entropy

Annex C. Alzheimer's Clinical & Demographic Dataset (Kaggle) — categories

Patient identifier

Field	Description
PatientID	Unique patient identifier (4751–6900)

Demographics

Field	Description
Age	Age (60–90)
Gender	0 = male, 1 = female
Ethnicity	0: Caucasian, 1: African American, 2: Asian, 3: Other
EducationLevel	0: none, 1: high school, 2: bachelor, 3: higher

Lifestyle

Field	Description
BMI	Body mass index (15–40)
Smoking	0 = no, 1 = yes
AlcoholConsumption	Weekly units (0–20)
PhysicalActivity	Hours/week (0–10)
DietQuality	0–10
SleepQuality	4–10

Medical history

Field	Description
FamilyHistoryAlzheimers	0 = no, 1 = yes
CardiovascularDisease	0 = no, 1 = yes
Diabetes	0 = no, 1 = yes
Depression	0 = no, 1 = yes

HeadInjury	0 = no, 1 = yes
Hypertension	0 = no, 1 = yes

Clinical measures

Field	Description
SystolicBP	90–180 mmHg
DiastolicBP	60–120 mmHg
CholesterolTotal	150–300 mg/dL
CholesterolLDL	50–200 mg/dL
CholesterolHDL	20–100 mg/dL
CholesterolTriglycerides	50–400 mg/dL

Cognitive function

Field	Description
MMSE	Mini-Mental State Examination (0–30, lower = more impairment)
FunctionalAssessment	0–10, lower = more impairment
MemoryComplaints	0 = no, 1 = yes
BehavioralProblems	0 = no, 1 = yes
ADL	Activities of daily living, 0–10, lower = more impairment

Clinical symptoms

Field	Description
Confusion	0 = absent, 1 = present
Disorientation	0 = absent, 1 = present
PersonalityChanges	0 = absent, 1 = present
DifficultyCompletingTasks	0 = absent, 1 = present
Forgetfulness	0 = absent, 1 = present

Label

Field	Description
Diagnosis	0 = no diagnosis, 1 = Alzheimer’s

Confidential info

Field	Description
DoctorInCharge	Always "XXXConfid"