Porównawcza analiza metod uczenia maszynowego w klasyfikacji chorób neurodegeneracyjnych

Autorzy: Tymur Huselnykov[1], Miraslau Alkhovik[2], Anhelina Mendohralo[3]

Data: 21.01.2025

Streszczenie

Celem projektu jest opracowanie modelu klasyfikacyjnego do przewidywania stanu poznawczego pacjentów na podstawie danych demograficznych, wyników testów neuropsychologicznych oraz miar volumetrycznych z MRI (OASIS-2). Wstępne przetwarzanie obejmuje imputację braków (mediana dla SES, usunięcie MMSE), kodowanie one-hot zmiennych kategorycznych oraz wyrównanie rozkładu klas metodą SMOTE. Jako klasyfikator wybrano Random Forest (200 drzew, class_weight='balanced'), który bez potrzeby skalowania dobrze radzi sobie z heterogenicznymi cechami i nieliniowościami. Model oceniono za pomocą 5-krotnej walidacji krzyżowej i zestawu testowego, osiągając macro-F1=0,93 oraz AUC-ROC=0,99. Dodatkowo porównano wydajność przed i po selekcji cech, co pokazało, że cztery kluczowe atrybuty (MR Delay, SES, MMSE, CDR) zapewniają wysoką interpretowalność przy minimalnym spadku skuteczności.

Słowa kluczowe: demencja, random forest, SMOTE, OASIS-2, selekcja cech, choroby neurodegeneracyjne

1 Wstęp

Choroby neurodegeneracyjne, takie jak choroba Parkinsona, Alzheimer i demencja, stanowią jedno z najpoważniejszych wyzwań współczesnej medycyny i opieki społecznej [1–3]. Postępująca utrata funkcji poznawczych i ruchowych negatywnie wpływa na jakość życia pacjentów oraz ich rodzin, a rosnąca liczba zachorowań w starzejących się populacjach nasila presję na systemy ochrony zdrowia [4]. Wczesne wykrycie zmian patologicznych może istotnie wydłużyć okres, w którym możliwe jest skuteczne leczenie wspomagające oraz interwencje rehabilitacyjne [5]. W ostatnich latach rozwój algorytmów uczenia maszynowego umożliwił automatyczne

rozpoznawanie wzorców w dużych i wielowymiarowych zbiorach danych pochodzących z badań neuroobrazowych, testów klinicznych oraz pomiarów fonicznych [6,7].

Niniejsza praca prezentuje porównawczą analizę wybranych metod klasyfikacyjnych – w tym drzewiastych modeli ensemble, metod wektorów nośnych (SVM), algorytmów gradient boosting oraz technik stackingowych – na trzech publicznie dostępnych zbiorach związanych z chorobami neurodegeneracyjnymi: UCI Parkinson's Telemonitoring [8], Alzheimer's Clinical & Demographic (Kaggle) [9] oraz OASIS-2 Dementia Prediction [10]. W toku eksperymentów zastosowano jednolite procedury wstępnej obróbki danych, takie jak imputacja braków, kodowanie zmiennych kategorycznych oraz oversampling metodą SMOTE [11], a następnie oceniono modele przy pomocy walidacji krzyżowej i niezależnych zbiorów testowych.

Celem badania jest nie tylko porównanie skuteczności poszczególnych algorytmów (metryki: accuracy, precision, recall, F1-score, AUC-ROC), lecz także zidentyfikowanie kompromisów między wydajnością predykcji a interpretowalnością modeli, co stanowi istotny aspekt wdrożeniowy w diagnostyce chorób neurodegeneracyjnych.

2 Materiały i metody

W badaniu wykorzystano trzy otwarte zbiory danych: UCI Parkinson's Telemonitoring [8], Alzheimer's Clinical & Demographic (Kaggle) [9] oraz OASIS-2 Dementia Prediction [10]. Po wczytaniu zestawy poddano ujednoliconemu wstępnemu przetwarzaniu:

- **Imputacja braków** – mediana dla zmiennej SES; usunięcie rekordów bez wartości MMSE.
- **Kodowanie kategoryczne** – one-hot dla wszystkich zmiennych nominalnych.
- **Standaryzacja** – dla modeli wrażliwych na skalę cech, np. SVM czy gradient boosting [6].

Dane podzielono na zbiory treningowe (80 %) i testowe (20 %) z zachowaniem proporcji klas (stratified split) [7], a w zbiorze treningowym rozkład klas wyrównano metodą SMOTE [11].

Jako klasyfikatory porównano:

1. SVM
2. Random Forest
3. Gradient boosting
4. Stacking ensembles

Modele trenowano przy użyciu 5-krotnej walidacji krzyżowej oraz oceniano na odrębnym zbiorze testowym z metrykami: accuracy, precision, recall, F1-score i AUC-ROC [7].

Dla zestawu OASIS-2 dodatkowo przeprowadzono selekcję cech na podstawie ważności uzyskanej z Random Forest, redukując liczbę predyktorów do czterech najbardziej istotnych [7].

2.1 Zbiory danych

Tabela 1 Zestaw metadanych

Dataset	**Źródło / link**	**Liczba próbek**	**Cechy**	**Rodzaj etykiet**	**Licencja**
Alzheimer disease dataset (Kaggle) [9]	Alzheimer (Kaggle)	2 149	PatientID, Age, Gender, Ethnicity, EducationLevel, BMI, Smoking, AlcoholConsumption, DietQuality, SleepQuality, FamilyHistoryAlzheimers, CardiovascularDisease, Diabetes, Depression, HeadInjury, Hypertension, SystolicBP, DiastolicBP, CholesterolTotal, CholesterolLDL, CholesterolHDL, CholesterolTriglycerides, MMSE, FunctionalAssessment, MemoryComplaints, BehavioralProblems, ADL, Confusion, Disorientation, PersonalityChanges, DifficultyCompletingTasks, Forgetfulness, Diagnosis, DoctorInCharge	Klasyfikacja binarna: `0` – brak diagnozy `1` – choroba Alzheimera	Creative Commons Attribution 4.0 International (CC BY 4.0)
DARWIN (Diagnosis Alzheimer WIth Handwriting) [11]	Darwin (UCI)	174	ID, air_time1, disp_index1, gmrt_in_air1, gmrt_on_paper1, max_x_extension1, ..., paper_time25, pressure_mean25, pressure_var25, total_time25	Klasyfikacja binarna: `H` – osoba zdrowa `P` – pacjent z zaburzeniem	Creative Commons Attribution 4.0 International (CC BY 4.0)
Augumented Alzheimer MRI Dataset [12]	Alzheimer MRI (Kaggle)	~40 000 zdjęć	Obrazy MRI mózgu w formacie JPEG/PNG, rozmiar 176×208–256×256 px. Przetwarzane jako macierze pikseli (RGB lub grayscale)	Klasyfikacja kategoryczna (4 klasy): • Non Demented • Very Mild Demented • Mild Demented	GNU Lesser General Public License 3.0

				• Moderate Demented	
Parkinson's (UCI, disease) [8]	Parkinson's (UCI)	197	MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP, MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA, NHR, HNR, RPDE, D2, DFA, spread1, spread2, PPE	Klasyfikacja binarna: • 0 – osoba zdrowa • 1 – osoba z chorobą Parkinsona	Creative Commons Attribution 4.0 International (CC BY 4.0)
Parkinon's (UCI, telemonitoring) [8]	Parkinson's (UCI)	5 875	• Demograficzne/ czasowe: subject#, age, sex, test_time •Głosowe: Jitter (%), Jitter (Abs), Jitter: RAP, Jitter: PPQ5, Jitter: DDP, Shimmer, Shimmer (dB), Shimmer: APQ3, Shimmer: APQ5, Shimmer: APQ11, Shimmer: DDA, NHR, HNR, RPDE, DFA, PPE	Regresja (dwie zmienne ciągłe): • motor_UPDRS – ocena objawów ruchowych • total_UPDRS – całkowita ocena stanu pacjenta	Creative Commons Attribution 4.0 International (CC BY 4.0)
Parkinson's Disease Classification [13]	Parkinson's (UCI)	756	754 wyekstrahowane cechy z sygnału audio (akustyczne, spektralne, statystyczne), np.: tqwt_kurtosisValue_dec_32, tqwt_kurtosisValue_dec_33, tqwt_kurtosisValue_dec_34, tqwt_kurtosisValue_dec_35	Klasyfikacja binarna: • 1 – osoba chora na Parkinsona • 0 – osoba zdrowa	Creative Commons Attribution 4.0 International (CC BY 4.0)
Dementia Prediction Dataset [10]	Dementia (Mendeley)	373	Subject ID, MRI ID, Visit, MR Delay, M/F, Hand, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF	Klasyfikacja kategoryczna (3 klasy): • Nondemented • Demented • Converted	CC BY-NC 3.0
A dataset of EEG recordings from: Alzheimer's disease, Frontotemporal dementia and Healthy subjects [14]	Alzheimer's + Dementia (OpenEuro)	88	Surowe EEG (czasowe sygnały z 19 elektrod) W wersji przetworzonej: 19 niezależnych komponentów (ICA), opcjonalnie zredukowane cechy (np. pasma, entropia, moc). → liczba cech zależna od wersji: sygnał czasowy lub 20–100+ cech wyodrębnionych	Klasyfikacja kategoryczna (3 klasy): • AD (Alzheimer's Disease) • FTD (Frontotemporal Dementia) • CN (Cognitively Normal)	CC0
Remote Healthcare Monitoring In Dementia [15]	Dementia (Zenodo)	2803 dni danych od 56 uczestników (średnio ~50 dni / uczestnika)	- Activity: location_name, timestamp, patient_id - Sleep: sleep_state, heart_rate, respiratory_rate, snoring - Physiology: device_type, value, unit - Demographics: sex, age_group	Klasyfikacja kategoryczna (6 klas zdarzeń klinicznych): • Agitation • Weight • Heart rate • Body temperature • Blood pressure • Body water	Creative Commons Attribution 4.0 International

2.2 Metody uczenia maszynowego

W badaniu wykorzystano trzy główne zbiory danych: zestaw Parkinson's z UCI, obejmujący 31 osób (23 chorych na chorobę Parkinsona i 8 zdrowych) oraz 195 nagrań głosowych, zbiór Alzheimer's z Kaggle, zawierający informacje o 2149 pacjentach (1389 bez diagnozy i 760 z chorobą Alzheimera) oraz Dementia Dataset (OASIS-2; 150 osób, 373 sesje MRI).

W przypadku danych Parkinsona zastosowano technikę SMOTE (k_neighbors=2, random_state=42) do zbilansowania zbioru, a następnie H2O AutoML (max_runtime_secs=600, balance_classes=False, seed=42) wybrało jako najlepszy model Stacked Ensemble (StackedEnsemble_BestOfFamily_4).

Dla danych Alzheimer'a najpierw użyto SMOTE (random_state=42) do uzupełnienia brakującej klasy, a jako docelowy model wybrano Gradient Boosting z biblioteki scikit-learn (standard scaler + GradientBoostingClassifier z random_state=42) trenowany na pełnym zbiorze 34 cech klinicznych i demograficznych. W wariancie z selekcją cech zastosowano SelectKBest (chi2, k=13) w ramach potoku MinMaxScaler → SelectKBest → GradientBoostingClassifier.

Podczas analizy zestawu Dementia Prediction Dataset (OASIS-2; 150 osób, 373 sesje MRI) wstępne przetwarzanie obejmowało imputację wartości brakujących: mediana dla zmiennej SES oraz usunięcie nielicznych braków w MMSE. Następnie zakodowano cechy kategoryczne („M/F", „Hand") metodą one-hot. Podział na zbiór uczący i testowy wykonano w stosunku 80 %/20 % z zachowaniem rozkładu klas (stratify=y, random_state=42). W zbiorze uczącym ponownie zastosowano SMOTE (k_neighbors=5, random_state=42), aby wyrównać liczbę próbek w klasie „Converted" (~10 %). Ostatecznie wybrano jako główny algorytm RandomForestClassifier (200 drzew, class_weight='balanced', random_state=42), ze względu na jego odporność na heterogeniczne cechy, brak konieczności skalowania i wbudowaną regularyzację. Dodatkowo doświadczalnie przetestowano selekcję cech na podstawie SelectFromModel(threshold='mean') w celu oceny, czy zmniejszenie wymiaru wejścia poprawia interpretowalność przy minimalnym spadku wydajności.

3 Eksperymenty i wyniki

3.1 Parkinson's Telemonitoring (UCI [8])

Zbiór zrównoważono metodą SMOTE (k_neighbors=2, random_state=42) [11], po czym w H2O AutoML (max_runtime_secs=600, balance_classes=False, seed=42) jako najlepszy model wybrano **StackedEnsemble_BestOfFamily_4**, osiągając AUC ≈ 0,999, accuracy 96,7 %, czułość 100%, oraz również uzyskano wartość precyzji równącej się 92,9 % [12]. Dla porównania w literaturze: Little et al. (QDA) uzyskali accuracy 91,8–95,4 %, Dutta et al. (ANN Levenberg–Marquardt) 95,89 % accuracy i 93,75 % precision, a Kumar et al. (Random Forest) 94,92 % accuracy, F1 ≈ 95 % i AUC = 1,00.

3.2 Alzheimer's Clinical & Demographic (Kaggle [9])

Nadreprezentowaną klasę uzupełniono SMOTE (random_state=42) [11], następnie trenowano **Gradient Boosting** (scikit-learn, StandardScaler + GradientBoostingClassifier(random_state=42)) na pełnym zestawie 34 cech. W wariancie ze selekcją cech zastosowano potok MinMaxScaler → SelectKBest(chi², k=13) → GradientBoostingClassifier [14], co dało AUC ≈ 0,96, accuracy 91 %, precision 95% i także uzyskano wartość metryki recall równącej się 88%. W literaturze: CNN (INFEB Journal 2024) osiągnęło accuracy 88,65 %, precision 88,84 %, recall 88,65 % i F1 = 88,62 %; z kolei Mahamud et al. (Voting LGBM+RF ze SMOTE) raportują 96,35 % accuracy, 92 % precision, 97 % recall i F1 = 95 %.

3.3 Dementia Prediction (OASIS-2 [10])

Wstępne przetwarzanie objęło imputację braków (mediana dla SES, usunięcie rekordów bez MMSE) [6], kodowanie one-hot zmiennych „M/F" i „Hand" oraz podział na zbiór uczący i testowy (80 %/20 %) ze stratifikacją (random_state=42) [7]. Zbiór treningowy wyrównano SMOTE (k_neighbors=5, random_state=42) [11], a jako główny algorytm wybrano **RandomForestClassifier** (200 drzew, class_weight='balanced', random_state=42) [7], z uwagi na jego odporność na heterogeniczne cechy i brak konieczności skalowania. Dodatkowo przetestowano selekcję cech z użyciem SelectFromModel(threshold='mean') [15]. Wyniki modelu na pełnym zestawie:

- AUC-ROC = 0,99
- Accuracy = 0,93
- Precision = 0,93
- Recall = 0,93
- F1-score = 0,93

Dla porównania w literaturze: Battineni et al. (SVM RBF) raportują accuracy = 68,75 % i precision = 64,18 %; Rawat et al. (stacking GBM+ANN) osiągnęli accuracy = 0,89; natomiast Vinayak et al. (XGBoost) uzyskali accuracy = 97,87 %.

Tabela 2. Wyniki eksperymentów

Lp.	**Dataset**	**Zadanie**	**Wyniki z literatury – model (odniesienie do bibliografii)**	**Model z zespołu (podstawowe parametry)**	**Wynik zespołu**
1.1*	Parkinson's (UCI)	Klasyfikacja choroby Parkinsona na podstawie cech głosu	• Little et al. 2007 (QDA): Accuracy 91,8 %–95,4 % • Dutta et al. 2018 (ANN Levenberg–Marquardt): Accuracy 95,89 %, Precision 93,75 % • Kumar et al. 2020 (Random Forest): Accuracy 94,92 %, F1 ≈ 95 %, AUC 1,0	StackedEnsemble (H2O AutoML, max_runtime_secs=600, SMOTE (k_neighbors=2), seed=42)	• AUC 0,999 • Accuracy 96,72 % • Precision 92,86 % • Recall 100 % • F1 96,30 %
1.2	Alzheimer's (Kaggle)	Klasyfikacja choroby Alzheimera na podstawie cech klinicznych i demograficznych	• INFEB Journal 2024 (CNN): Accuracy 88,65 %, Precision 88,84 %, Recall 88,65 %, F1 88,62 % • IJISRT 2024 (Ensemble + 13 cech): Accuracy 94,08 %, Precision 94,74 %, Recall 93,42 %, F1 94,07 %, AUC 94,08 % • Mahamud et al. 2025 (Voting LGBM + RF z SMOTE): Accuracy 96,35 %, Precision 92 %, Recall 97 %, F1 95 %, AUC ~ 0,97–0,98	Gradient Boosting (scikit-learn, random_state=42) z SMOTE i SelectKBest(k=13, chi2)	• AUC 0,96 • Accuracy 0,91 • Precision 0,95 • Recall 0,88 • F1 0,91
1.3	Dementia Prediction Dataset	Klasyfikacja demencji na podstawie danych demograficznych oraz wyników neuroobrazowych i testów poznawczych	• Battineni et al. 2019 (SVM RBF): Accuracy 68.75%, Precision 64.18% • Rawat et al. 2020 (Stacking GBM + ANN): Accuracy 0.89 • Vinayak et al. 2020 (XGBoost): Accuracy 97.87%	Random Forest (scikit-learn, n_estimators=200, class_weight='balanced') + SMOTE(k_neighbors=5)	• AUC 0.99 • Accuracy 0.93 • Precision 0.93 • Recall 0.93 • F1 0.93

* nr_osoby_wg_listy_autorów.nr_zbioru_danych

4 Dyskusja

Przeprowadzone eksperymenty wykazały, że techniki zespołowe i boostingowe konsekwentnie przewyższają modele bazowe we wszystkich trzech zadaniach klasyfikacji chorób neurodegeneracyjnych. Dla zbioru Parkinson's (UCI) najlepszy okazał się model Stacked Ensemble wygenerowany przez H2O AutoML (AUC ≈ 0,999, recall = 1,00), co przewyższa wyniki z literatury (najlepsze SVM/RF z AUC = 1,00 i accuracy do 95 %) – efekt automatycznego łączenia wielu algorytmów oraz wstępnego oversamplingu SMOTE. W zadaniu Alzheimer's (Kaggle) Gradient Boosting ze standaryzacją i selekcją 13 cech osiągnął AUC = 0,96 oraz F1 = 0,91, co jest porównywalne z wynikami Mahamud et al. (ensemble LGBM + RF) i przewyższa prostsze podejścia CNN. W predykcji demencji na danych OASIS-2 Random Forest po pełnej obróbce i SMOTE uzyskał AUC = 0,99 i F1 = 0,93, znacząco poprawiając wynik Battineni et al. (SVM RBF: accuracy ≈ 0,69) oraz dorównując XGBoost (accuracy ≈ 0,98).

Wszystkie modele odniosły korzyść z ujednoliconego preprocessing'u – imputacji medianą, kodowaniu one-hot i oversamplingu SMOTE – co skutecznie niwelowało problemy niewielkich zbiorów i nierównowagi klas. Selekcja cech (OASIS-2) pozwoliła zmniejszyć wymiar wejścia do czterech najistotniejszych atrybutów przy jedynie niewielkim obniżeniu metryk (AUC z 0,99 do 0,97), co wskazuje na możliwość uproszczenia modeli przy zachowaniu wysokiej skuteczności.

Głównym ograniczeniem pozostaje rozmiar i hybrydowy charakter zestawów danych – aby potwierdzić ogólność ustaleń, potrzebne są dalsze testy na zewnętrznych kohortach oraz analiza stabilności w warunkach długofalowych obserwacji pacjentów.

Wnioski

W przeprowadzonych eksperymentach okazało się, że metody zespołowe (stacking, boosting) oraz Random Forest, wspierane oversamplingiem SMOTE i staranną obróbką danych (imputacja, one-hot encoding), znacząco przewyższają tradycyjne SVM w zadaniach klasyfikacji chorób neurodegeneracyjnych. Stacked Ensemble osiągnął niemal idealne AUC dla Parkinsona, Gradient Boosting zapewnił wysoką precyzję i czułość przy rozpoznawaniu Alzheimera, a Random Forest z OASIS-2 uzyskał AUC=0,99 i F1=0,93 w predykcji demencji. Dodatkowo selekcja cech na podstawie ważności pozwoliła zredukować zbiór predyktorów do czterech kluczowych zmiennych przy niewielkim obniżeniu metryk, co ułatwia interpretację wyników. Przed wdrożeniem klinicznym niezbędna jest jednak dalsza walidacja na niezależnych kohortach oraz analiza stabilności modeli w warunkach długoterminowego monitorowania pacjentów.

5 Bibliografia

[1] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig,

“Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease,”

IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[2] D. Dutta and S. Banerjee,

“Classification of Parkinson’s disease using neural networks with Levenberg–Marquardt algorithm,”

Int. J. Comput. Appl., vol. 181, no. 35, pp. 1–5, 2018.

[3] S. Kumar and P. Sharma,

“Random forest based classification of Parkinson’s disease,”

Int. J. Mech. Eng. Technol., vol. 11, no. 4, pp. 145–155, 2020.

[4] G. Battineni, N. Chintalapudi, and F. Amenta,

"Machine learning in medicine: classification and prediction of dementia by support vector machines (SVM),"

Informatics in Medicine Unlocked, vol. 16, p. 100200, 2019.

[5] G. Battineni, F. Amenta, and N. Chintalapudi,

"Data for: MACHINE LEARNING IN MEDICINE: CLASSIFICATION AND PREDICTION OF DEMENTIA BY SUPPORT VECTOR MACHINES (SVM),"

Mendeley Data, V1, 2019, doi: 10.17632/tsy6rbc5d4.1.

[6] R. M. Rawat, M. Mithil, M. Akram, and S. S. Pradeep,

"Dementia detection using machine learning by stacking models,"

in Proc. 5th Int. Conf. Communication and Electronics Systems (ICCES),

Coimbatore, India, Jun. 2020, pp. 434–439, doi: 10.1109/ICCES48766.2020.9137852.

[7] S. S. Vinayak, A. Shahina, and A. N. Khan,

"Dementia prediction on OASIS dataset using supervised and ensemble learning techniques,"

Int. J. Eng. Adv. Technol. (IJEAT), vol. 10, no. 1, pp. 244–253, Oct. 2020.

[8] "Alzheimer disease dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/some/path. [Accessed: Jun. 20, 2025].

[9] “DARWIN (Diagnosis Alzheimer WIth Handwriting) dataset,” UCI Machine Learning Repository, 2024. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/SomePath. [Accessed: Jun. 20, 2025].

[10] “Augmented Alzheimer MRI Dataset,” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/some/otherpath. [Accessed: Jun. 20, 2025].

[11] “Parkinson’s Telemonitoring Dataset,” UCI Machine Learning Repository, 2024. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/AnotherPath. [Accessed: Jun. 20, 2025].

6 Załączniki

```
/.zip
 │
 ├── data/
 │    ├── Alzheimer/
 │    │    ├── data.csv
 │    │    ├── alzheimers_disease_data.csv
 │    │    └── Augmented Alzheimer_MRI_Dataset/
 │    │         ├── MildDemented/
 │    │         ├── ModerateDemented/
 │    │         ├── NonDemented/
 │    │         ├── VeryMildDemented/
 │    │         ├── alzheimers_classes_grid.png
 │    │         └── alzheimers_classes_grid_labeled.png
 │    │
 │    ├── Dementia/
 │    │    ├── dementia_dataset.csv
 │    │    ├── participants.tsv
 │    │    └── TIHM_datasets/
 │    │         ├── Activity.csv
 │    │         ├── Demographics.csv
 │    │         ├── Labels.csv
 │    │         ├── Physiology.csv
 │    │         └── Sleep.csv
 │    │
 │    └── Parkinsons/
 │         ├── pd_speech_features.csv
 │         └── ParkinsonsUCI/  (oryginalne pliki z UCI)
 │
 └── notebooks/
     ├── anhelina_mendohralo_dementia.ipynb
     ├── miraslau_alkhovik_alzheimer.ipynb
     └── tymur_huselnykov_parkinsons.ipynb
```

7 Aneksy

Tabela 1. Dementia Prediction Dataset (OASIS-2)

Atrybut	Opis
Subject ID	Unikalny identyfikator pacjenta
MRI ID	Unikalny identyfikator badania MRI
Group	Klasa docelowa: Nondemented (zdrowi), Converted (przejście MCI → demencja), Demented (demencja)
Visit	Numer wizyty klinicznej (kolejność pojawiania się wizyt)
MR Delay	Opóźnienie między dniem badania MRI a oceną kliniczną (dni)
M/F	Płeć pacjenta (M = mężczyzna, F = kobieta)
Hand	Dominująca ręka (Left lub Right)
Age	Wiek pacjenta (lata)
EDUC	Liczba lat wykształcenia
SES	Status społeczno-ekonomiczny (skala 1–5)
MMSE	Wynik testu Mini-Mental State Examination (0–30 pkt)
CDR	Clinical Dementia Rating (0 = brak, 0.5 = MCI, ≥ 1 = demencja)
eTIV	Szacowana całkowita objętość czaszki (Estimated Total Intracranial Volume, ml)
nWBV	Znormalizowana objętość mózgu (Normalized Whole Brain Volume, % objętości czaszki)
ASF	Atlas Scaling Factor (współczynnik skalowania atlasu mózgu)

Tabela 2. Parkinson's Telemonitoring Dataset (UCI)

Atrybut	Opis
name	Nazwa pacjenta (ASCII) i numer nagrania
MDVP:Fo(Hz)	Średnia podstawowa częstotliwość głosu
MDVP:Fhi(Hz)	Maksymalna częstotliwość podstawowa
MDVP:Flo(Hz)	Minimalna częstotliwość podstawowa
MDVP:Jitter(%)	Procentowa zmienność częstotliwości podstawowej
MDVP:Jitter(Abs)	Bezwzględna zmienność częstotliwości podstawowej
MDVP:RAP	Relative Average Perturbation – miara drgań częstotliwości
MDVP:PPQ	Period Perturbation Quotient – kolejna miara nieregularności
Jitter:DDP	Derivative of Difference of Periods – złożona miara zmienności
MDVP:Shimmer	Zmienność amplitudy sygnału głosowego
MDVP:Shimmer(dB)	Zmienność amplitudy wyrażona w decybelach
Shimmer:APQ3	Amplitude Perturbation Quotient – 3 okresy
Shimmer:APQ5	Amplitude Perturbation Quotient – 5 okresów
MDVP:APQ	Średnia nieregularność amplitudy
Shimmer:DDA	Derivative of Difference of Amplitude – bardziej złożona metryka
NHR	Noise-to-Harmonics Ratio – stosunek szumu do sygnału
HNR	Harmonics-to-Noise Ratio – stosunek sygnału do szumu
status	Status zdrowotny pacjenta: 1 – Parkinson, 0 – zdrowy
RPDE	Recurrence Period Density Entropy – miara nieliniowej złożoności
D2	Correlation Dimension – miara złożoności fazowej
DFA	Detrended Fluctuation Analysis – wykładnik fraktalny sygnału
spread1	Nieliniowa miara zmienności częstotliwości podstawowej
spread2	Kolejna nieliniowa miara zmienności częstotliwości podstawowej

Atrybut	Opis
PPE	Pitch Period Entropy – entropia zmienności częstotliwości podstawowej

Tabela 3. Alzheimer's Clinical & Demographic Dataset (Kaggle)

Kategoria	Nazwa	Opis
Identyfikator pacjenta	PatientID	Unikalny identyfikator pacjenta (4751–6900).
Dane demograficzne	Age	Wiek pacjenta (60–90 lat).
	Gender	Płeć: 0 = mężczyzna, 1 = kobieta.
	Ethnicity	Grupa etniczna (0: Caucasian, 1: African American, 2: Asian, 3: Other).
	EducationLevel	Poziom wykształcenia (0: brak, 1: szk. średnia, 2: licencjat, 3: wyższe).
Styl życia	BMI	Wskaźnik masy ciała (15–40).
	Smoking	Palenie tytoniu (0 = nie, 1 = tak).
	AlcoholConsumption	Tygodniowe spożycie alkoholu (0–20 jednostek).
	PhysicalActivity	Aktywność fizyczna w godzinach tygodniowo (0–10).
	DietQuality	Jakość diety (0–10).
	SleepQuality	Jakość snu (4–10).
Historia medyczna	FamilyHistoryAlzheimers	Wywiad rodzinny w kierunku Alzheimera (0 = nie, 1 = tak).
	CardiovascularDisease	Choroba układu krążenia (0 = nie, 1 = tak).
	Diabetes	Cukrzyca (0 = nie, 1 = tak).
	Depression	Depresja (0 = nie, 1 = tak).
	HeadInjury	Uraz głowy w wywiadzie (0 = nie, 1 = tak).
	Hypertension	Nadciśnienie (0 = nie, 1 = tak).

Kategoria	Nazwa	Opis
Pomiar kliniczny	SystolicBP	Ciśnienie skurczowe (90–180 mmHg).
	DiastolicBP	Ciśnienie rozkurczowe (60–120 mmHg).
	CholesterolTotal	Całkowity cholesterol (150–300 mg/dL).
	CholesterolLDL	Cholesterol LDL (50–200 mg/dL).
	CholesterolHDL	Cholesterol HDL (20–100 mg/dL).
	CholesterolTriglycerides	Trójglicerydy (50–400 mg/dL).
Funkcjonowanie poznawcze	MMSE	Wynik Mini-Mental State Examination (0–30, niższe = większe upośledzenie).
	FunctionalAssessment	Ocena funkcjonalna (0–10, niższe = większe upośledzenie).
	MemoryComplaints	Skargi na pamięć (0 = nie, 1 = tak).
	BehavioralProblems	Problemy behawioralne (0 = nie, 1 = tak).
	ADL	Ocena codziennych czynności (0–10, niższe = większe upośledzenie).
Objawy kliniczne	Confusion	Splątanie (0 = brak, 1 = obecne).
	Disorientation	Dezorientacja (0 = brak, 1 = obecna).
	PersonalityChanges	Zmiany osobowości (0 = brak, 1 = obecne).
	DifficultyCompletingTasks	Trudności w wykonywaniu zadań (0 = brak, 1 = obecne).
	Forgetfulness	Zapominanie (0 = brak, 1 = obecne).
Etykieta klasyfikacji	Diagnosis	Diagnoza choroby Alzheimera (0 = brak, 1 = obecność).
Informacja poufna	DoctorInCharge	Dane o lekarzu prowadzącym (zawsze: "XXXConfid").