

A Machine Learning-Based Framework for differential screening of Neurodegenerative Dementias using routine clinical data

Davide Vilella

University of Bari Aldo Moro at Pia Fondazione "Card. G. Panico"

Benedetta Tafuri

University of Salento

Daniele Urso

Daniele.urso@etu.unige.ch

University Hospital of Geneva

Eleonora Rollo

University of Bari Aldo Moro at Pia Fondazione "Card. G. Panico"

Stefano Giannoni-Luza

Clinica Viarnetto

Giuseppe Gigli

National Research Council (CNR-NANOTEC)

Bruno Tartaglione

University of Bari Aldo Moro

Giancarlo Logroscino

University of Bari Aldo Moro at Pia Fondazione "Card. G. Panico"

Research Article

Keywords: Machine learning, Alzheimer's disease, Lewy Body Dementia, Frontotemporal Dementia, Vascular Dementia, differential diagnosis

Posted Date: October 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10815316/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

The increasing prevalence of neurocognitive disorders (NCDs), including Alzheimer's disease (AD), vascular dementia (VaD), dementia with Lewy bodies (DLB), primary progressive aphasia (PPA), and behavioral variant frontotemporal dementia (bvFTD), highlights the need for accessible tools to support early diagnosis. Machine learning (ML) may enhance diagnostic accuracy by leveraging routinely collected clinical information. We evaluated an ML pipeline based exclusively on first-visit neurological assessments to support differential diagnosis of major NCDs.

Methods

We retrospectively analysed clinical data from 932 patients. A CatBoost multiclass classifier was trained using 43 routinely available variables, including demographic characteristics, neurological signs, and symptom chronology. Bayesian optimization (Optuna) was used to maximize the macro-averaged F1-score while addressing class imbalance. Model interpretability was assessed through SHAP (SHapley Additive exPlanations), enabling patient-level and global feature attribution. Performance was evaluated on a held-out test set using accuracy, precision, recall, and F1-score.

Results

The model achieved 73.8% overall accuracy, with a macro F1-score of 0.646 and a weighted F1-score of 0.743. Diagnostic performance was highest for AD (F1 = 0.821), followed by DLB (0.704), bvFTD (0.681), PPA (0.667), and VaD (0.357). SHAP analysis identified first symptom, age at diagnosis, and language symptom chronology as the most influential predictors, closely reflecting established clinical diagnostic criteria.

Conclusions

Symptom chronology emerged as a robust diagnostic discriminator, underscoring the clinical value of temporal disease trajectories. The prototype Streamlit interface demonstrated practical clinical deployment by providing probabilistic predictions with transparent explanations. These findings support clinically driven, biomarker-free ML models as scalable first-line decision-support tools, particularly in resource-limited settings, while emphasizing the need for prospective multicenter external validation before clinical implementation.

1. BACKGROUND

The proportion of elderly individuals in the global population is steadily increasing, with 1 in 11 people over the age of 65 in 2019, a figure projected to rise to 1 in 6 by 2050 according to the World Population Prospects 2019 (United Nations, 2019). Consequently, the prevalence of age-related conditions is rising worldwide. The Global Burden of Disease Study (GBD) estimates that the number of individuals with dementia will reach 152.8 million by 2050 [1]. Neurocognitive disorders (NCDs) comprise a wide spectrum of conditions, including Alzheimer's disease (AD) [2], vascular dementia (VaD) [3], dementia with Lewy bodies (DLB)[4], and frontotemporal lobar degeneration, such as the behavioural variant of frontotemporal dementia (bvFTD) [5] and primary progressive aphasia (PPA) [6].

The diagnosis of NCDs is challenging, as it requires specialized neurological expertise and a multidisciplinary approach. A decline from a previous level of cognitive functioning, accompanied by interference in daily activities, is necessary to establish a diagnosis of minor or major NCD [7]. Furthermore, biological and differential diagnosis often demands ancillary investigations, such as high-field magnetic resonance imaging (MRI) and fluid biomarker assessment. However, implementing such diagnostic strategies can be difficult even in medium- and high-income countries [8, 9], largely due to healthcare system-related factors [10], including insufficient access to tertiary care facilities, long waiting lists, and limited infrastructure to perform advanced diagnostic investigations [9, 11].

Underdetection of dementia remains a major issue in both low-income and high- to medium-income countries. Even in healthcare systems with advanced infrastructures, up to 80% of cases, particularly in the early stages, may go unrecognized according to previous studies and meta-analyses [12]. Factors contributing to this include the tendency of healthcare professionals to focus on overt symptoms, time constraints during consultations, insufficient dementia-specific training among primary care providers, and limited availability of diagnostic tools [10]. These barriers contribute to delays in detection, which hinder timely interventions essential for symptom management and for improving the quality of life of affected individuals.

In this context, machine learning (ML) has emerged as a transformative tool within diagnostic pathways, enabling the integration and analysis of heterogeneous biomedical data. Leveraging large datasets from neuroimaging, electronic health records, and clinical assessments, ML algorithms can uncover subtle and complex patterns beyond the capacity of conventional approaches. These models not only improve diagnostic accuracy but also facilitate earlier detection and more refined patient stratification, thereby supporting personalized treatment strategies. Moreover, ML holds promise for enhancing diagnostic pathways by processing clinical and anamnestic data collected during neurological consultations, potentially improving the early recognition of NCDs.

The clinical characterisation of NCD based on symptom profile and disease trajectory has gained increasing attention as a foundation for data-driven diagnostic support. Vöglein et al. demonstrated that the first reported symptom carries significant prognostic and diagnostic value across neurodegenerative diseases, with distinct first-symptom signatures differentiating AD, FTD, DLB and other syndromes [13]. These results support the clinical rationale for encoding symptom onset order as a structured feature in predictive models. More recently, Mekkes et al. applied natural language processing to large-scale electronic health records to identify clinically meaningful disease trajectories in neurodegenerative disorders, showing that temporal patterns of symptom accumulation can be automatically extracted from unstructured clinical data and used to characterise subtype-specific progression profiles [14]. These previous works define the key background underlying the present study, namely that the identity and temporal order of symptom onset are informative biomarkers accessible from routine clinical documentation, and that machine learning approaches operating on structured symptom data can capture diagnostically relevant patterns without reliance on costly neuroimaging or fluid biomarkers.

Building on these findings, we propose a supervised classification pipeline trained on clinical examination variables and symptom onset sequences from the first neurological examination to improve the differential screening of five neurocognitive disorders subtype and to explore its potential utility as a support tool in routine diagnostic pathways and in primary-care settings.

2. METHODS

2.1. Study population and data collection

We retrospectively collected clinical information from 932 subjects evaluated at the Center for Neurodegenerative Diseases and Aging Brain of the University of Bari Aldo Moro, at Pia Fondazione Cardinale G. Panico, Tricase, who received a final diagnosis of minor and major NCD due to AD, VaD, bvFTD, PPA and DLB. All subjects underwent a standardized diagnostic protocol encompassing clinical evaluation, neuropsychological assessment, high-field (3-Tesla) brain magnetic resonance imaging and, whenever possible, lumbar puncture for biomarker of neurodegeneration assay. The study was conducted in accordance with ethical standards of the Helsinki Declaration and approved by the local health trust's ethics committees. Written consent was obtained from all participants.

2.2. Features description and missing data handling

The predictive model was built upon 43 clinical features, partitioned into four domains to capture both the phenotypic presentation and the temporal progression of the diseases. The first domain comprised 24 ordinal variables representing the chronological sequence of symptom onset (prefixed as '*Symp_*'). The second group included 11 categorical indicators derived from the neurological examination and baseline clinical signs ('*NeurEx_*'), binarized as '*Presence/Absence*'. Five continuous variables were also considered: age at diagnosis, education, total symptoms count, and the specific rank-order of the initial and final symptoms. Finally, sex and the name of the first and last symptoms were included as categorical features. A comprehensive breakdown of all variables is provided in Supplementary Table 1.

The management of missing data was tailored to the nature of each feature subset. To evaluate how these strategies influenced the model's robustness, we performed a sensitivity analysis (see Supplementary Material). The following protocols were implemented:

1. Symptom Onset Chronology: Missingness in this subset exhibited high variance, from 18.7% for memory impairment to 98.7% for stereotypies. These were imputed with a value of 0. The absence of an onset rank was interpreted as the symptom never having manifested during the observation period.
2. Neurological Examination: For '*NeurEx_*' variables, missing values were treated as an explicit, informative category "*missing*". Data revealed that 34.9% of the cohort lacked the entire neurological block, rather than individual signs being omitted at random. This pattern of missingness was heterogeneously distributed across diagnostic labels, ranging from 23.5% in Vascular Dementia to 44.2% in bvFTD. Such a distribution confirms a 'Missing Not At Random (MNAR)' mechanism [15]. By encoding "*missing*" as a distinct level, we enabled the ML classifier to leverage the structured absence of the exam as a potential diagnostic signal, while its native handling of categorical strings prevented the imposition of an arbitrary numerical order.
3. Demographic and Nominal Features: Sex and symptom name variables with missing entries were similarly assigned an explicit "*missing*" label to preserve the information related to data absence.
4. Continuous Variables: Parameters such as age at diagnosis, years of education, and the number of symptoms were fully observed across the cohort; therefore, no imputation was required.

2.3. Model implementation and optimization

A supervised multiclass classification task was implemented involving five distinct clinical classes. We employed CatBoost (Categorical Boosting) [16], a gradient-boosted decision tree chosen for its architectural advantages in handling heterogeneous clinical data. Unlike standard boosting algorithms, CatBoost manages categorical features internally through an ordered target statistics approach, eliminating the need for preliminary one-hot or label encoding. This was particularly advantageous for our neurological examination variables, preserving the intrinsic nominal or ordinal relationships within the feature set. Furthermore, the algorithm's native capacity to treat missing values as informative categories was central to our methodology. As previously noted, missingness in the

neurological block was not a stochastic process but a clinical state representing an unperformed assessment. CatBoost treats the "missing" string label as a first-class value during tree splitting, allowing the model to learn diagnostic patterns associated with the absence of the neurological assessment. To address the inherent class imbalance (e.g. AD vs. minority classes), we leveraged the *auto_class_weights='Balanced'* parameter. This mechanism automatically computes inverse-frequency class weights, adjusting the loss function to penalize misclassifications of minority classes proportionally, without requiring manual oversampling or undersampling that could introduce data leakage in the cross-validation loop. Lastly, gradient boosted decision trees have consistently demonstrated state-of-the-art performance on structured clinical tabular data [17], outperforming deep learning approaches in low-sample-size frameworks [18].

The cohort was partitioned into an 80% training set and a 20% held-out test set via stratified random sampling (*random_state = 42*). This stratification ensured that the phenotypic prevalence of all five diagnostic categories remained consistent across both subsets. The test set remained strictly held-out throughout all model development and hyperparameter optimization steps. It was used exclusively for the evaluation of model performance.

Model refinement was conducted through Optuna [19], a Bayesian optimization framework utilizing the Tree-structured Parzen Estimator (TPE). By modeling the conditional probability of the objective function, TPE directs the search toward high-performing regions of the hyperparameter space with significantly higher efficiency than traditional grid or random search. The optimization was driven by the macro-averaged F1-score to ensure that the predictive accuracy was not dominated by the majority class, thus maintaining sensitivity toward minority classes. We executed 15 optimization trials, each validated through a 3-fold stratified cross-validation scheme with an early stopping patience of 50 rounds to mitigate overfitting.

The optimal configuration yielded by this process (*learning_rate = 0.043*, *depth = 5*, *l2_leaf_reg = 5.546*, *random_strength = 0.551*, *bagging_temperature = 0.623*) was utilized to train the final classifier on the aggregate training data. This final iteration incorporated a maximum of 1,000 boosting steps, monitored via an internal 10% validation split. The final model was then evaluated once on the held-out test set. Computational analysis was performed using Python 3.13.5, with the following library versions: CatBoost v1.2.8, Optuna v4.5.0, scikit-learn v1.6.1, and SHAP v0.49.1.

2.4. Explainability analysis

To assess the clinical interpretability of the model and validate the decision process, SHAP (SHapley Additive exPlanations) values [20] were computed using the TreeExplainer algorithm, which provides exact Shapley value decompositions for tree-based models. SHAP values quantify the marginal contribution of each feature to a given prediction, relative to the expected model output over the training distribution. For a multiclass classifier, SHAP produces one set of values per class per sample, allowing the analysis of feature contributions both globally (across all patients) and at the diagnostic-subtype level. Global feature importance reports mean absolute SHAP values averaged across all classes and all test-set patients, ranking features by overall discriminative contribution to the model. On the contrary, per-class SHAP summary plots display, for each diagnostic subtype, SHAP values as a beeswarm plot, with feature values color-coded, allowing identification of the direction and magnitude of each feature's contribution to the classification of a specific dementia subtype.

2.5. Prototype Clinical Decision Support Tool

To demonstrate the translational potential of the proposed pipeline, a proof-of-concept web application was developed using Streamlit (Python), allowing clinicians to interact with the trained CatBoost model through a structured graphical interface without requiring any programming expertise. The tool accepts the same input

variables used during model training: demographic data (sex, age at diagnosis, years of education), neurological examination findings (11 binary *NeurEx* variables), and the onset order of 24 symptoms (integer scale 1–7, where 0 indicates the symptom was not reported). The five derived temporal features (first and last symptom name and order, total number of symptoms) are computed automatically from the onset inputs, replicating the feature engineering applied during training and eliminating manual entry errors. Upon submission, the application generates two outputs: (i) the predicted diagnostic subtype with the associated class probability for each of the five subtypes, and (ii) a patient-level SHAP bar chart showing the 12 features with the highest contribution to the predicted class, with direction (positive bars supporting the classification, negative bars opposing it). This per-patient explainability layer is intended to support clinical reasoning by making the model's decision transparent to the end user. The application is intended strictly as a proof-of-concept screening aid. It has not undergone prospective clinical validation and is not approved for diagnostic use. Screenshots of the interface are provided in Supplementary Material.

2.6. Statistical analysis

For each clinical variable, a comparison was conducted across the five diagnostic groups. Continuous variables were described as mean $\pm$ standard deviation (SD) and compared using one-way ANOVA. Normality was verified using the Shapiro-Wilk test and, in case of non-normality, the Mann-Whitney U test was applied for pairwise comparisons. Categorical variables were described as absolute frequency and percentage and compared using the Pearson Chi-square test. The Bonferroni correction for multiple comparisons was applied for both global tests and pairwise comparisons. Results with a Bonferroni-corrected p-value < 0.05 were considered statistically significant. All analyses were conducted in Python 3.13.5.

Given the multiclass nature of the problem and the pronounced class imbalance, we evaluated the model performance with macro F1-score, weighted F1-score, Per-class F1-score, Precision and Recall (Sensitivity) on the held-out test set. The confusion matrix was additionally computed and visualized to characterize the pattern of inter-class misclassifications, providing clinically interpretable information on which diagnostic subtypes are most frequently confused by the model.

3. RESULTS

3.1. Cohort description

The dataset comprised 932 patients with a neurodegenerative dementia diagnosis, distributed across five diagnostic subtypes: AD ($n = 566$), DLB ($n = 118$), PPA ($n = 76$), VaD ($n = 68$), and bvFTD ($n = 104$). Significant age differences were found (Bonferroni-corrected p-value < 0.001): DLB (74.4 ± 7.1 years) and VaD (75.3 ± 7.7 years) had a later age of onset than bvFTD (66.5 ± 9.4 years) and PPA (68.3 ± 8.4 years) patients. Pairwise comparisons showed significant differences between bvFTD and both AD, DLB and VaD ($p < 0.001$). Over neurological signs, as expected, parkinsonian gait was significantly higher in the DLB group ($p < 0.001$), while early language impairment was almost pathognomonic for the PPA group, and behavioural disturbances was prevalent in bvFTD (Table 1). The analysis of the *symptom onset order variables* confirmed that memory impairment was the most frequent first symptom in AD, whereas language dysfunction appeared significantly earlier in the PPA group.

Table 1

Demographic and clinical characteristics of the study cohort by diagnostic subtype. Continuous variables are expressed as mean $\pm$ standard deviation (SD). Categorical variables are expressed as absolute frequency and percentage. Statistical comparisons: ANOVA for continuous variables; Pearson's Chi-squared test for categorical variables. p-values are Bonferroni-corrected for multiple comparisons across all tested variables. Significant results (p Bonferroni < 0.05) are highlighted in green.

	Overall (n = 932)	AD (n = 566)	DLB (n = 118)	PPA (n = 76)	VaD (n = 68)	bvFTD (n = 104)	Test statistic	p Bonferroni
	Mean ± SD or n (%)					ANOVA or χ²		
Demographic characteristics								
Age at diagnosis (years)	71.3 ± 8.4	71.5 ± 7.9	74.4 ± 7.1	68.2 ± 8.4	75.3 ± 7.7	66.5 ± 9.4*	F = 20.78	< 0.001
Education (years)	9.0 ± 6.3	9.1 ± 7.1	8.1 ± 4.9	10.1 ± 4.8	7.8 ± 3.9	9.6 ± 4.9	F = 2.10	1.000
Female sex, n (%)	505 (54.2%)	345 (61.0%)	51 (43.2%)	35 (46.1%)	22 (32.4%)	52 (50.0%)	χ² = 37.10	< 0.001
Neurological examination findings, n (%)†								
Memory impairment	468 (50.2%)	310 (54.8%)	58 (49.2%)	28 (36.8%)	41 (60.3%)	31 (29.8%)	χ² = 30.28	< 0.001
Language impairment	205 (22.0%)	99 (17.5%)	18 (15.3%)	42 (55.3%)*	19 (27.9%)	27 (26.0%)	χ² = 61.20	< 0.001
Behavioural disturbance	149 (16.0%)	68 (12.0%)	21 (17.8%)	8 (10.5%)	10 (14.7%)	42 (40.4%)*	χ² = 54.80	< 0.001
Parkinsonian gait	109 (11.7%)	38 (6.7%)	41 (34.7%)*	7 (9.2%)	11 (16.2%)	12 (11.5%)	χ² = 76.09	< 0.001
Psychosis	37 (4.0%)	12 (2.1%)	15 (12.7%)	0 (0.0%)	3 (4.4%)	7 (6.7%)	χ² = 33.99	< 0.001
Mood disorder	254 (27.3%)	150 (26.5%)	39 (33.1%)	15 (19.7%)	21 (30.9%)	29 (27.9%)	χ² = 4.80	1.000
Agnosia	34 (3.6%)	19 (3.4%)	4 (3.4%)	0 (0.0%)	2 (2.9%)	9 (8.7%)	χ² = 10.55	0.482
Anosognosia	67 (7.2%)	33 (5.8%)	12 (10.2%)	5 (6.6%)	4 (5.9%)	13 (12.5%)	χ² = 7.75	1.000
Symptom burden								
Number of symptoms	3.5 ± 2.1	3.2 ± 1.8	5.0 ± 2.2	3.1 ± 2.0	2.8 ± 1.7	4.6 ± 2.4	F = 33.76	< 0.001
First symptom at onset, n (%)‡								
Memory impairment	536 (57.5%)	409 (72.3%)*	42 (35.6%)	29 (38.2%)	38 (55.9%)	18 (17.3%)	χ² = 154.11	< 0.001
Language impairment	156 (16.7%)	74 (13.1%)	5 (4.2%)	57 (75.0%)*	10 (14.7%)	10 (9.6%)	χ² = 207.78	< 0.001

† Neurological examination available for all 932 patients. ‡ Onset variable computed from symptom order data (value = 1 indicates first reported symptom; missing values imputed as 0, indicating symptom not reported). AD = Alzheimer's disease; DLB = Dementia with Lewy Bodies; PPA = Primary Progressive Aphasia; VaD = Vascular Dementia; bvFTD = behavioral variant Frontotemporal Dementia.

Symptom burden and temporal onset profile across diagnostic subtypes are summarized in Fig. 1, which integrates pie charts of first and most recent symptom category, the distribution of total symptom count, and the onset order density curves for six clinically relevant symptoms.

The pie charts of the first and most recent reported symptoms confirm distinct subtype-specific profiles (Fig. 1 a-b). Memory impairment was the predominant first symptom in AD (32%), while in bvFTD behavioural/mood symptoms dominated (54%). PPA showed the most concentrated profile, with language impairment as the first symptom in 51% of patients. The last reported symptom distribution reflects disease progression rather than onset and reveals converging patterns across subtypes. Behavioural/mood symptoms became prominent across all classes by the most recent assessment, notably in bvFTD (42%) and AD (25%). DLB was uniquely characterised by the progressive emergence of hallucinations/psychosis (26%) and motor/parkinsonian features (23%) as the most recent symptoms, consistent with the natural history of the disease. The total number of reported symptoms differed significantly across subtypes (Fig. 1c, Kruskal-Wallis $H = 105.40$, $p < 0.001$). DLB patients presented the highest symptom burden (median = 5, IQR 3–6), followed by bvFTD (median = 4, IQR 3–6), while VaD showed the lowest (median = 3, IQR 2–4). Pairwise comparisons confirmed significant differences between AD and DLB ($p < 0.001$), AD and bvFTD ($p < 0.001$), DLB and PPA ($p < 0.001$), DLB and VaD ($p < 0.001$), PPA and bvFTD ($p < 0.001$), and VaD and bvFTD ($p < 0.001$) after Bonferroni correction. The density plots of onset order for the six clinically relevant symptoms further characterise subtype-specific temporal trajectories (Fig. 1d). As expected, language impairment appeared as the first symptom in 75% of PPA patients, behavioural/apathy features were the first symptom in 32.7% of bvFTD and REM sleep behaviour disorder (RBD) emerged as the first symptom in 20.3% of DLB patients, supporting its value as an early clinical marker of DLB.

3.2. Model performances

The CatBoost classifier was trained on 745 patients and evaluated on the held-out test set ($n = 187$), achieved an overall accuracy of 73.8%, a macro-averaged F1-score of 0.646, and a weighted F1-score of 0.743 (Table 2).

Table 2
Overall classification performance of the CatBoost model on the held-out test set.

Metric	Value	95% CI (approx.)	Chance level*
Accuracy	73.8%	—	60.8%
Macro F1-score	0.646	—	0.200
Weighted F1-score	0.743	—	—
5-fold CV F1 macro	0.495 ± 0.028	[0.467, 0.523]	0.200

**Expected performance of a random classifier. Macro F1 chance = $1/5 = 0.200$. CV = 5-fold stratified cross-validation on the training set ($n = 745$).*

The predictive model showed varying performance levels across the different diagnostic groups, as summarized in the confusion matrix (Fig. 2, Table 3). Alzheimer’s Disease achieved the most robust results, with a precision of

0.879 and an F1-score of 0.821. The model correctly identified 87 out of 113 test patients (77.0%). DLB exhibited the highest recall among the minority classes at 0.792, with 19 of 24 test patients correctly identified. Primary Progressive Aphasia classification reached an F1-score of 0.667, correctly classifying 11 of 15 patients (73.3%). A misclassification rate of 13.3% toward AD was observed. bvFTD patients were classified with an F1-score of 0.681, with 16 of 21 test patients (76.2%) accurately predicted. The most frequent confusion occurred with DLB (14.3%). Vascular Dementia proved to be the most challenging class for the algorithm, yielding a precision, recall, and F1-score of 0.357. Only 5 out of 14 test patients were correctly classified, with a significant portion (42.9%) misclassified as AD.

Table 3
Per-class classification performances of the CatBoost model on the held-out test set.

Class	N° test	Precision	Recall	F1-score	Correct predictions
AD	113	0.879	0.770	0.821	87/113 (77.0%)
DLB	24	0.633	0.792	0.704	19/24 (79.2%)
PPA	15	0.611	0.733	0.667	11/15 (73.3%)
VaD	14	0.357	0.357	0.357	5/14 (35.7%)
bvFTD	21	0.615	0.762	0.681	16/21 (76.2%)

3.3. Feature importance: SHAP-based analysis

To assess the clinical interpretability of the model, SHAP values were computed on the test set using the TreeExplainer algorithm. At the global level (Fig. 3), the three most influential features were first symptom (mean |SHAP| = 0.221), age at diagnosis (0.197), and language symptom onset order (Language onset order = 0.177). The top 5 features accounted for 81.6% of the total mean |SHAP| in the top-15 set, indicating a concentrated importance structure where a small subset of variables drives most of the classification signal.

The SHAP importance analysis (Fig. 4) also revealed distinct predictive "fingerprints" for each diagnostic subtype. In AD, the identity of the first reported symptom (SHAP: 0.1914) and the specific onset rank of memory impairment (0.1902) emerged as the most discriminative features. Interestingly, biological sex (0.1765) occupied the third rank of importance, exerting a higher influence in AD than in any other subtype. DLB showed a more heterogeneous importance profile where language onset order was the top feature (0.3277), acting as a negative discriminator due to its typical preservation in early DLB. The diagnosis was strongly characterized by visual hallucinations onset (0.2483) and age at diagnosis (0.2363). Motor indicators, such as bradykinesia onset (0.1771) and parkinsonian gait (0.1168), also provided significant contributions. Primary Progressive Aphasia exhibited the most concentrated importance profile in the entire study. The language onset order achieved a SHAP value of 0.5891, the highest recorded across all classes and features. Both qualitative clinical observations (NeurEx: Language, 0.2349) and the temporal rank of language deficits were critical, underscoring the pathognomonic role of early focal linguistic decline in this group. For VaD, the model's decisions were heavily weighted toward demographic and general progression markers rather than specific neurological signs. Age at diagnosis was the primary driver (0.3725), followed by the identity of the first symptom (0.2120). Notably, the total number of symptoms (0.1422) were more influential, suggesting a more diffuse and complex clinical presentation for vascular cases. Finally, bvFTD was characterized by a strong emphasis on behavioral assessment and age. Age at diagnosis (0.3199) and the first symptom's nominal identity (0.2832) were the leading predictors. The model's reliance on qualitative behavioral signs (NeurEx: Behaviour, 0.2602) and specific temporal markers like change in habits (0.1797) and irritability (0.1549) confirms that the

algorithm successfully captured the early socio-emotional and personality shifts that define the bvFTD phenotype. All detailed results are reported in Supplementary Table 2.

4. DISCUSSION

The central finding of this study is that the temporal trajectory of symptom presentation carries independent diagnostic signal beyond the simpler presence or absence of individual clinical features. The present study demonstrates that a gradient boosting classifier trained exclusively on routine clinical variables achieves promising performances in the five-class differential diagnosis of neurodegenerative dementias, without any reliance on neuroimaging or fluid biomarkers. These results are in line with previous findings and even more interesting when compared with the performance of biomarker-based approaches [13, 14]. Indeed, multimodal classifiers combining neuropsychological assessment, MRI, and CSF data have reported higher discriminative accuracy in comparable diagnostic scenarios, and the MUQUBIA algorithm, trained on structural MRI features, achieved a macro F1-score of 0.91 across four dementia subtypes, while FDG-PET-based SVM classifiers reached per-class F1-scores between 0.74 and 0.87 for AD, DLB, and FTD [21]. The performance gap between the present work and these approaches is therefore expected and interpretable: it reflects the cost of operating without imaging data, rather than a methodological limitation of the pipeline itself. Crucially, the present model operates on variables already available at any outpatient neurological visit, making the observed accuracy a pragmatically relevant baseline for a first-line screening tool rather than a diagnostic benchmark to be compared against specialist workups.

With respect to diagnostic subtypes, classification performances reflect the clinical separability. The high precision and F1-score for Alzheimer's Disease confirm the model's ability to recognize the classic amnesic profile. Moreover, the high DLB recall is a significant finding, especially given the known clinical difficulty in distinguishing DLB from AD during early stages [22]. In the case of PPA, the confusion with AD (13.3%) may be attributed to "logopenic" presentations or cases where language impairment coexists with early episodic memory deficits, blurring the syndromic boundaries. It must be noted, however, that the small sample size for this subtype limits the statistical stability of these specific metrics. Finally, the most critical diagnostic challenge remains Vascular Dementia with low F1-score and the high misclassification rate as AD suggesting that symptom chronology alone may be insufficient to distinguish these two entities. VaD often shares the same cognitive "signature" as AD, namely memory impairment and advanced age [23], and lacks the highly specific markers (the parkinsonian gait seen in DLB or the language impairment in PPA) that drive the model's discriminative power in other classes. Of note, the confusion between bvFTD and DLB (14.3%) can be explained by the overlap in behavioural symptoms [24, 25]. Both conditions frequently present overlapping neuropsychiatric and behavioural features, including apathy, agitation and disinhibition, as well as the occurrence of DLB-like features such as psychosis and parkinsonism in bvFTD, particularly in genetic cases; in this cohort, the similar age at onset further complicated differentiation based solely on clinical phenotyping [25, 26].

Across all five subtypes, the SHAP-derived importance ranking is strongly consistent with established clinical diagnostic criteria and the group-level statistical differences. Moreover, the explainability results demonstrate that the chronological sequence of symptom onset serves as a key diagnostic indicator, often providing more discriminative power than the simple presence of clinical signs. The prominence of memory-related features in AD classification aligns with the amnesic phenotype that characterizes its typical presentation. Interestingly, the model uses "negative discriminative markers" to refine its diagnosis: the presence/absence of parkinsonian gait serves as a key pivot to rule out DLB and support an AD classification. In DLB, the high Shapley contribution of visual hallucinations onset is particularly noteworthy, as it confirms the strong diagnostic specificity of this symptom. Furthermore, the language onset order acts as a negative predictor: late or absent language involvement effectively distinguishes DLB from PPA. The temporal pattern is most evident in PPA, where the role of early language

impairment explains the majority of the predictive signal. For VaD, the model's limited discriminative capacity reflects clinical reality as depicted in the Shapley distribution of features. Indeed, the SHAP values indicate that VaD lacks a uniquely dominant signature symptom but as expected, shows a significant overlap with the AD profile, suggesting that without neuroimaging markers, symptom chronology alone may not be sufficient to isolate purely vascular cases from mixed presentations. Finally, the fact that bvFTD showed the highest SHAP value for the identity of the first symptom highlights that the onset type is the fingerprint for this pathology. The model successfully captured the behavioural fingerprint of bvFTD, prioritizing personality changes, irritability, and disinhibition over cognitive deficits, thereby emulating the logic of a clinician during the differential diagnosis process.

Overall, the alignment between data-driven importance and clinical knowledge constitutes evidence of face validity for the model's learned representations and supports the potential clinical utility of the pipeline as a screening tool. Importantly, the model does not rely on a single feature for any subtype: even PPA, which shows the most concentrated profile, draws on multiple neurological and symptomatic variables. This multi-feature dependency mirrors the syndromic, multi-domain nature of neurodegenerative dementia diagnosis in clinical practice. The SHAP analysis further reveals a complementary structure: features that are highly informative for one subtype tend to have lower SHAP contributions for others, consistent with a model that has learned subtype-specific, non-redundant representations. This finding supports the interpretability and robustness of the classification approach proposed in this study.

As a further innovation, we developed a prototype Clinical Decision Support System (CDSS) based on the proposed pipeline as a part of a growing body of research on machine learning-assisted differential diagnosis of neurodegenerative dementias [27, 28]. Prospective validation of the PredictND [29], CDSS demonstrated that algorithmic support not only influenced clinicians' final diagnoses but also increased their confidence in the classification, while van Gils et al. extended this approach to guide stepwise diagnostic testing across multiple clinical scenarios using multimodal data including MRI and CSF biomarkers [30]. A key limitation of these tools, however, is their reliance on neuroimaging and laboratory data that require substantial infrastructure, restricting applicability outside tertiary centres. The present study avoids this dependency, relying exclusively on variables collectable during a routine neurological examination and clinical interview. READi-Dem [31] demonstrated that a competitive ML-based dementia screening platform could be built on brief, cost-effective assessments without specialist testing; the present work extends this approach to a five-class differential diagnosis problem using a richer clinical feature set that additionally captures the temporal sequence of symptom onset. This design is particularly relevant given that low- and middle-income countries currently achieve a dementia diagnostic rate of only 21%, far below the WHO 2025 target of 50%, and that biomarker-based criteria remain inaccessible in resource-limited settings where diagnosis necessarily relies on clinical evaluation alone [32]. Consistent with current recommendations, the limitations of ML systems must be clearly communicated to clinical end users before integration into clinical workflows, and the tool presented here is intended strictly as a proof-of-concept screening aid. It provides probabilistic output and patient-level SHAP explainability to support clinician decision-making and has not undergone prospective validation. Future work should address external multicentre validation, clinician usability evaluation, and assessment of impact on diagnostic confidence and time-to-referral in real-world outpatient settings [33].

Several limitations of the present study should be acknowledged. First, the dataset derives from a single clinical centre, which limits the generalisability of both the classification model and the cohort characterisation to other populations, healthcare systems, and diagnostic practices. External validation on independent multicentre cohorts is necessary before any clinical deployment can be considered. Second, the class imbalance between subtypes represents a structural constraint on model performance for minority classes. Despite the application of balanced

class weights, VaD remained the most difficult subtype to classify, reflecting not only the imbalance but also the genuine clinical overlap between VaD and AD in the absence of neuroimaging data. A prospective enrichment of minority classes in future data collection could partially address this limitation. Third, the design choice motivated by the goal of low-resource applicability, with the absence of neuroimaging, neuropsychological test scores, and fluid biomarkers, restricts the model's access to highly discriminative diagnostic information. While this trade-off is justified by the translational objective, it inherently bounds the achievable classification performance compared to multimodal approaches. Fourth, the SHAP analysis was conducted on numerically encoded data for visualisation purposes only, which limits the interpretability of SHAP values for nominal categorical variables such as first and last symptom name. Future work should explore SHAP extensions for native categorical features. Finally, the prototype clinical tool has not undergone usability evaluation with clinician end-users, prospective validation, or regulatory assessment. It should be regarded strictly as a methodological proof-of-concept and not as a clinical decision-making instrument.

5. CONCLUSION

This study provides evidence that structured clinical phenotyping, particularly the temporal sequence of symptom emergence, captures diagnostically meaningful information for the differential classification of neurocognitive disorders. A machine learning model based solely on routine clinical variables achieved clinically relevant performance across five dementia subtypes, while preserving interpretability through subtype-specific SHAP profiles aligned with established neurological syndromes.

By avoiding dependence on neuroimaging, fluid biomarkers, or specialist investigations, this approach may offer a scalable screening strategy for settings where advanced diagnostic infrastructure is limited, including LMICs and primary-care pathways, potentially reducing diagnostic delays. Prospective multicentre validation is required to determine its real-world impact on early detection, referral prioritisation, and diagnostic decision-making.

Declarations

Funding

This work has been supported with the funding of Regione Puglia within the national “Fund for Alzheimer's and Dementia 2021-2023” (Piano Regionale Demenze 2021/2023 - D.G.R. 1284/2022) and with the funding of Regione Puglia for Tecnopolo per la Medicina di Precisione. D.G.R. n. 2117 of 21.11.2018 (B84I18000540002)

Acknowledgements

We are deeply grateful to all the participants in the study for the collaboration.

Data Availability Statement

Datasets are available from the corresponding author upon reasonable request.

Competing interests

The Authors: Davide Vilella, Benedetta Tafuri, Daniele Urso, Eleonora Rollo, Stefano Giannoni-Luza, Giuseppe Gigli, Bruno Tartaglione and Giancarlo Logroscino have no potential conflict of interest to disclose and state that the manuscript is being submitted only to “*BMC Medical Informatics and Decision Making*” that it will not be submitted elsewhere while under consideration, that it has not been published elsewhere, and, should it be published in “*BMC*

Medical Informatics and Decision Making”, that it will not be published elsewhere—either in similar form or verbatim—without permission of the editors.

Authors' contributions

Conceptualization: Giancarlo Logroscino, Bruno Tartaglione, Davide Vilella, Daniele Urso, Benedetta Tafuri, Giuseppe Gigli, Stefano Giannoni-Luza

Methodology: Benedetta Tafuri, Davide Vilella, Daniele Urso

Formal Analysis: Benedetta Tafuri, Stefano Giannoni-Luza, Daniele Urso, Davide Vilella.

Data Curation: Davide Vilella, Daniele Urso, Eleonora Rollo.

Writing—original draft: Davide Vilella, Benedetta Tafuri

Writing—review and editing: Bruno Tartaglione, Eleonora Rollo, Daniele Urso, Giuseppe Gigli, Giancarlo Logroscino.

Funding acquisition: Giancarlo Logroscino, Giuseppe Gigli

All authors have read the manuscript and approved it for publication

Consent statement

Ethical approval was obtained from the regional ethics committee (Ethic Committee for Medical Research at ASL Lecce on May 25, 2017 (number 6)), and written informed consent was secured from all participants or their legal representatives. The study adhered to the principles outlined in the Declaration of Helsinki. Data confidentiality was maintained through anonymization and secure data storage.

References

1. GBD 2019 Dementia Forecasting Collaborators, Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019, *Lancet Public Health* 7 (2022) e105–e125. [https://doi.org/10.1016/S2468-2667\(21\)00249-8](https://doi.org/10.1016/S2468-2667(21)00249-8).
2. C.R. Jack, J.S. Andrews, T.G. Beach, T. Buracchio, B. Dunn, A. Graf, O. Hansson, C. Ho, W. Jagust, E. McDade, J.L. Molinuevo, O.C. Okonkwo, L. Pani, M.S. Rafii, P. Scheltens, E. Siemers, H.M. Snyder, R. Sperling, C.E. Teunissen, M.C. Carrillo, Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup, *Alzheimers Dement* 20 (2024) 5143–5169. <https://doi.org/10.1002/alz.13859>.
3. G.C. Román, T.K. Tatemichi, T. Erkinjuntti, J.L. Cummings, J.C. Masdeu, J.H. Garcia, L. Amaducci, J.M. Orgogozo, A. Brun, A. Hofman, Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International Workshop, *Neurology* 43 (1993) 250–260. <https://doi.org/10.1212/wnl.43.2.250>.
4. I.G. McKeith, B.F. Boeve, D.W. Dickson, G. Halliday, J.-P. Taylor, D. Weintraub, D. Aarsland, J. Galvin, J. Attems, C.G. Ballard, A. Bayston, T.G. Beach, F. Blanc, N. Bohnen, L. Bonanni, J. Bras, P. Brundin, D. Burn, A. Chen-Plotkin, J.E. Duda, O. El-Agnaf, H. Feldman, T.J. Ferman, D. Ffytche, H. Fujishiro, D. Galasko, J.G. Goldman, S.N. Gomperts, N.R. Graff-Radford, L.S. Honig, A. Iranzo, K. Kantarci, D. Kaufer, W. Kukull, V.M.Y. Lee, J.B. Leverenz, S. Lewis, C. Lippa, A. Lunde, M. Masellis, E. Masliah, P. McLean, B. Mollenhauer, T.J. Montine, E. Moreno, E. Mori, M. Murray, J.T. O'Brien, S. Orimo, R.B. Postuma, S. Ramaswamy, O.A. Ross, D.P. Salmon, A. Singleton, A. Taylor, A. Thomas, P. Tiraboschi, J.B. Toledo, J.Q. Trojanowski, D. Tsuang, Z. Walker, M. Yamada, K. Kosaka, Diagnosis and

- management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium, *Neurology* 89 (2017) 88–100. <https://doi.org/10.1212/WNL.0000000000004058>.
5. K. Rascovsky, J.R. Hodges, D. Knopman, M.F. Mendez, J.H. Kramer, J. Neuhaus, J.C. van Swieten, H. Seelaar, E.G.P. Dopper, C.U. Onyike, A.E. Hillis, K.A. Josephs, B.F. Boeve, A. Kertesz, W.W. Seeley, K.P. Rankin, J.K. Johnson, M.-L. Gorno-Tempini, H. Rosen, C.E. Prioleau-Latham, A. Lee, C.M. Kipps, P. Lillo, O. Piguet, J.D. Rohrer, M.N. Rossor, J.D. Warren, N.C. Fox, D. Galasko, D.P. Salmon, S.E. Black, M. Mesulam, S. Weintraub, B.C. Dickerson, J. Diehl-Schmid, F. Pasquier, V. Deramecourt, F. Lebert, Y. Pijnenburg, T.W. Chow, F. Manes, J. Grafman, S.F. Cappa, M. Freedman, M. Grossman, B.L. Miller, Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia, *Brain* 134 (2011) 2456–2477. <https://doi.org/10.1093/brain/awr179>.
 6. M.L. Gorno-Tempini, A.E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S.F. Cappa, J.M. Ogar, J.D. Rohrer, S. Black, B.F. Boeve, F. Manes, N.F. Dronkers, R. Vandenberghe, K. Rascovsky, K. Patterson, B.L. Miller, D.S. Knopman, J.R. Hodges, M.M. Mesulam, M. Grossman, Classification of primary progressive aphasia and its variants, *Neurology* 76 (2011) 1006–1014. <https://doi.org/10.1212/WNL.0b013e31821103e6>.
 7. P.S. Sachdev, D. Blacker, D.G. Blazer, M. Ganguli, D.V. Jeste, J.S. Paulsen, R.C. Petersen, Classifying neurocognitive disorders: the DSM-5 approach, *Nat Rev Neurol* 10 (2014) 634–642. <https://doi.org/10.1038/nrneurol.2014.181>.
 8. C.P. Ferri, K.S. Jacob, Dementia in low-income and middle-income countries: Different realities mandate tailored solutions, *PLoS Med* 14 (2017) e1002271. <https://doi.org/10.1371/journal.pmed.1002271>.
 9. D.A. Pérez-Martínez, A.-C. Jarry, G. Martino, G. Pesce, I. Pieniazek, J. Presto, P. Miernik, R. Casciano, K.S. Frederiksen, Uncovering barriers to early diagnosis of Alzheimer’s disease: a European perspective, *Alzheimers Dement (N Y)* 11 (2025) e70178. <https://doi.org/10.1002/trc2.70178>.
 10. L. Boise, D.L. Morgan, J. Kaye, R. Camicioli, Delays in the diagnosis of dementia: Perspectives of family caregivers, *American Journal of Alzheimer’s Disease* 14 (1999) 20–26. <https://doi.org/10.1177/153331759901400101>.
 11. C. Giebel, M.R. Readman, A. Godfrey, A. Gray, J. Carton, M. Polden, Geographical inequalities in dementia diagnosis and care: A systematic review, *Int Psychogeriatr* 37 (2025) 100051. <https://doi.org/10.1016/j.inpsyc.2025.100051>.
 12. D. Vilella, D. Urso, A. Valguarnera, G. Volpe, V. Gnani, E. Rollo, A. Giugno, M. Caggiula, B. Coluccia, A. Mauro, R. Barone, M. Accogli, M. Leopizzi, A. Introna, M. Musio, S. Giannoni-Luza, G. Logroscino, Underdetection of neurocognitive disorders in Southern Italy: Evidence from the Salento region, *J Prev Alzheimers Dis* 12 (2025) 100295. <https://doi.org/10.1016/j.tjpad.2025.100295>.
 13. J. Vöglein, I. Kostova, T. Arzberger, S. Roeber, P. Schmitz, M. Simons, V. Ruf, O. Windl, J. Herms, M. Dieterich, A. Danek, G.U. Höglinger, A. Giese, J. Levin, First symptom guides diagnosis and prognosis in neurodegenerative diseases—a retrospective study of autopsy proven cases, *Eur J Neurol* 28 (2021) 1801–1811. <https://doi.org/10.1111/ene.14800>.
 14. N.J. Mekkes, M. Groot, E. Hoekstra, A. de Boer, E. Dagkesamanskaia, S. Bouwman, S.M.T. Wehrens, M.K. Herbert, D.D. Wever, A. Rozemuller, B.J.L. Eggen, I. Huitinga, I.R. Holtman, Identification of clinical disease trajectories in neurodegenerative disorders with natural language processing, *Nat Med* 30 (2024) 1143–1153. <https://doi.org/10.1038/s41591-024-02843-9>.
 15. M.W. Heymans, J.W.R. Twisk, Handling missing data in clinical research, *J Clin Epidemiol* 151 (2022) 185–188. <https://doi.org/10.1016/j.jclinepi.2022.08.016>.

16. L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2018. https://proceedings.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html (accessed July 2, 2026).
17. A.Y. Yildiz, A. Kalayci, Gradient Boosting Decision Trees on Medical Diagnosis Over Tabular Data, in: *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, 2025: pp. 1–8. <https://doi.org/10.1109/ICAD65464.2025.11114069>.
18. L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data?, (2022). <https://doi.org/10.48550/arXiv.2207.08815>.
19. T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, New York, NY, USA, 2019: pp. 2623–2631. <https://doi.org/10.1145/3292500.3330701>.
20. S. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, (2017). <https://doi.org/10.48550/arXiv.1705.07874>.
21. S. De Francesco, C. Crema, D. Archetti, C. Muscio, R.I. Reid, A. Nigri, M.G. Bruzzone, F. Tagliavini, R. Lodi, E. D'Angelo, B. Boeve, K. Kantarci, M. Firbank, J.-P. Taylor, P. Tiraboschi, A. Redolfi, RIN – Neuroimaging Network, Differential diagnosis of neurodegenerative dementias with the explainable MRI based machine learning algorithm MUQUBIA, *Sci Rep* 13 (2023) 17355. <https://doi.org/10.1038/s41598-023-43706-6>.
22. K.A. Wyman-Chick, P. Chaudhury, E. Bayram, C. Abdelnour, E. Matar, S.Y. Chiu, D. Ferreira, C.A. Hamilton, P.C. Donaghy, F. Rodriguez-Porcel, J.B. Toledo, A. Habich, M.J. Barrett, B. Patel, A. Jaramillo-Jimenez, G.D. Scott, J.P.M. Kane, Differentiating Prodromal Dementia with Lewy Bodies from Prodromal Alzheimer's Disease: A Pragmatic Review for Clinicians, *Neurol Ther* 13 (2024) 885–906. <https://doi.org/10.1007/s40120-024-00620-x>.
23. J. Attems, K.A. Jellinger, The overlap between vascular disease and Alzheimer's disease--lessons from pathology, *BMC Med* 12 (2014) 206. <https://doi.org/10.1186/s12916-014-0206-2>.
24. M.S. Barker, M. Manoochchri, E.D. Huey, The Diagnostic Landscape of Behavioral Variant Frontotemporal Dementia, *R I Med J* (2013) 108 (2025) 17–22.
25. D.O. Claassen, J.E. Parisi, C. Giannini, B.F. Boeve, D.W. Dickson, K.A. Josephs, Frontotemporal dementia mimicking dementia with Lewy bodies, *Cogn Behav Neurol* 21 (2008) 157–163. <https://doi.org/10.1097/WNN.0b013e3181864a09>.
26. T. Xia, C. Li, A. Iverson, J. Spat-Lemus, A. Woroch, G. Naasan, Behavioral variant frontotemporal dementia with pathogenic variant in MAPT presenting as dementia with Lewy body disease, *Neurocase* 31 (2025) 23–28. <https://doi.org/10.1080/13554794.2024.2440548>.
27. B.G. Ephrem, A. Varghese, V. Sherimon, P. Gouda, Web Application for Deciphering Alzheimer's Disease Prediction and Progression with Explanations, *Alzheimers Dement* 20 (2025) e085960. <https://doi.org/10.1002/alz.085960>.
28. X. Xiao, Y. Li, Q. Wu, X. Liu, X. Cao, M. Li, J. Liu, L. Gong, X. Dai, Development and validation of a novel predictive model for dementia risk in middle-aged and elderly depression individuals: a large and longitudinal machine learning cohort study, *Alz Res Therapy* 17 (2025) 103. <https://doi.org/10.1186/s13195-025-01750-6>.
29. M. Bruun, K.S. Frederiksen, H.F.M. Rhodius-Meester, M. Baroni, L. Gjerum, J. Koikkalainen, T. Urhema, A. Tolonen, M. van Gils, T. Tong, R. Guerrero, D. Rueckert, N. Dyremose, B.B. Andersen, A.H. Simonsen, A. Lemstra, M. Hallikainen, S. Kurl, S.-K. Herukka, A.M. Remes, G. Waldemar, H. Soininen, P. Mecocci, W.M. van der Flier, J.

- Lötjönen, S.G. Hasselbalch, Impact of a Clinical Decision Support Tool on Dementia Diagnostics in Memory Clinics: The PredictND Validation Study, *Curr Alzheimer Res* 16 (2019) 91–101.
<https://doi.org/10.2174/1567205016666190103152425>.
30. A.M. van Gils, A. Tolonen, A.C. van Harten, S. Vigneswaran, F. Barkhof, L.N.C. Visser, J. Koikkalainen, S.-K. Herukka, S.G. Hasselbalch, P. Mecocci, A.M. Remes, H. Soininen, A.W. Lemstra, C.E. Teunissen, L. Jönsson, J. Lötjönen, W.M. van der Flier, H.F.M. Rhodius-Meester, Computerized decision support to optimally funnel patients through the diagnostic pathway for dementia, *Alzheimers Res Ther* 16 (2024) 256.
<https://doi.org/10.1186/s13195-024-01614-5>.
31. V. Klar, E. Thompson, M.S. Atay, P. Owoade, S. Toniolo, A. Dehsarvi, S. Rathee, READi-Dem: ML-powered, web-interface tool for Robust, Efficient, Affordable Diagnosis of Dementia, (2023) 2023.10.23.23297405.
<https://doi.org/10.1101/2023.10.23.23297405>.
32. K. Seeher, R. Cataldi, T. Dua, D. Kestel, Inequitable Access to Dementia Diagnosis and Care in Low-Resource Settings - A Global Perspective, *Clin Gerontol* 46 (2023) 133–137.
<https://doi.org/10.1080/07317115.2022.2054391>.
33. A. Atri, B.C. Dickerson, C. Clevenger, J. Karlawish, D. Knopman, P.-J. Lin, M. Norman, C. Onyike, M. Sano, S. Scanland, M. Carrillo, The Alzheimer’s Association clinical practice guideline for the diagnostic evaluation, testing, counseling, and disclosure of suspected Alzheimer’s disease and related disorders (DETeCD-ADRD): Validated clinical assessment instruments, *Alzheimers Dement* 21 (2025) e14335.
<https://doi.org/10.1002/alz.14335>.

Figures

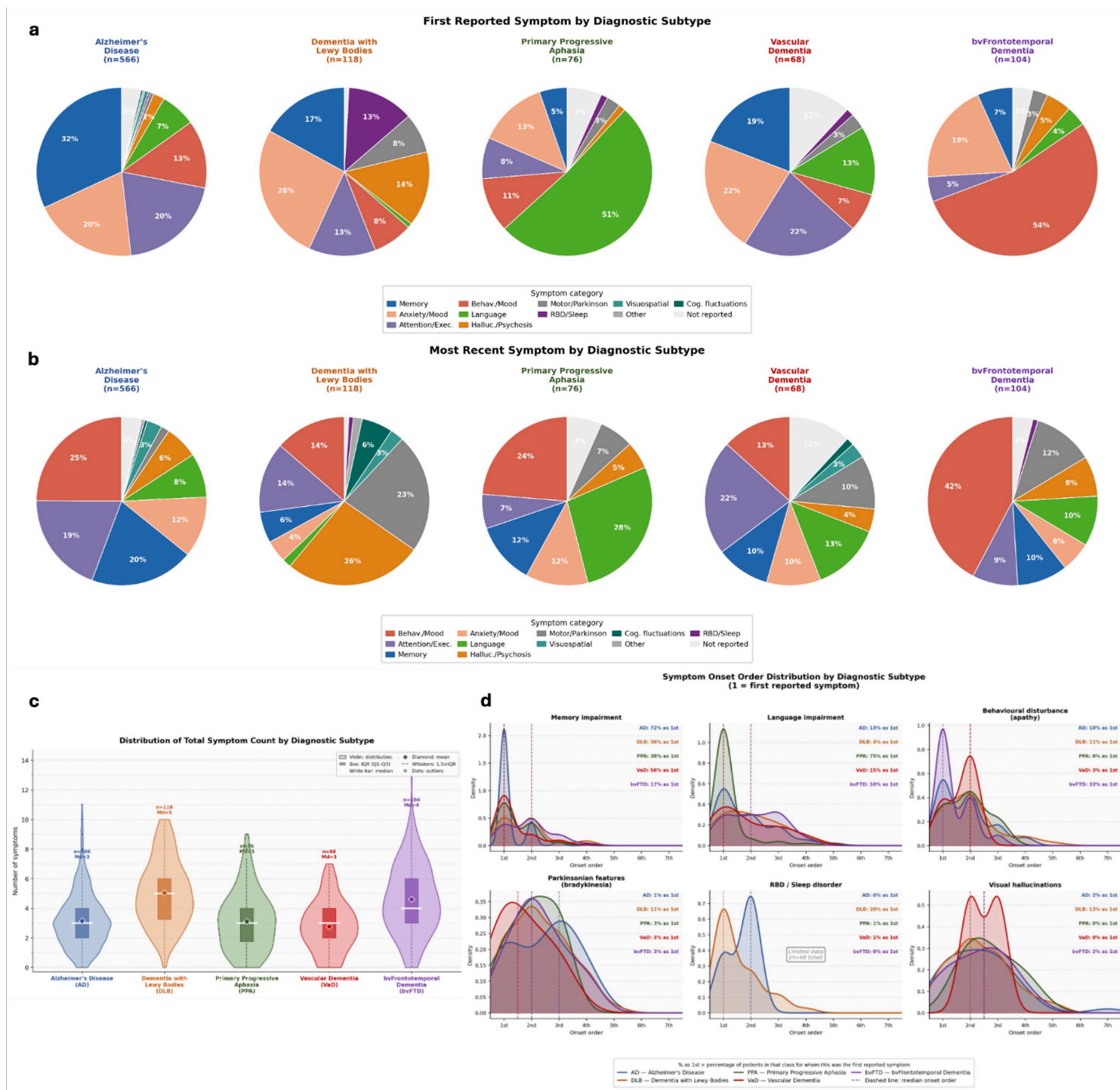


Figure 1

Complementary visualizations of the clinical presentation across the five diagnostic groups. (a) First reported symptom at disease onset for each subtype. (b) Most recent symptom reflecting the disease progression profile. (c) Distribution of the number of reported symptoms per patient, with the internal box indicating the interquartile range (IQR), the white bar the median, and the diamond the mean; whiskers extend to 1.5×IQR; individual outliers are shown as dots. (d) Onset order density plots for six clinically relevant symptoms (memory impairment, language impairment, behavioural disturbance/apathy, parkinsonian features (bradykinesia), REM sleep behaviour disorder (RBD), and visual hallucinations). Density curves for RBD, visual hallucinations, and bradykinesia in non-DLB groups

are based on limited observations ($n < 10$ in some classes) and should be interpreted with caution. *Abbreviations: IQR = interquartile range; KDE = kernel density estimate.*

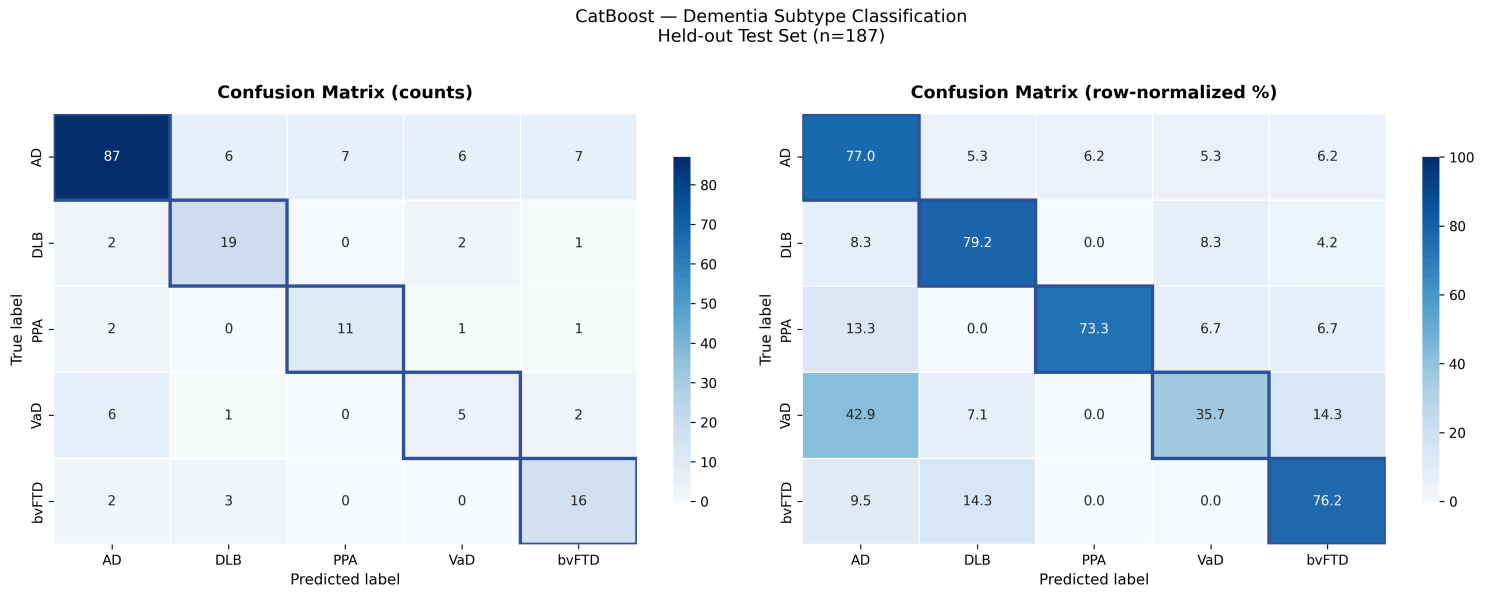


Figure 2

Confusion matrix on the held-out test set. Values on the left represent absolute counts, meanwhile on the right, diagonal values include row-normalized recall percentage.

Global Feature Importance Mean |SHAP value| across all diagnostic subtypes

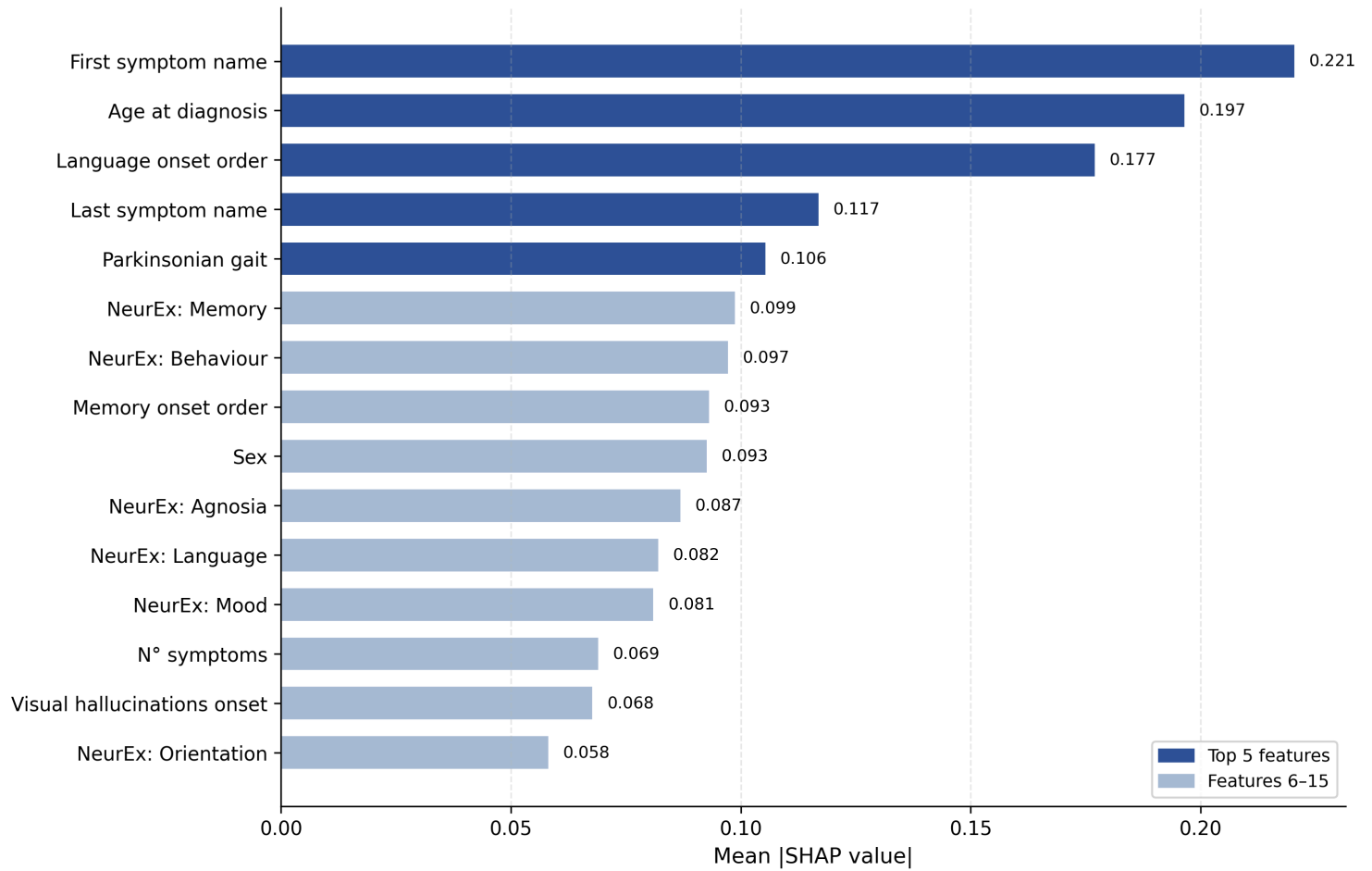


Figure 3

Global feature importance based on mean |SHAP value|

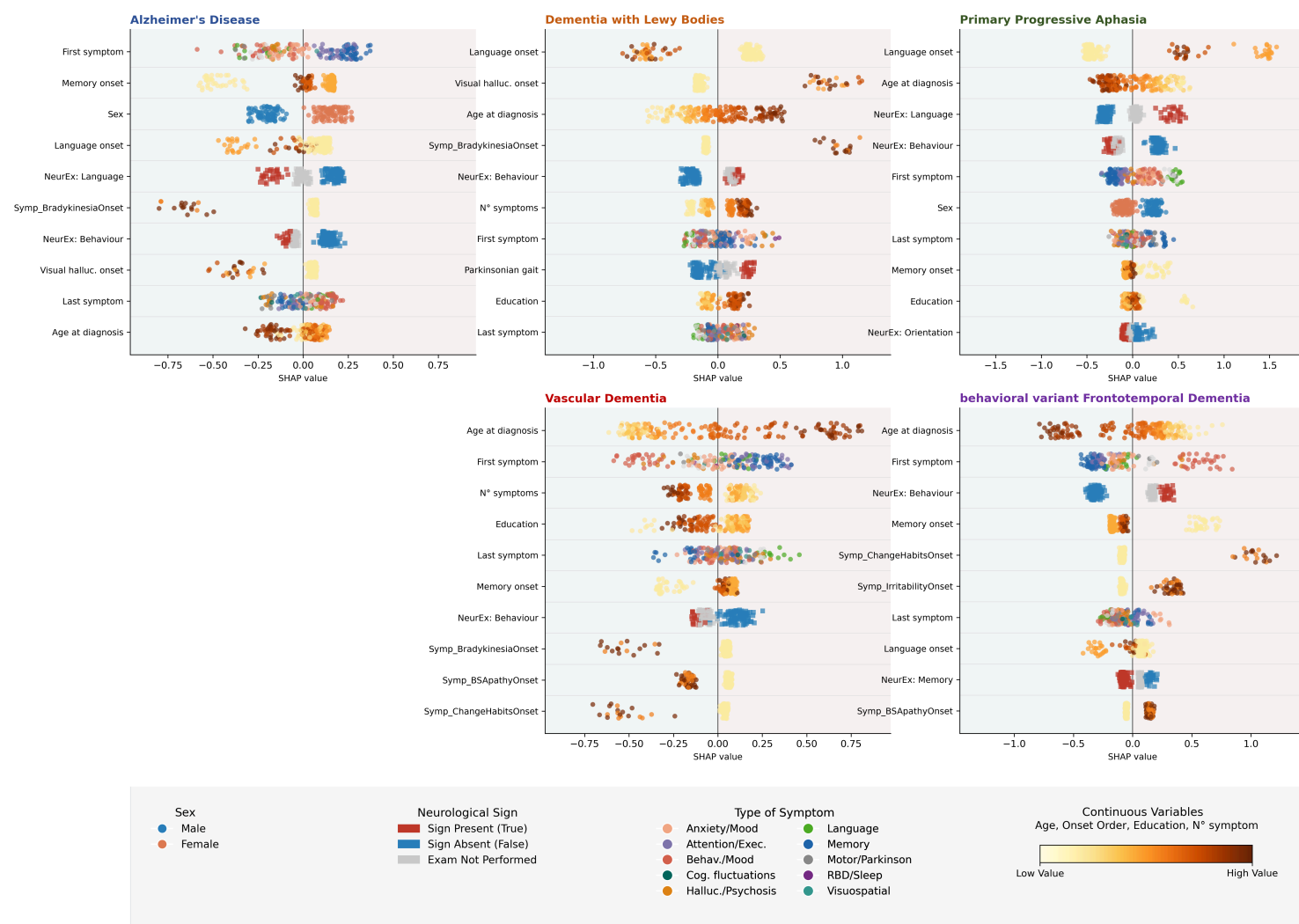


Figure 4

SHAP diverging dot plots per-class feature impact, top 10 features per diagnostic subtype. Each symbol represents one patient (test set, n=187). The x-axis reports the SHAP value: positive values (right) indicate that the feature pushes the model toward classifying the patient into that subtype; negative values (left) indicate the opposite. Marker shape: squares (■) = neurological examination variables (NeurEx); circles (●) = all other features. NeurEx color: red = sign present, blue = sign absent, grey = examination not performed. First/last symptom name color: dot color indicates the clinical symptom category (see legend per panel). Continuous variables (age, onset order, education, n° symptoms): color gradient from yellow (low value) to red (high value).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterials.docx](#)