On the Applicability and Utility of Machine Learning in the Field of Neuropsychology

Using the Example of Differential Dementia Diagnosis

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Philosophie (Dr. phil.)

der Ludwig-Maximilians-Universität

München

vorgelegt von

Clara Dominke

aus

Bad Kreuznach

2026

Erstgutachter/-in: Prof. Dr. Thomas Schenk

Zweitgutachter/-in: Prof. Dr. Markus Conci

Tag der mündlichen Prüfung: 09.02.2026

Acknowledgements

Viele Menschen haben diese Arbeit erst möglich gemacht, bei denen ich mich hiermit gerne von Herzen bedanken möchte: Ein besonderer Dank gilt Prof. Dr. Thomas Schenk und Prof. Dr. Thomas Jahn, für die herausragende fachliche Unterstützung, die zahlreichen Gespräche und fruchtbaren Diskussionen, ihr Vertrauen in meine Fähigkeiten und die Freiheiten, die sie mir bei der Erstellung dieser Arbeit gelassen haben.

Ein Dank gilt außerdem allen Mitgliedern der ehemaligen Arbeitsgruppe „Klinische und Experimentelle Neuropsychologie“ am Klinikum rechts der Isar der TU München sowie dem Lehrstuhl für Neuropsychologie an der LMU München. In beiden Arbeitsgruppen habe ich mich immer sehr wohl gefühlt und konnte mich fachlich weiterentwickeln.

Ich danke allen Patient:innen, die durch ihre Teilnahme an den jeweiligen – teilweise zeitlich intensiven - neuropsychologischen Untersuchungen, diese Arbeit ermöglicht haben.

Abbreviations

CV Cross - Validation

DEM Dementia

DAT Dementia of the Alzheimer's type

DEP Depression

HC Healthy Controls

ML Machine Learning

SVM Support Vector Machine

NB Naïve Bayes

RF Random Forest

LR Logistic Regression

GLMnet Lasso and Elastic-Net Regularized Generalized Linear Model

Abstract

Dementia (DEM) and Depression (DEP) represent the most prevalent neuropsychiatric disorders in individuals aged over 65 years. To ensure accurate treatment, correct differentiation between the two is of great importance. However, similar cognitive deficits complicate differential diagnosis in clinical practice. This work investigated for the first time the extent to which machine learning (ML) and different neuropsychological test batteries can improve differential diagnosis between DEM and DEP. The well-established CERAD-NAB, a compilation of individual tests (flexible battery approach) and the newly developed Cognitive Functions Dementia (CFD) test set were used. Accuracies for the differentiation between Alzheimer's Dementia (DAT) and DEP (Balanced accuracy of 87 %) as well as for the discrimination between DEP and DEM (Balanced accuracy of 80.8%) were high across all algorithms and test batteries used, suggesting that ML algorithms in combination with comprehensive neuropsychological test batteries can aid the differential diagnosis in clinical practice. More research and validation of our results in other samples is however needed for the implementation of such an algorithm in daily clinical routine.

Zusammenfassung

Demenz (DEM) und Depression (DEP) stellen die häufigsten neuropsychiatrischen Erkrankungen bei älteren Menschen über 65 Jahren dar. Um eine zufriedenstellende Behandlung zu gewährleisten, ist die richtige Differenzialdiagnostik zwischen beiden Erkrankungen von großer Bedeutung. Allerdings erschweren ähnlich anmutende kognitive Defizite die Differenzialdiagnose in der klinischen Praxis. In dieser Arbeit wurde erstmals untersucht, inwieweit Machine Learning (ML) und verschiedene neuropsychologische Testbatterien die Differentialdiagnose zwischen DEM und DEP verbessern können. Hierzu wurden die demenzspezifische Testbatterie CERAD-NAB, eine individuelle Zusammenstellung verschiedener neurokognitiver Tests (flexible battery approach) sowie das neu entwickelte Test-Set Cognitive Functions Dementia (CFD) angewandt. Die Genauigkeiten für die Unterscheidung zwischen Alzheimer-Demenz (DAT) und DEP (ausgewogene Genauigkeit von 87 %) sowie für die Unterscheidung zwischen DEP und DEM (ausgewogene Genauigkeit von 80.8 %) waren bei allen verwendeten Algorithmen und Testbatterien hoch, was darauf schließen lässt, dass ML-Algorithmen in Kombination mit umfassenden neuropsychologischen Testbatterien die Differentialdiagnose in der klinischen Praxis unterstützen können. Für die Implementierung eines solchen Algorithmus in der klinischen Praxis sind jedoch weitere Forschung und die Validierung der Ergebnisse in größeren Untersuchungsstichproben erforderlich.

1. General Introduction

1.1 Dementia

Due to an ever-rising life expectancy, the aged population is now at the highest peak in human history (Bloom, Canning & Lubet, 2015). The numbers of elderly people are predicted to further increase within the upcoming years (Nichols et al., 2019). Healthy aging results in declines of memory and executive functioning as well as changes in brain structure (Martins et al., 2024; Murman, 2015). Mild Cognitive Impairment (MCI), on the other hand, is considered a transitional stage between healthy aging and dementia (DEM). While concerned individuals report and exhibit memory impairments, activities of daily living can be undertaken normally (Bischkopf, Busse & Angermeyer, 2002). The prevalence of MCI among individuals over the age of 60 is estimated to be approximately 15 to 20% (Bai et al., 2022; Petersen, 2016).

The ageing of society has also brought about severe clinical syndromes such as DEM. It describes a syndrome, which refers to the deterioration of cognitive functioning (typically predominantly memory impairments) with respect to an earlier time, which interferes with activities of daily living. Varying pathophysiological processes can underlie DEM: While Alzheimer's disease (AD) affects approximately 50 to 70% of cases (Lobo et al., 2000; Rasmussen & Langerman, 2019), DEM can also arise from other distinct causes, which make up a great proportion of cases presented in clinics (Karantzoulis & Galvin, 2011). AD prominently involves the destruction of parieto-temporal regions, being primarily associated with a decline in memory and learning (Storey, Slavin & Kinsella, 2002). Memory impairments are indeed considered the hallmark feature of AD. Difficulties in object naming, executive functions and attention are yet also common (Musa et al., 2020; Weintraub, Wicklung & Salmon, 2012).

Lewy Body DEM accounts for approximately 5-10% of DEM cases, similarly to vascular DEM and frontotemporal DEM (Javeed et al., 2023). Lewy Body DEM is characterized by features of parkinsonism in combination with DEM and visual hallucinations (Brenowitz et al.,

2017). Fluctuations in cognitive functioning and alertness are common, and Lewy Body DEM is often associated with impaired executive functions, whereas memory performance and verbal fluency remain rather intact (Kemp et al., 2017). The most common risk factors for vascular DEM are heart disease, hypertension, smoking, and diabetes (Kivipelto et al., 2006). The cognitive impairments associated with vascular DEM vary considerably with the anatomical location of the insult. It can primarily affect cortical, subcortical or both regions. Small-vessel vascular DEM has been shown to be associated with executive dysfunctions and large-vessel vascular DEM with greater impairments in visuoconstruction tasks and language dysfunction (Ying et al., 2016). Thus, the variability in deficits complicated the description of relevant cognitive profiles (Braaten et al., 2006). Behavioral variant frontotemporal DEM (FTD) is the most common form of FTD and can be characterized by changes in behavior and emotions, a lack of empathy but also cognitive dysfunctions such as impaired executive functioning (Johnen & Bertoux, 2019).

An early and accurate diagnosis of DEM facilitates timely intervention, enables access to appropriate support services, helps maintain quality of life, and ensures the initiation of suitable treatment strategies (Rasmussen et al., 2019). Due to the complexity of neurodegenerative disorders and their overlapping symptomatology, accurate differential diagnosis in early stages of DEM yet often poses a problem to clinicians and frequently remains far from clear cut (Braaten et al., 2006; Karantzoulis et al., 2011; Porsteinsson et al., 2021). Hence, patients often present with diverse behavioral, affective, cognitive, and motor disturbances, impeding correct clinical diagnosis (Carrarini et. al., 2024; Karantzoulis et al., 2011). This matter of fact is further underlined by a missing association between subjective complaints, clinical impressions, and neuropsychological performance measurements (Burmester, Leathem & Merrick, 2016; Moritz, Ferahli & Naber, 2004). Patients presented in hospitals and memory clinics thus regularly represent a more heterogenous cognitive profile than what would be expected according to diagnostic criteria alone, often hindering correct clinical interpretations (Seelaar

et al., 2011). Obtaining differential diagnosis is however an imperative for accurate treatment and management of concerned individuals.

However, the difficulty lies not only in differentiating between different forms of DEM, but differentiation is also impeded by other psychiatric disorders, which can mimic the cognitive symptoms observed in patients with DEM (Tetsuka, 2021). One of the most challenging differential diagnoses in elderly is the distinction between DEM and depression (DEP), especially during the early stages of DEM (Kang et al., 2014; Leyhe et al., 2017; Zihl et al., 2010).

1.2 Depression and cognitive impairment

Depression (DEP) is a common psychiatric disorder, primarily characterized by depressed mood, sadness or anhedonia and other behavioral symptoms such as loss of appetite and weight, difficulties in making decisions or sleep disturbances (Cui et al., 2024). Together with DEM it is considered one of the most common psychiatric syndromes in people aged over 65 years: The prevalence of DEP in the elderly has been found to be very high, ranging between 13.3% to 23.6%, depending on the population investigated (Abdoli et al., 2022; Copeland et al., 2004). Even though cognitive dysfunctions mentioned in the diagnostic criteria of DEP only include concentration deficits and impaired decision making, DEP can also be characterized by attention deficits, difficulties in executive functioning and memory impairments (Linnemann & Lang, 2020; Morimoto, Kanellopoulos & Alexopoulos, 2014). Late-life depression has been shown to be particularly associated with cognitive dysfunctions (Mackin et al., 2014; Masse et al., 2021). Depending on the population investigated, over 25% of patients with DEP exhibit significant cognitive deficits and a considerably larger proportion demonstrate some sort of idiographic cognitive impairment (Gualtieri & Morgan, 2008; Tran et al., 2021). A recent review reported that patients with DEP performed significantly worse on all 16 investigated neuropsychological domains as compared to healthy controls (HC) with a large heterogeneity

between studies (Parkinson et al., 2020). Due to these profound deficits in cognitive functions, Kiloh founded the term *pseudodementia* in 1961 to describe cognitive dysfunctions associated with DEP (Kiloh, 1961). Thus, DEP can mimic typical cognitive symptoms of DEM (Tetsuka, 2021). This fact often complicates differential DEM diagnosis. The differentiation is further impeded by the fact that depressive symptoms or low mood are frequently observed in individuals with DEM as well (Kitching, 2015) and that the two disorders frequently co-occur as comorbid conditions (Berk et al., 2023; Novais & Starkstein, 2015).

Many researchers even suggest a relation between DEM and DEP: Some consider late-life DEP to be a risk factor for DEM (Byers & Yaffe, 2011; Cantón-Habas et al., 2020) or even a prodromal stage of it (Wiels, Baeken & Engelborghs, 2020), while others suggest DEP to be a response to the cognitive decline experienced by individuals suffering from DEM (Brzezińska et al., 2020) or found no significant relation at all (Brodaty et al., 2003).

Due to these difficulties, extensive research has been conducted to investigate the differences between the clinical syndromes: Neuropsychological research has investigated potential differences in the cognitive profiles between DEM and DEP and found a higher degree of impairment within measures of recognition memory (Barlet et al., 2023; Gasser, Salamin & Zumbach, 2018; Leyhe et al., 2017), visuoconstructional practice (Silva Dos Santos Durães et al., 2022) and verbal fluency tasks (Barlet et al., 2023) for DEM, but research was not consistent on these findings, partly reporting no significant differences between these groups at all (Barth et al., 2005; Christensen et al., 1997). Effect sizes of reported differences have additionally been found to be too small to help with the diagnosis of single cases in clinical practice (Lanza et al., 2020). Accordingly, accurate differentiation between the two diagnoses remains challenging due to overlapping features of cognitive dysfunction (Gasser et al., 2018). Nevertheless, neuropsychological assessments remain crucial for the early detection and differential diagnosis of DEM.

1.3 The importance of neuropsychological assessment in differential dementia diagnosis and its weaknesses

Neuropsychological assessments are considered the gold-standard in DEM diagnosis (Alzola et al., 2024). They are useful in the determination of a DEM syndrome by quantifying the cognitive deficits. Comprehensive neuropsychological assessments encompassing all relevant domains of cognitive functioning can provide valuable information on the severity and stage of cognitive impairment, as well as important indications regarding its underlying etiology (Weintraub, 2022). Neuropsychological variables like verbal memory measures (i.e., especially wordlist learning tests) and different language tests have shown great accuracy in predicting progression from MCI to DAT (Belleville et al., 2017; Gallucci et al., 2017; Duke Han et al., 2017). Compared to invasive and expensive techniques such as lumbar puncture, neuropsychological testing is inexpensive and easy to learn and administer by trained psychologists (Gurevich et al., 2017). Cognitive examinations are furthermore standardized regarding procedure and scoring, which aid to bias reduction and improve clinical interpretations (Hsieh et al., 2013). Both for the diagnosis and prognosis of DEM 10 years in advance of a transition to a clinical syndrome (Mistridis et al., 2015), the accuracy of specific neuropsychological tests has been found to be comparable to cerebrospinal fluid markers (Schmand, Huizenga & van Gool, 2010). Differences in performance on specific neuropsychological tasks between DEM in early stages and DEP are however usually too small to help clinicians with individual case diagnosis (Kang et al., 2014; Lanza et al., 2020) and some studies have found no significant differences between MCI and DEP (Zihl et al., 2010), making neuropsychological investigations alone often not sufficient to aid this differential diagnosis.

1.4 Machine Learning (ML) as a potential diagnostic aid

The term Machine Learning (ML) describes various methods enabling a specific algorithm to learn from presented data. What makes ML algorithms so interesting is the fact

that they, in contrast to traditional statistical methods, do not require any pre-specified hypotheses regarding the association between certain predictor variables and a prediction (Graham et al., 2020), but can instead detect complicated interactions between variables, eventually increasing classification accuracy (Dwyer et al., 2018; Graham et al., 2020). These algorithms are thereby increasingly used to support clinical decision making (Jiang et al., 2017).

One can differentiate between two main classes of ML algorithms: Supervised and unsupervised learning algorithms. While during unsupervised learning, the algorithm detects patterns within unlabeled data to find clusters within the data, during supervised learning, the algorithm learns from data which has previously been labeled. The output data is called "labels" while the input data is described by the term "features". Labels in our data were the different diagnostic groups, and the features consisted of data from different neuropsychological test batteries. Therefore, in the present work, only supervised learning algorithms were used.

The aim of ML is the creation of a model making the most accurate prediction of the target variable using the input data (Hastie et al., 2009). The performance of a supervised machine learning algorithm can be evaluated by comparing the labels it predicts to the actual diagnostic categories observed in the data. For validation of the performance of an algorithm, k-fold-cross-validation (CV) is normally used (Arlot & Celisse, 2010): It describes the division of a data set into k different training sets (i.e., on which the algorithm learns) and test sets (i.e., for which the algorithm predicts the labels based on the model established using the test set). The folds are disjoint and exhaustive**,** meaning each data point belongs to exactly one fold, and every data point is used in the validation process across the folds. There is no overlap between the folds, and cases are not drawn with replacement. Each fold is used once as the validation set, while the remaining folds serve as the training set. This ensures that all data points are utilized for both training and validation without repetition**.** This process of dividing the data into k subsets can be repeated multiple times and performance is then averaged across the different iterations to gain the performance measures. The training groups are more variable,

which results in more stable estimates (Hastie et al., 2009). Any preprocessing of the data also must be embedded into a CV-pipeline, to prevent information leakage from the test set, resulting in an over-optimistic performance estimate (Dwyer et al., 2018). The different ML algorithms were compared based on a benchmark design: thus, it was secured that they all underwent the same preprocessing steps and that the different algorithms were trained and tested on the same folds to ensure appropriate comparison.

Usually, accuracy (the percentage of correctly classified subjects), sensitivity (percentage of correctly classified cases in the group of interest) and specificity (percentage of correctly classified healthy subjects or patients with DEP in our study) are reported. Because the sample sizes in the present investigation were quite imbalanced, balanced accuracies (the accuracy balanced by the sample sizes of the two groups) were furthermore reported.

Hereinafter, I will describe the supervised ML algorithms that were also employed in the two studies.

1.4.1 Support Vector Machine (SVM)

SVM refers to a class of supervised learning algorithms, whereby the algorithm learns relations between pre-classified and labeled data (i.e., healthy controls vs. DEM for example) and different features. It can be used for classification and regression problems. It maximizes the margin (i.e., the maximum width) between the support vectors (i.e., the data points closest to the separating hyperplane) of the two groups within a high-dimensional space (Pereira, Mitchell & Botvinick, 2009). The mapping into the higher dimension is performed by using a kernel function, which transforms lower-dimensional input data into a higher dimensional feature space. By using the kernel function, the groups can be linearly separated in higher dimensional space, even in cases where a linear separation was originally not possible. SVM has been successfully used in numerous studies investigating psychiatric disorders (Orrù et al., 2012).

1.4.2 Naïve Bayes (NB)

Naïve Bayes is a classification algorithm based on the Bayes theorem. The Bayes theorem is a way of calculating conditional probabilities. Firstly, prior probabilities for each diagnostic class are calculated as well as the probability distribution for the features. Then, the NB calculates the class probability given a particular set of features. Within the NB algorithm, these features are assumed to be independent of each other, wherefore it is called "naïve". Even though, this assumption is rarely true in real-world-circumstances, it has been found to perform particularly well (Al-Aidaroos et al., 2010). In this work, the Gaussian NB was used, which assumes that the features follow a normal (Gaussian) distribution. This allows the effective handling of continuous data.

1.4.3 Random Forest

Random Forest describes an ensemble classifier (Breiman, 2001): Multiple decision trees are generated to classify an input's vector (Biau & Scornet, 2016). A decision tree iteratively partitions the feature space into subsets based on specific feature values to create a tree-like model of decisions, which is then used for classification. Each decision tree is built from a random subsample of the training set and the features. This process is repeated for the given number of trees. Predictions across the different decision trees are then averaged to get the final prediction for a given observation. By combining the individual predictions of the trees, it produces more accurate predictions than the individual trees alone and increases model generalization. It is widely used both in classification and regression problems.

1.4.4 Lasso and Elastic-Net Regularized Generalized Linear Models (GLMnet)

Lasso and Elastic-Net Regularized Generalized Linear Models (GLMnet) extend traditional logistic regression by using regularization methods, which prevent overfitting by adding penalty terms to the cost function of a given model to constrain the model's parameter

estimates. It performs feature selection and only keeps the most important features. Multicollinearity can be handled by only keeping one of the correlated features and eliminating the other feature (L1 Penalty). L2 penalty uses the squared magnitude of the coefficients. Both L1 and L2 penalties are used within the given algorithm. An advantage of this method is the capability of dealing with large number of input variables while reducing the complexity of the model (Friedman, Hastie & Tibshirani, 2010).

1.4.5 Logistic Regression

Logistic Regression is often used within medical and psychological research, if one deals with a qualitative dependent variable like diagnostic groups (Nick & Campbell, 2007). One can differentiate between multinominal logistic regression and binary logistic regression. Binary logistic regression is used when the outcome variable has two possible outcomes, while multinominal logistic regression is used when the target variable consists of more than two possible categories. In the following work, only binary logistic regressions were used to focus on the direct comparison between two groups at a time. Logistic regression assumes a linear relation between the predictors and the logit (the natural logarithm of the odds; Schober & Vetter, 2021).

Since most of the models described in the previous paragraph are hard to interpret on their own (i.e., so called “black box models”), additional methods must be applied to increase interpretability. The method used to ensure interpretability in this work was permutation feature importance: It is a measure for how much the performance of a model deteriorates when the values on a given feature are permutated by randomly shuffling the values of that feature across all samples so that the relation between the feature and the output is destroyed. The difference in prediction accuracy between the permutated version and the original version (i.e., without the permutation) is called the feature importance (Molnar, 2022).

1.5 Machine Learning in the context of dementia diagnosis

Machine Learning holds great promise in aiding the diagnosis of DEM. Previous studies have mostly investigated its potential by using imaging data such as Magnetic Resonance Imaging (MRI) or Diffusion Tensor Imaging (DTI) data to detect or predict DEM or to differentiate between Alzheimer's DEM and vascular DEM (Basaia et al., 2019; Castellazzi et al., 2020; Pellegrini et al., 2018). Other investigations tried to predict the progression to DEM incidence within a 2-years-interval using PET-Scans and cerebrospinal fluid markers with accuracies up to 90%, outperforming traditional methods (James et al., 2021).

The importance of neuropsychological features has also been proven however: Other studies have investigated the potential of neuropsychological and socio-demographic data in differentiating between DEM with Lewy Bodies and Parkinson's Disease DEM (Bougea et al., 2022) with promising results. Accuracies up to 100% were found within the differentiation of Alzheimer's DEM, Parkinson's disease DEM and Lewy Bodies DEM using electroencephalographic features (Garn et al., 2017).

ML models have also been shown to accurately classify the cognitive status of participants (no impairment, mild impairment, DEM) with accuracies up to 93.3%, using solely neurocognitive test results and demographic data (Kang et al., 2019) and to robustly generalize to other samples (Bachli et al., 2020). Neurocognitive data can further differentiate between individuals suffering from depression and individuals with other causes of cognitive impairment, without the addition of mood assessments (Mato-Abad et al., 2018), providing evidence for the applicability of neurocognitive data for differential DEM diagnosis. Substantial evidence on the applicability of ML in differentiating between DEM and DEP is however largely lacking, which highlights the necessity of this research to explore and evaluate the potential of ML in improving diagnostic accuracy in this complex neuropsychological context.

1.6 Objectives of the dissertation and methodological approach

This study aimed to examine the diagnostic utility of combining machine learning (ML) algorithms with comprehensive neuropsychological assessments for differentiating between dementia (DEM) and depression (DEP). Specifically, it investigated which of 3 neuropsychological test batteries—each differing in scope and structure—and which ML models yielded the highest classification accuracy. To achieve this, the following specific sub-objectives and research questions were addressed:

I) Evaluation of different neuropsychological test batteries: Three distinct neuropsychological test batteries were employed to investigate, which provided the most discriminative power for distinguishing between DEM and DEP in combination with ML

II) Three different classification tasks were performed to address distinct diagnostic challenges and to evaluate the robustness and generalizability of the machine learning models across varying clinical contexts. In the first study the differentiation between Alzheimer's DEM and DEP was analyzed: This comparison targeted a common diagnostic dilemma, as depressive symptoms in older adults can closely resemble early-stage Alzheimer's disease. In the second study, two different differentiation tasks took place: HC vs DEM of various etiologies. This classification aimed to evaluate the models' general ability to detect cognitive impairment across a broader range of dementia types, such as Alzheimer's, vascular, or frontotemporal dementia as compared to healthy controls. Secondly, DEM of various etiologies vs. DEP: By including multiple dementia subtypes, this task increased diagnostic complexity and reflected real-world clinical diversity. The objective was to assess whether ML models can differentiate DEP from various forms of dementia beyond Alzheimer's alone.

III) To increase interpretability of the algorithms, permutation-based feature importance was calculated for all neuropsychological features and classifications. This approach aimed to identify the most relevant neuropsychological features contributing to classification performance.

2. Experimental Studies

This dissertation was submitted as a cumulative thesis and is based on the following two papers that have been published in international peer-reviewed journals.

2.1 CERAD-NAB and flexible battery based neuropsychological differentiation of Alzheimer's dementia and depression using machine learning approaches

This study has been published as:

Dominke, C., Fischer, A. M., Grimmer, T., Diehl-Schmid, J., & Jahn, T. (2024). CERAD-NAB and flexible battery based neuropsychological differentiation of Alzheimer's dementia and depression using machine learning approaches. *Aging, Neuropsychology, and Cognition*, *31*(2), 221–248. https://doi.org/10.1080/13825585.2022.2138255

Author's contribution:

Clara Dominke: Conceptualization of the evaluation strategies, data analysis and manuscript writing.

AGING, NEUROPSYCHOLOGY, AND COGNITION
https://doi.org/10.1080/13825585.2022.2138255

CERAD-NAB and flexible battery based neuropsychological differentiation of Alzheimer's dementia and depression using machine learning approaches

Clara Dominke[a], Alina Maria Fischer[b], Timo Grimmer[b], Janine Diehl-Schmid[b,c] and Thomas Jahn[a,b]

[a]Division Clinical Neuropsychology, Department of Psychology, Ludwig-Maximilians-University, Munich, Germany; [b]School of Medicine, Department of Psychiatry and Psychotherapy, Technical University of Munich, Munich, Germany; [c]Centre for Geriatric Medicine, Kbo-Inn-Salzach-Klinikum, Wasserburg am Inn, Germany

ABSTRACT

Depression (DEP) and dementia of the Alzheimer's type (DAT) represent the most common neuropsychiatric disorders in elderly patients. Accurate differential diagnosis is indispensable to ensure appropriate treatment. However, DEP can yet mimic cognitive symptoms of DAT and patients with DAT often also present with depressive symptoms, impeding correct diagnosis. Machine learning (ML) approaches could eventually improve this discrimination using neuropsychological test data, but evidence is still missing. We therefore employed Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and conventional Logistic Regression (LR) to retrospectively predict the diagnoses of 189 elderly patients (68 DEP and 121 DAT) based on either the well-established Consortium to Establish a Registry for Alzheimer's Disease neuropsychological assessment battery (CERAD-NAB) or a flexible battery approach (FLEXBAT). The best performing combination consisted of FLEXBAT and NB, correctly classifying 87.0% of patients as either DAT or DEP. However, all accuracies were similar across algorithms and test batteries (83.0% – 87.0%). Accordingly, our study is the first to show that common ML algorithms with their default parameters can accurately differentiate between patients clinically diagnosed with DAT or DEP using neuropsychological test data, but do not necessarily outperform conventional LR.

ARTICLE HISTORY
Received 18 March 2022
Accepted 14 October 2022

KEYWORDS
Alzheimer's dementia; depression; neuropsychological assessment; machine learning; differential diagnosis; CERAD-NAB; flexible battery approach

Introduction

Alzheimer's disease is the leading cause of dementia in late life, accounting for approximately 60% of cases (Alzheimer's Disease International, 2019). Obtaining differential diagnosis of dementia is crucial to ensure appropriate treatment and correct clinical trial inclusion (Bruun et al., 2018; Horgan et al., 2020). In clinical practice, however, it is sometimes difficult to reliably differentiate dementia of the Alzheimer's type (DAT) from other common neurodegenerative or mental disorders, which are associated with

CONTACT Clara Dominke Clara.Dominke@psy.lmu.de
Supplemental data for this article can be accessed online at https://doi.org/10.1080/13825585.2022.2138255

considerably different disease courses but often very similar cognitive symptoms (Karantzoulis & Galvin, 2011; Leyhe et al., 2017; Tetsuka, 2021).

Depression (DEP), like DAT, is among the most common neuropsychiatric disorders observed in the elderly population (Liguori et al., 2018; Tetsuka, 2021). For older adults (age > 65 years), prevalence rates for both DEP and dementia have been found to be similarly high, reaching up to 16.3% and 14.9%, respectively (Andreas et al., 2018; Bacigalupo et al., 2018; Lobo et al., 1995; Steffens et al., 2009). However, among patients referred to memory clinics, prevalence rates are different: studies reported a 29.5% – 41.5% prevalence rate of DAT and 8.9% – 19.7% prevalence rate for depression among this population (Claus et al., 2016; Kao et al., 2019; Knapskog et al., 2014). The higher prevalence of DAT in comparison to DEP within memory clinics can be explained by the fact that patients are usually only referred to memory clinics if they exhibit noticeable cognitive impairment, automatically excluding DEP patients without severe cognitive impairments. Even though only difficulties in concentration and decision making have been included as cognitive diagnostic criteria for DEP, many studies have shown that DEP is regularly associated with difficulties in executive functioning, a decline in processing speed and deficits in terms of attention and also memory impairments (Butters et al., 2004; Linnemann & Lang, 2020; Steffens & Potter, 2008). These deficits are already noticeable during first-episode DEP and are often misinterpreted as early signs of dementia in elderly patients (Ahern & Semkovska, 2017; Beblo et al., 2011; Tetsuka, 2021).

As compared to younger adults, late-life DEP is associated with more severe cognitive impairments, including executive functions, verbal learning, and memory – a hallmark feature of DAT (Thomas et al., 2009). In clinical practice, this fact renders the correct differential diagnosis between DAT and DEP in elderly patients especially difficult (Liguori et al., 2018; Tetsuka, 2021). An accurate differentiation is however crucial, as cognitive impairments associated with DEP are often – but not always (Perini et al., 2019; Semkovska et al., 2019) – treatable whereas patients with DAT experience a worsening of cognitive functions regardless of the treatment provided (Beblo et al., 2011).

In the context of neuropsychological investigations, it has been found that patients with DAT usually exhibit higher rates of forgetting of initially recalled material and also show greater impairments in discriminability within recognition memory tasks as compared to patients with DEP (Braaten et al., 2006; Jahn et al., 2004; Leyhe et al., 2017). Patients with DEP, on the other hand, can show greater impairments in verbal fluency tasks as compared to patients in the early stages of DAT (Tetsuka, 2021). The performance on tasks assessing executive functions such as planning, inhibition or task-switching is usually equally impaired in both DAT and DEP, speaking against the use of these cognitive functions for differential diagnosis (Rushing et al., 2014). Severe difficulties in visuoconstruction and object naming are instead suggested to be specifically found within DAT (Lin et al., 2014; Weintraub et al., 2012; Wright & Persad, 2007). Nevertheless, the effect sizes of these performance differences on group levels usually tend to be too small to allow for an accurate clinical diagnosis on the individual subject level (Lanza et al., 2020).

Thus, even though the neuropsychological characteristics of both DAT and DEP have been studied extensively on group levels in the last decades, due to similarities in clinical presentations and overlapping symptoms, differential diagnosis on the individual case level in clinical practice remains challenging (Lanza et al., 2020; Leyhe et al., 2017). Besides overlapping cognitive symptoms, appropriate clinical differentiation between DAT and

DEP is additionally complicated by the fact that many individuals with DAT also report depressive symptoms (Tetsuka, 2021; Tonga et al., 2021). Not surprisingly, high rates of either false-positive or false-negative dementia diagnoses have consequently been reported in the past (Hunter et al., 2015; Jammeh et al., 2018). These difficulties with the transferability of group level results to individual clinical case diagnosis have led to inquiries regarding the implementation of new computational methods, other than traditional statistical inferential paradigms, to improve classification and thereby reliability of diagnosis on the single subject level.

Machine-learning (ML) algorithms have been proposed to constitute a promising method for improving both accuracy and confidence in diagnosis (Arbabshirani et al., 2017; Dwyer et al., 2018). ML algorithms automatically search for the optimal solution, detecting unforeseen relationships between predictors, without putting emphasize on a specific model as it is the case with conventional statistics (S. A. Graham et al., 2020; Dwyer et al., 2018). They are used in the context of dementia research using neuroimaging or electrophysiological data (Ahmed et al., 2018; Trambaiolli et al., 2011). Neuroimaging techniques are, however, not always available within primary care institutions and even if available, patients may reject these procedures, limiting the use of these methods in practice (Gurevich et al., 2017). Furthermore, even if these methods are available, due to differences in scanners and acquisition parameters among clinics, generalizability of results remains unsure (Abdulkadir et al., 2011). Accordingly, other measures are necessary, which are more frequently available and accepted by a broader group of patients.

Neuropsychological examinations, which are a central prerequisite to identify a dementia syndrome, can be administered in a standardized manner by qualified neuropsychologists (Gurevich et al., 2017). An extensive amount of evidence speaks in favor of the importance of neuropsychological examinations not only for the diagnosis of dementia, but also in terms of its potential for differential diagnosis (Begali, 2020; Mansoor et al., 2015; Wright & Persad, 2007). For correct diagnoses, the accuracy of specific neuropsychological tests has been found to be comparable to beta and tau in cerebrospinal fluid (CSF), structural magnetic resonance imaging as well as 18 F-fluorodeoxyglucose positron emission tomography (FDG-PET; Hansen et al., 2022; Schmand et al., 2010, 2011).

First studies have already provided preliminary evidence for the utility of combining neuropsychological test data with ML to facilitate dementia diagnosis: ML models have been shown to accurately classify the cognitive status of participants (no impairment, mild cognitive impairment, dementia) with accuracies of up to 93.3%, using solely neurocognitive test results and demographic data (Kang et al., 2019) and to robustly generalize to other samples (Bachli et al., 2020). Studies even report accuracies of up to 100% in classifying patients with probable Alzheimer's disease and healthy controls with age-related cognitive decline based on neuropsychological test results (Er et al., 2017). Gurevich et al. (2017) reported an accuracy of 89% in differentiating between DAT and no DAT using the well-known *Consortium to Establish a Registry for Alzheimer's Disease* neuropsychological assessment battery (CERAD-NAB; Welsh et al., 1994) and two newly developed additional tests (verbal comprehension and recollection). Cognitive test data can further differentiate between individuals suffering from Mild Cognitive Impairment

(MCI) with depression and MCI without depression (Mato-Abad et al., 2018), so that their combination with ML might be fruitful.

Nevertheless, evidence for the usefulness of combining ML and cognitive test data for the differential individual case diagnoses of DAT in clinical practice is far from clear cut: Firstly, most research has compared patients with dementia solely to healthy controls (Almubark et al., 2019; Er et al., 2017; Gupta & Kahali, 2020; Gurevich et al., 2017; Kang et al., 2019) and has not yet investigated the more complex and clinically more relevant differentiation between DAT and DEP. Secondly, most studies focusing on the use of ML in mental disorders have only used one specific ML technique and have not compared different algorithms (Shatte et al., 2019). However, there is evidence that, depending on the given research question, particular ML algorithms perform differently in terms of classification accuracy (Chen & Herskovits, 2010; Maroco et al., 2011; Zulfiker et al., 2021), speaking in favor of a comparison of different algorithms. Whether or not ML algorithms outperform traditional logistic regression – the most widely used statistical method for classification problems in psychiatric research (Hosmer et al., 2013) –, remains largely unanswered by previous investigations. Finally, it is still under debate which neuropsychological test strategy promises to be most accurate in identifying patients with dementia (Logie et al., 2015; Olson et al., 2021; Woodford & George, 2007) – a fixed battery approach with instruments such as the CERAD-NAB (Welsh et al., 1994; Wolfsgruber et al., 2014) or a more comprehensive flexible battery approach (FLEXBAT), comprising patient-tailored selections of cognitive tests from the much larger neuropsychological tool box. While the CERAD-NAB is both clinically and scientifically most widely used in the field of dementia diagnosis, the FLEXBAT approach is recommended by many experts to improve confidence in diagnosis for clinically challenging cases (Beck et al., 2014; Deutsche Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN) & Deutsche Gesellschaft für Neurologie (DGN), 2016; Pasternak & Smith, 2019).

Therefore, the aim of the present study was to investigate the performance of three common ML algorithms (Support Vector Machines, Naive Bayes and Random Forest) with their default parameters and a more traditional statistical method (Logistic Regression) in classifying patients as either DAT or DEP using neuropsychological test data. In this context we used two different neuropsychological test batteries: a comprehensive compilation of neuropsychological tests (FLEXBAT) and a more commonly used neuropsychological test battery (CERAD-NAB) to see which would perform better. Finally, we investigated which subtest variables of the better performing neuropsychological test battery were most important for the classification decision.

To our knowledge, this is the first investigation using ML algorithms to identify DAT and DEP based on neuropsychological test data. Due to the explorative nature of our investigation and a lack of similar investigations in the literature, there were no explicit hypotheses regarding which combination of algorithm and neuropsychological test battery would perform best, even though it was certainly expected that ML algorithms would generally outperform traditional logistic regression, since ML algorithms have already been shown to perform better than classical approaches for diverse classification and prediction problems (Dwyer et al., 2018; Liew et al., 2022; Yang et al., 2020). This superiority of ML can be explained by one of the main differences between traditional statistical approaches and ML – the fact that the focus of the former lies in the interference

of relationships between different variables, whereas the latter aims at the most accurate prediction (Orrù et al., 2020; Rajula et al., 2020).

Methods

Sample

The present sample was retrospectively selected from 1013 patients with suspected dementia treated between January 1998 and August 2008 at the Center for Cognitive Disorders, an outpatient clinic of the Department of Psychiatry and Psychotherapy of the Technical University Munich. The patients received a comprehensive cognitive assessment with the FLEXBAT at the hospital's Clinical and Experimental Neuropsychology Unit. Since this is a retrospective study, sample size was not specifically determined a priori. Of these patients, 747 had already received an initial CERAD-NAB at the Center for Cognitive Disorders before they were referred to the Clinical and Experimental Neuropsychology Unit. After excluding patients who presented for a repeated visit, 669 patients remained. Out of these, all patients older than 49 years with complete patient data available, a maximum time-interval of 59 days between the CERAD-NAB and the FLEXBAT, and a final diagnosis of either dementia of the Alzheimer's type (DAT; N = 126) or depression or dysthymia (DEP; N = 72) were selected. In addition, five patients with DAT and four patients with DEP were excluded, because they did not undergo the comprehensive California Verbal Learning Test (CVLT) as part of FLEXBAT, leaving 121 patients with DAT and 68 patients with DEP for the final analyses. Figure 1 shows the CONSORT diagram of the stepwise participant selection procedure.

Since this is a retrospective cohort, no standardized procedure was used for diagnostic assignment. Final clinical diagnoses, used herein as standard-of-truth, were however established by expert consensus of at least two psychiatrists based on all available information, i.e., patient's history, clinical neurological and neuropsychiatric assessments, and the global judgment of cognitive abilities following the neuropsychological examinations (CERAD and FLEXBAT). In addition, cranial MRI-scans (N = 106), FDG PET-scans (N = 85), and cerebrospinal biomarkers (Tau protein, Beta-Amyloid; N = 10) were available for some patients. The psychiatrists making the diagnosis are experts in the field of neurodegenerative disorders since the focus of the Center for Cognitive Disorders is on differential dementia diagnosis as well as differentiation between dementia and cognitive deficits caused by affective disorders in the elderly. Thus, the probability of the diagnoses being accurate can be considered high. Primary diagnoses of the 68 patients with DEP according to the International Classification of Diseases (ICD-10) were: F 32.0 = 20; F32.1 = 25; F32.2 = 2; F33.0 = 4; F33.1 = 9; F 33.2 = 1; F33.4 = 2 and F 34.1 = 5. Within this group, 16 patients also had a secondary diagnosis of F06.7 (Mild Cognitive Disorder). While according to ICD-10, this etiologically unspecific diagnosis should be made only in association with a specified physical disorder and should not be made in the presence of any mental or behavioral disorder classified with F10-F99, some psychiatrists nevertheless use F06.7 to indicate that a depressive patient exhibits cognitive impairments in addition to the predominant affective symptoms. Primary diagnoses of the 121 patients with DAT according to ICD-10 were: F00.0 = 45; F00.1 = 69; F00.2 = 7.

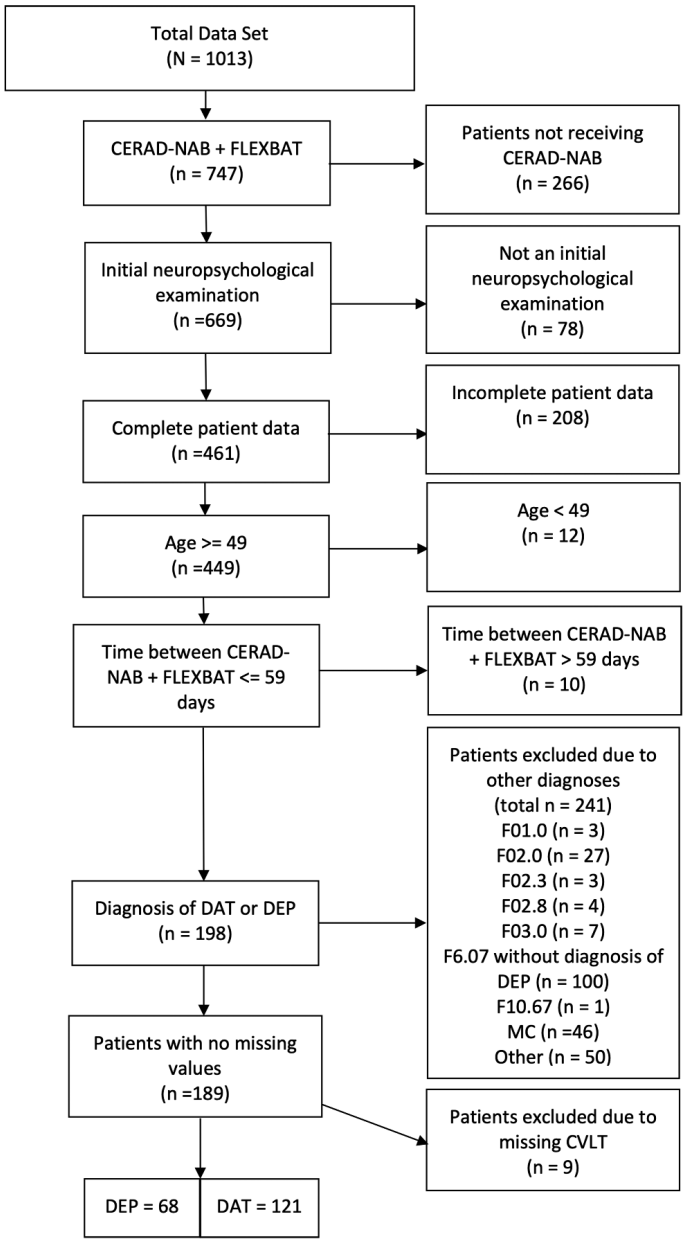


Figure 1. Summary of the selection process of the current sample from the total sample available including the specific selection criteria. CERAD-NAB = consortium to establish a registry for Alzheimer's disease neuropsychological assessment battery; CVLT = California verbal learning test; DAT = dementia of the Alzheimer type; DEP = depression; FLEXBAT = Flexible battery approach; MC = memory complainer.

There was no comorbid diagnosis of DAT and DEP for any of the patients. Demographic details of the patient sample are summarized in Table 1.

The DAT group was significantly older than the DEP group (t (187) = 5.21; $p < .0001$). Furthermore, performance on the Mini-Mental State Examination differed between the two groups. As expected, patients with DAT performed significantly worse on the MMSE

Table 1. Demographic characteristics of the patient samples.

		DAT	DEP	*p*
Number of patients	N	121	68	
Sex	% female	54	59	.50
Age (Years)	Mean (SD)	68.6 (8.8)	61.5 (9.3)	<.001
	Range	49–90	49–83	
Education (Years)	Mean (SD)	12.6 (3.1)	13.1 (3.5)	.29
	Range	7–20	7–20	
MMSE score	Mean (SD)	23.2 (3.4)	28.1 (1.7)	<.001
	Range	12–29	22–30	
BDI	N	66	67	
	Mean (SD)	9.8 (7.2)	18.6 (8.8)	
	Range	0–35	0–41	

BDI = Beck's Depression Inventory, DAT = Dementia of the Alzheimer type, DEP = Depression, MMSE = Mini-Mental State examination.

as the DEP group (t (185.06) = −13.33; $p < .001$). Scores on the Beck's Depression Inventory (BDI) were significantly lower in the DAT group than in the DEP group (t (131) = −6.31; $p < .0001$), indicating less severe depressive symptoms. There was no significant difference between the groups regarding sex or education.

within a first orienting Logistic Regression (LR) (Variance inflation factor VIF > 4; see Table S1 in the supplementary material for the results of the LR including the remaining variables after the exclusion as well as their VIF scores).

As mentioned above, a second neuropsychological examination at the Clinical and Experimental Neuropsychology Unit followed, called herein the FLEXBAT, thus realizing an individual compilation of a broader range of psychometric tests. This procedure corresponds to the requirement of the German S3 guidelines for a more elaborate neuropsychological investigation in diagnostically ambiguous cases (Deutsche Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN) & Deutsche Gesellschaft für Neurologie (DGN), 2016). While some experts decide on extending fixed test batteries such as the CERAD-NAB with additional measurement procedures (Beck et al., 2014; Schmid et al., 2014), our aim was to compare the CERAD-NAB with the tests from the FLEXBAT in terms of their differential diagnostic performance. Tests administered too rarely within the individually oriented FLEXBAT approach were excluded, leaving the following tests for data analysis:

The Aachener Aphasie Test (AAT) is a German test to investigate aphasia (Huber et al., 1983). Only results from the second subtest of the AAT (naming of pictures presenting objects, shapes, colors, situations, and actions) were used in the present investigation. The norms from the non-aphasic control patients were used for the generation of standardized values.

The California Verbal Learning Test (CVLT; German version: Niemann et al., 2008) is a comprehensive procedure to measure learning and memory of verbal material. A list of 16 words is read by an examiner five times in a fixed order. After each time the list is read, the subject is asked to recall all the words s/he remembers. Following the learning trials, an interference list is presented once (List B), including different words, which must also be recalled afterward. Then, the subject must recall the words from list A again without any help (short-delay free recall) and then indicate words from list A belonging to the four different semantic categories (short-delay cued recall). Following a break of at least 15 minutes, the long-delay free and cued-recall take place. Finally, the CVLT ends with a presentation of 44 words, including words from list A, words from list B, and completely new words. The subject is asked to indicate which words were part of list A. The American norms of the CVLT (Delis et al., 1987) had to be used for the generation of standardized values, since they – in contrast to the German norms – go beyond the age of 60 years. Since z-transformations based on German CVLT scores from healthy subjects aged 52 to 79 years (N = 52; 59.6% female; Szudrowicz, 2001) produced comparable results, use of the American norms seemed justifiable in view of the lack of alternatives.

Furthermore, a variant of the Trail Making Test A (Zahlen-Verbindungs-Test ZVT; Oswald & Roth, 1987) was used, which is a measure of CNS information processing speed that requires the participant to connect numbers from 1 to 90 as fast as possible in ascending order (four different stimulus matrices). For individuals over the age of 60 years, the geriatric version of the ZVT (Oswald & Fleischmann, 1999) was used together with its respective age norms.

Lastly, the Rey Complex Figure Test and Recognition Trial (RCFT; Meyers & Meyers, 1995) contains a complex geometric figure which must be copied (short and long delay free recalls omitted).

Variables from FLEXBAT obtained in the present study included: AATb: number of correctly named objects and situations (max. = 120), RCFT: figure copying score (max. = 36; RCFTcop) and ZVT: the mean performance time over four stimulus matrices (ZVTtot). Furthermore, in terms of the CVLT: number of correctly recalled words after first learning trial (max. = 16; CVLT1), number of correctly recalled words after fifth learning trial (max. = 16; CVLT5), total number of correctly recalled words (max. = 80; CVLTtot), number of words recalled from interference list B (max. = 16; CVLTb), short delayed free and cued recall (max. = 16 each; CVLTsf and CVLTsc), long delayed free and cued recall (max. = 16 each; CVLTlf and CVLTlc), number of intrusions (CVLTint), and from the recognition trial: hit rate (number of correctly identified words; max = 16; CVLThit), false positive responses (number of words not originally presented in list A, but falsely claimed to be part of A; max. = 28; CVLTfp), and discriminability (a measure defined by signal detection theory; max. = 100; CVLTdis). The variables CVLT1, CVLT5, CVLTsf, CVLTsc, CVLTlf and CVLTlc were further excluded, because there was evidence for multicollinearity within a first orienting Logistic Regression (LR) (VIF > 4; see Table S2 in the supplementary material for the results of the LR including the remaining variables after the exclusion as well as their VIF scores).

The two test procedures were performed independently by different test administrators at different locations (trained psychiatrists and clinical psychologists in the case of the CERAD-NAB; specialized clinical neuropsychologists in the case of the FLEXBAT). The average time interval between the CERAD-NAB, which was always administered first, and the FLEXBAT was 6 days (SD = 10).

Machine learning algorithms

The term machine learning (ML) describes various methods enabling a specific algorithm to learn statistical models from presented data and thereby improve predictions for new cases (Keith et al., 2019; Molnar, 2019). What makes ML algorithms so interesting is the fact that they, in contrast to traditional statistical methods, do not require any pre-specified hypotheses regarding the association between features and a given prediction, but can instead detect complicated interactions between variables, eventually improving classification accuracy (S. A. Graham et al., 2020). ML algorithms are robust to many predictors, even if these outnumber the quantity of observations, preventing a time-consuming preselection of variables (Countanche & Hallion, 2020). ML algorithms are thereby increasingly used to inform clinical decision making (Dwyer et al., 2018; Jiang et al., 2017).

One can differentiate between two classes of ML algorithms: Supervised and unsupervised learning algorithms. During unsupervised learning, the algorithm detects patterns within unlabeled data in the training phase, whereas within supervised learning, the algorithm learns from data which has previously been labeled with classes (diagnostic groups in this context) to identify the associations between features and classes. In the testing phase, the algorithm predicts the classes of previously unseen observations. The overall performance of the algorithm is determined by the amount of correctly classified cases. Since this was a retrospective study and clinical diagnoses were therefore already given, supervised learning algorithms were used. More detailed information about the functionality of ML and its use in clinical psychology and psychiatry can be found in Dwyer et al. (2018), Countanche and Hallion (2020), Keith et al. (2019) or Kubat (2015).

To investigate which algorithm would show the highest accuracy within the classification at hand, the performance of four algorithms using either CERAD-NAB or FLEXBAT data was investigated in the present study: Support Vector Machine (SVM) is a supervised learning algorithm, which maps the features (i.e., predictors) into a higher dimension and aims to separate the two outcome groups by means of creating a hyperplane (i.e., it does so by finding the maximal margin hyperplane on the training data, which accurately divides the two different classes; Cortes & Vapnik, 1995). The resulting hyperplane has the largest distance between the nearest training points of the different classes. Naïve Bayes (NB) is a classification algorithm based on the Bayes theorem and the theorem of total probability (Murphy, 2006), which assumes independence of predictors. It operates by counting the frequency of specific features within a given outcome class set. NB has been proven to be effective in medical diagnostics and even under circumstances where predictors are not completely independent of each other (Bhagya Shree & Sheshadri, 2018). Random Forest (RF) is an ensemble classifier, which generates many decision trees to classify a predictor's input vector (Biau & Scornet, 2016). These predictions are then averaged by the number of trees to obtain the final classification. The decision trees are built by randomly selecting a specific number of features to keep the different trees as uncorrelated as possible. RFs have been shown to not be susceptible to overfitting and insensitive to outliers (Byeon, 2020). Logistic Regression (LR), used herein for comparison with the ML approaches, models the relationship between one or more independent predictor variables and a categorical output variable by selecting parameters that maximize the probability of class membership. It is specifically designed to model probabilities and is often used as a starting point for binary classification. LR is indeed considered to be the first choice in psychiatric research whenever in a sample of patients, individual membership to one of two diagnostic classes or treatment outcomes is the problem at hand.

In the context of model comparison, the performance of these four algorithms was furthermore compared to that of a featureless learner (FL), which always predicted the most frequent class label without learning from the neuropsychological features.

The open-source software R 4.0.2 (R Core Team, 2019) was utilized for all statistical analyses. Figures and plots were created using the „ggplot2" package (Wickham, 2016). Specifically, the "mlr3" package and ecosystem (Lang et al., 2019) was used for the implementation of all ML algorithms. "mlr3" provides a unified interface for different ML algorithms and allows for the comparison of specific learning algorithms by means of benchmarking.

All algorithms were used with their default parameters as defined in mlr3. The "iml" package (Molnar et al., 2018) was implemented to assess the feature importance of the best performing combination of neuropsychological test battery and ML algorithm.

Classification procedure

Before the actual classification procedure, raw scores were transformed into z-scores for each of the neuropsychological variables based on the respective age, sex, and educational norms. Logistic regressions were then calculated for the CERAD-NAB and the FLEXBAT, to assess multicollinearity of the neuropsychological variables via the variance inflation factor (VIF). Peculiar variables where consequently excluded. The results of the

logistic regressions for both CERAD-NAB and FLEXBAT after the exclusion of these critical variables, including the VIF score of the remaining variables, is shown in Supplementary Table 1 and 2, respectively. Following the exclusion of peculiar variables within both neuropsychological test batteries, the CERAD-NAB consisted of 11 features, while the FLEXBAT consisted of 9 features.

Using the "mlr3" package, performance of individual ML algorithms (SVM, NB and RF) as well as LR, using the different neuropsychological test batteries was compared using a benchmark design. Since the sample sizes of the two groups were rather unbalanced in this investigation, which can lead to poor predictive performance for the minority class, we decided to use the SMOTE algorithm (Chawla et al., 2002), already implemented in mlr3, to oversample the minority class (i.e., the DEP group). SMOTE synthesizes new examples of the minority class by selecting a class instance by chance and finds its k-nearest minority class neighbors in feature space. SMOTE then draws a line between the two examples and creates a randomly selected point on that line between the neighbors. By doing so, as many newly synthesized examples as needed can be synthesized by SMOTE to align the group sizes. This is more effective than simply doubling already present cases of the minority group, since SMOTE creates new cases that resemble the existing cases within the minority group but are not identical. Following this procedure, the DEP group consisted of 136 cases, while the DAT group consisted of 121 cases. These synthesized cases were only used for training the model but not for the performance estimation.

Missing values on the neuropsychological test batteries (2.23% missing values in the whole dataset) were imputed by using multivariate imputation by chained equations (MICE, Van Buuren, 2018; Van Buuren & Groothuis-Oudshoorn, 2011). MICE is a multiple imputation method that imputes variables with missing values by means of creating multiple regressions models, conditional upon the other variables in the data. MICE was implemented in the R interface "NADIA" (Borowski, 2021), which allows for the usage of several advanced imputation methods as part of the "mlr3 pipelines" in the mlr3 ecosystem (Lang et al., 2019). We used the default imputation method of predictive mean matching (pmm) with 5 numbers of candidates of non-missing values to sample from. As recommended, 10 cycles were used, which generally leads to good convergence of the algorithm (Azur et al., 2011). We created 40 data sets, since the power has been shown to increase by imputing more datasets (J. W. Graham et al., 2007).

This process was conducted within every fold of the cross-validation (CV): Specifically, a repeated 5-fold CV pipeline was implemented to obtain an unbiased performance estimate in the data not previously seen during the training. Thus, the data was randomly divided into 5 different folds. Each fold was iteratively held back as a validation sample, while the other 4 folds served as the training set. This process was repeated 5 times, whereby each fold once served as the testing set and the rest as the training set. The random splitting of the data set into 5 folds was repeated 10 times. For each classifier, the mean performance over all folds was considered as the classification performance. The classification performance was primarily described by means of balanced accuracy (sensitivity + specificity/2), which accounts for this imbalance. Additionally, classification error (wrongly classified cases/N), sensitivity (proportion of positives (i.e., DAT) correctly classified as being positive), specificity (proportion of negatives (i.e., DEP) correctly classified as being negative), area under the ROC curve (AUC), the negative predictive value (true

negatives/(false negatives + true negatives)) and the positive predictive value (true positives/ (false positives + true positives)) were assessed.

To understand the relationship between individual features and the classification procedure, feature importance was calculated for the best performing combination of ML algorithm and neuropsychological test battery by means of permutation importance. Specifically, the feature importance was indicated by calculating the increase in classification error (CE) of the model after permutation of this specific feature.

The same analyses as for the whole sample (N = 257 for the training phase, N = 189 for the testing phase) were once again repeated using the DAT group (N = 121) and only those DEP patients without a secondary diagnosis of F06.7 (N = 52), to determine whether the exclusion of DEP patients with a secondary diagnosis of cognitive impairment would substantially improve accuracies. To deal with the class imbalance, the minority group of DEP patients without a diagnosis of F06.7 was again oversampled by using SMOTE (Chawla et al., 2002), resulting in 104 DEP cases and 121 DAT cases for training (total N = 125) and N = 173 for the testing phase. Another analysis was performed using age and the BDI score as sole predictors on the whole data set (N = 257), which were not used as features in the previous models, to investigate the specific value of neuropsychological examinations in the differential diagnosis of dementia by means of comparing it to the use of solely clinical measures. All relevant R scripts are available on request from the corresponding author.

Results

Neuropsychological results

In terms of the CERAD-NAB, patients with DEP performed within the normal range ($-1 < z < +1$) on most of the subtests. The group means within the verbal fluency task ($z = -1.02$), the MMSE ($z = -1.05$), the second trial learner set of the word list ($z = -1.15$) and the delayed verbal recall ($z = -1.03$) were marginally below average. The best performance was achieved within figure copying ($z = 0.13$). The DAT group on average showed performance below normal range in all subtests, despite the number of intrusions within the word learning subtest ($z = -0.99$). Within the FLEXBAT, patients with DEP performed within normal range on all subtests, despite the CVLT word list learning of list B ($z = -1.51$). The best performance was achieved within confrontation naming of the AAT ($z = 0.75$) and figure copying of the RCFT ($z = 0.29$). Patients with DAT showed impairments on all subtests, apart from the hit rate within the recognition trial of the CVLT ($z = -0.15$). For a depiction of the mean z- scores with 95% confidence intervals for each group on every subtest of both the CERAD-NAB and the FLEXBAT, see Figure 2.

Classification results

The classification performances of the four different algorithms and a featureless learner using either the CERAD-NAB or the FLEXBAT are summarized in Table 2.

As can be seen, balanced classification accuracies (BA) of the four algorithms were generally high and comparable for both neuropsychological test batteries. Using CERAD-NAB, RF had the best accuracy, while LR had the lowest. The highest sensitivity for the

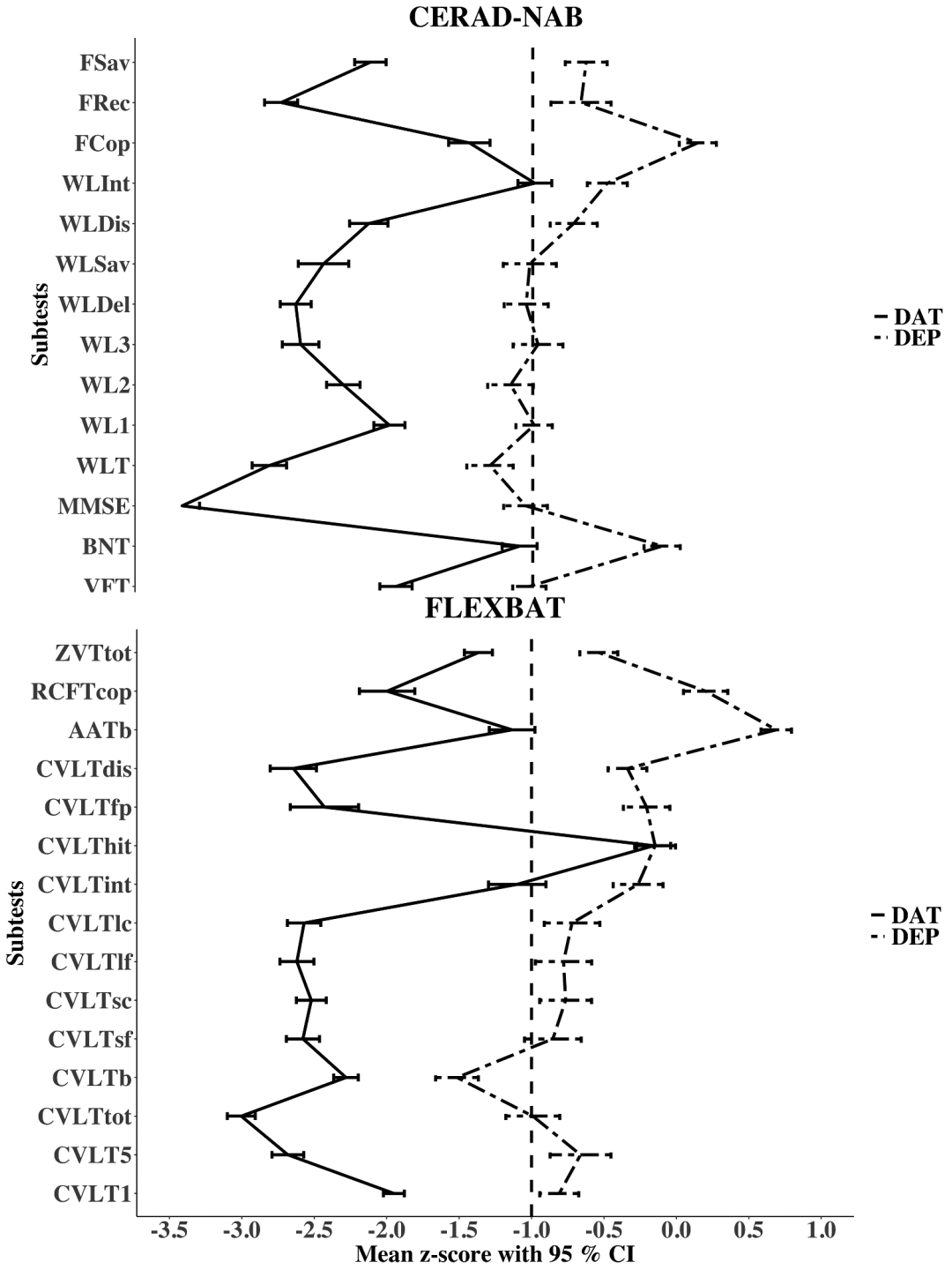
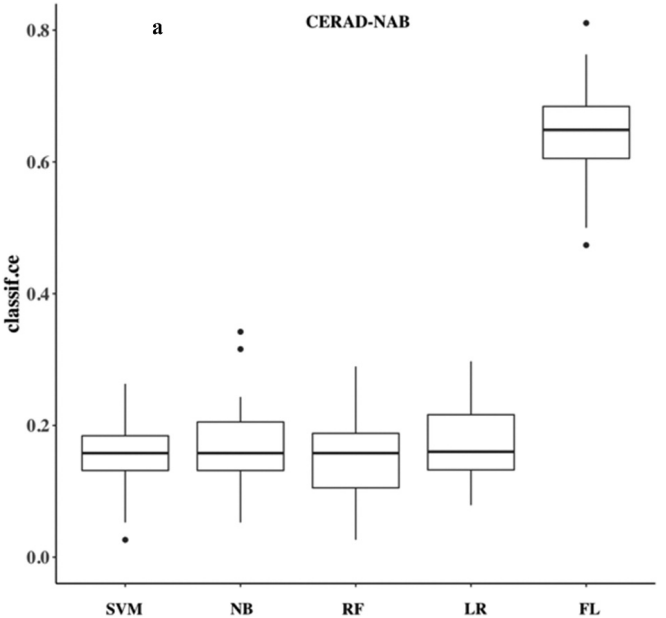


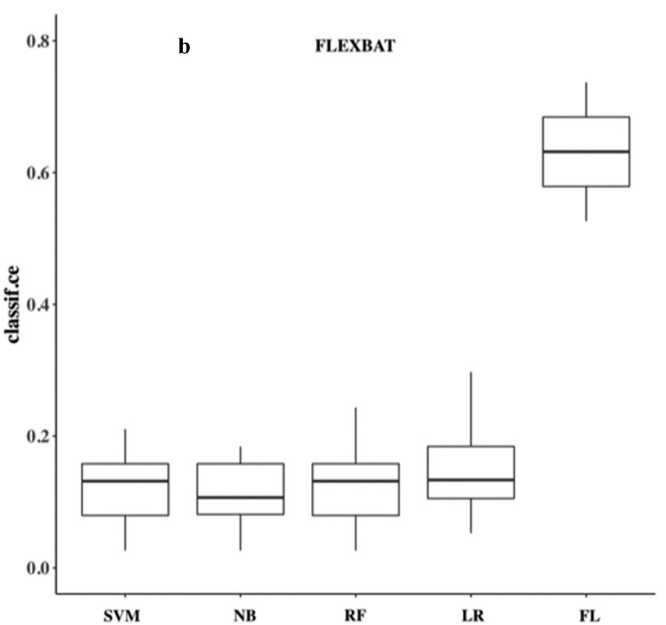
Figure 2. Comparison of the cognitive performance profile (z-values with 95% CI) of patients with depression (DEP) and patients with dementia of the Alzheimer type (DAT) for the subtests of the CERAD-NAB (A) and the subtests of the FLEXBAT-approach (B) used in the present investigation separately. A: BNT = Boston naming test, FCop = figure copying, FRec = Recall of the figures after a delay, MMSE = Mini mental state examination, VFT = Verbal fluency test, WLDel = recall of the wordlist after a delay, WLDis = word list discriminability recognition trial, WLInt = word list intrusions, WL1 = word list learning set first trial, WL2 = word list learning set second trial, WL3 = word list learning set third trial. B: AATb = Aachener Aphasie Test (Subtest "B," Confrontation Naming),

Table 2. Performance measurements for the three different ML algorithms (Support vector machine, Naïve Bayes, random forest), Featureless Lerner and conventional logistic regression separately for the two different neuropsychological test batteries (CERAD-NAB and FLEXBAT) in patients with dementia of the Alzheimer's type (N = 121) or depression (N = 68).

	CERAD					FLEXBAT				
Learners	SVM	NB	RF	LR	FL	SVM	NB	RF	LR	FL
Balanced Accuracy (%)	84.1	83.2	84.6	83.0	50.0	86.9	87.0	86.9	86.0	50.0
CE (%)	14.9	16.3	15.2	18.3	64.0	12.4	11.7	12.0	14.3	64.0
Sensitivity (%)	88.0	85.3	86.2	79.8	0	88.9	91.3	91.0	84.8	0
Specificity (%)	80.2	81.1	82.9	86.2	100	85.0	82.7	82.9	87.3	100
AUC	0.91	0.92	0.91	0.91	0.50	0.92	0.93	0.94	0.93	0.50
NPV	0.79	0.76	0.77	0.71	0.36	0.81	0.84	0.84	0.76	0.36
PPV	0.89	0.89	0.90	0.91	NA	0.92	0.91	0.91	0.92	NA

AUC = Area under the Receiver Operating Characteristic (ROC) Curve; CE = Classification Error; FL = Featureless Learner; LR = Logistic Regression NB = Naïve Bayes; NPV = Negative Predictive Value, PPV = Positive Predictive Value; RF = Random Forest; SVM = Support Vector Machine. Note that the FL showed a sensitivity of 0% and a specificity of 100% since it always simply predicted the most common class from the training data. Due to the oversampling of the DEP group in the training set, FL always predicted DEP in the test set even though the ratio of the group sizes in the test set was opposite to that within the training set.


a
CERAD-NAB
classif.ce
0.0
0.2
0.4
0.6
0.8
SVM
NB
RF
LR
FL


b
FLEXBAT
classif.ce
0.0
0.2
0.4
0.6
0.8
SVM
NB
RF
LR
FL

Using only BDI scores and age as predictors resulted in comparably lower balanced accuracies (67.6% – 79.5%). More information on the classification performance using only these clinical predictors can be found in Supplementary Table 4.

Feature Importance

Using the best performing combination of features from the FLEXBAT approach and the ML algorithm NB, the five most important features for the classification of DEP vs. DAT were the number of false positive responses in the recognition trail of the California Verbal Learning Test (CVLTfp), the hit rate in the CVLT (CVLT hit), the total number of intrusions (CVLTint), the total performance on the CVLT (CVLTtot) and the visuoconstruction performance in the RCFT (RCFTcop). Figure 4 depicts the feature importance within the present investigation by means of increase in classification error of the model after permutation of the specific feature over 5 repetitions.

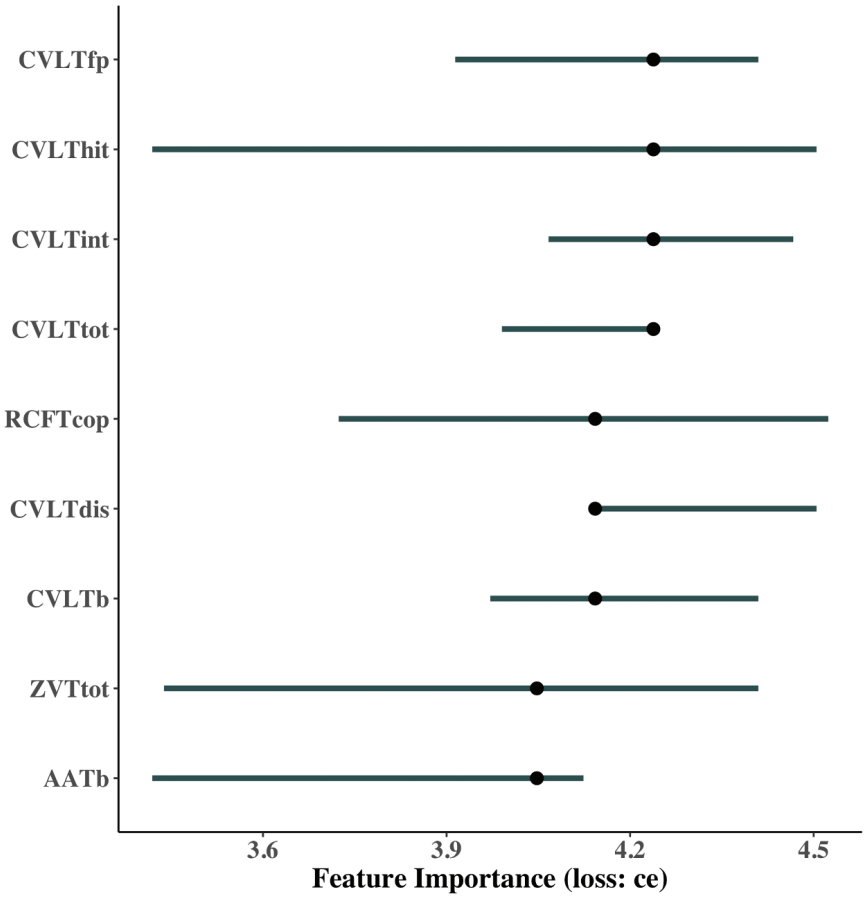


Figure 4. Importance of individual features of the FLEXBAT approach using the permutation-based importance (increase in classification error after permutation of that feature), depicted as the median loss over 5 repetitions with 95% CI. AATb = Aachener aphasie test (Subtest "B," Confrontation Naming), CVLTb = California Verbal Learning Test (learning set list B), CVLTdis = California verbal learning test (discriminability during the recognition trail), CVLTfp = California verbal learning test (false positives during recognition trail), CVLThit = California verbal learning test (hits during recognition trail), CVLTint = California verbal learning test (number of intrusions), CVLTtot = California verbal learning test (learning set total), RCFTcop = Rey complex figure test (Subtest Copying), ZVTtot = zahlen-verbindungs-test (total performance).

Discussion

Depression (DEP) and dementia of the Alzheimer's type (DAT) belong to the most common neuropsychiatric disorders observed in the elderly (Liguori et al., 2018; Tetsuka, 2021). Due to similarities in clinical presentations and overlapping cognitive and affective symptoms, the decision whether cognitive deficits of elderly patients firstly presenting to a psychiatric hospital are better explained by DAT or DEP can be extremely challenging (Karantzoulis & Galvin, 2011; Lanza et al., 2020; Leyhe et al., 2017). Since the disease courses differ considerably between DAT and DEP, the accurate differential diagnosis in clinical practice is of utmost importance (Tetsuka, 2021). The present study therefore investigated the potential of neuropsychological examinations in combination with machine learning (ML) to differentiate between DAT and DEP and to compare the classification accuracy of two different neuropsychological test batteries: the CERAD-NAB and a more comprehensive flexible battery approach (FLEXBAT). Three ML approaches (Support Vector Machine (SVM), Random Forests (RF), and Naïve Bayes (NB)) with their default parameters and conventional Logistic Regression (LR) were employed for the classification at hand and compared to a featureless learner. Due to the explorative nature of our study and the lack of similar investigations in the past, we could not form any concrete hypotheses on which combination of algorithm and neuropsychological test battery would perform best, but it was expected that ML algorithms would overall perform superior to traditional LR. This was presumed because of the focus of ML on prediction rather than interference and the fact that it has outperformed traditional statistical approaches in the context of diverse classification and prediction problems in the past (Dwyer et al., 2018; Liew et al., 2022; Orrù et al., 2020; Rajula et al., 2020; Yang et al., 2020).

However, classification results for both neuropsychological test batteries using either different ML algorithms or Logistic Regression (83.0–87.0% balanced accuracy) were similarly high. The highest classification accuracy in our analysis could be obtained using NB in combination with the more comprehensive FLEXBAT (87.0%). The most important features within this classification were the number of false positive responses in the recognition trail of the California Verbal Learning Test (CVLTfp), the hit rate in the CVLT (CVLT hit), the total number of intrusions (CVLTint), the total performance on the CVLT (CVLTtot) and the visuoconstruction performance in the RCFT (RCFTcop).

A balanced accuracy of 87.0% is noteworthy, given that all patients came to the memory clinic due to difficulties in memory and given that measures of symptom severity were not included within the present analysis. Using only those DEP patients without a secondary diagnosis of mild cognitive impairment resulted in slightly better classification performances, but accuracies were generally comparable to when these patients were included. Thus, the combination of neurocognitive data and ML seems to be useful to differentiate between DEP and DAT, even when DEP patients additionally showing clinically meaningful cognitive impairment are included. The results of this study are especially striking since no hyperparameter tuning was performed, to keep the ML algorithms as simple as possible with its default parameters, which can be used by any clinician familiar with R.

Since to our knowledge this is the first study to apply different ML algorithms to differentiate between DEP and DAT based on neuropsychological test data, it is

somewhat difficult to compare the present results to previous studies. However, the findings in terms of accuracy of classification are comparable to the findings of Gurevich et al. (2017), who predicted DAT vs. no DAT based on the CERAD-NAB and two additional tasks with accuracies of up to 89%. Accuracies within the present investigation are also comparable to Mato-Abad et al. (2018), who investigated whether mild cognitive impairment (MCI; defined as performance of one or more standard deviations below norms on neuropsychological tests in several cognitive areas) with depression and MCI without depression could be discriminated based on neuropsychological test data using artificial neural networks. As with the present results, 86% of cases were correctly classified within their investigation. Not surprisingly, our results are somewhat lower than in investigations differentiating between healthy controls and DAT (Er et al., 2017; Kang et al., 2019), reflecting the more complicated relationship between DEP and DAT.

Even though no demographic data was considered, accuracies were considerably higher than within studies using the MMSE, the Montreal Cognitive Assessment (MoCA) and the Korean Dementia Screening Questionnaire (KDSQ; Yim et al., 2020), suggesting that more comprehensive neuropsychological test batteries have an advantage in identifying the underlying cause for cognitive impairment. Indeed, when only age and BDI scores were used as predictors, we found lower balanced classification accuracies (67.6% – 79.5%).

The most important features within the classification using NB and the FLEXBAT approach (i.e., total performance, number of intrusions, hit rate, number of false positive responses within a verbal recognition task as well as figure copying) are in line with findings from other investigations (Braaten et al., 2006; Kang et al., 2019; Leyhe et al., 2017; Wright & Persad, 2007). This correspondence between existing medical knowledge and our most important features is especially noteworthy, since ML models might only be accepted by practitioners if they coincide with clinical expertise and previous research findings (Pazzani et al., 2001).

The usefulness of our approach in comparison to a clinical diagnostic decision based solely on anamnestic information and MMSE raw scores, as it has been traditionally done in clinical practice (and often continued to be done; Devenney & Hodges, 2017), is further underlined by the following example: For the DAT group, the MMSE ranged from 12 to 29. While it would be easy to classify someone with an MMSE of 12 as having DAT in clinical routine, the assignment of someone with a score of 26 would be much more questionable since DEP patients often show similar MMSE scores (range 22 to 30 in the present study). Using the NB and features from the FLEXBAT in our example, this difficulty in classification was reflected by the probability with which a specific case is considered to be part of one of the two classes (DEP and DAT), while the right assignment was still made in most of the cases: Specifically, a DAT patient with an MMSE of 12 was classified as having DAT, because according to the algorithm there was a 89.5% probability of this specific case belonging to the DAT group. A DAT patient with a comparably higher score on the MMSE (26) was also correctly classified as having DAT but with a lower probability of class membership (53%). Thus, the NB could correctly classify those cases that would probably pose the greatest challenges to clinicians in daily practice while simultaneously accounting for the difficulty of assignment.

One of the strengths of our study is clearly that the sample used herein actually resembles those patients treated by neuropsychologists in daily clinical practice. Some

patients had secondary diagnoses, which makes differential diagnosis difficult and most accurately resembles every day clinical challenges. This investigation generally provides preliminary evidence for the utility of ML and neuropsychology (especially the more comprehensive FLEXBAT) in the differentiation between DAT and DEP in assisting clinical decision-making, even though ML algorithms with their default parameters did not outperform Logistic Regression. Thus, based on our results, there was no evidence that a single ML algorithm performed significantly better than any other one, instead the use of all the algorithms investigated as well as traditional LR investigated seems feasible at this point.

The implications for clinical practice are twofold: Firstly, our study shows that more elaborate neuropsychological examinations have an advantage over symptom measures such as the BDI, implying that these procedures should be more widely used in daily clinical practice. Secondly, this study represents proof-of concept evidence that the combination of ML techniques with neuropsychological test data alone can efficiently differentiate between DEP and DAT. It represents the first step toward the long-term goal of creating a formal algorithm that can assist clinicians in individual case diagnosis by suggesting the most likely diagnosis based on the data available. Formal ML algorithms are for example, already used within the field of radiotherapy to help radiologists with the segmentation of CT or MRT data (Hosny et al., 2018). Yet, to create such a formal algorithm for neuropsychology and practically implement it, our model needs to be validated in diverse independent patient samples to investigate its generalizability.

Limitations

Firstly, our study was an explorative single-center study, resulting in relatively few patients investigated, yet other studies have used similar sample sizes (Almubark et al., 2019; Byeon, 2020), which might be appropriate for explorative purposes. The limited sample size may however reduce generalizability to other samples due to idiosyncrasies of participants from our center and might accordingly limit the predictions of our model for new cases (Dwyer et al., 2018; Schnack & Kahn, 2016). It also needs to be mentioned that imbalance of sample sizes can be a problem for ML algorithms, since they assume equal sample sizes. To deal with the imbalanced sample sizes, we used the oversampling method SMOTE, which has been shown to be very effective (Chawla et al., 2002; Qu et al., 2020). Even though the unequal prevalence rates of both DAT and DEP in memory clinics (Kao et al., 2019; Knapskog et al., 2014) is accurately reflected by the unequal group sizes, it might still be useful to align the sample sizes in prospective studies to improve algorithm performance and to circumvent the need for synthetic data generation. Again, as this study was conceptualized as a proof-of-concept, the sample size was considered sufficient.

Secondly, because of the ex post facto design of our study, the diagnoses of patients were not independent of the neuropsychological examination and did not follow a standardized diagnostic procedure, otherwise typical for prospective studies. This limitation characterizes many previous investigations too (Bachli et al., 2020; Kang et al., 2019; Williams et al., 2013) and might also be acceptable for explorative purposes. Usually, the psychiatrists responsible for the final diagnoses in this study relied on global judgments by the psychologists performing the neuropsychological examination, without inspection of the cognitive

performance profile or individual neuropsychological variables in detail by themselves. After all, the diagnostic procedure in our patients additionally included techniques other than neuropsychological examinations, such as history taking, neuroimaging, CSF biomarkers, and neurological examinations, although not in every single patient. Moreover, final diagnoses were always made following expert consensus of at least two experienced psychiatrists in a clinic specialized at differential dementia diagnosis, considering all available results from diverse investigation methods. The accuracy of psychiatric DAT and DEP diagnosis have been found to be relatively high, justifying the use of psychiatric diagnoses as the outcome measures (Beach et al., 2012; Reed et al., 2018; Snowden et al., 2011; Sommerlad et al., 2018).

A further drawback is that it could not be investigated whether differences in performance of the different classification algorithms are actually statistically significant, since this is currently not supported for non-independent tasks in mlr3 (GitHub – mlr-org /mlr3benchmark: Analysis and tools for benchmarking in mlr3 and beyond.). Nevertheless, confidence intervals for the classification errors (CE; Figure 3) suggest that accuracies were quite similar across algorithms.

Future research

Validation of our results within independent patient samples from cross-center approaches will be of major interest in the future. This is especially important for the creation of an empirically based formal algorithm that can be translated into clinical practice. ML algorithms need to be well-trained on massive datasets including numerous different individuals from various countries to make their predictions as generalizable as possible. Following this procedure, ML algorithms could then actually be used by clinicians to inform individual case diagnosis in clinical practice.

No regularization procedure was attempted since the goal of this investigation was to firstly determine a simple and easy to administer ML algorithm with its default parameters, to investigate whether this would already perform better than LR. It might nevertheless be worth to optimize hyperparameters prospectively, to see whether ML performance can further be boosted and consecutively eventually outperform traditional statistical methods such as LR. To further increase classification accuracy, future studies should also try even larger sets of neuropsychological features and combine them with neuroimaging data or cerebrospinal fluid markers. In their noteworthy study, Bachli and colleagues (2020) could predict behavioral variant frontotemporal dementia (bvFTD) and Alzheimer's dementia with over 90% accuracy, including scores from neuropsychological examinations and brain atrophy volumes, indicating that the combination of neuropsychological test scores with biomarkers might further boost classification performance. Finally, the differentiation between DAT and DEP in the elderly is not the only intricate diagnostic question clinicians have to deal with. The differentiation between different types of dementia is often just as difficult (Begali, 2020; Byeon, 2020). Future research should investigate whether the combination of neuropsychological examinations and ML algorithms is thereby equally effective.

Under ideal conditions, future studies would train and test their models on much larger sample sizes ($N > 200$; Dwyer & Koutsouleris, 2022) than used in this proof-of-concept investigation, focus more on the representativeness of the sample investigated by using the data of patients from different culturally distinct centers and externally validate their

results in an independent patient sample (Dwyer et al., 2018). The studies should also be planned prospectively and use biomarkers in addition to neuropsychological test data to improve classification accuracies. Ideally, the diagnosis of DAT would not only be based on biomarkers but additionally validated by postmortem brain examinations. Although this approach was far beyond our possibilities and would require plenty of material and human resources, this ideal design should be kept in mind when planning future studies.

Conclusion

We show that the combination of neuropsychological test data and ML algorithms with their default parameters can differentiate DAT and DEP in elderly patients with balanced accuracies of 87%. Our work thus provides preliminary evidence that ML approaches can inform clinical decision making in the context of the clinically challenging differential diagnosis between DAT and DEP using solely cognitive test data. Yet, common ML algorithms with their default parameters did not clearly outperform traditional statistical measures (i.e., Logistic Regression) within this specific classification task.

Acknowledgments

We gratefully acknowledge the work of past and present staff of the Center for Cognitive Disorders and the Clinical and Experimental Neuropsychology Unit at the Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar, Technical University of Munich.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID

Janine Diehl-Schmid http://orcid.org/0000-0002-7745-1382

References

Abdulkadir, A., Mortamet, B., Vemuri, P., Jack, C. R., Jr, Krueger, G., & Kloppel, S. (2011). Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. *Neuroimage*, *58*(3), 785–792. https://doi.org/10.1016/j.neuroimage.2011.06.029

Ahern, E., & Semkovska, M. (2017). Cognitive functioning in the first episode of major depressive disorder: A systematic review and meta-analysis. *Neuropsychology*, *31*(1), 52–72. https://doi.org/10.1037/neu0000319

Ahmed, M. R., Zhang, Y., Feng, Z., Lo, B., Omer, T., & Liao, H. (2018). Neuroimaging and machine learning for dementia diagnosis: Recent advancements and future prospects. *IEEE Reviews in Biomedical Engineering*, *12*, 19–33. https://doi.org/10.1109/RBME.2018.2886237

Almubark, I., Chang, L., Nguyen, T., Turner, R. S., & Jiang, X. (2019). *Early detection of Alzheimer's disease using patient neuropsychological and cognitive data and machine learning techniques*. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019, 5971–597.

Alzheimer's Disease International. (2019). *World Alzheimer report 2019: About Alzheimer's & dementia*. Retrieved August 03, 2021, from https://www.alz.co.uk/research/world-report-2019

Andreas, S., Schulz, H., Volkert, J., Dehoust, M., Sehner, S., Suling, A., Ausín, B., Canuto, A., Crawford, M., Da Ronch, C., Grassi, L., Hershkovitz, Y., Muñoz, M., Quirk, A., Rotenstein, O., Santos-Olmo, A. B., Shalev, A., Strehle, J., Weber, K., . . . Härter, M. (2018). Prevalence of mental disorders in elderly people: The European MentDis_ICF65+ study. *Journal of Psychiatry*, *210*(2), 125–131. https://doi.org/10.1192/bjp.bp.115.180463

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, *145*(Pt B), 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Bachli M Belen *et al* . (2020). Evaluating the reliability of neurocognitive biomarkers of neurodegenerative diseases across countries: A machine learning approach. NeuroImage, 208 116456 10.1016/j.neuroimage.2019.116456

Bachli, M. B., Sedeño, L., Ochab, J. K., Piguet, O., Kumfor, F., Reyes, P., Torralva, T., Roca, M., Cardona, J. F., Campo, C. G., Herrera, H., Slachevsky, A., Matallana, D., Manes, F., García, A. M., Ibáñez, A., & Chialvo, D. R. (2020). Evaluating the reliability of neurocognitive biomarkers of neurodegenerative diseases across countries: A machine learning approach. *Neuroimage*, *208*, 116456. https://doi.org/10.1016/j.neuroimage.2019.116456

Bacigalupo, I., Mayer, F., Lacorte, E., Di Pucchio, A., Marzolini, F., Canevelli, M., Di Fiandra, T., & Vanacore, N. (2018). A systematic review and meta-analysis on the prevalence of dementia in Europe: Estimates from the highest-quality studies adopting the DSM IV diagnostic criteria. *Journal of Alzheimer's Disease*, *66*(4), 1471–1481. https://doi.org/10.3233/JAD-180416

Beach, T. G., Monsell, S. E., Phillips, L. E., & Kukull, W. (2012). Accuracy of the clinical diagnosis of *Alzheimer* disease at national institute on aging *Alzheimer* disease centers, 2005-2010. *Journal of Neuropathology and Experimental Neurolology*, *71*(4), 266–273. https://doi.org/10.1097/NEN.0b013e31824b211b

Beblo, T., Sinnamon, G., & Baune, B. T. (2011). Specifying the neuropsychology of affective disorders: Clinical, demographic and neurobiological factors. *Neuropsychology Review*, *21*(4), 337–359. https://doi.org/10.1007/s11065-011-9171-0

Beck, I. R., Schmid, N. S., Berres, M., & Monsch, A. U. (2014). Establishing robust cognitive dimensions for characterization and differentiation of patients with Alzheimer's disease, mild cognitive impairment, frontotemporal dementia and depression. *International Journal of Geriatric Psychiatry*, *29*(6), 624–634. https://doi.org/10.1002/gps.4045

Begali, V. L. (2020). Neuropsychology and the dementia spectrum: Differential diagnosis, clinical management, and forensic utility. *NeuroRehabilitation*, *46*(2), 181–194. https://doi.org/10.3233/nre-192965

Bhagya Shree, S. R., & Sheshadri, H. S. (2018). Diagnosis of Alzheimer's disease using naive *Bayesian* classifier. *Neural Computing and Applications*, *29*(1), 123–132. https://doi.org/10.1007/s00521-016-2416-3

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Borowski, J. (2021). NADIA: NA data imputation algorithms. R package version 0.4.1. https://cran.r-project.org/web/packages/NADIA/NADIA.pdf

Braaten, A. J., Parsons, T. D., McCue, R., Sellers, A., & Burns, W. J. (2006). Neurocognitive differential diagnosis of dementing diseases: Alzheimer's Dementia, vascular dementia, frontotemporal dementia, and major depressive disorder. *International Journal of Neuroscience*, *116*(11), 1271–1293. https://doi.org/10.1080/00207450600920928

Bruun, M., Rhodius-Meester, H. F. M., Koikkalainen, J., Baroni, M., Gjerum, L., Lemstra, A. W., Barkhof, F., Remes, A. M., Urhemaa, T., Tolonen, A., Rueckert, D., van Gils, M., Frederiksen, K. S., Waldemar, G., Scheltens, P., Mecocci, P., Soininen, H., Lötjönen, J., Hasselbalch, S. G., & van der Flier, W. M. (2018). Evaluating combinations of diagnostic tests to discriminate different dementia types. *Alzheimers & Dementia (Amst)*, *10*(1), 509–518. https://doi.org/10.1016/j.dadm.2018.07.003

Butters, M. A., Whyte, E. M., Nebes, R. D., Begley, A. E., Dew, M. A., Mulsant, B. H., Zmuda, M. D., Bhalla, R., Meltzer, C. C., Pollock, B. G., Reynolds, C. F., 3rd, & Becker, J. T. (2004). The nature and determinants of neuropsychological functioning in late-life depression. *Archives of General Psychiatry*, *61*(6), 587–595. https://doi.org/10.1001/archpsyc.61.6.587

Byeon, H. (2020). Is the random forest algorithm suitable for predicting parkinson's disease with mild cognitive impairment out of parkinson's disease with normal cognition? *International Journal of Environmental Research and Public Health*, *17*(7), 2594. https://doi.org/10.3390/ijerph17072594

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, R., & Herskovits, E. H. (2010). Machine-learning techniques for building a diagnostic model for very mild dementia. *Neuroimage*, *52*(1), 234–244. https://doi.org/10.1016/j.neuroimage.2010.03.084

Claus, J. J., Staekenborg, S. S., Roorda, J. J., Stevens, M., Herderschee, D., van Maarschalkerweerd, W., Schuurmans, L., Tielkes, C. E. M., Koster, P., Bavinck, C., & Scheltens, P. (2016). Low prevalence of mixed dementia in a cohort of 2,000 elderly patients in a memory clinic setting. *Journal of Alzheimer's Disease*, *50*(3), 797–806. https://doi.org/10.3233/JAD-150796

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Coutanche, M. N., & Hallion, L. S. (2020). Machine learning for clinical psychology and clinical neuroscience. In A. G. C. Wright & M. N. Hallquist (Eds.), *The Cambridge Handbook of Research Methods in Clinical Psychology*. Cambridge University Press.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test (CVLT) adult version* (*Research edition* ed.). The Psychological Corporation/Harcourt Brace Jovanovich.

Deutsche Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN), & Deutsche Gesellschaft für Neurologie (DGN). (Eds.) (2016, January). *S3-leitlinie "Demenzen" [German S3 guidelines on dementia]* http://www.dgn.org/images/red_leitlinien/LL_2016/PDFs_Download/038013_LL_Demenzen_2016.pdf

Devenney, E., & Hodges, J. R. (2017). The mini-mental state examination: Pitfalls and limitations. *Practical Neurology*, *17*(1), 79–80. https://doi.org/10.1136/practneurol-2016-001520

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Dwyer, D., & Koutsouleris, N. (2022). Annual research review: Translational machine learning for child and adolescent psychiatry. *The Journal of Child Psychology and Psychiatry*, *63*(4), 421–443. https://doi.org/10.1111/jcpp.13545

Er, F., Iscen, P., Sahin, S., Çinar, N., Karsidag, S., & Goularas, D. (2017). Distinguishing age-related cognitive decline from dementias: A study based on machine learning algorithms. *Journal of Clinical Neuroscience*, *42*, 186–192. http://doi.org/10.1016/j.jocn.2017.03.021

Graham, S. A., Lee, E. E., Jeste, D. V., Van Patten, R., Twamley, E. W., Nebeker, C., Yamada, Y., Kim, H.-C., & Depp, C. A. (2020). Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry Research*, *284*, 112732. https://doi.org/10.1016/j.psychres.2019.112732

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Gupta, A., & Kahali, B. (2020). Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests. *Alzheimers & Dementia (NY)*, *6*(1), e12049. https://doi.org/10.1002/trc2.12049

Gurevich, P., Stuke, H., Kastrup, A., Stuke, H., & Hildebrandt, H. (2017). Neuropsychological testing and machine learning distinguish Alzheimer's disease from other causes for cognitive impairment. *Frontiers in Aging Neuroscience*, *9*(114). https://doi.org/10.3389/fnagi.2017.00114

Hansen, S., Keune, J., Küfner, K., Meister, R., Habich, J., Koska, J., Förster, S., Oschmann, P., & Keune, P. M. (2022). The congruency of neuropsychological and F18-FDG brain PET/CT diagnostics of Alzheimer's disease (AD) in routine clinical practice: Insights from a mixed neurological patients cohort. *BMC Neurology*, *22*(1), 83. https://doi.org/10.1186/s12883-022-02614-4

Horgan, D., Nobili, F., Teunissen, C., Grimmer, T., Mitrecic, D., Ris, L., Pirtosek, Z., Bernini, C., Federico, A., Blackburn, D., Logroscino, G., & Scarmeas, N. (2020). Biomarker testing: Piercing the fog of Alzheimer's and related dementias. *Biomedicine Hub*, *5*(3), 511233. https://doi.org/10.1159/00051123

Hosmer, D. W., Jr, Lemeshow, S., & Sturidvant, R. X. (2013). *Applied logistic regression (Vol.398)*. John Wiley & Sons.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

Huber, W., Poeck, K., Weniger, D., & Willmes, K. (1983). *Aachener Aphasie Test (AAT): Handanweisung*. Verlag für Psychologie Hogrefe.

Hunter, C. A., Kirson, N. Y., Desai, U., Cummings, A. K., Faries, D. E., & Birnbaum, H. G. (2015). Medical costs of Alzheimer's disease misdiagnosis among US Medicare beneficiaries. *Alzheimers & Dementia*, *11*(8), 887–895. https://doi.org/10.1016/j.jalz.2015.06.1889

Jahn, T., Theml, T., Diehl-Schmid, J., Grimmer, T., Heldmann, B., Pohl, C., & Lautenschlager, N. (2004). CERAD-NP und flexible battery approach in der neuropsychologischen differenzialdiagnostik demenz versus depression [CERAD-NP and flexible battery approach in the neuropsychological differential diagnosis of dementia versus depression]. *Zeitschrift für Gerontopsychologie & -psychiatrie*, *17*(2), 77–95. https://doi.org/10.1024/1011-6877.17.2.77

Jammeh, E. A., Carroll, C. B., Pearson, S. W., Escudero, J., Anastasiou, A., Zhao, P., Chenore, T., Zajicek, J., & Ifeachor, E. (2018). Machine-learning based identification of undiagnosed dementia in primary care: A feasibility study. *BJGP Open*, *2*(2), 101589. https://doi.org/10.3399/bjgpopen18X101589

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Kang, M. J., Kim, S. Y., Na, D. L., Kim, B. C., Yang, D. W., Kim, E.-J., Na, H. R., Han, H. J., Lee, J.-H., Kim, J. H., Park, K. H., Park, K. W., Han, S.-H., Kim, S. Y., Yoo, S. J., Yoo, B., Seo, S. W., Moon, S. Y., Yang, Y., & Youn, Y. C. (2019). Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data. *BMC Medical Informatics and Decision Making*, *19*(1), 231. https://doi.org/10.1186/s12911-019-0974-x

Kao, S.-L., Chen, S.-C., Li, -Y.-Y., & Lo, R. Y. (2019). Diagnostic diversity among patients with cognitive complaints: A 3-year follow-up study in a memory clinic. *International Journal of Geriatric Psychiatry*, *34*(12), 1900–1906. https://doi.org/10.1002/gps.5207

Karantzoulis, S., & Galvin, J. E. (2011). Distinguishing Alzheimer's disease from other major forms of dementia. *Expert Review of Neurotherapeutics*, *11*(11), 1579–1591. https://doi.org/10.1586/ern.11.155

Keith, J., Williams, M., Taravath, S., & Lecci, L. (2019). A clinician's guide to machine learning in neuropsychological research and practice. *Journal of Pediatric Neuropsychology*, *5*(4), 177–187. https://doi.org/10.1007/s40817-019-00075-1

Knapskog, A.-B., Barca, M. L., & Engedal, K. (2014). Prevalence of depression among memory clinic patients as measured by the Cornell scale of depression in dementia. *Aging & Mental Health*, *18*(5), 579–587. https://doi.org/10.1080/13607863.2013.827630

Kubat, M. (2015). Artificial Neural Networks. In: *An Introduction to Machine Learning* Springer. https://doi.org/10.1007/978-3-319-20010-1_5

Lang, M., Binder, M., Richter, J., Schratz, P. F. P., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903. https://doi.org/10.21105/joss.01903

Lanza, C., Sejunaite, K., Steindel, C., Scholz, I., & Riepe, M. W. (2020). Cognitive profiles in persons with depressive disorder and Alzheimer's disease. *Brain Communications*, *2*(2), fcaa206. https://doi.org/10.1093/braincomms/fcaa206

Leyhe, T., Reynolds, C. F., Melcher, T., Linnemann, C., Klöppel, S., Blennow, K., Zetterberg, H., Dubois, B., Lista, S., & Hampel, H. (2017). A common challenge in older adults: Classification, overlap, and therapy of depression and dementia. *Alzheimer's & Dementia*, *13*(1), 59–71. https://doi.org/10.1016/j.jalz.2016.08.007

Liew, B. X. W., Kovacs, F. M., Rügamer, D., & Royuela, A. (2022). Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain. *European Spine Journal: Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, *31*(8), 2082–2091. https://doi.org/10.1007/s00586-022-07188-w

Liguori, C., Pierantozzi, M., Chiaravalloti, A., Sancesario, G. M., Mercuri, N. B., Franchini, F., Schillaci, O., & Sancesario, G. (2018). When cognitive decline and depression coexist in the elderly: CSF biomarkers analysis can differentiate Alzheimer's disease from Late-life depression. *Frontiers in Aging Neuroscience*, *10*, 38. https://doi.org/10.3389/fnagi.2018.00038

Lin, C.-Y., Chen, T.-B., Lin, K.-N., Yeh, Y.-C., Chen, W.-T., Wang, K.-S., & Wang, P.-N. (2014). Confrontation naming errors in Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, *37*(1–2), 86–94. https://doi.org/10.1159/000354359

Linnemann, C., & Lang, U. E. (2020). Pathways connecting late-life depression and dementia. *Frontiers in Pharmacology*, *11*, 279. https://doi.org/10.3389/fphar.2020.00279

Lobo, A., Saz, P., Marcos, G., Día, J. L., & De-la-Cámara, C. (1995). The prevalence of dementia and depression in the elderly community in a Southern European population. The Zaragoza study. *Archives of General Psychiatry*, *52*(6), 497–506. https://doi.org/10.1001/archpsyc.1995.03950180083011

Logie, R., Parra, M., & Sala, S. (2015). From cognitive science to dementia assessment. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 81–91. https://doi.org/10.1177/2372732215601370

Mansoor, Y., Jastrzab, L., Dutt, S., Miller, B. L., Seeley, W. W., & Kramer, J. H. (2015). Memory profiles in pathology or biomarker confirmed Alzheimer disease and frontotemporal dementia. *Alzheimer's Disease and Associated Disorders*, *29*(2), 135–140. https://doi.org/10.1097/wad.0000000000000062

Maroco, J., Silva, S., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, *4*(1), 299. https://doi.org/10.1186/1756-0500-4-299

Mato-Abad, V., Jiménez, I., García-Vázquez, R., Aldrey, M. J., Rivero, D., Cacabelos, P., Andrade-Garda, J., Pías-Peleteiro, J. M., & Rodríguez-Yáñez, S. (2018). Using artificial neural networks for identifying patients with mild cognitive impairment associated with depression using neuropsychological test features. *Applied Sciences*, *8*(9), 1629. https://doi.org/10.3390/app8091629

Meyers, J. E., & Meyers, K. R. (1995). *Rey complex tableure test and recognition trial: Professional manual*. Psychological Assessment Resources.

Molnar, C. (2019). *Interpretable machine learning - A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Molnar, C., Bischl, B., Casalicchio, G. (2018). iml: An R package for interpretable machine learning." *Journal of Open Source Software*, *3*(26), 786. https://doi.org/10.21105/joss.00786

Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, *18*(60), 1–8.

Niemann, H., Sturm, W., Thoene-Otto, A., & Willmens, K. (2008). *Der California Verbal Learning Test CVLT*. Pearsson-Assessment.

Olson, L. T., Smerbeck, A., Figueroa, C. M., Raines, J. M., Szigeti, K., Schretlen, D., & Benedict, R. H. B. (2021). Preliminary validation of the global neuropsychological assessment in Alzheimer's disease and healthy volunteers. *Assessment*, 1073191121991221. https://doi.org/10.1177/1073191121991221

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology*, *10*, 2970. https://doi.org/10.3389/fpsyg.2019.02970

Oswald, W. D., & Fleischmann, U. M. (1999). *Nürnberger-Alters-Inventar (NAI)* (4th ed.). Verlag für Psychologie Hogrefe.

Oswald, W. D., & Roth, E. (1987). *Der Zahlen-Verbindungs-Test (ZVT); ein sprachfreier Intelligenz-Test zur Messung der "kognitiven Leistungsgeschwindigkeit"; Handanweisung*. Verlag für Psychologie Hogrefe.

Pasternak, E., & Smith, G. (2019). Cognitive and neuropsychological examination of the elderly. *Handbook of Clinical Neurology*, *167*, 89–104. https://doi.org/10.1016/B978–0-12-804766–8.00006–6

Pazzani, M. J., Mani, S., & Shankle, W. R. (2001). Acceptance of rules generated by machine learning among medical experts. *Methods of Information in Medicine*, *40*(5), 380–385. https://doi.org/10.1055/s-0038-1634196

Perini, G., Ramusino, M. C., Sinforiani, E., Bernini, S., Petrachi, R., & Costa, A. (2019). Cognitive impairment in depression: Recent advances and novel treatments. *Neuropsychiatry Disease and Treatment*, *15*, 1249–1258. https://doi.org/10.2147/NDT.S199746

Qu, W., Balki, I., Mendez, M., Valen, J., Levman, J., & Tyrrell, P. N. (2020). Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray. *International Journal of Computer Assisted Radiology and Surgery*, *15*(12), 2041–2048. https://doi.org/10.1007/s11548-020-02260-6

Rajula, H., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina (Kaunas, Lithuania)*, *56*(9), 455. https://doi.org/10.3390/medicina56090455

R Core Team (2019) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reed, G., Sharan, P., Rebello, T. J., Keeley, J. W., Medina-Mora, M. E., Gureje, O., Ayuso-Mateos, J. L., Kanba, S., Khoury, B., Kogan, C. S., Krasnow, V. N., Maj, M., de Jesus Mari, J., Stein, D. J., Zhao, M., Akiyama, Andrews, T., Asevedo, E., Chedour, M., … Pike, K. M. (2018). The ICD-11 developmental field study of reliability of diagnoses of high-burden mental disorders: Results among adult patients in mental health settings of 13 countries. *World Psychiatry*, *17*(2), 174–186. https://doi.org/10.1002/wps.20524

Rushing, N. C., Sachs-Ericsson, N., & Steffens, D. C. (2014). Neuropsychological indicators of preclinical Alzheimer's disease among depressed older adults. *Aging, Neuropsychology, and Cognition*, *21*(1), 99–128. https://doi.org/10.1080/13825585.2013.795514

Schmand, B., Eikelenboom, P., & van Gool, W. A., & Alzheimer's Disease Neuroimaging Initiative. (2011). Value of neuropsychological tests, neuroimaging, and biomarkers for diagnosing Alzheimer's disease in younger and older age cohorts. *Journal of the American Geriatric Society*, *59* (9), 1705–1710. https://doi.org/10.1111/j.1532-5415.2011.03539.x

Schmand, B., Huizenga, H. M., & van Gool, W. A. (2010). Meta-analysis of CSF and MRI biomarkers for detecting preclinical Alzheimer's disease. *Psychological Medicine*, *40*(1), 135–145. https://doi.org/10.1017/S0033291709991516

Schmid, N. S., Ehrensperger, M. M., Berres, M., Beck, I. R., & Monsch, A. U. (2014). The extension of the German CERAD neuropsychological assessment battery with tests assessing subcortical, executive and frontal functions improves accuracy in dementia diagnosis. *Dementia and Geriatric Cognitive Disorders Extra*, *4*(2), 322–334. https://doi.org/10.1159/000357774

Schnack, H. G., & Kahn, R. S. (2016). Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Frontiers in Psychiatry*, *7*, 50. https://doi.org/10.3389/fpsyt.2016.00050

Semkovska, M., Quinlivan, L., O'Grady, T., Johnson, R., Collins, A., O'Connor, J., Knittle, H., Ahern, E., & Gload, T. (2019). Cognitive function following a major depressive episode: A systematic review and meta-analysis. *Lancet Psychiatry*, *6*(10), 851–861. https://doi.org/10.1016/S2215-0366(19)30291-3

Shatte, A., Hutchinson, D., & Teague, S. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1–23. https://doi.org/10.1017/S0033291719000151

Snowden, J. S., Thompson, J. C., Stopford, C. L., Richardson, A. M. T., Gerhard, A., Neary, D., & Mann, D. M. A. (2011). The clinical diagnosis of early-onset dementias: Diagnostic accuracy and clinicopathological relationships. *Brain*, *134*(9), 2478–2492. https://doi.org/10.1093/brain/awr189

Sommerlad, A., Perera, G., Singh-Manoux, A., Lews, G., Stewart, R., & Livingston, G. (2018). Accuracy of general hospital dementia diagnoses in England: Sensitivity, specificity, and predictors of diagnostic accuracy 2008-2016. *Alzheimer's & Dementia*, S1552526018300669. https://doi.org/10.1016/j.jalz.2018.02.012

Steffens, D. C., Fisher, G. G., Langa, K. M., Potter, G. G., & Plassman, B. L. (2009). Prevalence of depression among older Americans: The aging, demographics and memory study. *International Psychogeriatrics*, *21*(5), 879–888. https://doi.org/10.1017/S1041610209990044

Steffens, D. C., & Potter, G. G. (2008). Geriatric depression and cognitive impairment. *Psychological Medicine*, *38*(2), 163–175. https://doi.org/10.1017/s003329170700102x

Szudrowicz, K. (2001). Profilanalyse kognitiver Leistungen bei Patienten mit leichter Alzheimer Demenz oder Depression und einer gesunden Kontrollgruppe im Alter von 51 bis 80 als Beitrag der Differentialdiagnose von Depression und beginnender Demenz vom Alzheimer Typ [Thesis, Ludwig-Maximilian University Munich]

Tetsuka, S. (2021). Depression and dementia in older adults: A neuropsychological review. *Aging and Disease*, *12*(8), 1920–1934. https://doi.org/10.14336/AD.2021.0526

Thalmann, B., Monsch, A., Schneitter, M., Bernasconi, F., & Staehelin, H. (2000). The cerad neuropsychological assessment battery (Cerad-NAB) A minimal data set as a common tool for German-speaking Europe. *Neurobiology of Aging*, *21*(1), 30. https://doi.org/10.1016/S0197-4580(00)82810-9

Thomas, A. J., Gallagher, P., Robinson, L. J., Porter, R. J., Young, A. H., Ferrier, I. N., & O'Brien, J. T. (2009). A comparison of neurocognitive impairment in younger and older adults with major depression. *Psychological Medicine*, *39*(5), 725–733. https://doi.org/10.1017/S0033291708004042

Tonga, J. B., Benth, J. Š., Arnevik, E. A., Werheid, K., Korsnes, M. S., & Ulstein, I. D. (2021). Managing depressive symptoms in people with mild cognitive impairment and mild dementia with a multicomponent psychotherapy intervention: A randomized controlled trail. *International Psychogeriatrics*, *33*(3), 217–231. https://doi.org/10.1017/S1041610220000216

Trambaiolli, L. R., Lorena, A. C., Fraga, F. J., Kanda, P. A., Anghinah, R., & Nitrini, R. (2011). Improving Alzheimer's disease diagnosis with machine learning techniques. *Clinical EEG and Neuroscience*, *42*(3), 160–165. https://doi.org/10.1177/155005941104200304

Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Weintraub, S., Wicklund, A. H., & Salmon, D. P. (2012). The neuropsychological profile of Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*, *2*(4), a006171–a006171. https://doi.org/10.1101/cshperspect.a006171

Welsh, K. A., Butters, N., Mohs, R. C., Beekly, D., Edland, S., Fillenbaum, G., & Heyman, A. (1994). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part V. A normative study of the neuropsychological battery. *Neurology*, *44*, 609–614. https://doi.org/10.1212/wnl.44.4.609

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer. https://ggplot2.tidyverse.org

Williams, J. A., Weakley, A., Cool, D. J., & Schmitter-Edgecombe, M. (2013). Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In *Workshops at the twenty-seventh AAAI conference on artificial intelligence*, (Bellevue, WA), 71–76.

Wolfsgruber, S., Jessen, F., Wiese, B., Stein, J., Bickel, H., Mösch, E., Weyerer, S., Werle, J., Pentzek, M., Fuchs, A., Köhler, M., Bachmann, C., Riedel-Heller, S. G., Scherer, M., Maier, W., & Wagner, M., & AgeCoDe study group. (2014). The CERAD neuropsychological assessment battery total score detects and predicts Alzheimer disease dementia with high diagnostic accuracy. *American Journal of Geriatric Psychiatry*, *22* (10), 1017–1028. https://doi.org/10.1016/j.jagp2012.08.021

Woodford, H. J., & George, J. (2007). Cognitive assessment in the elderly: A review of clinical methods. *Qjm: An International Journal of Medicine*, *100*(8), 469–484. https://doi.org/10.1093/qjmed/hcm051

Wright, S. L., & Persad, C. (2007). Distinguishing between depression and dementia in older persons: Neuropsychological and neuropathological correlates. *Journal of Geriatric Psychiatry and Neurology*, *20*(4), 189–198. https://doi.org/10.1177/0891988707308801

Yang, J. H., Park, J. H., Jang, S. H., & Cho, J. (2020). Novel method of classification in knee osteoarthritis: Machine learning application versus logistic regression model. *Annals of Rehabilitation Medicine*, *44*(6), 415–427. https://doi.org/10.5535/arm.20071

Yim, D., Yeo, T. Y., & Park, M. H. (2020). Mild cognitive impairment, dementia, and cognitive dysfunction screening using machine learning. *Journal of International Medical Research*, *48*(7), 300060520936881. https://doi.org/10.1177/0300060520936881

Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T., & Uddin, M. S. (2021). An in-depth analysis of machine learning approaches to predict depression. *Current Research in Behavioral Science*, *2*, 10044. https://doi.org/10.1016/j.crbeha.2021.100044

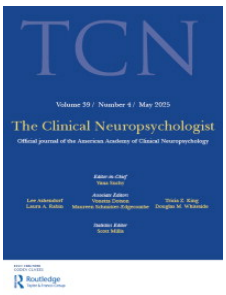
2.2 Differentiating patients with Dementia from patients with Depression and Healthy Controls using Machine Learning and the Cognitive-Functions-Dementia (CFD) testset

This study has been published as:

Dominke, C., Schenk, T., & Jahn, T. (2025). Differentiating patients with dementia from patients with depression and healthy controls using the Cognitive Functions Dementia (CFD) test set and machine learning. *The Clinical neuropsychologist*, 1–29. Advance online publication. https://doi.org/10.1080/13854046.2025.2513446

Author's contribution:

Clara Dominke: Conceptualization of the evaluation strategies, participation at data acquisition, data analysis and manuscript writing.



The Clinical Neuropsychologist

ISSN: 1385-4046 (Print) 1744-4144 (Online) Journal homepage: www.tandfonline.com/journals/ntcn20

Differentiating patients with dementia from patients with depression and healthy controls using the cognitive Functions dementia (CFD) test set and machine learning

Clara Dominke, Thomas Schenk & Thomas Jahn

To cite this article: Clara Dominke, Thomas Schenk & Thomas Jahn (03 Jun 2025): Differentiating patients with dementia from patients with depression and healthy controls using the cognitive Functions dementia (CFD) test set and machine learning, The Clinical Neuropsychologist, DOI: 10.1080/13854046.2025.2513446

To link to this article: https://doi.org/10.1080/13854046.2025.2513446

THE CLINICAL NEUROPSYCHOLOGIST
https://doi.org/10.1080/13854046.2025.2513446

Differentiating patients with dementia from patients with depression and healthy controls using the cognitive Functions dementia (CFD) test set and machine learning

Clara Dominke[a], Thomas Schenk[a] and Thomas Jahn[a,b]

[a]Chair of Clinical Neuropsychology, Department of Psychology, Ludwig-Maximilians-University, Munich, Germany; [b]School of Medicine, Department of Psychiatry and Psychotherapy, Technical University of Munich, Munich, Germany

ABSTRACT

Objective: Due to similarities in cognitive impairments shown in early states of dementia (DEM) and depression (DEP), accurate differentiation between the two in elderly remains challenging in clinical practice. Using Machine Learning (ML) algorithms in addition to the gold standard of neuropsychological assessment could help the differentiation between healthy controls (HC) and DEM, as well as between DEM and DEP by providing a diagnostic rule for clinicians. **Methods**: We used four different ML algorithms (SVM: Support Vector Machine, NB: Gaussian Naïve Bayes, RF: Random Forest, GLMnet: Lasso and Elastic-Net Regularized Generalized Linear Models) and logistic regression (LR) to differentiate between HC ($n=407$), patients with DEM ($n=131$) and patients with DEP ($n=145$) using features from the tablet-based neuropsychological test battery Cognitive Functions Dementia (CFD). We also investigated whether the type of input data (i.e. raw scores vs. sociodemographically adjusted raw scores or T-scores) influences the classification accuracy. **Results**: Using raw data from the CFD and the GLMnet algorithm, we could accurately differentiate between DEM vs. HC with accuracies ranging up to 94.0%. Similarly, we could classify DEM and DEP with accuracies up to 80.8% using the Naïve Bayes algorithm and raw scores. Measures for verbal memory, word fluency and processing speed showed the highest feature importance within these classifications, highlighting their importance for differential diagnosis. **Conclusions**: We provide preliminary evidence that ML algorithms in combination with the CFD can aid clinicians in the differential diagnosis of HC and DEM, as well as DEM and DEP by providing a decision-making aid.

ARTICLE HISTORY
Received 30 January 2025
Accepted 26 May 2025
Published online 03 June 2025

KEYWORDS
Dementia; depression; neuropsychological assessment; machine learning; differential diagnosis; CFD

CONTACT Clara Dominke Clara.Dominke@campus.lmu.de; c.dominke1@web.de Chair of Clinical Neuropsychology, Department of Psychology, Ludwig-Maximilians-University, Munich, Leopoldstr. 13, 80802 Munich, Germany.

Introduction

Due to a pronounced demographic shift, we are witnessing a substantial increase in the number of elderly people within our society (Süle et al., 2019). Numbers are predicted to further increase within the upcoming years (Nichols et al., 2019). Normal aging is already typically associated with declining processing speed, memory, and executive functions accompanied by structural and functional changes in the brain (Murman, 2015). Furthermore, two clinical syndromes appear frequently in older age: dementia (DEM) and depression (DEP). Prevalence rates for both have been found to be very high in adults aged over 65 years, ranging up to 16% (Andreas et al., 2017). Within people referred to memory clinics, the prevalence is suggested to be even higher (Kao et al., 2019; Knapskog et al., 2014).

Dementia is an umbrella term used to describe pronounced cognitive deterioration and behaviors which impair daily functioning of affected individuals. There are many different pathophysiological mechanisms that can underlie dementia (Carrarini et al., 2024). Depression is primarily characterized by depressive mood and loss of interest or anhedonia (Sözeri-Varma, 2012). Recent research suggests a relationship between DEM and DEP but evidence is far from clear cut concerning the exact type of connection shared: Some research suggests that depression in late life constitutes a risk factor for dementia (Cantón-Habas et al., 2020), others consider it being a prodromal stage of dementia and again others found no significant connection between the two (Ly et al., 2021; Wiels et al., 2020).

Due to the predominance of cognitive dysfunctions associated with DEM, neuropsychological examinations have been considered the gold standard in differentiating between these diagnostic groups (Alzola et al., 2024; Wright & Persad, 2007). While DEM can be characterized by a large range of cognitive deficits, late-life DEP has been found to be associated with severe cognitive impairments including deficits in attention, memory, and executive functioning (Barlet et al., 2023; Butters et al., 2004; Linnemann & Lang, 2020; Steffens & Potter, 2008). Depressive symptoms are furthermore often found in people suffering from dementia (Kuring et al., 2018), further complicating differentiation based solely on diagnostic instruments such as the Geriatric Depression Scale (GDS; Montorio & Izal, 1996) alone. Due to these similarities in clinical symptoms, the differentiation of these two disorders often remains challenging. Accurate differentiation is however pivotal to ensure appropriate psychological and/or pharmacological treatment. Therefore, a large amount of neuropsychological research has been conducted investigating the exact differences in cognitive impairment (Silva Dos Santos Durães et al., 2022; Wunner et al., 2022) between the two disorders. Even though differences in the degree of impairment between DEP and DEM have been found in recognition memory (Leyhe et al., 2017) and verbal fluency tasks (Barlet et al., 2023; Tetsuka, 2021) as well as visuoconstruction (Silva Dos Santos Durães et al., 2022), these differences are usually too small to allow a reliable distinction between DEM and DEP on the level of individual patients (Lanza et al., 2020). This fact emphasizes the need for additional support options in daily clinical decision-making.

Machine learning (ML) techniques are increasingly adopted to facilitate clinical practice and specifically improve diagnostic accuracy and efficiency. Applications range

from cardiovascular imaging (Sanchez-Martinez et al., 2021), the differentiation of different types of Parkinson's (Vaccaro et al., 2021), the prediction of schizophrenia and bipolar-disorder (Montazeri et al., 2023) to the differential diagnosis of dementia (Bougea et al., 2022; Castellazzi et al., 2020; Martin et al., 2023; Myszczynska et al., 2020). ML identifies patterns and relations of input and output data are then stored within a model, which can be validated and tested on new datasets (test set; Scott et al., 2019). ML is extensively used to improve both diagnosis and disease trajectory predictions (Dwyer et al., 2018; Myszczynska et al., 2020). It has the potential to recognize subtle patterns and interactions of different variables, which might be difficult for clinicians to detect. By doing so, ML can improve both the sensitivity and specificity of diagnoses, acting as a valuable decision support system, particularly in complex or ambiguous cases. It can potentially help in making more accurate and data-driven diagnostic choices by providing insights that assist clinicians in distinguishing between conditions with similar symptoms, thereby reducing the risk of misdiagnosis.

The intersection of ML and neuropsychology could thus possibly constitute a promising frontier in addressing the challenges in the differential diagnosis between DEM and DEP. Numerous studies have already shown the potential of combining ML and neuropsychological test data in differentiating between healthy controls (HC) and patients with probable Alzheimer's disease (Almubark et al., 2019; Er et al., 2017; Gupta & Kahali, 2020; Gurevich et al., 2017) or between different types of DEM (Garcia-Gutierrez et al., 2021; Carrarini et al., 2024; Maito et al., 2023; Matias-Guiu et al., 2021; Piccolino et al., 2025). Other investigators have focused on the prediction of the conversion of HC to those with mild cognitive impairment (MCI) or from MCI to DEM using ML algorithms and - among other parameters -neuropsychological test data (Battista et al., 2020; Franciotti et al., 2023; Prabhakaran et al., 2024; Tuena et al., 2024) with satisfactory results. Evidence on the applicability of ML in the differentiation between DEM and DEP is however almost missing.

A previous study of our research group (Dominke et al., 2024) suggests that ML and neuropsychological tests are well suited to distinguish between DEP and DEM using the CERAD-NAB (Fillenbaum & Mohs, 2023) or by using parameters gained with a flexible battery approach as input features. It was also shown however, that ML did not outperform traditional logistic regression. Previous research has not yet focused on using neuropsychological data from novel test batteries such as the test set Cognitive Functions Dementia (CFD; Jahn & Hessler, 2023), which might be better suited for differential diagnosis since it was specifically developed for the early detection of dementia. Furthermore, examining the impact of different types of input data for the ML diagnostic system (i.e. raw scores, adjusted raw-scores, T-scores) on the ML system's diagnostic accuracy might be useful, has not been done before, but might yield relevant insights for clinicians. Such an analysis might indicate which of the available parameters provide a better basis for their own diagnostic process. Another potential shortcoming of our last study was the fact that we used data from just one clinical site. It is therefore not clear whether and to what extent our findings generalize to other sites and other clinical samples (Dwyer et al., 2018).

In this study, we investigated the potential of different ML algorithms and data from a neuropsychological toolbox designed with respect to the DSM-5 diagnostic

criteria for DEM (CFD) to differentiate between HC and DEM, as well as between DEM and DEP. A direct comparison between HC and DEP was not performed due to the primary clinical challenge of distinguishing dementia from depression rather than depression from healthy aging. We compared the performance of four different ML algorithms (Lasso and Elastic-Net Regularized Generalized Linear Models, Support Vector Machine, Gaussian Naïve Bayes, Random Forest) to that of traditional logistic regression *via* a benchmark design and used different types of input data (i.e. raw scores, adjusted raw scores, and T-scores) from the CFD to find the best performing combination between ML algorithm and data type. To validate the performance results, we used the model—trained on a well-matched Propensity Score Matching (PSM) dataset—to predict cases from the same overall sample that were not included in the training set (i.e. not part of the PSM dataset) and therefore unseen by the algorithm during training. These specific machine learning algorithms were selected for their ability to handle non-linear relationships between features and the target variable, their capacity to manage high-dimensional data, and their minimal data preprocessing requirements. Due to the relatively small sample size in our dataset, more complex models like deep neural networks (DNN) were not used, since they require larger data sets and significant tuning to avoid overfitting (Cheng et al., 2025). Other methods such as Gradient Boosted Machines (GBM) and XGBoost were also considered but excluded because they do not only often suffer from poor interpretability, but they tend to require extensive hyperparameter tuning and can be computationally expensive, especially with our relatively small dataset (Florek & Zagdański, 2023). Furthermore, these models can sometimes overfit if not carefully regularized, which could lead to less generalizable results. Similarly, methods like k-Nearest Neighbors (k-NN) were not used due to their poor scalability in the presence of noise and sensitivity to feature correlations, which could make the results unstable. Additionally, Linear Discriminant Analysis (LDA) was excluded, as it makes more assumptions (i.e. normal distribution of each class) about the underlying data as compared to Logistic Regression (LR). The selected methods in this study offer a degree of interpretability and have been successfully applied in both medical and psychological research. While ensemble methods are often used to improve performance by combining multiple models, we chose not to incorporate them due to their added complexity and computational demands, which might not provide significant advantages in this context. Different types of input data were used to investigate whether feature engineering would improve the diagnostic accuracy of the models. The general aim was to explore to what extent machine learning can help clinical neuropsychologists with the differential diagnosis between HC, patients with depression and dementia by finding an optimal diagnostic rule and providing insight into the most important neuropsychological variables used for these classifications at hand. Given the exploratory nature of this investigation and the lack of similar studies in the past, we did not have explicit hypotheses regarding the best combination of algorithm and data type. However, based on previous research (Leyhe et al., 2017; Silva Dos Santos Durães et al., 2022; Tetsuka, 2021), we anticipated that neuropsychological variables related to recognition memory, verbal fluency, and visuoconstruction would play a significant role.

Table 1. Demographics of the sample ($N = 683$).

		DEM	DEP	HC	*p*
Number	N	131	145	407	
Sex	% female	50.4%	57.2%	60.0%	.16
Age (Years)	Mean (SD)	74.3 (8.9)[a]	67.2 (9.7)[b]	67.8 (9.9)[b]	<.001
	Range	51-93	50-85	50-94	
Education					<.001
	Low	39[a]	30[b]	67[c]	
	Middle	68[a]	62[b]	151[c]	
	High	24[a]	53[b]	189[c]	
GDS	Mean (SD)	3.3 (3.3)[a]	7.5 (3.8)[b]	1.6 (2.0)[c]	<.001
	Range	0 – 14	0 – 15	0 – 10	
MMSE	Mean (SD)	23.68 (3.71)[a]	27.96 (1.99)[b]	28.57 (1.33)[b]	<.001
	Range	12-30	17-30	22-30	
	n	127	137	379	

Note: DEM = Dementia; DEP = Depression; HC = Healthy Controls; GDS = Geriatric Depression Scale; MMSE = Mini Mental Status Examination; Education: Low refers to people with no information on any school leaving certificate or people without any educational qualifications as well as people who went to special schools or who finished secondary school; Middle refers to people who have completed technical school or vocational training; High refers to people with a high school diploma or a university degree; SD = Standard Deviation. Superscript letters indicate significant differences between groups ($p < .05$). Groups that share the same superscript letter do not differ significantly from each other.

were blinded to the results of the CFD for both conditions, ensuring the independence of the diagnoses. As can be seen in Table 1, the groups differed significantly regarding age ($p < .001$), education ($p < .001$) and (as expected) in the total score on the Geriatric Depression Scale (GDS; Montorio & Izal, 1996) as a measure of depressive symptomatology ($p < .001$).

Cognitive functions dementia (CFD)

The CFD (Jahn & Hessler, 2023) is a computerized neuropsychological test battery presented on a tablet. Even though the tests can all be presented digitally, the system is not fully automated and requires the presence of an examiner. The examiner plays a crucial role in demonstrating and explaining the individual tasks using examples. The practice phase, during which the respondent completes one or more practice items independently, begins only after the demonstration phase. It is specifically designed for the early detection and differential diagnosis of dementia according to the neurocognitive domains of dementia as described in DSM-5. The duration of the complete test set is approximately 65 min. The tests from the CFD toolbox demonstrate strong reliability across various cognitive domains, as evidenced by high Cronbach's α. For instance, verbal fluency tasks (semantic and lexical) exhibit reliability coefficients ranging from 0.71 to 0.75, while learning and recall measures (AWLT) show excellent reliability with values between 0.77 and 0.91. The reaction time measures in tasks assessing alertness and divided attention also demonstrate high reliability, with Cronbach's α ranging from 0.94 to 0.97. Similarly, the CORSI span task and the TMT-L working time parts maintain strong internal consistency, with coefficients between 0.83 and 0.92. The CFD consists of the following individual tests:

The AWLT (Auditory Word List Learning Test) assesses verbal memory and consists of a list of 12 words presented four times either by the examiner (verbally) or *via* audio recordings. The words were read by the examiner if the subjects had difficulty in understanding the audio recordings. Immediately after each trial and again after delays of 5 min (short delay) and 20 min (long delay), participants are instructed to recall freely as many words as possible. In the final test a recognition task is employed. The 12 previously presented words are combined with 12 new words (distractors). Participants are asked to indicate those words which were part of the original learning list. The main variables are the sum of correctly recalled words across all four trials (AWLTleg; learning ability), the number of correctly remembered words after the short delay (AWLTkme; short-term delayed recall) as well as after the long delay (AWLTlme; long-term delayed recall). Recognition is assessed by the logarithmic odds-ratio of true-positive to false-positive and constitutes the fourth main variable of the AWLT (AWLTwdi).

The CORSI (Corsi-Block-Tapping-Test) assesses the visual-spatial working memory. Nine different dices are presented on the screen, which are tapped in a specific order by an animated hand. The subject must tap the dices in the reverse order. The length of the tapping sequences increases with test progression. The longest sequence in which at least two of the three given tasks (all with the same length span) are correctly reproduced is called the immediate block span backwards and constitutes the main variable of the CORSI (CORSIlubs).

The TMT-L (Trail Making Test; Langensteinbach Version) consists of two different parts. Part A is used to assess processing speed. The subject must connect different numbers (25 in total) on the screen in an ascending order as fast as possible. The time required to connect all numbers is the main variable of Part A (TMTbta). Part B assesses cognitive flexibility and consists of both numbers and letters. The subject is instructed to connect both numbers (1-13) and letters (A-L) alternately in an ascending and alphabetical order. The main outcome variable is the time required to connect all letters and numbers in an ascending order (TMTbtb). Errors on the TMT are handled automatically by the system: If a participant attempts to connect an incorrect number (in Part A) or an incorrect number-letter combination (in Part B), the system prevents the connection from being made. The task does not progress until the correct sequence is selected, ensuring that participants must follow the correct order to continue. Within the Langensteinbach- Version, both test parts are the same length. The numbers furthermore lie within a visual angle that allows foveal recognition: Performance in TMT-B is thereby not influenced by increased visual search efforts and the processing times of A and B are more similar to each other.

The VISCO (Visual Construction Test) assesses visuoconstruction and consists of different target shapes. The subject is given a set of equilateral triangles and must move and rotate those triangles to produce the indicated target shape. Subjects must complete each shape within 60 s. As the test progresses, the number of triangles required to recreate the shape increases and thus item difficulty increases. The task stops whenever the subject does not manage to assemble the correct shape for three trials in a row. The main variable is the number of correctly reconstructed shapes (Viscovisco). Extra points are awarded for fast solutions (shape completed in 30 s or less).

The WAF (Perception and Attention Functions): Alertness: Its aim is to assess the basic attentional function to respond quickly to new events. Black circles are presented on white background. In one condition, the circles are preceded by an auditory warning tone. Subjects are asked to press a button as quickly as possible as soon as the target stimulus (black circle) appears on the screen. Reaction time of the button-press response is measured and separate average reaction times (arithmetic mean of logarithmized reaction times) are calculated for the condition without a warning tone (WAFA1lmrtr1) and the condition with a warning tone (WAFA2lmrtzr2).

The WAF: Divided Attention: Its aim is to assess the ability of processing more than one information at a time. The subject is presented with dark circles, squares and either high or low tones and must press the response button as quickly as possible whenever one of the target stimulus types (square or a high tone) appears twice in a row. The mean of all reaction times across both stimuli channels is used as the main variable (WAFGlmrtc2).

The WIWO (Vienna Verbal Fluency Test) assesses verbal fluency and consists of a semantic and a phonemic task. In the semantic version, subjects are asked to name for a specific category (e.g. animals) as many items as possible within a time interval of two minutes. The spoken responses are recorded by the system and can later be reviewed by the examiner to verify whether the correct number of words has been scored by him/her. In the lexical task, they are given a letter and asked to list as many words as possible that start with that letter (again an interval of two minutes

is provided). The number of words provided by the subject for the lexical fluency task (WIWOlexwof) and the semantic fluency task (WIWOsemwof) are used as the main dependent measures for the WIWO.

The WOBT (Vienna Object Naming Test) assesses the naming of visually presented objects. Images of different objects are shown, and the subject must name these objects correctly. Whenever subjects cannot name the object or provide the wrong name, examiners provide first a lexical and if necessary, a semantic clue. The main variable is the number of objects that were named without assistance (WOBTrbu).

In total, the CFD provides 132 variables for a fine-grained description and a complete profile of an individual's cognitive abilities. For the current study, we selected a subset of 14 variables from the broader test battery. This decision was made to reduce redundancy and multicollinearity among variables, enhance interpretability of the results, and avoid unnecessary complexity in the analyses. Our aim was to focus on the most informative and non-overlapping measures (i.e. the main variables) to facilitate clearer insights into group differences. More detailed information on the CFD can be found in Jahn and Hessler (2023).

Machine learning algorithms

The open-source software R 4.0.2 (R Core Team, 2019) was utilized for all the statistical analyses. In particular, the "mlr3" package and ecosystem (Lang et al., 2019) was used for the implementation of different ML algorithms. Since we already knew the final clinical diagnoses within our study sample, we focused on supervised ML algorithms. They explore the relationship between the features and a given output label (diagnostic group). Specifically, we compared the performance of Support Vector Machine (SVM), Gaussian Naïve Bayes (NB), Random Forest (RF), Lasso and Elastic-Net Regularized Generalized Linear Models (GLMnet) as well as traditional logistic regression (LR).

SVM maps the features (the neuropsychological variables in our case) into a higher dimension and creates a hyperplane to separate the outcome groups. It does so by finding the maximal margin hyperplane on the training data, correctly dividing the groups (Cortes & Vapnik, 1995). Random Forest (RF) is an ensemble learning method that combines multiple decision trees to improve classification accuracy. Each tree is built by using a random subset of both data and features at each decision point. Thereby diversity among the decision trees is ensured. For classification tasks, the model aggregates all the predictions from the individual trees, using majority voting (Biau & Scornet, 2016). This helps in reducing overfitting. Gaussian Naive Bayes (NB) is a probabilistic machine learning method based on Bayes' theorem and the law of total probability. It calculates the likelihood of specific features given a particular class outcome, if each feature follows a Gaussian (normal) distribution within each class. A key assumption of Gaussian Naive Bayes, like standard Naive Bayes, is the independence of predictors. However, studies have shown that it can still perform effectively even if the independence assumption is violated (Bhagya Shree & Sheshadri, 2018), making it robust for many practical applications, including those with continuous data. GLMnet is an advanced statistical modeling approach, extending traditional logistic regression by using regularization techniques to improve classification accuracy. Lasso (L1 regularization) and Elastic-Net regularization (combination of L1 and L2

regularization) procedures are used, which prevent overfitting by adding a penalty term to the model's cost function. Thus, some coefficients are shrunken down to zero, reducing the complexity of the model by excluding less important features. These models can therefore deal with many features (Friedman et al., 2010). Logistic regression models the relationship between predictor variables and a categorical output variable by selecting parameters that maximize the likelihood of observing the true class labels in the training data. It can be considered the gold standard in clinical research in case of any classification problem. Performance of the methods was compared to that of a featureless learner (FL), which can be considered the simplest prediction method and should always be outperformed by any meaningful model. It always predicts the most frequent diagnostic group without using the neuropsychological features. The classification will be explained in more detail now.

Data processing and classification procedure

Since our sample sizes were unequal (i.e. much more HC as compared to DEM and DEP) and since we wanted to investigate whether the performance of the ML algorithms would differ depending on the specific data type used, four different data sets were analyzed: Raw scores, raw scores after parallelizing the groups following Prospensity Score Matching (PSM), adjusted raw scores and T-scores. Adjusted raw scores were calculated using regression-based analysis: Test performance was adjusted for age, gender, and education as well as to account of bivariate interactions between those variables. More information on regression-based raw score adjustment can be found in Berres et al. (2008). PSM is a statistical matching technique used to balance the covariates between two groups to reduce bias and therefore improve causal interference (Austin, 2011). Prospensity scores were estimated using logistic regressions for group membership using the covariates age, gender, and education as well as their interactions. Following this procedure, 88 DEM, 88 DEP and 88 HC were selected based on a matched PSM score. The remaining sample ($N = 419$) was used as a test set for the model trained on the PSM data set to get a better performance estimate on this previously unseen sample. For the other data types, the previously described numbers of subjects per group were analyzed (DEM = 131; DEP = 145; HC = 407), resulting in imbalanced sample sizes. T-scores are standardized scores with a mean of 50 and a standard deviation of 10.

Due to the imbalance of the sample sizes, the SMOTE algorithm (Chawla et al., 2002) was used to deal with this problem. It firstly identifies the minority class (i.e. the class that has significantly less subjects as compared to the other one). For each data point in the minority class, the k nearest neighbors of the same class is being calculated based on the Euclidean distance. A random neighbor is then selected from the k nearest neighbors, followed by the creation of a synthetic data point along the line between the original point and the selected neighbor. As a result, it leads to more reliable and accurate predictions for all diagnosis groups. This approach allows for a fairer comparison and analysis of the different categories, even though it does not guarantee complete fairness, and some residual bias may remain. As shown in Appendix Table A1, the DEM vs. HC classification was also calculated without SMOTE.

Using the whole dataset, 1.47% of values were missing across all the 14 main variables of the CFD. Within the DEM group 5.70% of values were missing, the DEP group had 1.79% missing values and the HC had 0% missing values. Depending on the diagnostic group, missing values for individual variables of the CFD ranged between 0% and 25.95% (WAFA2lmrtr2 in den DEM group). A detailed overview of the percentage of missing values for each variable based on the diagnostic group can be found in Table 2.

To impute missing values, we used the missForest algorithm (Stekhoven & Bühlmann, 2012), which is a machine-learning based imputation algorithm, predicting missing values based on the observed variables within the dataframe using a random forest. This process was done for all data types. missForest can deal with non-linear interactions between variables and impute categorical as well as numerical variables. It finishes the predictions if the stopping criterion is met: Every time the difference between the newly imputed data matrix and the previously imputed data matrix increased, the iterations are stopped. It is assumed that as changes between iterations become small, the algorithm approaches a stable solution. Using observed values, which were held out of training, an out-of-bag error estimate of the model can be calculated using the Normalized Mean Squared Error. We used $N = 500$ trees, 20 as the maximum number of iterations before stopping (if the stopping criterion is not met before) and used variable wise imputation.

Following this preprocessing procedure, the actual classification was performed with a benchmark design: All ML algorithms were used with their default parameters as predefined in mlr3 and trained on the same data.

We used a 5-fold Cross-validation (CV) -pipeline: Data was randomly divided into 5 different folds. (i.e. partitions of the whole dataset). Each fold was iteratively once held back as a validation sample, while the other 4 folds served as the training set. This process was repeated 10 different times. The validation strategy was used to obtain an unbiased performance estimate. This strategy was used to calculate balanced classification accuracies ((sensitivity + specificity)/2), sensitivity, specificity, and the respective confidence intervals for each of the three parameters. We additionally calculated F1 scores and their respective confidence intervals to obtain a harmonic mean of precision and recall. Precision is the proportion of true positive predictions among all cases predicted as positive, while recall refers to the proportion of all true positive cases the model correctly identified as positive. F1 is defined as $2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$. It ranges from 0—1, with higher values indicating better balance between false positives and false negatives. For each classifier, the mean performance over all folds was considered as the main indicator of classification performance. Additionally, the performance results of the trained ML model in predicting the unseen cases ($N = 419$) were calculated for the PSM dataset.

We then also calculated the feature importance for every model using the permutation feature importance of "iml" package (Molnar et al., 2018). This method helps to understand which features are most important for a given model to perform well: A low importance implies that the feature is not necessary for a given model to perform well, while a high importance implies that the feature is necessary. Feature importance not only enhances the interpretability of ML predictions but also provides valuable insights into which neuropsychological variables hold the greatest relevance for the specific differential diagnostic question. Clinicians are thereby enabled to prioritize the

Table 2. Percentage of missing values by CFD variable and diagnostic group ($N = 407$ for the Healthy Controls; $N = 145$ for patients with Depression, $N = 131$ for patients with Dementia).

Group	AWLTleg	AWLTkme	AWLTlme	AWLTwdi	CORSIubs	TMTbta	TMTbtb	VISCOvisco	WAFA1lmrtr1	WAFA2lmrtr2	WAFGlmrtc2	WIWOsemwof	WIWOlexwof	WOBTrbu
DEM	0.76	0.76	2.29	3.05	3.82	1.53	8.40	20.61	1.53	25.95	13.74	0	0.76	2.29
DEP	0.00	0.00	0.00	2.07	1.38	0.00	0.69	6.21	0.69	7.59	5.52	0	0.69	2.07
HC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0.00	0.00

Note: DEM = Dementia; DEP = Depression; HC = Healthy Controls; AWLT = Auditory Word List Learning Test; AWLTleg = sum of correctly recalled words across all four trials (learning ability); AWLTkme = the number of correctly remembered words after the short delay (short-term delayed recall); AWLTlme = long-term delayed recall; AWLTwdi = logarithmic odds-ratio of true-positive to false-positive in the recognition part; CORSI = Corsi-Block-Tapping-Test, CORSIlubs immediate block span backwards; TMT-L (Trail Making Test; Langensteinbach Version; TMTbta = time required to connect all numbers; TMTbtb = time required to connect all letters and numbers in an ascending order; VISCO = Visual Construction Test; Viscovisco = number of correctly reconstructed shapes; WAF = Perception and Attention Functions); WAFA1lmrtr1 = average reaction times (arithmetic mean of logarithmized reaction times) for the condition without a warning tone; WAFA2lmrtzr2 = average reaction times (arithmetic mean of logarithmized reaction times) for the condition with a warning; WAFGlmrtc2 = mean of all reaction times across both stimuli channels; WIWO = Vienna Verbal Fluency Test; WIWOlexwof = number of words provided by the subject for the lexical fluency task; WIWOsemwof = number of words in the semantic fluency task; WOBT = Vienna Object Naming Test; WOBTrbu = number of correctly named objects without assistance.

most important variables when the full neuropsychological assessment may not be feasible due to patient motivation or endurance, which supports its integration into clinical decision-making. Firstly, the performance of the model is assessed without any changes. Then, for each feature separately, the values are randomly permuted, which destroys the relationship between the feature and the outcome (the diagnostic group). The model performance is then again measured with the given feature being permuted and compared to the unmodified one. The difference between the two performance measures (i.e. an increase in classification error) is then considered as the importance of that given feature. We report the three most important features for each best performing combination of ML algorithm and different data type.

Results

Results of the CFD across diagnostic groups

Table 3 depicts the main effects of an ANOVA for the different variables of the CFD. Post-Hoc-analyses using the Tukey method revealed that there were significant differences (p <.05) across all groups in every of the 14 main variables of the CFD, except for the difference between HC and DEP on the WAFA2lmrtr2 (p = .08), which was not significant. The DEM group always performed significantly worse than both the HC group and the DEP group. The DEP group performed significantly worse than the HC group on every subtest of the CFD (p <.05), except for the WAFA2lmrtr2.

Table 3. Results of the ANOVA for the 14 main variables of the CFD.

Variable	Dementia n	Dementia mean ± SD	Depression n	Depression mean ±SD	Controls n	Controls mean ± SD	F	p
AWLTleg.T	130	32.03 ± 8.57[a]	145	43.46 ± 10.88[b]	407	50.31 ± 9.83[c]	173.58	<.001
AWLTkme.T	130	31.69 ± 7.54[a]	145	43.09 ± 12.36[b]	407	49.90 ± 9.78[c]	165.98	<.001
AWLTlme.T	128	32.59 ± 8.20[a]	145	42.83 ± 12.02[b]	407	49.79 ± 9.65[c]	149.75	<.001
AWLTwdi.T	127	31.64 ± 10.95[a]	142	43.88 ± 12.63[b]	407	50.87 ± 10.30[c]	152.19	<.001
CORSIubs.T	126	38.67 ± 10.69[a]	143	43.40 ± 11.38[b]	407	50.16 ± 9.98[c]	66.96	<.001
TMTbta.T	129	38.84 ± 10.30[a]	145	45.03 ± 9.97[b]	407	49.36 ± 10.34[c]	53.38	<.001
TMTbtb.T	120	36.03 ± 10.07[a]	144	42.66 ± 10.01[b]	407	50.64 ± 10.45[c]	104.92	<.001
VISCOvisco.T	104	39.10 ± 8.04[a]	136	45.26 ± 9.84[b]	407	49.86 ± 10.51[c]	50.98	<.001
WAFA1lmrtr1.T	129	42.19 ± 10.53[a]	144	45.12 ± 10.57[b]	407	50.13 ± 10.10[c]	34.30	<.001
WAFA2lmrtr2.T	97	43.22 ± 10.20[a]	134	47.56 ± 12.56[b]	407	49.89 ± 10.52[b]	15.06	<.001
WAFAGlmrtc2.T	113	39.39 ± 9.86[a]	137	43.59 ± 11.61[b]	407	51.49 ± 10.36[c]	71.14	<.001
WIWOsemwof.T	131	32.30 ± 9.14[a]	145	41.09 ± 10.60[b]	407	50.63 ± 10.24[c]	176.85	<.001
WIWOlexwof.T	130	37.45 ± 10.94[a]	144	44.67 ± 10.99[b]	407	49.98 ± 9.85[c]	75.37	<.001
WOBTrbu.T	128	35.80 ± 10.71[a]	142	44.08 ± 10.92[b]	407	49.25 ± 10.37[c]	81.01	<.001

Note: All degrees of freedom (df) in the analyses are 2; AWLT = Auditory Word List Learning Test; AWLTleg = sum of correctly recalled words across all four trials (learning ability); AWLTkme = the number of correctly remembered words after the short delay (short-term delayed recall); AWLTlme = long-term delayed recall; AWLTwdi = logarithmic odds-ratio of true-positive to false-positive in the recognition part; CORSI = Corsi-Block-Tapping-Test, CORSIlubs immediate block span backwards; TMT-L (Trail Making Test; Langensteinbach Version, TMTbta = time required to connect all numbers; TMTbtb = time required to connect all letters and numbers in an ascending order; VISCO = Visual Construction Test, Viscovisco = number of correctly reconstructed shapes; WAF = Perception and Attention Functions): WAFA1lmrtr1 = average reaction times (arithmetic mean of logarithmized reaction times) for the condition without a warning tone; WAFA2lmrtzr2 = average reaction times (arithmetic mean of logarithmized reaction times) for the condition with a warning; WAFGlmrtc2 = mean of all reaction times across both stimuli channels; WIWO = Vienna Verbal Fluency Test; WIWOlexwof = number of words provided by the subject for the lexical fluency task; WIWOsemwof = number of words in the semantic fluency task; WOBT = Vienna Object Naming Test; WOBTrbu = number of correctly named objects without assistance. Superscript letters indicate significant differences between groups (p < .05). Groups that share the same superscript letter do not differ significantly from each other.

Table 4. Classification results for patients with dementia ($N = 131$) vs. healthy controls ($N = 407$) across the different Machine Learning algorithms and data types used.

	FL	GLM	SVM	Ranger	NB	LR
Raw scores	BAC = 50% [50.0 – 50.0] SEN = 0% [0 – 0] SPE = 100% [100 – 100]	BAC = 94.0% [88.6 – 98.0] SEN = 91.2% [81.6 – 100] SPE = 96.7% [91.6 – 100] F1 = .90 [.84 – .97]	BAC = 93.5% [88.7 – 97.2] SEN =90.6% [81.5 – 99.4] SPE = 96.4% [91.4 – 100] F1 = .90 [.85 – .96]	BAC = 93.5% [86.7 – 97.6] SEN = 90.4% [77.9 – 100] SPE = 96.7% [92.6 – 100] F1 = .90 [.83 – .96]	BAC = 92.5% [86.5 – 97.2] SEN = 91.2% [80.9 - 100] SPE = 93.9% [86.1 – 98.4] F1 = .87 [.77 – .94]	BAC = 93.5% [88.6 – 98.0] SEN = 91.2% [81.5 – 100] SPE = 95.8% [90.3 – 99.7] F1 = .89 [.83 – .95]
PSM & raw scores	BAC = 50.0% [50.0 – 50.0] SEN = 53.0% [0 – 100] SPE = 47.0% [0 – 100]	BAC = 93.5% [85.3 – 100] SEN = 93.3% [78.9 – 100] SPE = 93.7% [86.9 – 100] F1 = .93 [.85 – .99]	BAC = 92.8% [85.8 – 97.9] SEN = 93.1% [80.2 – 100] SPE = 92.5% [83.3 – 100] F1 = .92 [.85 – .97]	BAC = 93.4% [84.7 – 100] SEN = 93.6% [81.3 – 100] SPE = 93.2% [82.2 – 100] F1 = .93 [.85 – 1.00]	BAC = 90.5% [80.4 – 97.2] SEN = 88.9% [75.8 – 100] SPE = 92.0% [78.0 – 100] F1 = .90 [.79 – .97]	BAC = 90.6% [81.2 – 97.2] SEN = 90.9% [72.2 – 100] SPE = 90.3% [74.6 – 100] F1 = .90 [.78 – .97]
T-scores	BAC = 50 % [50.0 – 50.0] SEN = 0% [0 – 0] SPE = 100 % [100 – 100]	BAC = 92.8% [86.5 – 98.2] SEN = 90.7% [81.1 – 100] SPE = 94.8% [90.0 – 98.8] F1 = .88 [.80 – .94]	BAC = 90.4% [81.2 – 97.4] SEN = 85.1% [70.8 – 100] SPE = 95.7% [90.5 – 100] F1 = .85 [.78 – .92]	BAC = 91.6% [86.4 – 96.8] SEN = 87.0% [76.8 – 99.2] SPE = 96.2% [91.0 – 100] F1 = .87 [.78 – .94]	BAC = 90.5% [84.3 – 96.1] SEN = 88.8% [78.9 – 100] SPE = 92.2% [86.9 – 98.5] F1 = .83 [.76 – .90]	BAC = 91.6% [85.7 – 97.9] SEN = 89.2 % [80.9 – 100] SPE = 94.0% [88.1 – 98.8] F1 = .86 [.78 – .93]
Adjusted raw scores	BAC = 50% [50.0 – 50.0] SEN = 0% [0 – 0] SPE = 100% [100 – 100]	BAC = 93.0% [87.1 – 98.3] SEN = 90.9% [80.2 - 100] SPE = 95.2% [91.5 – 98.4] F1 = .88 [.81 – .94]	BAC = 90.7% [83.4 – 96.3] SEN = 85.9% [72.1 – 99.3] SPE = 95.5% [91.1 – 98.8] F1 = .85 [.77 – .93]	BAC = 91.6% [85.5 – 97.4] SEN = 86.7% [74.6 – 100] SPE = 96.5% [92.8 – 98.9] F1 = .87 [.79 – .94]	BAC = 91.2% [84.3 – 95.9] SEN = 90.3% [76.9 – 100] SPE = 92.1% [87.4 – 96.3] F1 = .84 [.75 – .92]	BAC = 91.9% [85.3 – 98.0] SEN = 89.5% [79.2 – 100] SPE = 94.3% [90.3 – 97.5] F1 = .87 [.81 – .92]

FL = Featureless Learner; GLM = GLMnet = Lasso and Elastic-Net Regularized Generalized Linear Models; SVM = Support Vector Machine; Ranger = Random Forest; NB = Naïve Bayes; LR = Logistic Regression; BAC = Balanced Accuracy; SEN = sensitivity; SPE = specificity; PSM = Prospensity Score Matching. Note that for the PSM & raw scores $N = 88$ per group; F1 scores represent the harmonic mean of precision and recall.

Table 5. Classification results for patients with dementia ($N = 131$) vs. depression ($N = 145$) across the different machine learning algorithms and data types used.

	FL	GLM	SVM	Ranger	NB	LR
Raw scores	BAC = 50% [50.0 – 50.0] SEN = 94% [0 – 100] SPE = 6% [0 – 100]	BAC = 80.7% [70.1 – 89.6] SEN = 89.8% [72.2 – 100] SPE = 71.7% [58.7 – 87.5] F1 = .81 [.75 – .90]	BAC = 79.3% [69.7 – 87.2] SEN = 88.2% [72.2 – 96.7] SPE = 70.4% [55.0 – 83.9] F1 = .81 [.71 – .88]	BAC = 80.8% [72.0 – 87.7] SEN = 88.7% [72.4 – 100] SPE = 72.9% [60.8 – 83.9] F1 = .81 [.73 – .88]	BAC = 80.8% [71.6 – 88.4] SEN = 86.3% [74.3 – 96.2] SPE = 75.3% [63.3 – 83.8] F1 = .81 [.73 – .90]	BAC = 80.3% [72.1 – 87.4] SEN = 87.7% [67.4 – 100] SPE = 72.9% [57.6 – 89.7] F1 = .81 [.73 – .89]
PSM & raw scores	BAC = 50.0% [50.0 – 50.0] SEN = 44.0% [0 – 100] SPE = 56.0% [0 –100]	BAC = 79.2% [69.1 – 88.6] SEN = 80.0% [63.5 – 94.9] SPE = 78.4% [53.3 – 93.8] F1 = .79 [.67 – .91]	BAC = 79.1% [68.7 – 88.2] SEN = 81.7% [62.4 – 94.6] SPE = 76.5% [54.7 – 92.2] F1 = .78 [.69 – .91]	BAC = 80.5% [68.6 – 89.4] SEN = 84.4% [66.7 – 100] SPE = 76.6% [54.9 – 93.5] F1 = .80 [.72 – .92]	BAC = 79.9% [67.4 – 90.6] SEN = 81.7% [67.1 – 98.9] SPE = 78.2% [55.1 – 93.5] F1 = .79 [.67 – .92]	BAC = 78.2% [63.9 – 89.1] SEN = 78.1% [55.9 – 93.8] SPE = 78.2% [54.9 – 93.7] F1 = .78 [.63 – .89]
T-scores	BAC = 50% [50.0 – 50.0] SEN = 92% [0 – 100] SPE = 8% [0 – 100]	BAC = 75.6% [62.7 – 84.7] SEN = 87.1% [49.4 – 99.2] SPE = 64.1% [40.3 – 80.3] F1 = .76 [.63 – .85]	BAC = 75.7% [65.4 – 85.4] SEN = 83.8% [54.5 – 96.1] SPE = 67.5% [50.4 – 84.1] F1 = .77 [.65 – .85]	BAC = 76.1% [66.8 – 85.7] SEN = 83.9% [66.8 – 95.4] SPE = 68.3% [53.2 – 85.1] F1 = .77 [.65 – .84]	BAC = 76.8% [65.9 – 87.1] SEN = 82.9% [71.6 – 95.7] SPE = 70.8% [51.2 – 85.4] F1 = .76 [.65 – .85]	BAC = 76.8% [64.7 – 88.0] SEN = 86.6% [49.4 – 99.2] SPE = 67.0% [47.6 – 87.7] F1 = .77 [.65 – .87]
Adjusted raw scores	BAC = 50% [50.0 – 50.0] SEN = 100% [100 – 100] SPE = 0% [0 – 0]	BAC = 75.4% [61.5 – 85.2] SEN = 89.1% [76.6 – 100] SPE = 61.6% [41.7 – 80.0] F1 = .77 [.65 – .87]	BAC = 77.5% [63.8 – 85.4] SEN = 88.1% [75.7 – 100] SPE = 66.8% [48.4 – 81.4] F1 = .77 [.64 – .85]	BAC =76.9% [65.1 – 85.6] SEN = 85.9% [70.9 – 99.0] SPE = 68.0% [52.1 – 80.9] F1 = .77 [.65 – .86]	BAC = 76.7% [64.2 – 86.0] SEN = 82.1% [64.0 – 95.8] SPE = 71.3% [57.1 – 84.0] F1 = .77 [.66 – .86]	BAC = 76.1% [62.9 – 85.0] SEN = 88.2% [72.2 – 100] SPE = 64.1% [45.3 – 80.9] F1 = .76 [.68 – .87]

FL = Featureless Learner; GLM = GLMnet = Lasso and Elastic-Net Regularized Generalized Linear Models; SVM = Support Vector Machine; Ranger = Random Forest; NB = Naïve Bayes; LR = Logistic Regression; BAC = Balanced Accuracy; SEN = sensitivity; SPE = specificity; PSM = Prospensity Score Matching. Note that for the PSM & raw scores $N = 88$ per group; F1 scores represent the harmonic mean of precision and recall.

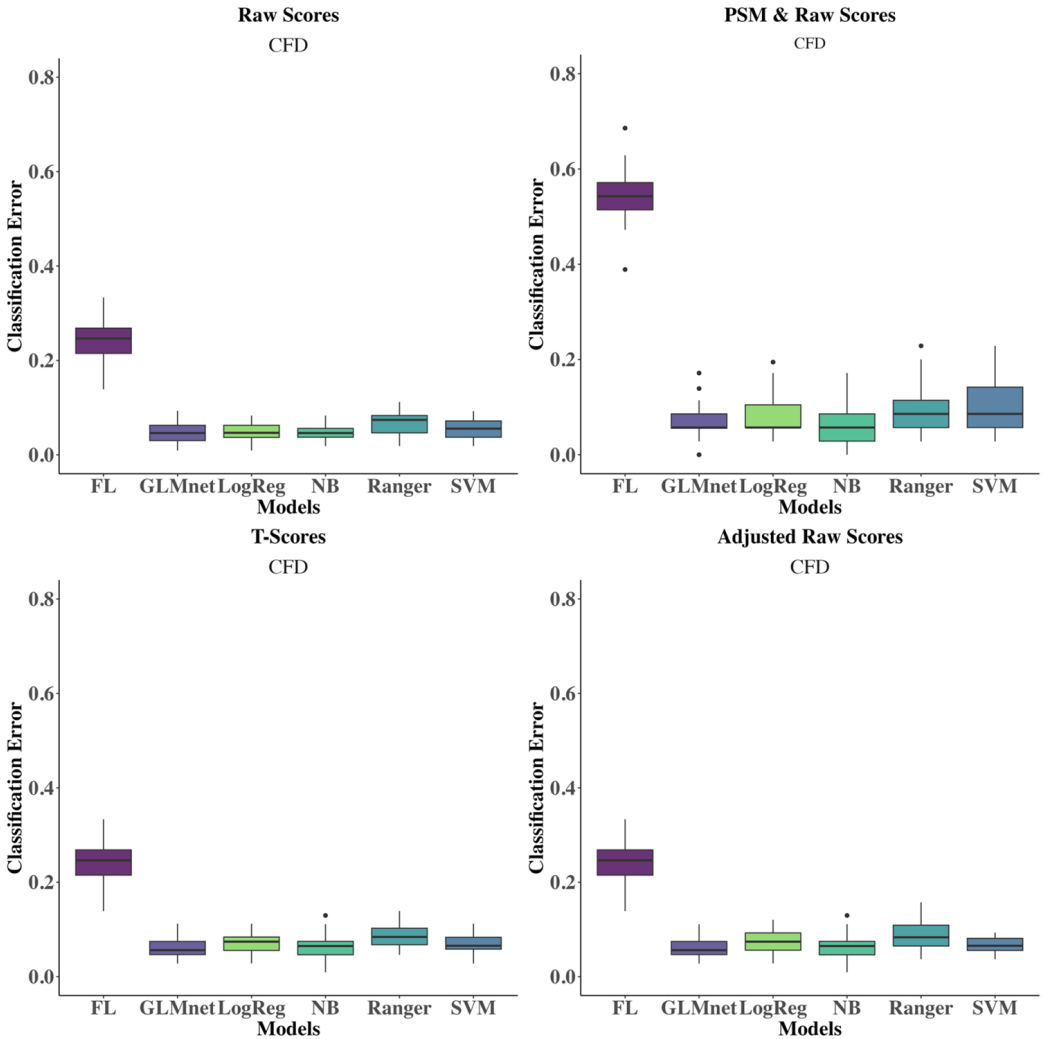


Figure 1. Dementia (DEM; $N = 131$) vs. Healthy Controls (HC; $N = 407$): Classification errors with their confidence intervals for all Machine Learning algorithms and different data types used; FL = Featureless Learner; GLM = GLMnet = Lasso and Elastic-Net Regularized Generalized Linear Models; SVM = Support Vector Machine; Ranger = Random Forest; NB = Naïve Bayes; LR = Logistic Regression; PSM = Prospensity Score Matching.

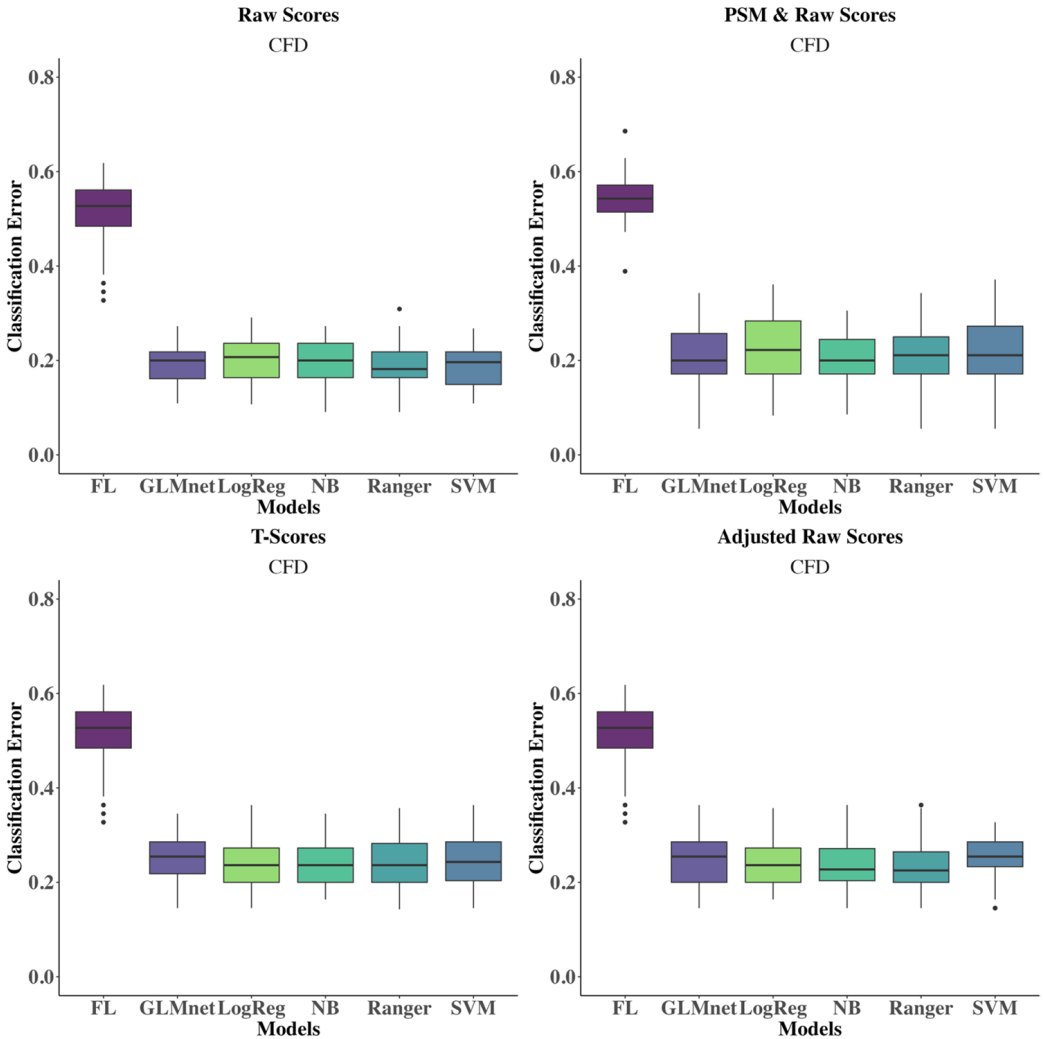


Figure 2. Dementia (DEM; $N = 131$) vs. Depression (DEP; $N = 145$); CFD = Cognitive Functions Dementia: Classification errors with their confidence intervals for all Machine Learning algorithms and different data types used; FL = Featureless Learner; GLM = GLMnet = Lasso and Elastic-Net Regularized Generalized Linear Models; SVM = Support Vector Machine; Ranger = Random Forest; NB = Naïve Bayes; LR = Logistic Regression; PSM = Prospensity Score Matching.

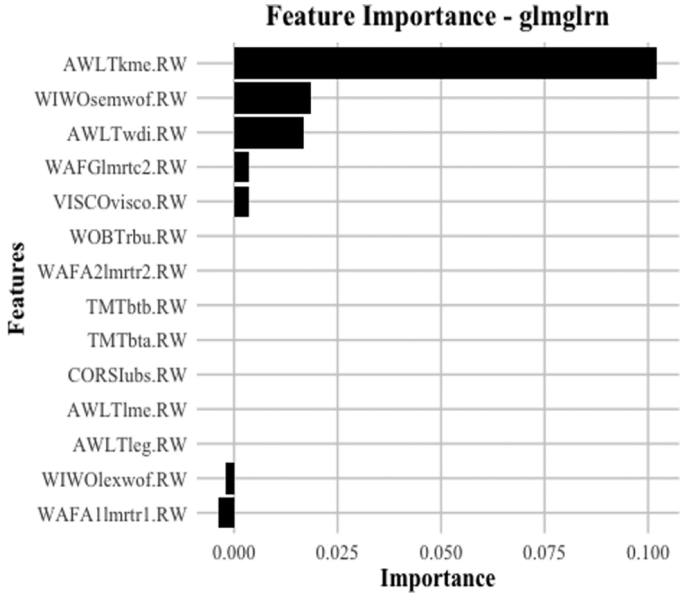


Figure 3. Dementia (DEM; $N = 145$) vs. Healthy Controls (HC; $N = 407$): Feature importance of the individual neuropsychological main variables for the given classification with Lasso and Elastic-Net Regularized Generalized Linear Models (GLMnet). The model performance with the given feature being permuted and is compared to the unmodified one. The difference between the two performance measures (i.e. an increase in classification error) is then considered as the importance of that given feature; AWLT = Auditory Word List Learning Test; AWLTleg = sum of correctly recalled words across all four trials (learning ability); AWLTkme = the number of correctly remembered words after the short delay (short-term delayed recall); AWLTlme = long-term delayed recall; AWLTwdi = logarithmic odds-ratio of true-positive to false-positive in the recognition part; CORSI = Corsi-Bloc k-Tapping-Test, CORSIlubs immediate block span backwards; TMT-L (Trail Making Test; Langensteinbach Version; TMTbta = time required to connect all numbers; TMTbtb = time required to connect all letters and numbers in an ascending order; VISCO = Visual Construction Test, Viscovisco = number of correctly reconstructed shapes; WAF = Perception and Attention Functions): WAFA1lmrtr1 = average reaction times (arithmetic mean of logarithmized reaction times) for the condition without a warning tone; WAFA2lmrtzr2 = average reaction times (arithmetic mean of logarithmized reaction times) for the condition with a warning; WAFGlmrtc2 = mean of all reaction times across both stimuli channels; WIWO = Vienna Verbal Fluency Test; WIWOlexwof = number of words provided by the subject for the lexical fluency task; WIWOsemwof = number of words in the semantic fluency task; WOBT = Vienna Object Naming Test; WOBTrbu = number of correctly named objects without assistance; RW = Raw Scores.

of GLMnet and raw data. The most important features were short-delayed verbal memory (AWLTkme.RW), semantic word fluency (WIWOsemwof.RW) and recognition memory (AWLTwdi.RW). In Figure 4, the feature importance for the comparison of DEM and DEP using the best performing combination of Naïve Bayes and raw data is shown. The most important features were short-delayed verbal memory (AWLTkme. RW), processing speed of the TMT-A (TMTbta.RW) and long-delayed verbal memory (AWLTlme.RW). In the following paragraph, the feature importance (the three most important features) for the best performing ML algorithm for every data type individually will be reported separately for each classification:

For HC vs. DEM: Using PSM and raw scores, the GLMnet algorithm performed best. The most important features were short-delayed verbal memory (AWLTkme.RW), semantic

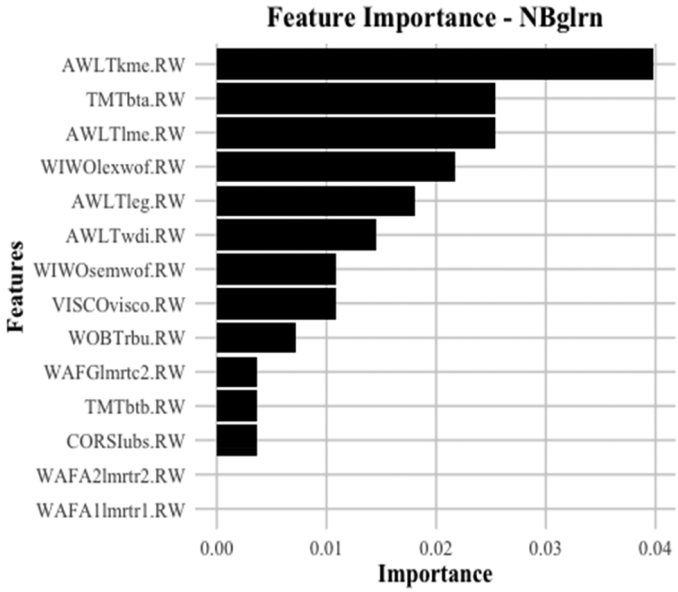


Figure 4. Dementia (DEM) vs. Depression (DEP): Feature importance of the individual neuropsychological main variables for the given classification with Naïve Bayes. The model performance with the given feature being permuted and is compared to the unmodified one. The difference between the two performance measures (i.e. an increase in classification error) is then considered as the importance of that given feature; AWLT = Auditory Word List Learning Test; AWLTleg = sum of correctly recalled words across all four trials (learning ability); AWLTkme = the number of correctly remembered words after the short delay (short-term delayed recall); AWLTlme = long-term delayed recall; AWLTwdi = logarithmic odds-ratio of true-positive to false-positive in the recognition part; CORSI = Corsi-Block-Tapping-Test, CORSIlubs immediate block span backwards; TMT-L (Trail Making Test; Langensteinbach Version, TMTbta = time required to connect all numbers; TMTbtb = time required to connect all letters and numbers in an ascending order; VISCO = Visual Construction Test; Viscovisco = number of correctly reconstructed shapes; WAF = Perception and Attention Functions): WAFA1lmrtr1 = average reaction times (arithmetic mean of logarithmized reaction times) for the condition without a warning tone, WAFA2lmrtzr2 = average reaction times (arithmetic mean of logarithmized reaction times) for the condition with a warning; WAFGlmrtc2 = mean of all reaction times across both stimuli channels; WIWO = Vienna Verbal Fluency Test; WIWOlexwof = number of words provided by the subject for the lexical fluency task; WIWOsemwof = number of words in the semantic fluency task; WOBT = Vienna Object Naming Test; WOBTrbu = number of correctly named objects without assistance; RW = Raw Scores.

word fluency (WIWOsemwof.RW) and divided attention in the WAFG task (WAFGlmrtc2.RW). Based on T-Scores, the best performing algorithm was again the GLMnet with short-delayed verbal memory performance (AWLTkme.T), recognition memory (AWLTwdi.T) and semantic word fluency (WIWOsemwof.T) as the most important features. Using adjusted raw scores, again the GLMnet performed best with semantic word fluency (WIWOsemwof.AdRoh), short-delayed verbal memory performance (AWLTkme.AdRoh) and recognition memory (AWLTwdi.AdRoh) being of greatest importance.

For the DEM vs. DEP comparison, using PSM and raw scores, the Random Forest performed best. The most important features were processing speed of the TMT-A (TMTbta.RW), recognition memory (AWLTwdi.RW) and divided attention in the WAFG task (WAFGlmrtc2.RW). Based on T-scores, the Naïve Bayes perfomed best with short-delayed verbal memory (AWLTkme.T), long-delayed verbal memory (AWLTleg.T)

and recognition memory (AWLTwdi.T) as the most important features. Using adjusted raw scores, the support vector machine ranked first. The most important features were recognition memory (AWLTwdi.AdRoh), short-delayed verbal memory (AWLTkme.AdRoh) and alertness (WAFA2lmrtr2.AdRoh).

Discussion

The aim of our study was to explore to what extent machine learning can help clinical neuropsychologists with the differential diagnosis between healthy controls (HC) and patients with dementia (DEM) as well as between DEM and patients with depression (DEP) by finding an optimal diagnostic rule and providing insight into the most important neuropsychological variables used for these classifications at hand.

We could show that the CFD in combination with ML could reliably differentiate DEM both from HC and from DEP. As expected, discriminative accuracy was greater for the HC vs. DEM comparison than for the DEM vs. DEP classification: The highest performance in distinguishing HC from DEM was achieved using the GLMnet algorithm and raw scores (BAC = 94.0%) as input data. This was also reflected by a high F1 Score (.90), as a measure of the model's ability to balance both precision and recall. In contrast, the lowest accuracy for this comparison was observed with the Gaussian Naïve Bayes classifier when using either raw scores and PSM or T-scores as input (BAC = 90.5%). Accuracies were yet not significantly different across algorithms. This finding is different to a previous study comparing RF and SVM performance on neuropsychological test data to classify the cognitive status of patients (Gupta & Kahali, 2020), which found a superiority of RF as compared to SVM. It is, however, like our previous study, in which no difference between algorithms was found (Dominke et al., 2024).

For the differentiation between DEM and DEP, the RF and the NB performed best when using raw scores as input data (BA = 80.8%), while the lowest accuracy was achieved using the GLMnet and adjusted raw scores (BA = 74.5%). In terms of F1 scores, both RF and NB again showed strong performance with F1 scores of 0.81 [0.73 – 0.90] for raw scores, indicating a good balance between precision and recall.

We could furthermore show that the model obtained using PSM data was able to predict unseen cases with a high degree of accuracy (BAC for the HC vs. DEM ranged up to 92.2% and to 87.6% for the DEM vs. DEP classification using the RF), speaking for the model's strong predictive power and its generalizability across different datasets. This suggests that the model is not only effective in the specific training context but also has the potential to be applied to new, unseen data, enhancing its utility in real-world clinical settings. The ability to transfer the model's performance to different cohorts reinforces its robustness and supports its broader applicability in neuropsychological diagnostics.

Thus, our findings suggest that all algorithms perform equally good within the differential diagnosis of HC, DEM and DEP using neuropsychological features. It might be noteworthy however for clinical neuropsychologists, depending on the specific data available to select the best performing ML algorithm as a diagnostic support system: Using raw scores, the GLMnet led to some of the best results, particularly when differentiating between HC and DEM. This suggests that, despite the lack of significant differences, GLMnet with raw scores may still be a reliable

option for improving classification accuracy in certain diagnostic scenarios. The fact that raw scores worked so well could be explained by their untransformed nature, eventually capturing subtle individual differences. Another possible explanation for the strong performance could be the relatively homogeneous age groups included in our study, consisting exclusively of older adults. In such a sample, the influence of age-related variance may be reduced, especially since there were pronounced cognitive deficits.

One potential explanation for the lack of significant differences in accuracy among the algorithms is that the neuropsychological features used, while informative, may not have been complex enough to reveal distinct patterns that would differentiate the groups in a way that would highlight the strengths of individual algorithms. Given that only 14 variables were included in the analysis, the models might not have had access to a sufficiently rich dataset, limiting their ability to uncover more complex distinctions. When the feature set is relatively small, even advanced algorithms may perform similarly, as there are fewer data patterns to exploit.

Nevertheless, it is important to emphasize that the classification performance remained high across all algorithms when distinguishing between DEM, HC, and DEP groups. This indicates that, while the algorithms did not show significant performance variation, they were still effective in accurately classifying the groups based on the available features. These findings suggest that the neuropsychological data were sufficiently informative to allow for reliable classification, even with a limited number of variables. Thus, neuropsychologists can select from a range of algorithms based on their clinical needs and the available data, making these tools versatile for different diagnostic scenarios. ML models, as used in this study, could be integrated into routine diagnostics. Neuropsychologists could rely on these algorithms to minimize diagnostic uncertainties and enhance diagnostic efficiency. Future research incorporating a broader array of features may help reveal more nuanced differences in algorithm performance and further enhance diagnostic accuracy. In future research, expanding the feature set with more complex or diverse data (e.g. neuroimaging, genetic markers) may help uncover more significant distinctions in algorithm performance.

Across different ML algorithms, our accuracies where higher than within a study by Almubark et al. (2020), who investigated the potential of neuropsychological tests such as the Mini-Mental State Examination in combination with Alzheimer's Disease Assessment Scale—Cognitive subscale and the Logical Memory subtest of the Wechsler Memory Scale in differentiating between patients with Alzheimer's dementia and healthy controls (Accuracy of the RF was around 76%). This supports the claim that CFD provides a superior tool for dementia diagnosis. The accuracy found in this study is noteworthy, given that it is comparable to accuracies reported in studies on HC and DEM (of Alzheimer's type) that used more and more costly parameters (e.g. including evidence from MRI, CSF and PET examinations, see Martin et al., 2023; Zhang et al., 2011). In fact, other studies using MR data or a combination of neuroimaging data and other biological data with ML found lower accuracies when trying to differentiate between HC and DEM (Gray et al., 2013; Vemuri et al., 2008). Previous research also speaks in favor of using neuropsychological assessment and ML for accurate diagnosis of Dementia with Lewy Bodies and Parkinsons's Disease Dementia (Bougea et al., 2022). A study by

Wang and colleagues found a better accuracy of neuropsychological assessments as compared to imaging data in differentiating between DEM of the Alzheimer's type and behavioral variant frontotemporal DEM (Wang et al., 2016). Since neuropsychological examinations are much more economical and non-invasive, this result underlines the particular importance a careful neuropsychological examination should have in dementia diagnosis. Across algorithms, the most important features for the classification between HC and patients with DEM in our study were related to tasks assessing verbal memory and semantic word fluency, which is in line with previous research (Henry et al., 2004; Li et al., 2023; Salmon & Bondi, 2009), suggesting that impairments in these cognitive functions constitute important early signs of dementia and should be accordingly prioritized. Clinicians should focus on these neuropsychological variables when trying to differentiate HC and patients with DEM.

Directly comparing our results of the DEM vs. DEP differentiation to other research is not possible, since to our knowledge our own previous study is the only one that used neuropsychological data to classify patients as either DEM or DEP: Surprisingly, our accuracies regarding the classification of DEM and DEP were lower in this study as compared to our previous one, in which we used the CERAD-NAB or a flexible battery approach (i.e. a combination of multiple individual tests) to differentiate between DEP and DEM (i.e. 80.8% vs. 87.0%, respectively). This finding might be explained by the fact that in the current study not only patients with Alzheimer's disease were included in the group of patients with DEM, but also patients with vascular dementia and other types of dementia. The DEM group within this study was therefore significantly more heterogeneous than in the previous one. Such a heterogenous set of patients reflects, however, the reality in German memory clinics much better. Accordingly, the findings of our current study are therefore probably more applicable to and relevant for most clinical settings.

Although there was no significant difference in performance between NB and the other algorithms across different data types, NB with raw scores tends to perform particularly well in the context of neuropsychological data. This could be explained by the fact that NB leverages the assumption that features are conditionally independent and likely follow a Gaussian normal distribution. Raw scores, being the direct, untransformed output of neuropsychological assessments, frequently approximate a normal distribution, making them a natural fit for Gaussian Naïve Bayes.

The most important features for the best performing algorithm in the comparison between patients with DEP and DEM were variables from the verbal memory test AWLT and from the TMT-A assessing processing speed. A key finding was the importance oft he recognition memory performance on the AWLT. This result is consistent with previous work suggesting that cognitive effort may be a critical factor in depression, with performance improving when support mechanisms, such as recognition-based tasks, are employed. In contrast, such a pattern of improvement is typically not observed in individuals with dementia (Leyhe et al., 2017). The most important features within this study are in line with recent neuropsychological research, suggesting important performance differences within these cognitive subdomains between the two diagnostic groups (Barlet et al., 2023; Ashendorf et al., 2008; Toda et al., 2022). Clinicans should focus on these variables when trying to differentiate between patients with DEP and DEM. This insight allows clinicians to prioritize these key variables,

especially when time or testing capacity is limited, thereby enhancing the efficiency and focus of the diagnostic process.

Interestingly, over 25% of the dementia group had missing data for the WAFA2lmrtr2 task, which may indicate that this particular task or component was especially challenging for these patients. While it was not identified as one of the most important features by our algorithms, the amount of missing data itself might warrant further investigation.

A strength of the present study is the multi-center approach used to collect the database for the present analyzes, which—together with an increased number of participants - increases the generalizability of our results and reduces the chances of overfitting. However, it must be noted that the sample examined here included only people from Germany and Austria and therefore does not allow any statements about how the results can be generalized to other populations outside of Germany and Austria.

Limitations

This study used a much larger sample size as compared to our previous investigation (Dominke et al., 2024) and other comparable investigations (Almubark et al., 2019; Gurevich et al., 2017; Wang et al., 2016), increasing generalizability of results to other memory clinics within Germany. A limitation of the present investigation is however the lack of ethnic diversity within our sample, restricting the generalizability of our results to a broader, culturally more diverse population. Furthermore, diagnosis was based on expert consensus of at least two experience psychiatrists. Accordingly, we have no pathological confirmation on the dementia diagnosis *via* postmortem brain examinations, ensuring correct dementia diagnosis.

While our study did not include patients with a dual diagnosis of DEP and DEM, the cognitive impairments observed in the DEP group could, over time, evolve into more pronounced deficits that may lead to diagnoses such as MCI or other forms of DEM. Thus, the CFD in combination with ML do not replace the need for continuous neuropsychological monitoring, especially in cases where depression-related cognitive deficits may eventually evolve into more severe forms of cognitive decline over time. In the age of clinical biomarkers for amyloid pathology, these results can be used by clinicians as a complementary tool to guide further investigation. Neuropsychological testing combined with machine learning models can help identify early cognitive changes that may warrant additional biomarker testing, such as amyloid PET scans, to assess for neurodegenerative processes. While biomarkers can provide more direct evidence of amyloid pathology, our approach highlights key cognitive features that clinicians may prioritize during initial assessments, particularly when biomarkers are not yet available or when the clinical picture is still uncertain.

Finally, the DEM group consisted of multiple underlying diseases, making the dementia group more heterogenous and thereby eventually complicating differential diagnosis. The individual number of subjects with diagnosis other than Alzheimer's dementia were however too small to create separate diagnostic groups (i.e. 18 with a diagnosis of vascular dementia (ICD-10: F01) and 25 with another diagnosis (ICD-10: F02-F03)).

Future actions

First, future research should try to validate our results on independent patient samples from different centers. It will be especially important to train the model on data from patients with different socio-demographic and different ethnic backgrounds, to make it generalizable to other settings. Our study focused exclusively on cognitive measures. Studies suggest that inclusion of other biomarkers and gait assessment might increase the accuracy with which DEM and HC and DEM and DEP can be distinguished (Gurevich et al., 2017; Tuena et al., 2024). Future studies using ML to design the most accurate diagnostic protocol should therefore also incorporate such non-cognitive measures in their input data set.

The present study focused on clearly delineated diagnostic groups in order to enable robust comparisons of cognitive profiles across individuals with DEP, DEM and HC. Thus, hard classification boundaries were employed. However, we acknowledge that especially in older populations, overlapping symptoms between depressive and neurocognitive disorders are common, and comorbidity or diagnostic uncertainty can complicate group assignment. While strict group definitions were necessary to ensure internal validity and statistical clarity, this approach may limit ecological validity. Future research should therefore consider methodological strategies that account for the spectrum-like nature of these conditions. Probabilistic classification approaches such as latent class analysis (Schreiber, 2017) for unsupervised ML or using soft class boundaries in supervised ML (Liu et al., 2011), where probabilistic predictions are made, could be used to reflect diagnostic uncertainty and allow for gradient or mixed group membership in future research. These methods offer the potential to model transitional or ambiguous clinical states more accurately than categorical approaches.

Conclusion

In the current research we were able to demonstrate that ML algorithms and data from the neuropsychological test battery CFD can accurately differentiate between DEM and HC as well as between DEM and DEP with accuracies ranging up to 94.0% and 80.8%, respectively. Our work also highlights the importance of using measures for verbal memory and word fluency as well as processing speed within neuropsychological assessments for the classification problems at hand. We could therefore provide evidence that ML algorithms in combination with the CFD can aid clinicians by providing both a decision aid and insight into the most important neuropsychological variables used for these classifications.

Acknowledgements

We would like to thank all centers involved in the underlying multicenter study.

Disclosure statement

Thomas Jahn receives fees from software license sales from Schuhfried GmbH, the manufacturer of the Vienna Test System (WTS).

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Almubark, I., Chang, L.-C., Nguyen, T., Turner, R. S., & Jiang, X. (2019). Early detection of Alzheimer's disease using patient neuropsychological and cognitive data and machine learning techniques. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 5971–5973).

Almubark, I., Chang, L.-C., Shattuck, K. F., Nguyen, T., Turner, R. S., & Jiang, X. (2020). A 5-min cognitive task with deep learning accurately detects early Alzheimer's disease. *Frontiers in Aging Neuroscience*, *12*, 603179. https://doi.org/10.3389/fnagi.2020.603179

Alzola, P., Carnero, C., Bermejo-Pareja, F., Sánchez-Benavides, G., Peña-Casanova, J., Puertas-Martín, V., Fernández-Calvo, B., & Contador, I. (2024). Neuropsychological assessment for early detection and diagnosis of dementia: current knowledge and new insights. *Journal of Clinical Medicine*, *13*(12), 3442. https://doi.org/10.3390/jcm13123442

Andreas, S., Schulz, H., Volkert, J., Dehoust, M., Sehner, S., Suling, A., Ausín, B., Canuto, A., Crawford, M., Da Ronch, C., Grassi, L., Hershkovitz, Y., Muñoz, M., Quirk, A., Rotenstein, O., Santos-Olmo, A. B., Shalev, A., Strehle, J., Weber, K., … Härter, M. (2017). Prevalence of mental disorders in elderly people: The European MentDis_ICF65+ study. *The British Journal of Psychiatry: The Journal of Mental Science*, *210*(2), 125–131. https://doi.org/10.1192/bjp.bp.115.180463

Ashendorf, L., Jefferson, A. L., O'Connor, M. K., Chaisson, C., Green, R. C., & Stern, R. A. (2008). Trail Making Test errors in normal aging, mild cognitive impairment, and dementia. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *23*(2), 129–137. https://doi.org/10.1016/j.acn.2007.11.005

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Barlet, B. D., Hauson, A. O., Pollard, A. A., Zhang, E. Z., Nemanim, N. M., Sarkissians, S., Lackey, N. S., Stelmach, N. P., Walker, A. D., Carson, B. T., Flora-Tostado, C., Reszegi, K., Allen, K. E., & Viglione, D. J. (2023). Neuropsychological performance in Alzheimer's disease versus late-life depression: A systematic review and meta-analysis. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *38*(7), 991–1016. https://doi.org/10.1093/arclin/acad036

Battista, P., Salvatore, C., Berlingeri, M., Cerasa, A., & Castiglioni, I. (2020). Artificial intelligence and neuropsychological measures: The case of Alzheimer's disease. *Neuroscience and Biobehavioral Reviews*, *114*, 211–228. https://doi.org/10.1016/j.neubiorev.2020.04.026

Berres, M., Zehnder, A., Bläsi, S., & Monsch, A. U. (2008). Evaluation of diagnostic scores with adjustment for covariates. *Statistics in Medicine*, *27*(10), 1777–1790. https://doi.org/10.1002/sim.3120

Bhagya Shree, S. R., & Sheshadri, H. S. (2018). Diagnosis of Alzheimer's disease using Naive Bayesian Classifier. *Neural Computing and Applications*, *29*(1), 123–132. https://doi.org/10.1007/s00521-016-2416-3

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bougea, A., Efthymiopoulou, E., Spanou, I., & Zikos, P. (2022). A novel machine learning algorithm predicts dementia with Lewy bodies versus Parkinson's disease dementia based on clinical and neuropsychological scores. *Journal of Geriatric Psychiatry and Neurology*, *35*(3), 317–320. https://doi.org/10.1177/0891988721993556

Butters, M. A., Whyte, E. M., Nebes, R. D., Begley, A. E., Dew, M. A., Mulsant, B. H., Zmuda, M. D., Bhalla, R., Meltzer, C. C., Pollock, B. G., Reynolds, C. F., 3rd., & Becker, J. T. (2004). The

nature and determinants of neuropsychological functioning in late-life depression. *Archives of General Psychiatry*, *61*(6), 587–595. https://doi.org/10.1001/archpsyc.61.6.587

Cantón-Habas, V., Rich-Ruiz, M., Romero-Saldaña, M., Carrera-González, M., & del, P. (2020). Depression as a risk factor for dementia and Alzheimer's disease. *Biomedicines*, *8*(11), 457. https://doi.org/10.3390/biomedicines8110457

Carrarini, C., Nardulli, C., Titti, L., Iodice, F., Miraglia, F., Vecchio, F., & Rossini, P. M. (2024). Neuropsychological and electrophysiological measurements for diagnosis and prediction of dementia: A review on Machine Learning approach. *Ageing Research Reviews*, *100*, 102417. https://doi.org/10.1016/j.arr.2024.102417

Castellazzi, G., Cuzzoni, M. G., Cotta Ramusino, M., Martinelli, D., Denaro, F., Ricciardi, A., Vitali, P., Anzalone, N., Bernini, S., Palesi, F., Sinforiani, E., Costa, A., Micieli, G., D'Angelo, E., Magenes, G., & Gandini Wheeler-Kingshott, C. A. M. (2020). A Machine Learning Approach for the Differential Diagnosis of Alzheimer and Vascular Dementia Fed by MRI Selected Features. *Frontiers in neuroinformatics, 14*, 25. https://doi.org/10.3389/fninf.2020.00025

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cheng, Y., Petrides, K. V., & Li, J. (2025). Estimating the minimum sample size for neural network model fitting—A Monte Carlo simulation study. *Behavioral Sciences*, *15*(2), 211. https://doi.org/10.3390/bs15020211

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dominke, C., Fischer, A. M., Grimmer, T., Diehl-Schmid, J., & Jahn, T. (2024). CERAD-NAB and flexible battery based neuropsychological differentiation of Alzheimer's dementia and depression using machine learning approaches. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *31*(2), 221–248. https://doi.org/10.1080/13825585.2022.2138255

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Er, F., Iscen, P., Sahin, S., Çinar, N., Karsidag, S., & Goularas, D. (2017). Distinguishing age-related cognitive decline from dementias: A study based on machine learning algorithms. *Journal of Clinical Neuroscience: Official Journal of the Neurosurgical Society of Australasia*, *42*, 186–192. https://doi.org/10.1016/j.jocn.2017.03.021

Fillenbaum, G. G., & Mohs, R. (2023). CERAD (Consortium to Establish a Registry for Alzheimer's Disease) neuropsychology assessment battery: 35 years and counting. *Journal of Alzheimer's Disease: JAD*, *93*(1), 1–27. https://doi.org/10.3233/JAD-230026

Florek, P., & Zagdański, A. (2023). Benchmarking state-of-the-art gradient boosting algorithms for classification. Preprint at https://doi.org/10.48550/arXiv.2305

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Franciotti, R., Nardini, D., Russo, M., Onofrj, M., & Sensi, S. L. (2023). Comparison of machine learning-based approaches to predict the conversion to Alzheimer's disease from mild cognitive impairment. *Neuroscience*, *514*, 143–152. https://doi.org/10.1016/j.neuroscience.2023.01.029

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Garcia-Gutierrez, F., Delgado-Alvarez, A., Delgado-Alonso, C., Díaz-Álvarez, J., Pytel, V., Valles-Salgado, M., Gil, M. J., Hernández-Lorenzo, L., Matías-Guiu, J., Ayala, J. L., & Matias-Guiu, J. A. (2021). Diagnosis of Alzheimer's disease and behavioural variant frontotemporal dementia with machine learning-aided neuropsychological assessment using feature engineering and genetic algorithms. *International Journal of Geriatric Psychiatry*, *37*(2). https://doi.org/10.1002/gps.5667

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., & Rueckert, D. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, *65*, 167–175. https://doi.org/10.1016/j.neuroimage.2012.09.065

Gupta, A., & Kahali, B. (2020). Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests. *Alzheimer's & Dementia (New York, N. Y.)*, *6*(1), e12049. https://doi.org/10.1002/trc2.12049

Gurevich, P., Stuke, H., Kastrup, A., Stuke, H., & Hildebrandt, H. (2017). Neuropsychological testing and machine learning distinguish Alzheimer's disease from other causes for cognitive impairment. *Frontiers in Aging Neuroscience*, *9*, 114. https://doi.org/10.3389/fnagi.2017.00114

Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia*, *42*(9), 1212–1222. https://doi.org/10.1016/j.neuropsychologia.2004.02.001

Jahn, T., & Hessler, J. B. (2023). *Manual cognitive functions dementia (Version 2 – Revision 6)*. Schuhfried GmbH.

Kao, S.-L., Chen, S.-C., Li, Y.-Y., & Lo, R. Y. (2019). Diagnostic diversity among patients with cognitive complaints: A 3-year follow-up study in a memory clinic. *International Journal of Geriatric Psychiatry*, *34*(12), 1900–1906. https://doi.org/10.1002/gps.5207

Knapskog, A.-B., Barca, M. L., & Engedal, K. (2014). Prevalence of depression among memory clinic patients as measured by the Cornell Scale of Depression in Dementia. *Aging & Mental Health*, *18*(5), 579–587. https://doi.org/10.1080/13607863.2013.827630

Kuring, J. K., Mathias, J. L., & Ward, L. (2018). Prevalence of depression, anxiety and PTSD in people with dementia: A systematic review and meta-analysis. *Neuropsychology Review*, *28*(4), 393–416. https://doi.org/10.1007/s11065-018-9396-2

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903. https://doi.org/10.21105/joss.01903

Lanza, C., Sejunaite, K., Steindel, C., Scholz, I., & Riepe, M. W. (2020). Cognitive profiles in persons with depressive disorder and Alzheimer's disease. *Brain Communications*, *2*(2), fcaa206. https://doi.org/10.1093/braincomms/fcaa206

Leyhe, T., Reynolds, C. F., Melcher, T., Linnemann, C., Klöppel, S., Blennow, K., Zetterberg, H., Dubois, B., Lista, S., & Hampel, H. (2017). A common challenge in older adults: Classification, overlap, and therapy of depression and dementia. *Alzheimer's*. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *13*(1), 59–71. https://doi.org/10.1016/j.jalz.2016.08.007

Li, K.-Y., Chien, C.-F., Huang, T.-W., & Yang, Y.-H. (2023). The use of verbal and nonverbal memory tests for Alzheimer's disease screening in Taiwan Chinese. *American Journal of Alzheimer's Disease and Other Dementias*, *38*, 15333175231201036. https://doi.org/10.1177/15333175231201036

Linnemann, C., & Lang, U. E. (2020). Pathways connecting late-life depression and dementia. *Frontiers in Pharmacology*, *11*, 279. https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2020.00279 https://doi.org/10.3389/fphar.2020.00279

Liu, Y., Zhang, H. H., & Wu, Y. (2011). Hard or soft classification? Large-margin unified machines. *Journal of the American Statistical Association*, *106*(493), 166–177. https://doi.org/10.1198/jasa.2011.tm10319

Ly, M., Karim, H. T., Becker, J. T., Lopez, O. L., Anderson, S. J., Aizenstein, H. J., Reynolds, C. F., Zmuda, M. D., & Butters, M. A. (2021). Late-life depression and increased risk of dementia: A longitudinal cohort study. *Translational Psychiatry*, *11*(1), 147. https://doi.org/10.1038/s41398-021-01269-y

Maito, M. A., Santamaría-García, H., Moguilner, S., Possin, K. L., Godoy, M. E., Avila-Funes, J. A., Behrens, M. I., Brusco, I. L., Bruno, M. A., Cardona, J. F., Custodio, N., García, A. M., Javandel, S., Lopera, F., Matallana, D. L., Miller, B., Okada de Oliveira, M., Pina-Escudero, S. D., Slachevsky, A., … Ibañez, A. (2023). Classification of Alzheimer's disease and frontotemporal dementia using routine clinical and cognitive measures across multicentric underrepresented samples: A cross sectional observational study. *The Lancet Regional Health - Americas*, *17*, 100387. https://doi.org/10.1016/j.lana.2022.100387

Martin, S. A., Townend, F. J., Barkhof, F., & Cole, J. H. (2023). Interpretable machine learning for dementia: A systematic review. *Alzheimer's*. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *19*(5), 2135–2149. https://doi.org/10.1002/alz.12948

Matias-Guiu, J. A., García-Gutiérrez, F., Pytel, V., Hernández-Lorenzo, L., Delgado-Álvarez, A., Valles-Salgado, M., Delgado-Alonso, C., Cabrera-Martín, M. N., Matias-Guiu, J., & Ayala, J. L. (2021). Machine learning for neuropsychological assessment of Alzheimer's disease and behavioral variant frontotemporal dementia. *Alzheimer's & Dementia*, *17*, e053036.

Molnar, C., Casalicchio, Giuseppe., & Bischl, B. (2018). iml: An R package for Interpretable Machine Learning. *Journal of Open Source Software*, *3*(26), 786. https://doi.org/10.21105/joss.00786

Montazeri, M., Montazeri, M., Bahaadinbeigy, K., Montazeri, M., & Afraz, A. (2023). Application of machine learning methods in predicting schizophrenia and bipolar disorders: A systematic review. *Health Science Reports*, *6*(1), e962. https://doi.org/10.1002/hsr2.962

Montorio, I., & Izal, M. (1996). The geriatric depression scale: A review of its development and utility. *International Psychogeriatrics*, *8*(1), 103–112. https://doi.org/10.1017/S1041610296002505

Murman, D. L. (2015). The impact of age on cognition. *Seminars in Hearing*, *36*(3), 111–121. https://doi.org/10.1055/s-0035-1555115

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M. B., Neil, D., Saffari, A., Mead, R., Hautbergue, G. M., Holbrook, J. D., & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews. Neurology*, *16*(8), 440–456. https://doi.org/10.1038/s41582-020-0377-8

Nichols, E., Szoeke, C. E. I., Vollset, S. E., Abbasi, N., Abd-Allah, F., Abdela, J., Aichour, M. T. E., Akinyemi, R. O., Alahdab, F., Asgedom, S. W., Awasthi, A., Barker-Collo, S. L., Baune, B. T., Béjot, Y., Belachew, A. B., Bennett, D. A., Biadgo, B., Bijani, A., Bin Sayeed, M. S., … Murray, C. J. L. (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *18*(1), 88–106. https://doi.org/10.1016/S1474-4422(18)30403-4

Petersen, R. C. (2016). Mild cognitive impairment. *Continuum (Minneapolis, Minn.)*, *22*(2 Dementia), 404–418. https://doi.org/10.1212/CON.0000000000000313

Piccolino, A. L., Piccolino, A. R., & Piccolino, S. G. (2025). Distinguishing Alzheimer's disease from other dementias using pattern profile analysis in the Meyers Neuropsychological Battery: An exploratory study. *Applied Neuropsychology. Adult*, *32*(4), 1087–1102. https://doi.org/10.1080/23279095.2023.2236742

Prabhakaran, D., Grant, C., Pedraza, O., Caselli, R., Athreya, A. P., & Chandler, M. (2024). Machine learning predicts conversion from normal aging to mild cognitive impairment using medical history, APOE genotype, and neuropsychological assessment. *Journal of Alzheimer's Disease: JAD*, *98*(1), 83–94. https://doi.org/10.3233/JAD-230556

R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rosas, A. G., Stögmann, E., & Lehrner, J. (2022). Neuropsychological prediction of dementia using the neuropsychological test battery Vienna – A retrospective study. *Brain Disorders*, *5*, 100028. https://doi.org/10.1016/j.dscb.2021.100028

Salmon, D. P., & Bondi, M. W. (2009). Neuropsychological assessment of dementia. *Annual Review of Psychology*, *60*(1), 257–282. https://doi.org/10.1146/annurev.psych.57.102904.190024

Sanchez-Martinez, S., Camara, O., Piella, G., Cikes, M., González-Ballester, M. Á., Miron, M., Vellido, A., Gómez, E., Fraser, A. G., & Bijnens, B. (2021). Machine learning for clinical decision-making: Challenges and opportunities in cardiovascular imaging. *Frontiers in Cardiovascular Medicine*, *8*, 765693. https://doi.org/10.3389/fcvm.2021.765693

Scott, I. A., Cook, D., Coiera, E. W., & Richards, B. (2019). Machine learning in clinical practice: Prospects and pitfalls. *The Medical Journal of Australia*, *211*(5), 203–205.e1. https://doi.org/10.5694/mja2.50294

Schreiber, J. B. (2017). Latent class analysis: An example for reporting results. *Research in Social & Administrative Pharmacy: RSAP*, *13*(6), 1196–1201. https://doi.org/10.1016/j.sapharm.2016.11.011

Silva Dos Santos Durães, R., Emy Yokomizo, J., Saffi, F., Castanho de Almeida Rocca, C., & Antonio de, P. S. (2022). Differential diagnosis findings between Alzheimer's disease and major depressive disorder: A review. *Psychiatry and Clinical Psychopharmacology*, *32*(1), 80–88. https://doi.org/10.5152/pcp.2022.21133

Sözeri-Varma, G. (2012). Depression in the elderly: Clinical features and risk factors. *Aging and Disease*, *3*(6), 465–471.

Steffens, D. C., & Potter, G. G. (2008). Geriatric depression and cognitive impairment. *Psychological Medicine*, *38*(2), 163–175. https://doi.org/10.1017/S003329170700102X

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Süle, A., Miljković, N., Polidori, P., & Kohl, S. (2019). Position paper on an ageing society. *European Journal of Hospital Pharmacy: Science and Practice*, *26*(6), 354–356. https://doi.org/10.1136/ejhpharm-2019-001910

Tetsuka, S. (2021). Depression and dementia in older adults: A neuropsychological review. *Aging and Disease*, *12*(8), 1920–1934. https://doi.org/10.14336/AD.2021.0526

Toda, A., Nagami, S., Katsumata, A., & Fukunaga, S. (2022). Verification of trail making test in elderly people with behavioral and psychological symptoms of dementia. *Ageing International*, *47*(3), 491–502. https://doi.org/10.1007/s12126-021-09424-y

Tuena, C., Pupillo, C., Stramba-Badiale, C., Stramba-Badiale, M., & Riva, G. (2024). Predictive power of gait and gait-related cognitive measures in amnestic mild cognitive impairment: A machine learning analysis. *Frontiers in Human Neuroscience*, *17*, 1328713. https://doi.org/10.3389/fnhum.2023.1328713

Vaccaro, M. G., Sarica, A., Quattrone, A., Chiriaco, C., Salsone, M., Morelli, M., & Quattrone, A. (2021). Neuropsychological assessment could distinguish among different clinical phenotypes of progressive supranuclear palsy: A Machine Learning approach. *Journal of Neuropsychology*, *15*(3), 301–318. https://doi.org/10.1111/jnp.12232

Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., Boeve, B. F., Petersen, R. C., & Jack, C. R. J. (2008). Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *NeuroImage*, *39*(3), 1186–1197. https://doi.org/10.1016/j.neuroimage.2007.09.073

Wang, J., Redmond, S. J., Bertoux, M., Hodges, J. R., & Hornberger, M. (2016). A comparison of magnetic resonance imaging and neuropsychological examination in the diagnostic distinction of Alzheimer's disease and behavioral variant frontotemporal dementia. *Frontiers in Aging Neuroscience*, *8*, 119. https://doi.org/10.3389/fnagi.2016.00119

Wiels, W., Baeken, C., & Engelborghs, S. (2020). *Depressive symptoms in the elderly—An early symptom of dementia? A systematic review*.

Wright, S. L., & Persad, C. (2007). Distinguishing between depression and dementia in older persons: Neuropsychological and neuropathological correlates. *Journal of Geriatric Psychiatry and Neurology*, *20*(4), 189–198. https://doi.org/10.1177/0891988707308801

Wunner, C., Schubert, A., Gosch, M., & Stemmler, M. (2022). Differential diagnosis of MCI, dementia and depression—A comparison of different cognitive profiles. *Psych*, *4*(2), 187–199. https://doi.org/10.3390/psych4020016

Zhang, D., Wang, Y., Zhou, L., Yuan, H., & Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, *55*(3), 856–867. https://doi.org/10.1016/j.neuroimage.2011.01.008

Appendix

Table A1. Classification results for patients with dementia ($N = 131$) vs. healthy controls ($N = 407$) across the different Machine Learning algorithms and data types used without using SMOTE.

	FL	GLM	SVM	Ranger	NB	LR
Raw scores	BAC = 50.0% [50.0 – 50.0] SEN = 0% [0 – 0] SPE = 100% [100 – 100]	BAC = 92.4% [87.3 – 97.5] SEN = 86.8% [77.0 – 96.3] SPE = 97.9% [93.2 – 100] F1 = .90 [.83 – .96]	BAC = 93.2% [89.0 – 97.8] SEN =88.6% [78.7 – 99.2] SPE = 97.8% [93.4 – 100] F1 = .90 [.84 – .97]	BAC = 93.2% [88.8 – 97.8] SEN = 88.6% [79.5 – 96.3] SPE = 97.8% [92.9 – 100] F1 = .90 [.84 – .97]	BAC = 92.7% [88.4 – 97.2] SEN = 91.1% [83.0 – 100] SPE = 94.4% [89.4 – 98.8] F1 = .87 [.77 – .94]	BAC = 92.5% [86.6 – 97.5] SEN = 88.0% [75.6 – 96.4] SPE = 97.1% [92.0– 100] F1 = .89 [.81 – .96]
T-scores	BAC = 50 % [50.0 – 50.0] SEN = 0% [0 – 0] SPE = 100 % [100 – 100]	BAC = 90.1% [83.4 – 97.4] SEN = 83.0% [68.9 – 96.3] SPE = 97.2% [93.1 – 100] F1 = .86 [.78 – .96]	BAC = 89.3% [82.3 – 95.4] SEN = 80.9% [65.9 – 95.6] SPE = 97.6% [93.7 – 100] F1 = .85 [.77 – .94]	BAC = 89.8% [81.2 – 97.1] SEN = 81.9% [63.8 – 96.2] SPE = 97.8% [94.8 – 100] F1 = .86 [.76 – .95]	BAC = 91.0% [86.0 – 97.5] SEN = 88.8% [77.2 – 100] SPE = 93.3% [86.0 – 98.4] F1 = .84 [.76 – .92]	BAC = 90.8% [84.4 – 97.6] SEN = 85.0 % [70.4 – 100] SPE = 96.5% [90.8 – 100] F1 = .86 [.80 – .93]
Adjusted raw scores	BAC = 50% [50.0 – 50.0] SEN = 0% [0 – 0] SPE = 100% [100 – 100]	BAC = 90.5% [84.3 – 97.3] SEN = 83.6% [69.7 – 99.2] SPE = 97.4% [93.9 – 100] F1 = .87 [.80 – .95]	BAC = 89.0% [82.3 – 95.4] SEN = 80.3% [65.5 – 95.2] SPE = 97.7% [94.9 – 100] F1 = .85 [.78 – .93]	BAC = 89.8% [80.1 – 97.2] SEN = 81.9% [62.4 – 96.2] SPE = 97.7% [93.7 – 100] F1 = .86 [.74 – .96]	BAC = 91.3% [86.5– 96.7] SEN = 89.4% [77.2 – 100] SPE = 93.2% [86.2 – 98.7] F1 = .85 [.76 – .92]	BAC = 91.6% [86.1 – 97.4] SEN = 86.6% [73.5 – 100] SPE = 96.6% [91.5 – 100] F1 = .87 [.82 – .93]

FL = Featureless Learner; GLM = GLMnet = Lasso and Elastic-Net Regularized Generalized Linear Models; SVM = Support Vector Machine; Ranger = Random Forest; NB = Naïve Bayes; LR = Logistic Regression; BAC = Balanced Accuracy; SEN = sensitivity; SPE = specificity; F1 scores represent the harmonic mean of precision and recall. Note that for the PSM & raw scores ($N = 88$ per group), SMOTE was not applied, as the dataset was already balanced through matching. Therefore the results are directly comparable to those reported in Table 4.

3. General Discussion

In the following chapter, I will first summarize the two research studies and then outline how they advance existing research on ML and neuropsychology in the differential diagnosis between DEM and DEP. Subsequently, I will focus on the limitations of the present studies and give directions for future research. Finally, a concluding summary will be provided.

3.1 Summary of studies

In the present dissertation, two studies were conducted to investigate the potential of ML algorithms in combination with neuropsychological assessments in the differentiation between DEM and DEP.

In the first study, I compared the performance of two different neuropsychological test batteries (i.e., a flexible battery approach consisting of multiple individually chosen neuropsychological tests vs. the well-established CERAD-NAB) and four differential ML algorithms (Support Vector Machine, Naïve Bayes, Random Forest, and conventional binary logistic regression) in the differentiation between DEP (N = 68) and Alzheimer's DEM (N= 121). Accuracies of the predictive models ranged between 83.0 – 87.0 %. All algorithms performed better than chance. There was yet no significant difference in classification performance across algorithms. Features related to verbal memory as well as figure copying of the Rey Complex Figure Test were most important for the classifications at hand as assessed via permutation feature importance. This was the first study showing that ML in combination with neuropsychological test data from the CERAD-NAB or a flexible battery approach can accurately classify patients as either having DEP or Alzheimer's DEM.

In the second study, I used a larger sample size and the newly developed Cognitive Functions Dementia (CFD) test set as input data, which was originally developed to improve early DEM diagnosis. In contrast to the first study, data was assessed across 12 different centers

in Germany and Austria. I compared more machine learning algorithms: Support Vector Machine, Naïve Bayes, Random Forest, GLMnet and traditional binary logistic regression in differentiating between DEP (N = 145) and a group of patients with DEM (N = 131) with different underlying pathophysiological processes as well as healthy controls (HC; N = 407).

Accuracies for the differentiation between DEM and HC reached up to 94.0%, for the differentiation between DEM and DEP 80.8%. All algorithms performed better than chance. Again, there was no significant differences in classification performance across algorithms. The most important features for the classification at hand were again variables related to verbal memory and tasks assessing verbal fluency as well as processing speed. This study provided further evidence for the usefulness of neuropsychological testing in differentiating between DEP and DEM. There was no significant difference between the different ML algorithms used.

These results indicate that the neuropsychological data provided enough valuable information to enable accurate classification, even with a limited number of variables. The results of our analyses highlight which neuropsychological variables—particularly those related to verbal memory, verbal fluency, and processing speed—are especially effective in differentiating between DEM and DEP. This enables neuropsychologists to tailor their test batteries to focus on these diagnostically relevant cognitive functions, thereby enhancing the accuracy of their assessments

3.2 Implications of the research

This research is the first to show that different comprehensive neuropsychological assessments such as CERAD-NAB, CFD or a flexible battery approach in combination with different ML algorithms can inform the clinical differentiation between DEM and DEP with accuracies ranging up to 87%. This work expands upon previous research, which already demonstrated the potential of automatic classification of MCI and Alzheimer's DEM using neuropsychological data (Battista et al., 2020; Kang et al., 2019), the differentiation of

Alzheimer's DEM from other sources for cognitive impairment based on cognitive tests (Gurevich et al., 2017), the potential of neuropsychological data in predicting the progression from MCI to Alzheimer DEM (Gallucci et al., 2018), the differentiation between DEM with Lewy Bodies and Parkinson's Disease DEM (Bougea et al., 2022) or between Alzheimer's DEM and frontotemporal DEM (Garcia-Gutierrez et al., 2021) using neuropsychological variables. This work shows that neuropsychological data can also help in the complex differentiation between DEM and DEP, which has been called one of the "knottiest problems of differential diagnosis" (Lezak et al., 2012). The most important features within our classifications were related to verbal memory tasks within both studies. The importance of these measures in differential diagnosis is consistent with previous research, that suggest that they might also be able to distinguish between different types of DEM and resemble the most early predictors for Alzheimer's DEM (Braaten et al., 2006; Bussè et al., 2017; Garcia Rosas, Stögmann & Lehrner, 2022) and for the conversion from MCI to Alzheimer's DEM in depressed individuals (Potter et al., 2013). This is especially important, given that ML algorithms have been shown to be more accepted by clinicians if they coincide with their professional expertise (Pazzani et al., 2001). In the second investigation, variables related to verbal fluency were also found to be important both for the differentiation between HC and DEM and DEM and DEP. A large effect size of the differences of DEM and DEP has been found within this cognitive domain in previous research, even though effect sizes varied between different investigations (de Araujo et al., 2011; Barlet et al., 2023).

Knowledge of the key neuropsychological variables is also important for another reason: It has been shown that some patients tend to quit neuropsychological examinations because they feel overwhelmed. Missing values in DEM research are therefore common (Schmid et al., 2014). Starting an assessment with tasks assessing the most important variables for the classification at hand therefore seems reasonable. This research can inform clinicians in the

way of presenting tasks assessing verbal memory in the first part of an examination due to their relevance for differential diagnosis.

The study also demonstrates that more comprehensive neuropsychological assessments offer a distinct advantage over simpler tools like the Beck Depression Inventory or basic cognitive measures such as the Mini-Mental State Examination (MMSE). Machine learning methods, as shown in the first study, can more accurately classify challenging cases that traditional methods may struggle with, even when MMSE scores are similar across different conditions. The results of our classification are furthermore comparable to investigations using more elaborate procedures such as combinations of MRI, CFS markers and PET to differentiate between HC and Alzheimer's DEM (Martin et al., 2023). Thus, comprehensive neuropsychological assessments should be implemented into daily clinical practice since they can give valuable hints regarding the diagnosis of investigated patients. This is encouraging given the fact that not all clinicians have access to more elaborate imaging procedures and some patients might not want to undergo invasive procedures such as lumbar puncture.

It also must be mentioned that the research shows that different ML algorithms did not outperform LR and that there were no significant differences between them in terms of classification accuracy. These finding contrasts other investigations showing that individual algorithms outperformed others: Maroco et al. (2011) for example showed that RF performed significantly better than other algorithms in differentiating between MCI and DEM. Miah and colleagues (2021) similarly found a superiority of SVM and RF in the classification of DEM based on clinical features from open access repositories. Large meta-analyses on the diagnosis of cognitive impairment and DEM based on neuroimaging, neuropsychological or electrophysiological data yet similarly found no evidence for a superiority of a given ML algorithm (Carrarini et al., 2024; Pellegrini et al., 2018).

There were furthermore no significant differences in terms of classification accuracy between the CERAD-NAB and the flexible battery approach used in the first study, suggesting

that both test batteries are equally effective in the differentiation between Alzheimer's DEM and DEP. A direct comparison to the CFD is not possible, since we used a different sample for the second study. Classification accuracy was a bit less in the second study, which can however be attributed to the composition of the more heterogenous DEM group and not the CFD itself.

The present findings not only advance the theoretical understanding of ML applications in dementia diagnostics but also have concrete implications for day-to-day neuropsychological practice. Given that variables related to verbal memory, verbal fluency, and processing speed consistently emerged as the most informative features, neuropsychologists should prioritize these domains for differential diagnosis between DEM and DEP. Moreover, the integration of individual patient test profiles into machine learning models offers a promising avenue to improve diagnostic accuracy beyond traditional interpretive methods. By feeding comprehensive neuropsychological data into validated ML algorithms, clinicians may benefit from data-driven, objective decision support that complements clinical judgment. This approach can be especially valuable in ambiguous cases where clinical symptoms overlap, increasing confidence in differential diagnosis.

3.3 Limitations

Even though the sample size within the first study is comparable to similar research studies (Almubark et al., 2019; Byeon, 2020) and our subjects resemble the patients seen in everyday practice, a limitation of the first study is the small sample size investigated, which limits the generalizability of results to other samples due to idiosyncrasies of patients (Dwyer et al., 2018). The samples in both studies furthermore only consisted of german-speaking subjects. Results might not be transferable to non-German speaking patients or individuals with a more ethnically diverse background. More research on potential differences in classification performance depending on the subjects investigated is needed.

Due to the explorative design of our studies, the diagnostic groups within this work were furthermore based on the diagnoses given by the respective psychiatrist, which also included results from imaging data and sometimes CSF markers. Diagnoses were however not validated by autopsy. Post-mortem brain examinations yet remain the gold-standard for definitive diagnosis (Nichols et al., 2023; Suemoto & Leite, 2023). Similarly, it would be interesting to see whether ML algorithms in combination with neuropsychological features perform similar or even outperform trained psychiatrists and neuropsychologists in their diagnostic decisions. A finding which has been shown for the classification of hip fractures (Murphy et al., 2022), breast cancer detection (Becker et al., 2017) or the classification of age-related macular degeneration (Burlina et al., 2018). The procedure of post-mortem brain examinations is however beyond the possibilities of most research groups, resulting in a great number of investigations using the diagnosis given by a psychiatrist and biomarkers as the standard of truth (see Almubark et al., 2020; Bachli et al., 2020; Gurevich et al., 2017). Yet, this ideal design should be kept in mind when planning future studies.

Within the second study, the DEM group consisted of patients with different underlying pathophysiological processes, making the group rather heterogenous. This procedure was chosen to increase the sample size of the DEM group and to investigate how neuropsychological data could be able to differentiate between DEM with different underlying pathophysiology's and DEP. This could yet explain the lower classification performance as compared to the first study. Patients with vascular DEM for example are known to exhibit heterogenous cognitive deficits based on the specific brain region affected (Braaten et al., 2006) and their cognitive profiles often differ from those observed in Alzheimer's DEM (Garcia-Gutierrez et al., 2021). Under ideal conditions, additionally larger sample sizes for all different types of DEM would have been investigated and compared to one another as well as to the cognitive profile associated with DEP. This approach was however far beyond the possibilities and would require much more time for recruiting and testing of all subjects.

3.4 Future Research

Future research should try to address the limitations of the present work: First, it will be necessary to validate our models on independent external samples from different centers to see whether they perform equally well. This step is crucial before one can think about implementation in clinical practice. To generalize to other samples, future investigations should focus on training the models on patients with various ethical and socio-demographic backgrounds to represent all people affected by DEM or DEP.

It will furthermore be of interest to see how well the CFD differentiates between different types of DEM when using ML algorithms. Our sample of DEM patients within our second study consisted of different types of DEM. The sample size of individual diagnoses was however too small to distinguish them individually. In their exploratory study, Piccolino et al. (2025) have used pattern profile analysis and the Meyer's Neuropsychological Battery to differentiate between Lewy Body DEM, Alzheimer's DEM, vascular DEM and Parkinson's DEM with accuracies up to 88%. Garcia-Gutierrez and colleagues (2021) have found neuropsychological assessment and ML to differentiate between Alzheimer's DEM and frontotemporal DEM with accuracies over 84%, which is especially noteworthy given the fact that these two disorders often show overlapping cognitive dysfunction profiles. Other reviews have come to similar conclusions (Carrarini et al., 2024). This finding is encouraging for future research investigating the potential of neuropsychological test batteries in differential DEM diagnosis in general. Since MCI – similarly to DEP - is characterized by less pronounced cognitive deficits as in DEM, the differentiation between MCI and DEP is much more challenging and traditional measures of neuropsychological investigations such as the clock-drawing test have been shown to fail at differentiating between the two (Wunner et al., 2022). It would be interesting to investigate the potential of ML within this specific differentiation in future studies.

It could also be interesting to record behavioral observations during the neuropsychological assessment in a standardized manner and use them as an additional feature in the ML algorithm: For example, it has been shown that patients with DEP often give answers such as “I don’t know” and give up easily while patients with DEM often give wrong answers and deny cognitive deficits (Tetsuka, 2021). Compared to controls, patients with DEP also show lower performance motivation and more negative attitudes towards neuropsychological testing, suggesting a mediating role of these factors in the cognitive deficits observed during assessment (Moritz et al., 2017). It is quite possible that, in addition to the objective differences in performance in neuropsychological tests, these observations provide valuable information for the diagnostic process.

Similarly, recent investigations on the capability of ML techniques in identifying the cognitive status of subjects based on spontaneous speech have demonstrated promising results (García-Gutiérrez et al., 2024; Liu et al., 2020; Yang et al., 2022). Integrating these findings with our neuropsychological features and relevant biomarkers could potentially enhance diagnostic accuracy and support more informed clinical decision-making.

4. Conclusion

Taken together, both experimental studies presented in this work provide robust evidence supporting the integration of ML algorithms with comprehensive neuropsychological assessments to improve differential diagnosis between DEM and DEP. This research thereby constitutes a pioneering effort to expand the existing body of knowledge on the applicability of ML techniques in DEM diagnosis by demonstrating their efficacy not only in detecting cognitive decline but also in accurately distinguishing DEM from DEP, a distinction that is often clinically challenging due to the overlapping symptomatology.

While these findings underscore the promise of ML and neuropsychological assessments withing this differential diagnostic question, it remains unclear which specific algorithm or model architecture yield optimal performance, highlighting the need for future systematic comparative evaluations.

Importantly, this work also shows that variables related to verbal memory and word fluency as well as processing speed seem to be especially useful in the differentiation between DEM and DEP and should be integral to every neuropsychological assessment in daily practice. Before clinical implementation of the ML algorithms for individual case diagnosis, the models require validation and training on larger and more heterogeneous data sets form multi-center cohorts to ensure generalizability and robustness.

From a practical standpoint, neuropsychologists are encouraged to prioritize testing of verbal memory, verbal fluency, and processing speed, as these domains provide the most diagnostically relevant information. In addition, incorporating individual patient test profiles into ML-based decision-support tools may enhance diagnostic accuracy and aid clinical judgment, particularly in complex or ambiguous cases. However, such tools should complement, not replace, expert clinical evaluation until further validated for routine use.

5. References

Abdoli, N., Salari, N., Darvishi, N., Jafarpour, S., Solaymani, M., Mohammadi, M., & Shohaimi, S. (2022). The global prevalence of major depressive disorder (MDD) among the elderly: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *132*, 1067–1073. https://doi.org/10.1016/j.neubiorev.2021.10.041

Al-Aidaroos, K. M., Bakar, A. A., & Othman, Z. (2010). Naive Bayes variants in classification learning. *2010 international conference on information retrieval & knowledge management (CAMP)*, 276–281.

Almubark, I., Chang, L.-C., Shattuck, K. F., Nguyen, T., Turner, R. S., & Jiang, X. (2020). A 5-min Cognitive Task With Deep Learning Accurately Detects Early Alzheimer's Disease. *Frontiers in Aging Neuroscience*, *12*, 603179. https://doi.org/10.3389/fnagi.2020.603179

Alzola, P., Carnero, C., Bermejo-Pareja, F., Sánchez-Benavides, G., Peña-Casanova, J., Puertas-Martín, V., Fernández-Calvo, B., & Contador, I. (2024). Neuropsychological Assessment for Early Detection and Diagnosis of Dementia: Current Knowledge and New Insights. *Journal of Clinical Medicine*, *13*(12). https://doi.org/10.3390/jcm13123442

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistic Surveys*, *4*, 40–79.

Bachli, M. B., Sedeño, L., Ochab, J. K., Piguet, O., Kumfor, F., Reyes, P., Torralva, T., Roca, M., Cardona, J. F., Campo, C. G., Herrera, E., Slachevsky, A., Matallana, D., Manes, F., García, A. M., Ibáñez, A., & Chialvo, D. R. (2020). Evaluating the reliability of neurocognitive biomarkers of neurodegenerative diseases across countries: A machine learning approach. *NeuroImage*, *208*, 116456. https://doi.org/10.1016/j.neuroimage.2019.116456

Bai, W., Chen, P., Cai, H., Zhang, Q., Su, Z., Cheung, T., Jackson, T., Sha, S., & Xiang, Y.-T. (2022). Worldwide prevalence of mild cognitive impairment among community dwellers

aged 50 years and older: A meta-analysis and systematic review of epidemiology studies. *Age and Ageing*, *51*(8), afac173. https://doi.org/10.1093/ageing/afac173

Barlet, B. D., Hauson, A. O., Pollard, A. A., Zhang, E. Z., Nemanim, N. M., Sarkissians, S., Lackey, N. S., Stelmach, N. P., Walker, A. D., Carson, B. T., Flora-Tostado, C., Reszegi, K., Allen, K. E., & Viglione, D. J. (2023). Neuropsychological Performance in Alzheimer's Disease versus Late-Life Depression: A Systematic Review and Meta-Analysis. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *38*(7), 991–1016. https://doi.org/10.1093/arclin/acad036

Barth, S., Schönknecht, P., Pantel, J., & Schröder, J. L. (2005). Neuropsychologische Profile in der Demenzdiagnostik: Eine Untersuchung mit der CERAD-NP-Testbatterie. *Fortschritte Der Neurologie Psychiatrie*, *73*, 568–576.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., & Filippi, M. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage. Clinical*, *21*, 101645. https://doi.org/10.1016/j.nicl.2018.101645

Battista, P., Salvatore, C., Berlingeri, M., Cerasa, A., & Castiglioni, I. (2020). Artificial intelligence and neuropsychological measures: The case of Alzheimer's disease. *Neuroscience and Biobehavioral Reviews*, *114*, 211–228. https://doi.org/10.1016/j.neubiorev.2020.04.026

Becker, A. S., Marcon, M., Ghafoor, S., Wurnig, M. C., Frauenfelder, T., & Boss, A. (2017). Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. *Investigative Radiology*, *52*(7), 434–440. https://doi.org/10.1097/RLI.0000000000000358

Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V., Croteau, J., & Consortium for the Early Identification of Alzheimer's disease-Quebec. (2017). Neuropsychological Measures that Predict Progression from Mild Cognitive Impairment to Alzheimer's type dementia in

Older Adults: A Systematic Review and Meta-Analysis. *Neuropsychology Review*, *27*(4), 328–353. https://doi.org/10.1007/s11065-017-9361-5

Berk, M., Köhler-Forsberg, O., Turner, M., Penninx, B. W. J. H., Wrobel, A., Firth, J., Loughman, A., Reavley, N. J., McGrath, J. J., Momen, N. C., Plana-Ripoll, O., O'Neil, A., Siskind, D., Williams, L. J., Carvalho, A. F., Schmaal, L., Walker, A. J., Dean, O., Walder, K., … Marx, W. (2023). Comorbidity between major depressive disorder and physical diseases: A comprehensive review of epidemiology, mechanisms and management. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *22*(3), 366–387. https://doi.org/10.1002/wps.21110

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bischkopf, J., Busse, A., & Angermeyer, M. C. (2002). Mild cognitive impairment—A review of prevalence, incidence and outcome according to current approaches. *Acta Psychiatrica Scandinavica*, *106*(6), 403–414. https://doi.org/10.1034/j.1600-0447.2002.01417.x

Bloom, D. E., Canning, D., & Lubet, A. (2015). Global Population Aging: Facts, Challenges, Solutions & Perspectives. In *Daedalus* (Bd. 144, Nummer 2, S. 80–92).

Bougea, A., Efthymiopoulou, E., Spanou, I., & Zikos, P. (2022). A Novel Machine Learning Algorithm Predicts Dementia With Lewy Bodies Versus Parkinson's Disease Dementia Based on Clinical and Neuropsychological Scores. *Journal of Geriatric Psychiatry and Neurology*, *35*(3), 317–320. https://doi.org/10.1177/0891988721993556

Braaten, A. J., Parsons, T. D., McCue, R., Sellers, A., & Burns, W. J. (2006). Neurocognitive differential diagnosis of dementing diseases: Alzheimer's Dementia, Vascular Dementia, Frontotemporal Dementia, and Major Depressive Disorder. *The International Journal of Neuroscience*, *116*(11), 1271–1293. https://doi.org/10.1080/00207450600920928

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brenowitz, W. D., Keene, C. D., Hawes, S. E., Hubbard, R. A., Longstreth, W. T. J., Woltjer, R. L., Crane, P. K., Larson, E. B., & Kukull, W. A. (2017). Alzheimer's disease neuropathologic change, Lewy body disease, and vascular brain injury in clinic- and community-based samples. *Neurobiology of Aging*, *53*, 83–92. https://doi.org/10.1016/j.neurobiolaging.2017.01.017

Brodaty, H., Luscombe, G., Anstey, K. J., Cramsie, J., Andrews, G., & Peisah, C. (2003). Neuropsychological performance and dementia in depressed patients after 25-year follow-up: A controlled study. *Psychological Medicine*, *33*(7), 1263–1275. https://doi.org/10.1017/s0033291703008195

Brzezińska, A., Bourke, J., Rivera-Hernández, R., Tsolaki, M., Woźniak, J., & Kaźmierski, J. (2020). Depression in Dementia or Dementia in Depression? Systematic Review of Studies and Hypotheses. *Current Alzheimer Research*, *17*(1), 16–28. https://doi.org/10.2174/1567205017666200217104114

Burlina, P., Joshi, N., Pacheco, K. D., Freund, D. E., Kong, J., & Bressler, N. M. (2018). Utility of Deep Learning Methods for Referability Classification of Age-Related Macular Degeneration. *JAMA Ophthalmology*, *136*(11), 1305–1307. https://doi.org/10.1001/jamaophthalmol.2018.3799

Burmester, B., Leathem, J., & Merrick, P. (2016). Subjective Cognitive Complaints and Objective Cognitive Function in Aging: A Systematic Review and Meta-Analysis of Recent Cross-Sectional Findings. *Neuropsychology Review*, *26*(4), 376–393. https://doi.org/10.1007/s11065-016-9332-2

Bussè, C., Anselmi, P., Pompanin, S., Zorzi, G., Fragiacomo, F., Camporese, G., Di Bernardo, G. A., Semenza, C., Caffarra, P., & Cagnin, A. (2017). Specific Verbal Memory Measures May Distinguish Alzheimer's Disease from Dementia with Lewy Bodies. *Journal of Alzheimer's disease : JAD*, *59*(3), 1009–1015. https://doi.org/10.3233/JAD-170154

Byeon, H. (2020). Is the Random Forest Algorithm Suitable for Predicting Parkinson's Disease with Mild Cognitive Impairment out of Parkinson's Disease with Normal Cognition? *International Journal of Environmental Research and Public Health*, *17*(7). https://doi.org/10.3390/ijerph17072594

Byers, A. L., & Yaffe, K. (2011). Depression and risk of developing dementia. *Nature Reviews. Neurology*, *7*(6), 323–331. https://doi.org/10.1038/nrneurol.2011.60

Cantón-Habas, V., Rich-Ruiz, M., Romero-Saldaña, M., & Carrera-González, M. D. P. (2020). Depression as a Risk Factor for Dementia and Alzheimer's Disease. *Biomedicines*, *8*(11), 457. https://doi.org/10.3390/biomedicines8110457

Carrarini, C., Nardulli, C., Titti, L., Iodice, F., Miraglia, F., Vecchio, F., & Rossini, P. M. (2024). Neuropsychological and electrophysiological measurements for diagnosis and prediction of dementia: A review on Machine Learning approach. *Ageing Research Reviews*, *100*, 102417. https://doi.org/10.1016/j.arr.2024.102417

Castellazzi, G., Cuzzoni, M. G., Cotta Ramusino, M., Martinelli, D., Denaro, F., Ricciardi, A., Vitali, P., Anzalone, N., Bernini, S., Palesi, F., Sinforiani, E., Costa, A., Micieli, G., D'Angelo, E., Magenes, G., & Gandini Wheeler-Kingshott, C. A. M. (2020). A Machine Learning Approach for the Differential Diagnosis of Alzheimer and Vascular Dementia Fed by MRI Selected Features. *Frontiers in Neuroinformatics*, *14*, 25. https://doi.org/10.3389/fninf.2020.00025

Christensen, H., Griffiths, K., Mackinnon, A., & Jacomb, P. (1997). A quantitative review of cognitive deficits in depression and Alzheimer-type dementia. *Journal of the International Neuropsychological Society: JINS*, *3*(6), 631–651.

Copeland, J. R. M., Beekman, A. T. F., Braam, A. W., Dewey, M. E., Delespaul, P., Fuhrer, R., Hooijer, C., Lawlor, B. A., Kivela, S.-L., Lobo, A., Magnusson, H., Mann, A. H., Meller, I., Prince, M. J., Reischies, F., Roelands, M., Skoog, I., Turrina, C., deVries, M. W., & Wilson,

K. C. M. (2004). Depression among older people in Europe: The EURODEP studies. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *3*(1), 45–49.

Cui, L., Li, S., Wang, S., Wu, X., Liu, Y., Yu, W., Wang, Y., Tang, Y., Xia, M., & Li, B. (2024). Major depressive disorder: Hypothesis, mechanism, prevention and treatment. *Signal Transduction and Targeted Therapy*, *9*(1), 30. https://doi.org/10.1038/s41392-024-01738-y

de Araujo, N. B., Barca, M. L., Engedal, K., Coutinho, E. S. F., Deslandes, A. C., & Laks, J. (2011). Verbal fluency in Alzheimer's disease, Parkinson's disease, and major depression. *Clinics*, *66*(4), 623–627. https://doi.org/10.1590/S1807-59322011000400017

Duke Han, S., Nguyen, C. P., Stricker, N. H., & Nation, D. A. (2017). Detectable Neuropsychological Differences in Early Preclinical Alzheimer's Disease: A Meta-Analysis. *Neuropsychology Review*, *27*(4), 305–325. https://doi.org/10.1007/s11065-017-9345-5

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gallucci, M., Di Battista, M. E., Battistella, G., Falcone, C., Bisiacchi, P. S., & Di Giorgi, E. (2018). Neuropsychological tools to predict conversion from amnestic mild cognitive impairment to dementia. The TREDEM Registry. *Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition*, *25*(4), 550–560. https://doi.org/10.1080/13825585.2017.1349869

García-Gutiérrez, F., Alegret, M., Marquié, M., Muñoz, N., Ortega, G., Cano, A., De Rojas, I., García-González, P., Olivé, C., Puerta, R., García-Sanchez, A., Capdevila-Bayo, M., Montrreal, L., Pytel, V., Rosende-Roca, M., Zaldua, C., Gabirondo, P., Tárraga, L., Ruiz, A., … Valero, S. (2024). Unveiling the sound of the cognitive status: Machine Learning-based

speech analysis in the Alzheimer's disease spectrum. *Alzheimer's Research & Therapy*, *16*(1), 26. https://doi.org/10.1186/s13195-024-01394-y

García-Gutièrrez, F., Delgado-Alvarez, A., Delgado-Alonso, C., Díaz-Álvarez, J., Pytel, V., Valles-Salgado, M., Gil, M. J., Hernández-Lorenzo, L., Matías-Guiu, J., Ayala, J. L., & Matias-Guiu, J. A. (2021). Diagnosis of Alzheimer's disease and behavioural variant frontotemporal dementia with machine learning-aided neuropsychological assessment using feature engineering and genetic algorithms. *International Journal of Geriatric Psychiatry*, *37*(2). https://doi.org/10.1002/gps.5667

Garcia Rosas, A. G., Stögmann, E., & Lehrner, J. (2022). Neuropsychological prediction of dementia using neuropsychological test battery Vienna – A retrospective study. *Brain Disorders,* 5, 100028. https://doi.org/10.1016/j.dscb.2021.100028

Garn, H., Coronel, C., Waser, M., Caravias, G., & Ransmayr, G. (2017). Differential diagnosis between patients with probable Alzheimer's disease, Parkinson's disease dementia, or dementia with Lewy bodies and frontotemporal dementia, behavioral variant, using quantitative electroencephalographic features. *Journal of Neural Transmission (Vienna, Austria: 1996)*, *124*(5), 569–581. https://doi.org/10.1007/s00702-017-1699-6

Gasser, A.-I., Salamin, V., & Zumbach, S. (2018). Dépression de la personne âgée ou maladie d'Alzheimer prodromique: Quels outils pour le diagnostic différentiel ? *L'Encéphale*, *44*(1), 52–58. https://doi.org/10.1016/j.encep.2017.03.002

Graham, S. A., Lee, E. E., Jeste, D. V., Van Patten, R., Twamley, E. W., Nebeker, C., Yamada, Y., Kim, H.-C., & Depp, C. A. (2020). Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry Research*, *284*, 112732. https://doi.org/10.1016/j.psychres.2019.112732

Gualtieri, C. T., & Morgan, D. W. (2008). The frequency of cognitive impairment in patients with anxiety, depression, and bipolar disorder: An unaccounted source of variance in clinical

trials. *The Journal of Clinical Psychiatry*, *69*(7), 1122–1130. https://doi.org/10.4088/jcp.v69n0712

Gurevich, P., Stuke, H., Kastrup, A., Stuke, H., & Hildebrandt, H. (2017). Neuropsychological Testing and Machine Learning Distinguish Alzheimer's Disease from Other Causes for Cognitive Impairment. *Frontiers in Aging Neuroscience*, *9*, 114. https://doi.org/10.3389/fnagi.2017.00114

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Bd. 2). Springer.

Hsieh, S., Schubert, S., Hoon, C., Mioshi, E., & Hodges, J. R. (2013). Validation of the Addenbrooke's Cognitive Examination III in frontotemporal dementia and Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, *36*(3–4), 242–250. https://doi.org/10.1159/000351671

James, C., Ranson, J. M., Everson, R., & Llewellyn, D. J. (2021). Performance of Machine Learning Algorithms for Predicting Progression to Dementia in Memory Clinic Patients. *JAMA Network Open*, *4*(12), e2136553. https://doi.org/10.1001/jamanetworkopen.2021.36553

Javeed, A., Dallora, A. L., Berglund, J. S., Ali, A., Ali, L., & Anderberg, P. (2023). Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. *Journal of Medical Systems*, *47*(1), 17. https://doi.org/10.1007/s10916-023-01906-7

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Johnen, A., & Bertoux, M. (2019). Psychological and Cognitive Markers of Behavioral Variant Frontotemporal Dementia-A Clinical Neuropsychologist's View on Diagnostic Criteria and Beyond. *Frontiers in Neurology*, *10*, 594. https://doi.org/10.3389/fneur.2019.00594

Kang, H., Zhao, F., You, L., Giorgetta, C., D, V., Sarkhel, S., & Prakash, R. (2014). Pseudo-dementia: A neuropsychological review. *Annals of Indian Academy of Neurology*, *17*(2), 147–154. https://doi.org/10.4103/0972-2327.132613

Kang, M. J., Kim, S. Y., Na, D. L., Kim, B. C., Yang, D. W., Kim, E.-J., Na, H. R., Han, H. J., Lee, J.-H., & Kim, J. H. (2019). Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data. *BMC medical informatics and decision making*, *19*, 1–9.

Karantzoulis, S., & Galvin, J. E. (2011). Distinguishing Alzheimer's disease from other major forms of dementia. *Expert Review of Neurotherapeutics*, *11*(11), 1579–1591. https://doi.org/10.1586/ern.11.155

Kemp, J., Philippi, N., Phillipps, C., Demuynck, C., Albasser, T., Martin-Hunyadi, C., Schmidt-Mutter, C., Cretin, B., & Blanc, F. (2017). Cognitive profile in prodromal dementia with Lewy bodies. *Alzheimer's Research & Therapy*, *9*(1), 19. https://doi.org/10.1186/s13195-017-0242-1

Kiloh, L. G. (1961). Pseudo-dementia. *Acta Psychiatrica Scandinavica*, *37*(4), 336–351. https://doi.org/10.1111/j.1600-0447.1961.tb07367.x

Kitching, D. (2015). Depression in dementia. *Australian Prescriber*, *38*(6), 209–2011. https://doi.org/10.18773/austprescr.2015.071

Kivipelto, M., Ngandu, T., Laatikainen, T., Winblad, B., Soininen, H., & Tuomilehto, J. (2006). Risk score for the prediction of dementia risk in 20 years among middle aged people: A longitudinal, population-based study. *The Lancet. Neurology*, *5*(9), 735–741. https://doi.org/10.1016/S1474-4422(06)70537-3

Lanza, C., Sejunaite, K., Steindel, C., Scholz, I., & Riepe, M. W. (2020). Cognitive profiles in persons with depressive disorder and Alzheimer's disease. *Brain Communications*, *2*(2), fcaa206. https://doi.org/10.1093/braincomms/fcaa206

Leyhe, T., Reynolds, C. F. 3rd, Melcher, T., Linnemann, C., Klöppel, S., Blennow, K., Zetterberg, H., Dubois, B., Lista, S., & Hampel, H. (2017). A common challenge in older adults: Classification, overlap, and therapy of depression and dementia. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *13*(1), 59–71. https://doi.org/10.1016/j.jalz.2016.08.007

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5. Aufl.). Oxford University Press.

Linnemann, C., & Lang, U. E. (2020). Pathways Connecting Late-Life Depression and Dementia. *Frontiers in Pharmacology*, *11*, 279. https://doi.org/10.3389/fphar.2020.00279

Liu, L., Zhao, S., Chen, H., & Wang, A. (2020). A new machine learning method for identifying Alzheimer's disease. *Simulation Modelling Practice and Theory*, *99*, 102023. https://doi.org/10.1016/j.simpat.2019.102023

Lobo, A., Launer, L. J., Fratiglioni, L., Andersen, K., Di Carlo, A., Breteler, M. M., Copeland, J. R., Dartigues, J. F., Jagger, C., Martinez-Lage, J., Soininen, H., & Hofman, A. (2000). Prevalence of dementia and major subtypes in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. *Neurology*, *54*(11 Suppl 5), S4-9.

Mackin, R. S., Nelson, J. C., Delucchi, K. L., Raue, P. J., Satre, D. D., Kiosses, D. N., Alexopoulos, G. S., & Arean, P. A. (2014). Association of age at depression onset with cognitive functioning in individuals with late-life depression and executive dysfunction. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry*, *22*(12), 1633–1641. https://doi.org/10.1016/j.jagp.2014.02.006

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks,

support vector machines, classification trees and random forests. *BMC research notes*, *4*, 299. https://doi.org/10.1186/1756-0500-4-299

Martin, S. A., Townend, F. J., Barkhof, F., & Cole, J. H. (2023). Interpretable machine learning for dementia: A systematic review. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, *19*(5), 2135–2149. https://doi.org/10.1002/alz.12948

Martins, V. F., Peyré-Tartaruga, L. A., Haas, A. N., Kanitz, A. C., Martinez, F. G., & Gonçalves, A. K. (2024). Observational evidence of the association between physical and psychological determinants of aging with cognition in older adults. *Scientific reports*, *14*(1), 12574. https://doi.org/10.1038/s41598-024-58497-7

Masse, C., Vandel, P., Sylvestre, G., Noiret, N., Bennabi, D., Mauny, F., Puyraveau, M., Barsznica, Y., Dartevelle, J., Meyer, A., Binetruy, M., Lavaux, M., Ryff, I., Giustiniani, J., Magnin, E., Galmiche, J., Haffen, E., & Chopard, G. (2021). Cognitive Impairment in Late-Life Depression: A Comparative Study of Healthy Older People, Late-Life Depression, and Mild Alzheimer's Disease Using Multivariate Base Rates of Low Scores. *Frontiers in Psychology*, *12*, 724731. https://doi.org/10.3389/fpsyg.2021.724731

Mato-Abad, V., Jiménez, I., García-Vázquez, R., Aldrey, J. M., Rivero, D., Cacabelos, P., Andrade-Garda, J., Pías-Peleteiro, J. M., & Rodríguez-Yáñez, S. (2018). Using artificial neural networks for identifying patients with mild cognitive impairment associated with depression using neuropsychological test features. *Applied Sciences*, *8*(9), 1629.

Miah, Y., Prima, C. N. E., Seema, S. J., Mahmud, M., & Shamim Kaiser, M. (2021). Performance Comparison of Machine Learning Techniques in Identifying Dementia from Open Access Clinical Datasets. In F. Saeed, T. Al-Hadhrami, F. Mohammed, & E. Mohammed (Hrsg.), *Advances on Smart and Soft Computing* (S. 79–89). Springer Singapore.

Mistridis, P., Krumm, S., Monsch, A. U., Berres, M., & Taylor, K. I. (2015). The 12 Years Preceding Mild Cognitive Impairment Due to Alzheimer's Disease: The Temporal

Emergence of Cognitive Decline. *Journal of Alzheimer's Disease: JAD*, *48*(4), 1095–1107. https://doi.org/10.3233/JAD-150137

Molnar, C. (2022). *Interpretable Machine Learning—A Guide for Making Black Box Models Explainable* (2. Aufl.). https://christophm.github.io/interpretable-ml-book

Morimoto, S. S., Kanellopoulos, D., & Alexopoulos, G. S. (2014). Cognitive Impairment in Depressed Older Adults: Implications for Prognosis and Treatment. *Psychiatric Annals*, *44*(3), 138–142. https://doi.org/10.3928/00485713-20140306-05

Moritz, S., Ferahli, S., & Naber, D. (2004). Memory and attention performance in psychiatric patients: Lack of correspondence between clinician-rated and patient-rated functioning with neuropsychological test results. *Journal of the International Neuropsychological Society: JINS*, *10*(4), 623–633. https://doi.org/10.1017/S1355617704104153

Moritz, S., Stöckert, K., Hauschildt, M., Lill, H., Jelinek, L., Beblo, T., Diedrich, S., & Arlt, S. (2017). Are we exaggerating neuropsychological impairment in depression? Reopening a closed chapter. *Expert Review of Neurotherapeutics*, *17*(8), 839–846. https://doi.org/10.1080/14737175.2017.1347040

Murman, D. L. (2015). The Impact of Age on Cognition. *Seminars in Hearing*, *36*(3), 111–121. https://doi.org/10.1055/s-0035-1555115

Murphy, E. A., Ehrhardt, B., Gregson, C. L., von Arx, O. A., Hartley, A., Whitehouse, M. R., Thomas, M. S., Stenhouse, G., Chesser, T. J. S., Budd, C. J., & Gill, H. S. (2022). Machine learning outperforms clinical experts in classification of hip fractures. *Scientific Reports*, *12*(1), 2058. https://doi.org/10.1038/s41598-022-06018-9

Musa, G., Slachevsky, A., Muñoz-Neira, C., Méndez-Orellana, C., Villagra, R., González-Billault, C., Ibáñez, A., Hornberger, M., & Lillo, P. (2020). Alzheimer's Disease or Behavioral Variant Frontotemporal Dementia? Review of Key Points Toward an Accurate Clinical and Neuropsychological Diagnosis. *Journal of Alzheimer's Disease: JAD*, *73*(3), 833–848. https://doi.org/10.3233/JAD-190924

Nichols, E., Merrick, R., Hay, S. I., Himali, D., Himali, J. J., Hunter, S., Keage, H. A. D., Latimer, C. S., Scott, M. R., Steinmetz, J. D., Walker, J. M., Wharton, S. B., Wiedner, C. D., Crane, P. K., Keene, C. D., Launer, L. J., Matthews, F. E., Schneider, J., Seshadri, S., … Vos, T. (2023). The prevalence, correlation, and co-occurrence of neuropathology in old age: Harmonisation of 12 measures across six community-based autopsy studies of dementia. *The Lancet. Healthy Longevity*, *4*(3), e115–e125. https://doi.org/10.1016/S2666-7568(23)00019-3

Nichols, E., Szoeke, C. E. I., Vollset, S. E., Abbasi, N., Abd-Allah, F., Abdela, J., Aichour, M. T. E., Akinyemi, R. O., Alahdab, F., Asgedom, S. W., Awasthi, A., Barker-Collo, S. L., Baune, B. T., Béjot, Y., Belachew, A. B., Bennett, D. A., Biadgo, B., Bijani, A., Bin Sayeed, M. S., … Murray, C. J. L. (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *18*(1), 88–106. https://doi.org/10.1016/S1474-4422(18)30403-4

Nick, T. G., & Campbell, K. M. (2007). Logistic regression. *Methods in Molecular Biology (Clifton, N.J.)*, *404*, 273–301. https://doi.org/10.1007/978-1-59745-530-5_14

Novais, F., & Starkstein, S. (2015). Phenomenology of Depression in Alzheimer's Disease. *Journal of Alzheimer's Disease: JAD*, *47*(4), 845–855. https://doi.org/10.3233/JAD-148004

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1140–1152. https://doi.org/10.1016/j.neubiorev.2012.01.004

Parkinson, W. L., Rehman, Y., Rathbone, M., & Upadhye, S. (2020). Performances on individual neurocognitive tests by people experiencing a current major depression episode: A systematic review and meta-analysis. *Journal of Affective Disorders*, *276*, 249–259. https://doi.org/10.1016/j.jad.2020.07.036

Pazzani, M. J., Mani, S., & Shankle, W. R. (2001). Acceptance of rules generated by machine learning among medical experts. *Methods of information in medicine*, *40*(5), 380–385.

Pellegrini, E., Ballerini, L., Hernandez, M. D. C. V., Chappell, F. M., González-Castro, V., Anblagan, D., Danso, S., Muñoz-Maniega, S., Job, D., Pernet, C., Mair, G., MacGillivray, T. J., Trucco, E., & Wardlaw, J. M. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, *10*, 519–535. https://doi.org/10.1016/j.dadm.2018.07.004

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, *45*(1 Suppl), S199-209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Petersen R. C. (2016). Mild Cognitive Impairment. *Continuum (Minneapolis, Minn.)*, *22*(2 Dementia), 404–418. https://doi.org/10.1212/CON.0000000000000313

Piccolino, A. L., Piccolino, A. R., & Piccolino, S. G. (2025). Distinguishing Alzheimer's disease from other dementias using pattern profile analysis in the Meyers Neuropsychological Battery: An exploratory study. *Applied neuropsychology. Adult*, *32*(4), 1087–1102. https://doi.org/10.1080/23279095.2023.2236742

Porsteinsson, A. P., Isaacson, R. S., Knox, S., Sabbagh, M. N., & Rubino, I. (2021). Diagnosis of Early Alzheimer's Disease: Clinical Practice in 2021. *The Journal of Prevention of Alzheimer's Disease*, *8*(3), 371–386. https://doi.org/10.14283/jpad.2021.23

Potter, G. G., Wagner, H. R., Burke, J. R., Plassman, B. L., Welsh-Bohmer, K. A., & Steffens, D. C. (2013). Neuropsychological predictors of dementia in late-life major depressive disorder. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry*, *21*(3), 297–306. https://doi.org/10.1016/j.jagp.2012.12.009

Rasmussen, J., & Langerman, H. (2019). Alzheimer's Disease – Why We Need Early Diagnosis. *Degenerative Neurological and Neuromuscular Disease*, *9*, 123–130. https://doi.org/10.2147/DNND.S228939

Schmand, B., Huizenga, H. M., & van Gool, W. A. (2010). Meta-analysis of CSF and MRI biomarkers for detecting preclinical Alzheimer's disease. *Psychological Medicine*, *40*(1), 135–145. https://doi.org/10.1017/S0033291709991516

Schmid, N. S., Ehrensperger, M. M., Berres, M., Beck, I. R., & Monsch, A. U. (2014). The Extension of the German CERAD Neuropsychological Assessment Battery with Tests Assessing Subcortical, Executive and Frontal Functions Improves Accuracy in Dementia Diagnosis. *Dementia and Geriatric Cognitive Disorders Extra*, *4*(2), 322–334. https://doi.org/10.1159/000357774

Schober, P., & Vetter, T. R. (2021). Logistic Regression in Medical Research. *Anesthesia and Analgesia*, *132*(2), 365–366. https://doi.org/10.1213/ANE.0000000000005247

Seelaar, H., Rohrer, J. D., Pijnenburg, Y. A. L., Fox, N. C., & van Swieten, J. C. (2011). Clinical, genetic and pathological heterogeneity of frontotemporal dementia: A review. *Journal of Neurology, Neurosurgery, and Psychiatry*, *82*(5), 476–486. https://doi.org/10.1136/jnnp.2010.212225

Silva Dos Santos Durães, R., Emy Yokomizo, J., Saffi, F., Castanho de Almeida Rocca, C., & Antonio de, P. S. (2022). Differential Diagnosis Findings Between Alzheimer's Disease and Major Depressive Disorder: A Review. *Psychiatry and Clinical Psychopharmacology*, *32*(1), 80–88. https://doi.org/10.5152/pcp.2022.21133

Storey, E., Slavin, M. J., & Kinsella, G. J. (2002). Patterns of cognitive impairment in Alzheimer's disease: Assessment and differential diagnosis. *Frontiers in Bioscience: A Journal and Virtual Library*, *7*, e155-184. https://doi.org/10.2741/A914

Suemoto, C. K., & Leite, R. E. P. (2023). Autopsy studies are key to identifying dementia cause. *The Lancet. Healthy Longevity*, *4*(3), e94–e95. https://doi.org/10.1016/S2666-7568(23)00022-3

Tetsuka, S. (2021). Depression and Dementia in Older Adults: A Neuropsychological Review. *Aging and Disease*, *12*(8), 1920–1934. https://doi.org/10.14336/AD.2021.0526

Tran, T., Milanovic, M., Holshausen, K., & Bowie, C. R. (2021). What is normal cognition in depression? Prevalence and functional correlates of normative versus idiographic cognitive impairment. *Neuropsychology*, *35*(1), 33–41. https://doi.org/10.1037/neu0000717

Weintraub, S. (2022). Neuropsychological Assessment in Dementia Diagnosis. *Continuum (Minneapolis, Minn.)*, *28*(3), 781–799. https://doi.org/10.1212/CON.0000000000001135

Weintraub, S., Wicklund, A. H., & Salmon, D. P. (2012). The neuropsychological profile of Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*, *2*(4), a006171. https://doi.org/10.1101/cshperspect.a006171

Wiels, W., Baeken, C., & Engelborghs, S. (2020). Depressive Symptoms in the Elderly-An Early Symptom of Dementia? A Systematic Review. *Frontiers in Pharmacology*, *11*, 34. https://doi.org/10.3389/fphar.2020.00034

Wunner, C., Schubert, A., Gosch, M., & Stemmler, M. (2022). Differential Diagnosis of MCI, Dementia and Depression—A Comparison of Different Cognitive Profiles. *Psych*, *4*(2), 187–199. https://doi.org/10.3390/psych4020016

Yang, Q., Li, X., Ding, X., Xu, F., & Ling, Z. (2022). Deep learning-based speech analysis for Alzheimer's disease detection: A literature review. *Alzheimer's Research & Therapy*, *14*(1), 186. https://doi.org/10.1186/s13195-022-01131-3

Ying, H., Jianping, C., Jianqing, Y., & Shanquan, Z. (2016). Cognitive variations among vascular dementia subtypes caused by small-, large-, or mixed-vessel disease. *Archives of Medical Science: AMS*, *12*(4), 747–753. https://doi.org/10.5114/aoms.2016.60962

Zihl, J., Reppermund, S., Thum, S., & Unger, K. (2010). Neuropsychological profiles in MCI and in depression: Differential cognitive dysfunction patterns or similar final common pathway disorder? *Journal of Psychiatric Research*, *44*(10), 647–654. https://doi.org/10.1016/j.jpsychires.2009.12.002