REVIEW

Machine learning and neuroimaging in neurodegenerative disease diagnosis: a systematic review of voxel-based and surface-based morphometry approaches based on PRISMA framework

Emrah Gürlek · İnci Zaim Gökbay

Received: 18 May 2025 / Accepted: 4 February 2026 / Published online: 4 March 2026

Abstract

This PRISMA-guided review systematically evaluated the application of machine learning (ML) and deep learning (DL) methodologies to structural Magnetic Resonance Imaging (MRI), specifically utilizing Voxel-Based Morphometry (VBM) and Surface-Based Morphometry (SBM) for ND classification. Our analysis encompassed twelve studies published between 2017 and 2024, focusing on key parameters such as participant demographics, utilized datasets, preprocessing steps, and applied algorithms. To assess the methodological robustness of the included studies, a 20-point Quality Assessment (QA) score was developed, covering data rigor, model development, and validation transparency. The findings indicated significant potential, with algorithms such as Support Vector Machine (SVM) and Extreme Learning Machine (ELM) demonstrating high accuracy, reaching up to 0.96, in the classification of Alzheimer's and Parkinson's disease. However, the QA analysis revealed a significant negative correlation ($r = -0.32$) between reporting transparency and performance, suggesting an 'optimistic bias' in studies with lower methodological rigor. Despite these promising results, the field faced significant challenges, primarily due to small sample sizes—for instance, a limited cohort of $n = 30$ for Lewy Body Dementia (LBD)—and a notable lack of consistency in preprocessing techniques, as well as a total absence of public source code sharing (0%), which collectively restricted the generalizability of the models. In conclusion, while the existing research yielded encouraging outcomes, the successful transition to clinical adoption necessitated further innovation in algorithmic development coupled with the incorporation of substantially larger cohorts. To effectively bridge the existing gap between research findings and practical clinical application, the review strongly advocated for the establishment of standardized processing pipelines and a greater commitment to open-data initiatives.

Keywords Neurodegenerative diseases · Surface-based morphometry (SBM) · Voxel-based morphometry (VBM) · Machine learning · Neural networks · Deep learning · Systematic literature review (SLR) · PRISMA framework

Abbreviations

AD	Alzheimer's disease
ADNI	Alzheimer's disease neuroimaging initiative
ALS	Amyotrophic lateral sclerosis
AAL	Automated anatomical labeling
AUC-ROC	Area under the receiver operating characteristic curve
bvFTD	Behavioral variant frontotemporal dementia
CAD	Computer-aided diagnostic
CAT	Computational anatomy toolbox
CNN	Convolutional neural network
CSF	Cerebrospinal fluid
CT	Computerized tomography
CV	Cross-validation
DARTEL	Diffeomorphic anatomical registration through exponentiated lie algebra
DICOM	Digital imaging and communications in medicine
DL	Deep learning
DLB	Dementia with Lewy bodies
DT	Decision tree
DTI	Diffusion tensor imaging
ELM	Extreme learning machine
FDG-PET	Fluorodeoxyglucose positron emission tomography
fMRI	Functional magnetic resonance imaging
FTD	Frontotemporal dementia
FTLD	Frontotemporal lobar degeneration
GAN	Generative adversarial network
GM	Gray matter
GLCM	Gray-level Co-occurrence matrix
HD	Huntington's disease
HC	Healthy controls
LBD	Lewy body dementia
LR	Logistic regression
MCI	Mild cognitive impairment
MIV	Mean impact value
ML	Machine learning
MNI	Montreal neurological institute
MRI	Magnetic resonance imaging
NC	Normal cognition/Normal controls
ND	Neurodegenerative diseases
NIfTI	Neuroimaging informatics technology initiative
OASIS	Open access series of imaging studies
PCA	Principal component analysis
PD	Parkinson's disease
PET	Positron emission tomography
PPMI	Parkinson's progression markers initiative
PSP	Progressive supranuclear palsy
QA	Quality assessment
qEEG	Quantitative electroencephalogram

RF	Radio frequency
ROI	Region of interest
SBM	Surface-based morphometry
SD	Standard deviation
sMCI	Stable mild cognitive impairment
sMRI	Structural magnetic resonance imaging
SPECT	Single-photon emission computerized tomography
SPM	Statistical parametric mapping
SVM	Support vector machine
TIV	Total intracranial volume
VBM	Voxel-based morphometry
WM	White matter

1 Introduction

Neurodegenerative diseases (ND) are chronic, progressive conditions marked by the gradual loss of neurons in motor, sensory, or cognitive systems. The term "neurodegeneration" refers to the deterioration of nerve cells. It describes the process of losing structure or function in tissues or organs. In the strict sense, neurodegeneration refers to any disease primarily affecting neurons [1], including the loss of neurons or their specific substructures [2]. Frontotemporal dementia (FTD), Alzheimer's disease (AD), dementia with Lewy bodies (DLB), Parkinson's disease (PD), and Huntington's disease (HD) are the prevailing neurodegenerative disorders [3].

Dementia is the fifth leading cause of death worldwide, and its prevalence is projected to triple by 2050 due to population growth and aging. Globally, the number of individuals affected by dementia has risen by 117% between 1990 and 2016, largely attributed to population aging [4]. Alzheimer's disease is the most common type of dementia, accounting for 70% of cases. The condition develops gradually in the brain, with pathological changes occurring up to 10–20 years before the onset of symptoms like memory impairment and difficulties with daily activities. Due to this extended 'silent' disease phase, AD is typically diagnosed after irreversible brain tissue loss has occurred [5]. Parkinson's disease is the second most prevalent neurodegenerative disorder after AD [6, 7]. In recent decades, the burden of neurodegenerative conditions, including PD and AD, has risen. Past studies have suggested that PD may have the fastest-growing prevalence among various neurological disorders [8, 9].

ND has multiple risk factors such as, aging, genetics, environmental exposure, and defective protein aggregation [1–3]. Determining the underlying causes and mechanisms driving neurodegeneration is an active area of ongoing scholarly inquiry [10, 11]. ND poses significant challenges and can become a substantial burden, as their etiologies remain elusive and effective curative treatments have yet to be developed. Current therapeutic approaches predominantly focus on alleviating the associated symptoms [12], as there are no known cures or treatments to halt the progression of ND [13]. Early diagnosis is a critical aspect of managing ND at present. Biomedical imaging modalities, including magnetic resonance imaging (MRI), positron emission tomography (PET), computerized tomography (CT), and single-photon emission computerized tomography (SPECT), are widely employed. Among these, MRI is one of the most commonly utilized and powerful imaging techniques [14, 15]. Structural MRI analysis techniques, such as Voxel-Based Morphometry (VBM) and Surface-Based Morphometry (SBM), are widely employed to assess brain changes. VBM specifically measures regional variations in brain volume and tissue concentration using mostly T1-weighted volumetric MRI. The images are segmented into the primary tissue components using tools like Statistical Parametric Mapping (SPM) and Computational Anatomy Toolbox (CAT). Additionally, for evaluating cortical thickness and surface area, applications like FreeSurfer are utilized alongside SPM and CAT. These methods aim to obtain cleaner data and mitigate distortions,

such as motion artifacts, using techniques like motion correction, scalp and skull stripping, normalization, and smoothing.

Advances in MRI technology, data storage, and machine learning (ML) methods have led to more studies classifying neurological conditions. These ML methods show promise for improving the diagnosis and monitoring of ND. These include the automated or semi-automated detection of disease onset, characterization of disease state, improvement of differential diagnosis, quantification of disease progression, and tracking of medication effects [16–18]. The application of ML algorithms has the potential to substantially advance these critical diagnostic and monitoring tasks.

This review aims to investigate the impact of ML methods on the early diagnosis of dementia, Alzheimer's disease, and Parkinson's disease, utilizing VBM and SBM imaging techniques.

This systematic review comprehensively evaluated the role of machine learning (ML) and deep learning (DL) methods in neurodegenerative disease classification using structural MRI. By synthesizing evidence from 12 key studies (Appx. B), we addressed critical gaps in the existing literature through:

1. A comparative analysis of methodological approaches, highlighting variations in preprocessing pipelines (Fig. 6), template usage (Fig. 4), and software tools (Fig. 5) across studies;
2. A performance benchmarking framework, quantifying the accuracy (Fig. 17 in Appendix C) and AUC-ROC (Fig. 18 in Appendix D) of ML/DL models for Alzheimer's, Parkinson's, and FTD subtypes;
3. An evaluation of clinical applicability, emphasizing challenges related to dataset heterogeneity (Fig. 4), sample size imbalances (Fig. 3), and validation protocols;
4. Evidence-based recommendations for standardizing MRI-based ML studies, including the need for:

- Multimodal data integration to overcome single-modality limitations
- Transparent reporting of preprocessing steps (Fig. 6)
- Rigorous cross-validation to address overfitting risks in small cohorts (Fig. 3).

In contrast to previous research that focused on specific conditions or techniques, our comprehensive analysis compared the application of machine learning across a range of neurodegenerative disorders, including Alzheimer's disease, Parkinson's disease, and frontotemporal dementia. By adopting this broader perspective and utilizing systematic literature review and the PRISMA framework, we were able to identify methodological inconsistencies and performance trends that can inform future investigations. This, in turn, will guide the development of more reproducible and clinically relevant models for the early detection and monitoring of these debilitating neurological conditions.

The highly constrained scope of our review–requiring the concurrent application of both VBM and SBM derived structural features alongside ML/DL classification for ND diagnosis–resulted in a focused, yet limited, final cohort ($n = 12$), underscoring the gap in comprehensive methodological reporting in this emerging field.

The rest of the paper is organized as follows: Sect. 2 presents the study background of MRI and ML methods in early diagnosis of ND, and related literature reviews. Section 3 describes our research methods, while Sect. 4 presents our results. Section 5 discusses the findings. Section 6 refers to the limitations of the study and Sect. 7 concludes the paper by presenting the implications of the findings and suggesting future research directions.

2 Background and previous review studies

2.1 Background

This section outlines the classification process of MRI images using Voxel-Based Morphometry and Surface-Based Morphometry techniques in conjunction with machine learning methods. These approaches aim to enable

the early diagnosis of neurodegenerative disorders, including Alzheimer's disease, Parkinson's disease, and frontotemporal dementia. The subsequent part summarizes the existing literature on this subject matter.

2.1.1 MRI in early diagnosis of neurodegenerative diseases (ND)

Early diagnosis of ND has greatly improved with the integration of MRI and ML methods. MRI is a non-invasive imaging technique that uses magnetic fields and radio frequency (RF) waves to visualize the internal structures of the body in detail. It is widely used to examine the structural and functional properties of brain tissue in particular. MRI images should be taken in three dimensions (3D) including axial (horizontal), sagittal (vertical) and coronal (anterior-posterior). This is necessary for accurate volume measurements. In addition, it is recommended that the gaps between the layers be less than 1.5 mm [19]. This allows the images to have higher resolution and therefore more precise analyses.

MRI imaging and MRI formats

Magnetic resonance imaging is a powerful diagnostic tool that provides high-resolution visualization of brain structure and function. During an MRI scan, the participant is placed within a strong magnetic field, causing the magnetic moments of hydrogen protons in the body's water and fat molecules to align. The scanner then emits RF pulses, which the protons absorb and subsequently emit. These emitted signals are detected and localized by the scanner, allowing it to construct detailed images depicting the density and spatial distribution of water and fat within the brain. By manipulating the parameters of the RF pulse sequences, a wide range of imaging contrasts can be obtained, rendering MRI a highly versatile imaging modality for neurodegenerative research [14, 20].

Voxel-Based Morphometry is a technique that quantifies volumetric changes in brain structure at the voxel level. Specifically, it is employed to detect alterations in the volumes of gray matter, white matter, and cerebrospinal fluid. VBM plays a crucial role in the early identification of neurodegenerative disorders, as these conditions are frequently characterized by a decrease in overall brain volume or localized volume loss [19]. The VBM analysis pipeline involves aligning all brain images to a common reference space and then performing statistical evaluations on a voxel-wise basis.

Surface-Based Morphometry is a method that analyzes morphological features on the cortical surface of the brain. This approach is particularly useful for detecting changes in cortical thickness. SBM is valuable in the early diagnosis of ND, as these disorders are often associated with a reduction in cortical thickness [21]. The SBM process involves reconstructing the brain's surface and then conducting statistical analyses based on the surface-level data.

Magnetic resonance imaging data can be stored and analyzed in various file formats, including the Digital Imaging and Communications in Medicine standard (DICOM) and the Neuroimaging Informatics Technology Initiative (NIfTI) format. DICOM is a widely used standard that typically preserves the raw image data. In contrast, the NIfTI format is particularly suited for neuroimaging studies, as it facilitates the processing and analysis of three-dimensional (3D) brain images [22–24]. The distinctive features of these [25] image formats enable different applications and computational analyses tailored to the requirements of neurodegenerative disorder research, such as quantifying brain structure changes or functional abnormalities associated with neurodegeneration.

2.1.2 Employing machine learning techniques to enhance early diagnosis of neurodegenerative disorders using MRI images based on VBM and SBM techniques

This section outlines the key steps in applying machine learning techniques to enhance early diagnosis of neurodegenerative disorders using VBM and SBM MRI images. It begins by describing the data sources, including publicly available databases like ADNI and custom datasets, which provide MRI scans and clinical information. Next, it highlights the importance of preprocessing MRI images to address distortions and standardize data through steps like noise reduction and spatial normalization. The use of brain atlases (e.g., MNI, AAL) for

volumetric analysis and feature extraction is then discussed, followed by an overview of classification methods such as SVM, Random Forest, and deep learning models. Finally, the section emphasizes the evaluation of performance using metrics like accuracy and AUC-ROC, underscoring the need for robust validation techniques to ensure reliable diagnostic outcomes.

Data sources

Publicly available and de-identified databases, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS), are extensively utilized in research. These databases house vast collections of MRI scans and associated clinical data. Additionally, researchers often construct custom datasets derived from single-center or multi-center sources. These datasets may incorporate demographic attributes like age and gender, as well as supplementary information such as cognitive and behavioral assessment scores [26, 27].

Preprocessing of MRI images

Preprocessing MRI images is essential when registering them to brain atlases, as it addresses distortions arising from physiological factors like respiration, pulse, and head motion during image acquisition. This involves applying techniques such as noise reduction, extraction of brain tissue from skull and skin, and spatial normalization to a common reference frame. These preprocessing steps serve to minimize morphological variations, enabling the subsequent application of voxel-based morphometric analysis [28].

Atlases and volumetric analysis

Atlases such as Montreal Neurological Institute (MNI152), Automated Anatomical Labeling (AAL), Harvard-Oxford, Neuromorphometrics and LBPA40 are widely used for VBM analysis. These atlases divide the brain into different numbers of regions and calculate gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) volumes for each region. For SBM analysis, quantitative values such as cortical thickness, surface area and GM/WM separation regions are calculated using Desikan-Killiany and Destrieux atlases [29–32].

Classification and prediction methods

The application of machine learning methods is a fundamental approach to diagnosing neurodegenerative disorders. This process typically involves dataset preparation, model training, prediction and evaluation, performance assessment, and cross-validation steps.

In dataset preparation, quantitative values obtained from MRI images and/or clinical information are compiled into a dataset. This dataset is usually divided into training and test subsets, with the training dataset used for model learning and the test dataset used to evaluate the model's performance.

During model training, the selected ML algorithm is trained on the training dataset. The model learns the patterns within the dataset and creates classification rules. For example, the Support Vector Machine algorithm performs classification by separating the data points with a hyperplane.

In prediction and evaluation, the trained model makes predictions on the test dataset, based on the rules it has learned. For instance, the features extracted from an MRI image are analyzed by the model to predict whether the image is indicative of a neurodegenerative disorder.

In performance assessment, the model's predictions are compared with the diagnoses made by physicians, and a comparison table is prepared. Performance metrics, such as accuracy and Area Under the Receiver Operating Characteristic Curve (AUC-ROC), are calculated from this table. Accuracy reflects the model's correct classification rate, while AUC-ROC evaluates the classification performance at different threshold values.

Finally, the cross-validation method is employed to improve the model's generalization performance. In this approach, the dataset is divided into multiple subsets, and a different subset is used as the test set in each iteration. This process evaluates the model's performance on diverse data and provides more reliable results.

The pipeline for employing machine learning techniques in neurodegenerative diseases is illustrated in Fig. 1, which outlines five key stages: (i) 3D rendering of 2D MRIs, (ii) MRI preprocessing (incorporating VBM/SBM techniques), (iii) Atlas based parcellation, ROI masks and feature extraction/selection, (iv) data preparation, (v) model training, and performance analysis.

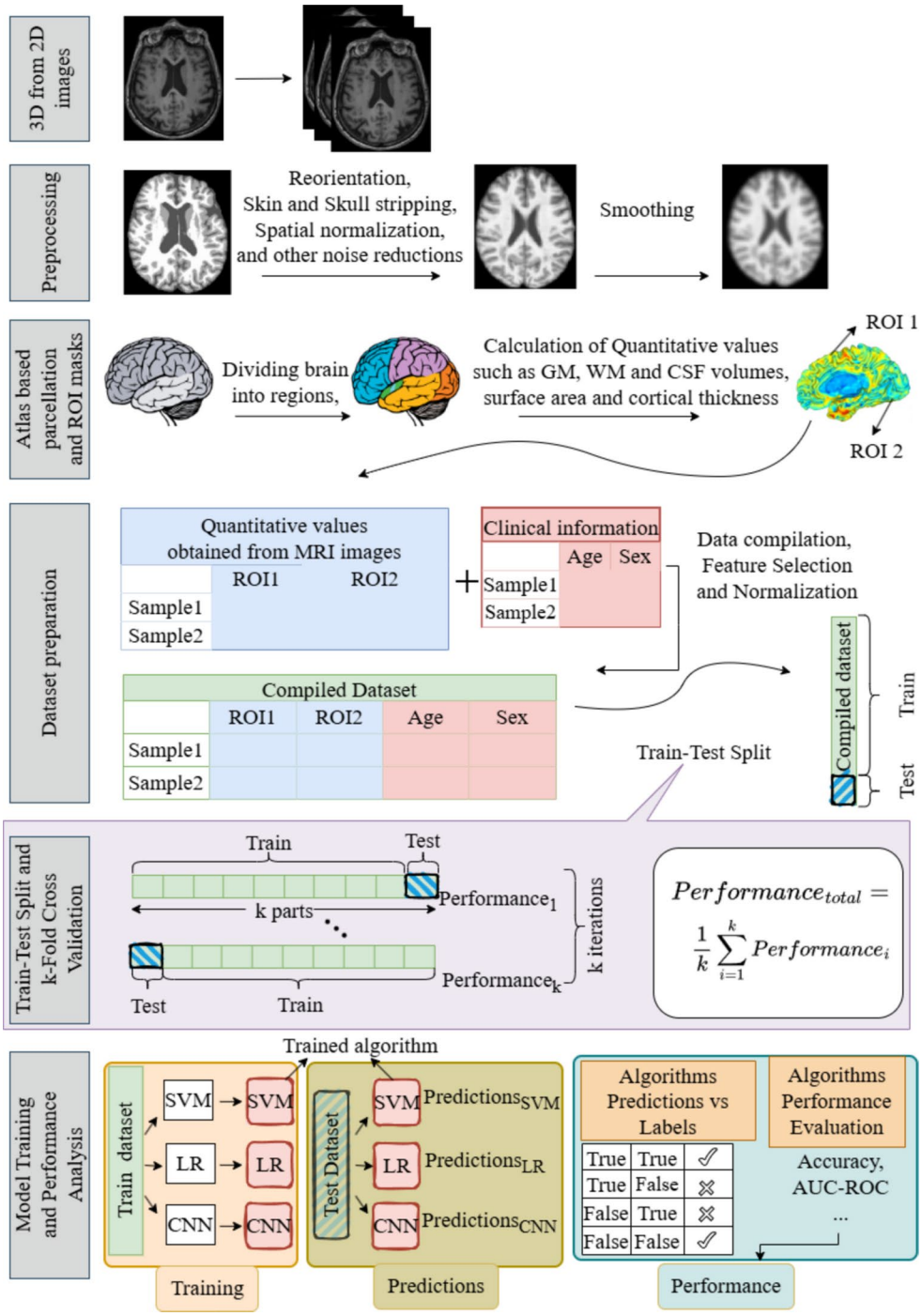


Fig. 1 Steps of machine learning techniques employment in early diagnosis of neurodegenerative disorders using MRI images based on VBM and SBM techniques (author's ownership)

2.1.3 Commonly used machine learning algorithms in early diagnosis of ND

Machine Learning algorithms used in the early diagnosis of ND vary according to the characteristics of the dataset and the analysis goals. SVM is an effective classification method, especially in high-dimensional datasets. SVM performs classification by separating data points with a hyperplane. This method is widely used in the diagnosis of ND because it performs well on both linear and non-linear datasets [33]. Random Forest is an ensemble method created by combining multiple decision trees. This method is especially effective in large and complex datasets. Random Forest can be used in both classification and regression problems and reduces the risk of overfitting [34]. Deep learning is a method that is especially effective in large datasets and complex models. Convolutional Neural Networks (CNN) are one of the deep learning methods and are widely used especially in the analysis of image data. CNNs provide automatic analysis of MRI images in the diagnosis of ND [35]. Logistic regression is a simple but effective method used especially in binary classification problems. This method is widely used especially in small datasets and feature selection processes [36].

2.1.4 Performance metrics and evaluation

Metrics such as accuracy and AUC-ROC are commonly used to evaluate the performance of classifiers. Accuracy indicates the correct classification rate of the model and is calculated by the Eq. 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where *TP*, *TN*, *FP*, and *FN* represent the true positive, true negative, false positive, and false negative values, respectively, obtained from the classification matrix. AUC-ROC evaluates the classification performance of the model at different threshold values and ranges from 0 to 1. The closer the AUC value is to 1, the better the performance of the model.

These performance evaluation methods play an important role in the early diagnosis of neurodegenerative disorders. However, larger and more diverse datasets can further improve model performance. Additionally, comparing different brain atlases and preprocessing techniques may provide more accurate and reliable results [37–40].

2.2 Previous review studies

Neuroimaging techniques are crucial for diagnosing and classifying neurodegenerative diseases. The use of MRI, PET, and other bioimaging techniques is growing in the diagnosis and prognostic evaluation of conditions like ALS, Alzheimer's, frontotemporal dementia, and progressive supranuclear palsy. Previous systematic reviews have highlighted the strengths and weaknesses of existing methodologies in this context.

According to the review on ALS, the key limitation of current neuroimaging methods is their reliance on group-based analyses, which lack sufficient sensitivity for individual diagnosis [41]. To address this, the authors recommend conducting large-scale, multicenter studies that integrate diverse imaging techniques [41]. Similarly, a review on PSP emphasizes that midbrain atrophy alone is insufficient for diagnosis and highlights the need for additional biomarkers, such as voxel-based morphometry (VBM), globus pallidus atrophy, and frontal lobe atrophy [42].

Comprehensive review studies on the classification of Alzheimer's disease emphasize the importance of MRI-based methods for early diagnosis. A comparison of techniques such as ROI-based analyses, VBM, wavelet transform, and cortical surface analyses showed that wavelet transform-based feature extraction yielded the highest accuracy [43]. Furthermore, studies predicting the transition between MCI and Alzheimer's highlight the entorhinal cortex as a key indicator and suggest that cortical thickness measurements may be more sensitive than volumetric analyses [44].

Regarding Frontotemporal Dementia (FTD), large-scale reviews have evaluated structural and functional biomarkers, noting that white matter deterioration (detected with DTI) often precedes gray matter loss, making it a critical tool for early diagnosis [45]. The integration of structural (VBM, DTI), functional (FDG-PET, fMRI), and molecular imaging is increasingly seen as vital for distinguishing between diseases such as Alzheimer's and FTD [46].

2.3 Limitations of existing literature and contributions of this review

While previous systematic reviews have highlighted the importance of advanced imaging, they often focus on specific diseases or modalities, leaving gaps in the comparative understanding of machine learning (ML) performance across the spectrum of neurodegeneration. Our review offered critical advancements over the existing literature by expanding the scope of data sources, disorders covered, and by employing a rigorous, systematic methodology based on the PRISMA framework.

2.3.1 Expanded scope and synthesis of disorders

Prior reviews have generally focused on a limited subset of conditions. For instance, studies by Garg et al. [43] and Leandrou et al. [44] primarily centered on the conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease, excluding other major neurodegenerative disorders. In contrast, our work analyzed the application of ML across a wide range of disorders, offering a unique, cross-disease synthesis.

Similarly, unlike the review by Bede and Hardiman [41], which focused solely on ALS and limited its search to PubMed, our systematic review incorporated multiple indices, significantly broadening the literature base. Furthermore, while Sakurai et al. [42] limited their PSP review to midbrain atrophy, our review sought to include studies encompassing all brain regions to provide a holistic view of structural changes.

2.3.2 Methodological insights and algorithm comparison

Unlike works by Ranganathan et al. [46] and Meeter et al. [45], which primarily described structural and functional imaging techniques, our review explicitly goes beyond description. We provided an in-depth comparison of ML algorithm performance across the studied disorders, synthesizing algorithm efficacy as a key, novel contribution.

Moreover, addressing the common limitation of "insufficient comparative analysis" found in earlier works, our study utilized the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) model. This ensures a transparent and reproducible approach to literature search and data extraction, distinguishing this work from many non-systematic reviews.

In summary, our paper advances the field by providing a PRISMA-compliant systematic review that synthesizes ML applications across the full spectrum of major neurodegenerative disorders. Moving beyond disease-specific comparisons, we critically analyze methodological variations and performance outcomes to establish a roadmap for developing robust, clinically viable diagnostic tools.

3 Review methodology

This review was conducted in accordance with the methodological guidelines and procedures outlined by [47]. The review process involved several key stages, including the development of a comprehensive review protocol, the establishment of inclusion and exclusion criteria, a keyword-driven search of relevant bibliographic databases, critical appraisal of the identified studies, data extraction, and data synthesis. This section provides a detailed account of each of these steps and the research methods employed.

3.1 Protocol development

We developed a protocol for the systematic review by following Kitchenham's guidelines and procedures [47]. This protocol outlined the research questions, search strategy, criteria for study inclusion and exclusion, data extraction approach, and methods of synthesis.

We utilized the PRISMA checklist to strengthen the review by enhancing its transparency, comprehensiveness, and rigor [48, 49]. By aligning with PRISMA guidelines, we followed best practices, mitigated bias, and enabled readers to critically evaluate the review process and findings (Fig. 2).

3.2 Data sources and search strategy

To conduct a comprehensive literature review, we searched several prominent electronic databases in February 2025, including ISI-Web of Science, IEEE Xplore, PubMed, ScienceDirect—Elsevier, and Scopus. In the first step, we formulated the research questions and identified keywords aligned with the research question and inclusion/exclusion criteria. Table 1 presents the keywords along with the rationale for their selection.

The initial search strategy required at least one keyword from each subgroup in the first group to match the title, abstract, or keywords. A summary of the databases searched and the optimized query versions used for each is provided in Table 2.

The searches yielded a total of 15 studies in PubMed, 4 in IEEE Xplore, 3 in ScienceDirect, and 5 in Web of Science. These articles were then imported into Endnote, a reference management software. Upon further examination, some records were found to be duplicates within their respective databases and were subsequently removed. Prior to this deduplication across all databases, the total number of studies had been 27, which was then reduced to 22.

3.3 Inclusion and exclusion criteria

Following the initial search, the stringent inclusion criteria, particularly the requirement for studies to report empirical findings utilizing both VBM and SBM structural features in combination with classification algorithms (ML/DL) and reporting quantitative metrics (Accuracy/AUC-ROC), severely limited the final pool. The resulting $n = 12$ studies included reflect the highly specific and demanding methodological intersection of our research question, rather than a failure to identify relevant literature.

Table 3 shows the inclusion and exclusion criteria that were adopted to define the publications that came within the scope of the review.

3.4 Citation management, retrieval, and inclusion decisions

All the relevant citations ($N = 27$) were entered into and stored using Endnote reference management software, after which duplicate citations were removed, with 22 citations remaining. Figure 2 shows the PRISMA 2020 flow chart and the number of papers identified at each stage.

The study's design and planning processes were carried out by the corresponding author. Following the research questions determined by both researchers, database searches were conducted using the relevant keywords, and the PRISMA standard was then applied in the subsequent stages. The PRISMA framework's process was implemented in two stages: Stage 1 involved identification and screening, while Stage 2 focused on finalizing the included articles and defining the final studies to be incorporated into the systematic literature review database. During the initial identification stage, the first author (EG) carefully reviewed all citations to determine their relevance to this systematic review. This process involved examining the titles, keywords, and abstracts, and excluding any studies that were not focused on the key topic of interest - the use of MRI-based machine learning techniques for the classification of neurodegenerative disorders. If the first author had any doubts about the

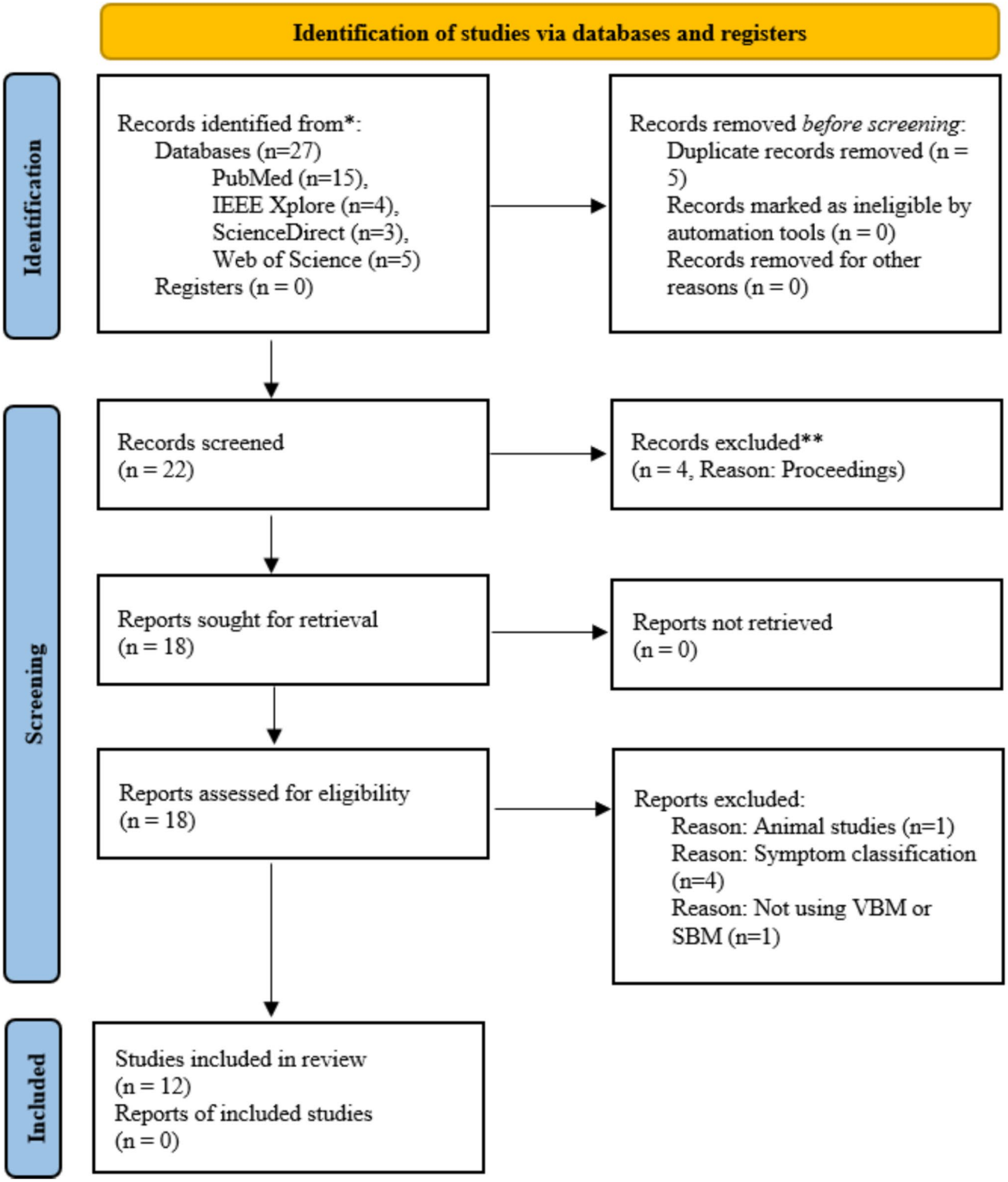


Fig. 2 The methodological guidelines and procedures outlined by Kitchenham [47], and PRISMA 2020 flow chart [48, 49]

Table 1 Keyword determination steps

Disease focused	Neurodegenerative diseases
Imaging techniques	SBM, VBM
Models used for analysis	Machine learning, deep learning, neural network
Performance measurement	Accuracy, AUC-ROC
Studies published between	Jan 2014–2025

Table 2 Database's and queries used for research

Database	Query
PubMed	(((“neurodegenerative diseases”[MeSH Terms]) AND ((accuracy) OR ("auc-roc"))) AND (((“neural network”) OR (“deep learning”)) OR (“machine learning”))) AND (("surface based morphometry") OR ("voxel based morphometry"))
IEEE	(((“All Metadata”:"voxel based morphometry") OR (“All Metadata”:"surface based morphometry")) AND ((“All Metadata”:“neural network”) OR (“All Metadata”:“machine learning”) OR (“All Metadata”:“deep learning”)) AND ((“All Metadata”:“neurodegenerative diseases”)))
ScienceDirect	Title, abstract, keywords: “neurodegenerative diseases” AND (accuracy OR "auc-roc") AND (“neural network” OR “deep learning” OR “machine learning”) AND ("surface based morphometry" OR "voxel based morphometry")
Web of science	TS = (“neurodegenerative diseases”) AND TS = ((accuracy) OR ("auc-roc")) AND TS = (("surface based morphometry") OR ("voxel based morphometry")) AND TS = (((“neural network”) OR (“deep learning”)) OR (“machine learning”))

Table 3 Inclusion/exclusion criteria

Inclusion	Exclusion
Peer-reviewed journal articles	Conference papers, abstracts, extended abstracts, books, book chapters, reviews, meta-analysis studies, editorials, and commentaries were excluded
Studies presenting empirical findings on the treatment, diagnosis, or monitoring of mental health conditions	Studies that were purely theoretical or review articles without empirical findings were also excluded
Focused on populations with neurodegenerative diseases.	Animal experimental studies
Utilizing physiological data such as SBM and/or VBM	Studies predicting disease symptoms, neurocognitive or behavioral test scores, or any clinical findings
	Studies focusing on psychiatric or neurological diseases other than neurodegenerative diseases
Employing machine learning, deep learning, artificial neural networks, or convolutional neural networks as the methodology	Studies that do not report AUC-ROC or accuracy values, even if ML/DL classification is performed
	Studies that analyze only statistical relationships or differences between imaging data and clinical scores (e.g., MMSE, MoCA) but do not present an explicit classification or prediction model
	Studies conducting quantitative analysis other than VBM or SBM (e.g. texture or shape analysis)
	Studies utilizing only non-MRI imaging techniques (e.g., PET, DTI, CT, SPECT)
	Studies based on queries other than those listed in Table 2
Articles presented in English and/or Turkish	If full-text access is unavailable, access will be provided through the databases of TUBITAK, Istanbul University, Kırşehir Ahi Evran University, University of Kent

inclusion or exclusion of a particular study, the corresponding author (IZG) was consulted, and both researchers re-checked the title, keywords, and abstract to reach a final decision. Occasionally, the study details presented in the title and abstract were not entirely clear, in which case the article was retained for further screening in the subsequent stage. Following the initial identification process, 18 studies were selected for a detailed review during the screening phase.

After completing the first stage of screening, these 18 studies then progressed to the second stage, which involved finalizing the included articles, extracting data, and synthesizing the findings of this review. Of these 18 studies, 12 were ultimately included in the final synthesis.

3.5 Data extraction

The authors implemented a systematic data extraction process to comprehensively document the details of the 12 studies and address the research objectives. The first author extracted the data using a predefined form (Table 4), and the second author cross-checked the extracted information to ensure consistency and accuracy throughout the data collection phase.

3.6 Data synthesis

To synthesize the results, we examined key aspects including the distribution of participants, the datasets used, the templates and software employed, the preprocessing steps undertaken, and the performance of the applied algorithms.

3.6.1 Quantitative synthesis and statistical analysis

To provide a robust meta-summary of the reviewed literature, a quantitative synthesis was performed for both Accuracy and AUC-ROC metrics. Given the variability in cohort sizes across the included studies, we calculated the sample-size weighted mean ($\bar{x}_w$) for each disease category using the Eq. 2:

Table 4 Question-driven data extraction template

Research question	Extracted data field
1. Study overview	
Study identifier	Study ID
Bibliographic reference	Author, Year, Journal, DOI
Is this study peer-reviewed and empirical?	Publication type (Journal/Conference) study design (PRISMA-guided SLR, experimental)
2. Participant and disease focus	
What neurodegenerative diseases (NDs) are studied?	Disease types (AD/PD/FTD/LBD/ALS) Subtypes (e.g., PD-MCI, bvFTD)
How are participant groups defined?	Inclusion/exclusion criteria (MMSE, MoCA scores) Demographic data (age, gender, education)
3. MRI data and preprocessing	
Which datasets are used?	Dataset source (ADNI/PPMI/Custom) sample size per group
What preprocessing steps are applied?	Skull stripping, normalization, smoothing software: SPM, FreeSurfer, DARTEL atlas: MNI, AAL, custom
4. Machine learning pipeline	
Which algorithms are compared?	Traditional ML: SVM, random forest, logistic regression Deep learning: CNN (3D/ResNet), ELM
How is model performance validated?	Validation method: k-fold CV, LOOCV metrics: accuracy, AUC-ROC, sensitivity
5. Results and clinical relevance	
What are the key performance outcomes?	Best Accuracy/AUC per disease Comparison across algorithms
Are findings clinically applicable?	Limitations: small samples, scanner variability generalizability assessment
6. Quality assessment	
Does the study address potential biases?	Sample size justification cross-validation rigor
Is the methodology reproducible?	Code/data availability preprocessing transparency

$$\bar{x}_w = \frac{\sum_{i=1}^{n}(w_i x_i)}{\sum_{i=1}^{n} w_i} \tag{2}$$

where x_i represents the performance metric (Accuracy or AUC-ROC) of an individual experiment and w_i represents the total sample size (N) of that specific experiment.

Furthermore, to assess the methodological consistency, we analyzed the inter-study and intra-study variability. For each study, the arithmetic mean and standard deviation (SD) of all reported algorithms were calculated. A 'Global Mean' and a 'Global Standard Deviation' (Global SD) were computed across all reviewed studies to establish a benchmark for global performance trends. This allowed for the identification of studies that significantly deviated from the expected performance range.

3.7 Quality assessment and reporting transparency scoring

To provide the supporting numeric data requested by the reviewer and to rigorously assess the methodological robustness of the included studies ($n = 12$), we developed a 20-point Quality Assessment Score. This score is derived from established guidelines in systematic reviews (PRISMA-S) and specific best practices for neuroimaging Machine Learning (ML/DL) research.

The 20 criteria are grouped into three main domains:

- I. Data and Preprocessing Rigor (Max: 9 points),
- II. Model Development and Analysis Strength (Max: 6 points),
- III. Validation and Reproducibility Transparency (Max: 5 points).

Each criterion is scored as binary (1 for Yes/Reported/Applied, 0 otherwise). Partially satisfied (e.g., lack of clarity in *recon-all* parameters or controlling for only a subset of TIV/Age/Sex in Q9 scored as 0.5. A detailed breakdown of these scoring criteria and their rationale is presented in the Table 5.

4 Results

4.1 Overview of the studies

In all studies, inclusion and exclusion criteria are uniquely defined based on specific factors such as neurological, neuropsychiatric, or neurocognitive test scores, medical history, age or image quality etc. These criteria determine whether participants are included or excluded from the study. For instance, in the study by Nemoto et al. [50], 56 participants with Lewy body dementia (LBD) and Alzheimer's disease (AD) were evaluated, with a subset of 26 participants (13 LBD, 13 AD) used for validation (Table 6). Normal controls (NC) were included solely as reference values for computational purposes. Rajagopalan and Pioro [51] utilized T2-weighted images for diagnostic purposes, while T1-weighted images were employed for voxel-based morphometry (VBM), consistent with most other studies. However, Zhang et al. [40] did not specify this distinction. Additionally, MRI data in Nemoto et al. [50] were acquired using 11 different scanners. Cognitive tests were administered to control groups in many studies, with healthy individuals classified as having normal cognition (NC) or healthy controls (HC), depending on the study's terminology.

Figure 3, a stacked bar chart showing the distribution of participants across studies, reveals that studies have largely focused on Alzheimer's (AD) and Parkinson's (PD) diseases, with other neurodegenerative disorders (ALS, LBD) being underrepresented. The study by Krüger et al. [52] included the largest sample size with 9063 participants (most of them are undiagnosed ("unknown") cases), while the control groups (HC) were balanced with the patient groups in most studies. The chart highlights the prominence of early-stage disease subtypes, such as PD-MCI and AD.

Table 5 Scoring criteria and their rationale of quality assessment

No.	Group	Criteria	Rationale
1.	I. Data and preprocessing rigor (max: 9 points)	Sample size adequate ($N > 100$ total participants)?	Ensures statistical validity and helps control the risk of overfitting in small samples
2.		DICOM/NIFTI format clearly reported?	Transparency regarding the original data format
3.		Tools/software (SPM, FreeSurfer, FSL, etc.) specified?	Essential for technical reproducibility and controlling for potential software bias during preprocessing
4.		Skull stripping/brain extraction applied?	Fundamental for accurate VBM/SBM analyses and preventing the influence of non-brain tissues
5.		Motion correction applied during preprocessing?	Crucial for minimizing artifacts in vulnerable cohorts (e.g., PD, AD)
6.		Smoothing kernel (FWHM) size explicitly reported?	Key for methodological transparency and reproducibility
7.		Normalization method clearly described?	Transparency regarding the algorithms used to align brains into a common space
8.		Template used specified?	Required for defining the target space (MNI, Talairach, etc.)
9.		Statistical control for TIV/Age/Sex performed?	Critical for isolating disease-specific change from individual and demographic variation
10.	II. Model development and analysis strength (max: 6 points)	Segmentation method clearly described?	Transparency of the algorithm used to extract grey/white matter features
11.		Multi-modality: Additional imaging data (PET/fMRI/DTI) used besides structural MRI?	Rewards the use of complementary biological information for enhanced diagnostic power
12.		Dual modality: Features from both VBM and SBM integrated?	Aligns with the review's core objective, rewarding the integration of two structural methods
13.		Multiple classification algorithms tested?	Important for strengthening model selection and showing algorithmic robustness
14.		Algorithm hyperparameters (e.g., Kernel, C, etc.) fully reported?	Essential for the reproducibility of traditional ML models
15.		DL architecture clarity: If DL is used, is the architecture (layers, activation, optimizer) fully reported?	Critical for the reproducibility of complex Deep Learning models
16.	III. Validation and reproducibility transparency (max: 5 points)	Cross-Validation (CV) method clearly specified?	Fundamental statistical step to prove the model is not suffering from overfitting
17.		External/Independent validation cohort used?	The strongest evidence for model generalizability in clinical settings
18.		Code sharing (GitHub, etc.) made publicly available?	The highest measure of scientific transparency for method verification
19.		Data sharing (Zenodo, etc.) made publicly available?	Highest measure of transparency for third-party result verification
20.		Risk of bias assessment clearly discussed?	Assesses the study's critical thinking regarding its own methodological limitations

Figure 4a shows that the majority of studies, 58.3%, utilized custom datasets, while the Alzheimer's Disease Neuroimaging Initiative and Parkinson's Progression Markers Initiative datasets were used in only 16.7% and 8.3% of studies respectively. Additionally, two studies combined custom and ADNI data. Figure 4b demonstrates a steady increase in publications from 2017 to 2024, with peak productivity observed in 2020–2021 and 2024, reflecting the growing research interest in applying machine learning to neurodegenerative diseases. Figure 4c indicates the dominance of the MNI template in neuroimaging studies, with limited adoption of alternative templates. Figure 4d highlights a research focus disparity, with Alzheimer's disease dominating over rarer conditions like amyotrophic lateral sclerosis and Lewy body dementia.

Table 6 Summary of included studies: participant demographics, dataset sources, and clinical findings

Title	Year	Core collection*	Dataset	Ethical approval	Participant demographics	Focused ND	Clinical findings
Zhang et al. [53]	2020	Clinical neurology |Neurosciences	Custom dataset PD-NC = 35 PD-MCI = 58HC = 20	Yes	Age,gender, education	PD	MMSE, MoCA, SDMT, Digit Span Test, Verbal Fluency Test, Clock Drawing Test, Visual reproduction and Logical memory, Similarities Test, Category Fluency Test, Block Design Test, UPDRS III, H-Y, HAMA, HAMD, age of onset, duration, medications, smoking and drinking history, past medical history (hypertension, diabetes mellitus, tumor and surgery), and family history
Zhang et al. [54]	2021	Radiology, Nuclear Medicine and Medical Imaging	Custom dataset PD-MCI = 36 PD-NC = 35	Yes	Age, gender	PD	H-Y, UPDRS III, LED, MoCA, and MMSE, disease duration
Zhang et al. [40]	2019	Neurosciences	ADNI AD=58 HC=94	No need	Age, gender, education	AD	MMSE, CDR-SB, ADAS, and CMgl, history of mental illness, history of neurological disease, head, eyes, ears, nose and throat disease, cardiovascular, respiratory, hepatic, dermatologic–connective tissue, musculoskeletal, endocrine–metabolic, gastrointestinal, hematopoietic–lymphatic, renal–genitourinary, allergies or drug sensitivities, alcohol abuse, drug abuse, smoking, history of malignancy, major surgical procedures, nausea, vomiting, diarrhea, constipation, abdominal discomfort, sweating, dizziness, low energy, bcdrowsy, blurred vision, headache, dry mouth, shortness of breath, coughing, palpitations, chest pain, urinary discomfort, urinary frequency, ankle swelling, musculoskeletal pain, rash, insomnia, depressed mood, crying, elevated mood, wandering, fall
Solana-Lavalle and Rosas-Romero [55]	2021	Medical Informatics |Computer Science, Theory and Methods |Computer Science, Interdisciplinary Applications |Engineering, Biomedical	PPMI PD male =231(78 1.5-T, 153 3-T) HC male =95(23 1.5-T, 72 3-T) PD female=117(36 1.5-T, 81 3-T) HC female= 64(26 1.5-T, 38 3-T)	No need	Gender	PD	
Rajagopalan and Pioro [51]	2024	Clinical Neurology |Neurosciences	Custom dataset HC= 14 ALS-CST+ = 21 ALS-CST- = 27	Yes	Age, gender	ALS	ALSFRS-R, symptom duration, DPR, EES

Table 6 (continued)

Title	Year	Core collection*	Dataset	Ethical approval	Participant demographics	Focused ND	Clinical findings
Nemoto et al. [50]	2021	Neuroimaging |Radiology, Nuclear Medicine and Medical Imaging |Clinical Neurology	Custom dataset LBD=101 AD=69 HC= 38	Yes	Age, gender	AD,LBD	MMSE
Meyer et al. [56]	2017	Neuroimaging	Custom dataset bvFTD= 52 HC= 52	Yes	Age, gender	FTD	neurological examination, neuropsychological testing, case history
Krüger et al. [52]	2024	Radiology, Nuclear Medicine and Medical Imaging	Custom Dataset Training Dataset: 8945 Diagnoses unknown Test Dataset: MCI=22 AD=18 PCA=11 bvFTD=20 svPPA=10 HC=37	Yes	Age	AD, FTD	
Ikemitsu et al. [57]	2024	Radiology, Nuclear Medicine and Medical Imaging	ADNI HC=200 early-stage AD = 200	No need	Age	AD	
Huang et al. [58]	2023	Neurosciences	Custom dataset MCI=17 HC=13 ADNI MCI=94 HC=60	Yes	Age, gender	AD	MMSE, MoCA
Egger et al. [59]	2020	Neurosciences	Custom dataset (including ADNI) AD=30 FTD=30 LBD=30 HC=30	Yes	Age, gender, education	AD, FTD, LBD	MMSE (for 6 patients MoCA was evaluated and converted to MMSE), Beck's Depression Inventory II
Bachli et al. [60]	2020	Neurosciences |Radiology, Nuclear Medicine and Medical Imaging |Neuroimaging	Custom dataset bvFTD=57 AD=29 HC=116	Yes	Age, gender, education	AD, FTD	ACE and IFS, volume of atrophy

PD: Parkinson's Disease, PD-NC: Parkinson's Disease Normal Cognitive, MCI: Mild Cognitive Impairment, HC: Healthy Control, AD: Alzheimer's Disease, ALS: amyotrophic lateral sclerosis, ALS-CST+: UMN(upper motor neuron)-predominant ALS patients with corticospinal tract hyperintensity, ALS-CST+: UMN (upper motor neuron)-predominant ALS patients without corticospinal tract hyperintensity, LBD: Lewy Bodies Dementia, bvFTD: Behavioral variant frontotemporal dementia, PCA: Posterior cortical atrophy, FTD: Frontotemporal dementia, FTLD: Frontotemporal lobar degeneration, svPPA: semantic variant primary progressive aphasia

ADNI: the Alzheimer's Disease Neuroimaging Initiative, PPMI: the Parkinson's Progression Markers Initiative

MMSE: Mini-Mental State Examination, MoCA: Montreal Cognitive Assessment, SDMT: Symbol Digit Modalities Test, UPDRS III: Unified Parkinson's Disease Rating Scale part III, H-Y: Hoehn-Yahr scale, HAMA: Hamilton Anxiety Rating Scale, HAMD: Hamilton Depression Rating Scale, LED: levodopa equivalent dose, ADAS: the Alzheimer's Disease Assessment Scale, CDR-SB: Clinical Dementia Rating Sum of Boxes, CMgl: cerebral glucose metabolism, ALSFRS-R: the Revised ALS Functional Rating Scale, DPR: disease progression rate, EES: El Escorial score, ACE: the Addenbrooke's cognitive examination, IFS: the INECO Frontal Screening

*https://mjl.clarivate.com/

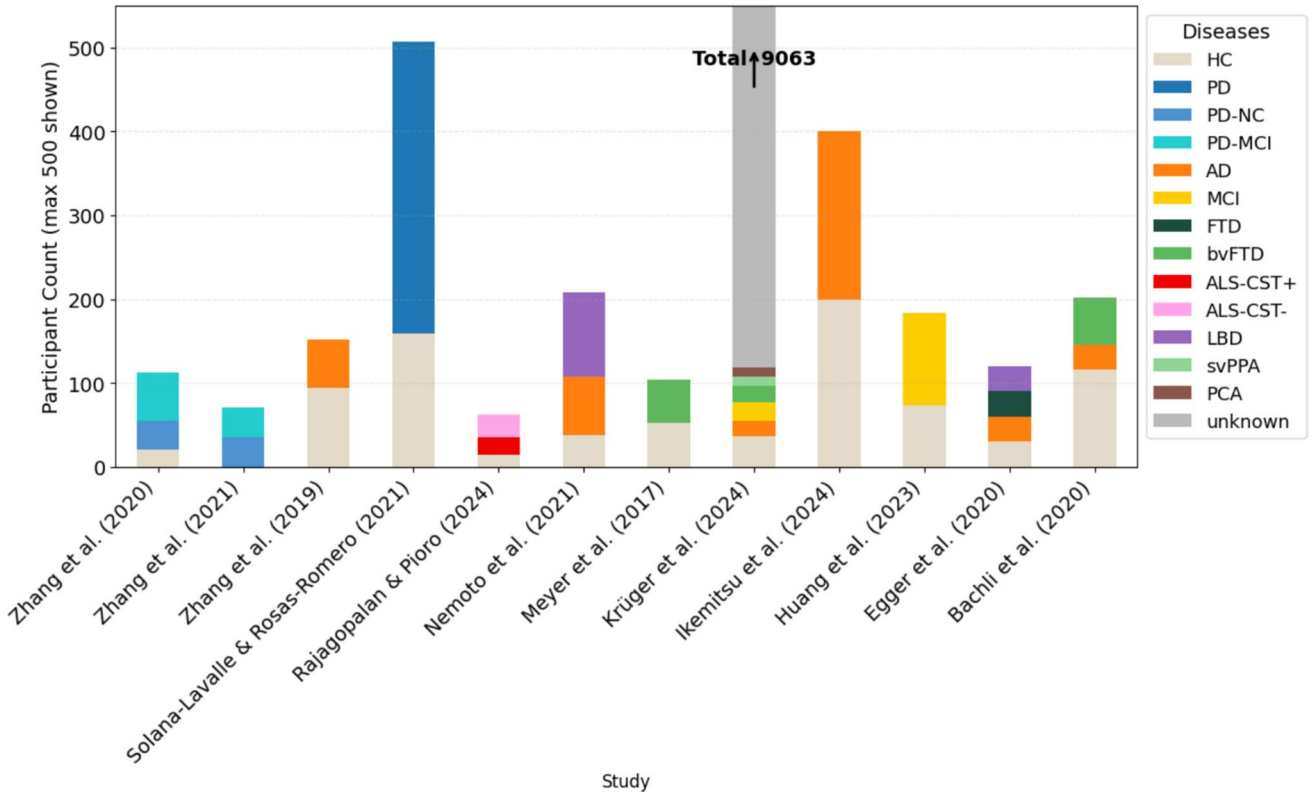


Fig. 3 Participant distribution across studies

4.2 Methodological quality assessment results

The methodological quality of the included studies was evaluated using a custom 20-point scale divided into three domains: Data and Preprocessing Rigor (9 points), Model Development and Analysis Strength (6 points), and Validation and Reproducibility Transparency (5 points). The mean overall quality score was 11.46 ($SD = 1.98$), ranging from a minimum of 7.0 to a maximum of 14.0 (Table 7).

4.2.1 Data and preprocessing rigor (Group I)

Studies achieved a mean compliance of 63.9% in this domain. While most studies excelled in reporting the normalization template (91.7%) and describing the normalization method (87.5%), critical preprocessing steps were frequently underreported. Notably, only 29.2% of studies explicitly mentioned applying motion correction, and only 31.8% clearly described the skull-stripping/brain extraction process. Although 83.3% of the studies utilized an adequate sample size ($N > 100$), the raw data format (DICOM/NIFTI) was specified in only 25.0% of the papers.

4.2.2 Model development and analysis strength (Group II)

The mean compliance for model development was 57.6%. The majority of studies (95.8%) clearly described their segmentation methods, and 75.0% incorporated multi-modal data (e.g., fMRI, DTI, or clinical scores) to enhance diagnostic power. However, reproducibility in machine learning workflows remains a challenge; only 58.3% fully reported algorithm hyperparameters, and among studies using Deep Learning, only 33.3% provided sufficient clarity regarding the architecture (layers, activation functions).

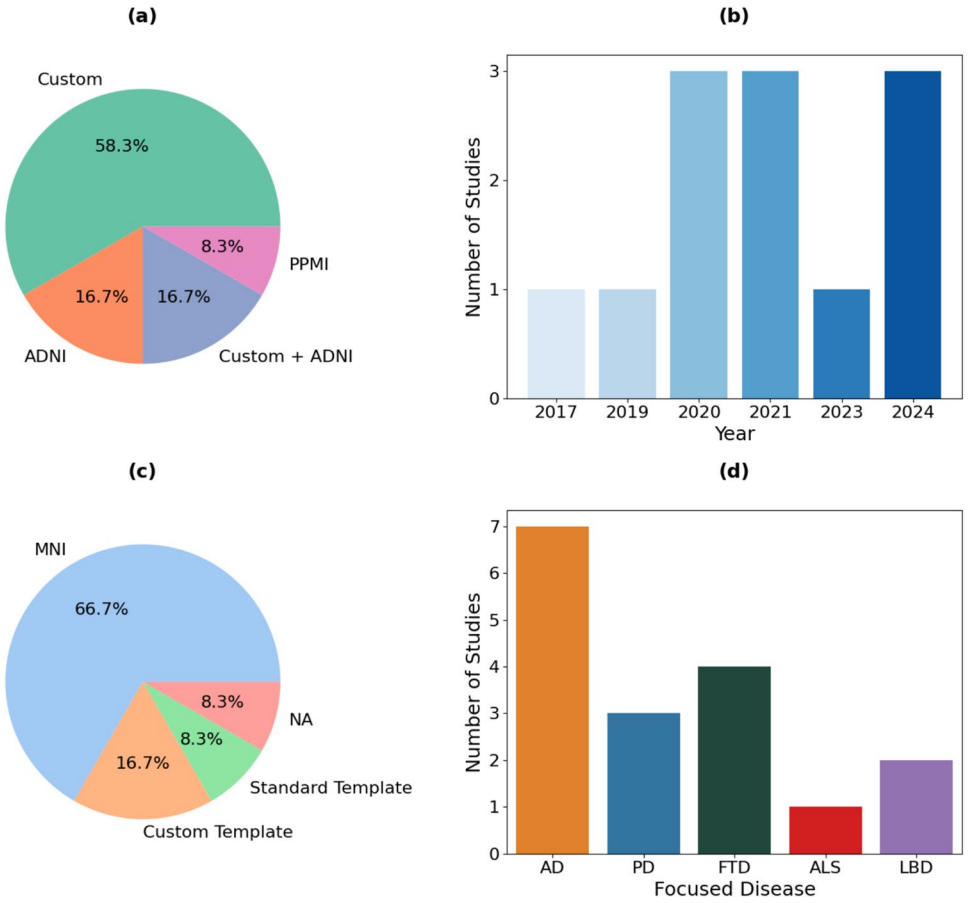


Fig. 4 **a** Distribution of dataset used in studies. **b** Number of studies published each year. **c** Template usage rates. **d** Distribution of neurodegenerative diseases in studies

4.2.3 Validation and reproducibility transparency (Group III)

This domain showed the lowest adherence, with a mean compliance of 45.0%. While 83.3% of studies clearly specified their cross-validation method (e.g., k-fold), external validity and open science practices were critically lacking. Only 25.0% of studies tested their models on an independent external validation cohort. Most alarmingly, 0% of the reviewed studies made their source code publicly available, and only 33.3% provided access to their datasets. This lack of transparency significantly hinders the reproducibility of the reported findings.

4.2.4 Summary of quality assessment

Overall, while studies demonstrated reasonable competence in describing core image processing steps (normalization, segmentation), there is a systematic failure in reporting reproducibility metrics (code sharing, hyperparameter details) and robustness checks (motion correction, external validation). These findings underscore the urgent need for the standardized reporting guidelines proposed in the following section.

4.3 Preprocessing and image registration methods

During preprocessing, several studies employed "recon-all instructions" [53, 54], though the specific details of their implementation were not always clear. In cases where details were provided, they were noted in the Table 8; otherwise, they were marked as "Maybe." [53] also included magnetic field homogeneity correction, automatic topological correction, and bilateral cerebral cortex reconstruction during preprocessing. For image standardization, MNI coordinates were derived from Talairach coordinates [53, 54] or using the DARTEL algorithm [50, 58,

https://doi.org/10.1007/s00521-026-12022-1

Table 7 Quality assessment scores

Study	I. Data and preprocessing									II. Model Dev.						III. Validation					S1	S2	S3	Total
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20				
Zhang et al. [53]	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	9	4	1	14
Zhang et al. [54]	0	1	1	0.5	0.5	0.5	0.5	1	1	0.5	1	1	0	1	0	1	0	0	0	1	6	3.5	2	11.5
Zhang et al. [40]	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	3	3	7
Solana-Lavalle and Rosas-Romero [55]	1	0	0	0	0	1	1	1	0.5	1	0	1	1	1	0	1	0	0	1	1	4.5	4	3	11.5
Rajagopalan and Pioro [51]	0	0	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	0	1	6	4	2	12
Nemoto et al. [50]	1	0	1	0	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	6	5	2	13
Meyer et al. [56]	1	0	1	0	0	1	1	1	1	1	1	0	0	0	0	1	1	0	1	1	6	2	4	12
Krüger et al. [52]	1	0	1	0	0	1	1	1	0.5	1	0	0	0	1	1	0	1	0	0	1	5.5	3	2	10.5
Ikemitsu et al. [57]	1	0	1	0	0	0	1	1	0	1	0	1	1	1	1	1	0	0	1	1	4	5	3	12
Huang et al. [58]	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	0	0	1	8	4	2	14
Egger et al. [59]	1	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	7	2	0	9
Bachli et al. [60]	1	0	1	0	0	1	1	1	1	1	1	0	0	0	0	1	1	0	0	1	6	2	3	11

Note: Scoring: 1: Fully satisfied; 0.5: Partially satisfied (e.g., lack of clarity in *recon-all* parameters or controlling for only a subset of TIV/Age/Sex in Q9); 0: Not satisfied/reported

S1, S2, S3: Subtotal scores for Data, Model, and Validation sections, respectively

Table 8 Methodological overview of studies: software tools and templates

Title	Template	Tools
Zhang et al. [53]	MNI	MRIcron, FreeSurfer
Zhang et al. [54]	MNI	MRIcron, FreeSurfer
Zhang et al. [40]	NA	NA
Solana-Lavalle and Rosas-Romero [55]	Custom template	NA
Rajagopalan and Pioro [51]	MNI	BET, FAST, FLIRT, FNIRT, FreeSurfer, ImageJ
Nemoto et al. [50]	MNI	SPM12, DARTEL
Meyer et al. [56]	MNI	SPM8
Krüger et al. [52]	Custom template	SPM12, DARTEL
Ikemitsu et al. [57]	A standard template	FreeSurfer
Huang et al. [58]	MNI	SPM8, DARTEL
Egger et al. [59]	MNI	SPM12, DARTEL
Bachli et al. [60]	MNI	SPM12, DARTEL

NA: Not available

MNI: Montreal Neurologic Institute

BET: Brain Extraction Tool, FMRIB: Functional Magnetic Resonance Imaging of the Brain, FAST: FMRIB Automated Segmentation Tool, FLIRT: FMRIB Linear Image Registration Tool, FNIRT: FMRIB Non Linear Registration Tool, DARTEL: Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra, SPM: Statistical Parametric Mapping

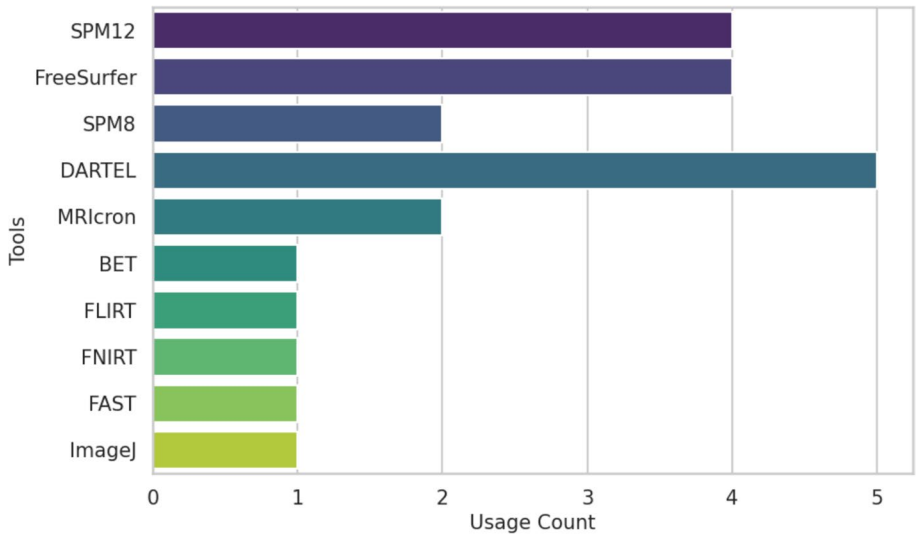


Fig. 5 Software used

60] or non-linear transformations [56, 59]. Alignment and registration of images to a standard anatomical axis are essential for comparative analysis. Solana-Lavalle and Rosas-Romero [55] detailed the standard affine registration process, emphasizing optimization using Gradient Descent or Levenberg-Marquardt methods. However, Zhang et al. [40] did not explicitly mention alignment or registration. Rajagopalan and Pioro [51] used the Jacobian of the warp field for VBM analysis and spherical atlas registration for cortical thickness analysis. Similarly, Ikemitsu et al. [57] registered images to a standard brain spherical atlas.

Krüger et al. [52] investigated potential differences between MRI scanners by analyzing individual scans. They created a normal database using 8945 MRI scans from 65 different scanners, with 25–150 healthy individuals per scanner (totaling 2150 scans). A test set included 118 individuals, with 26 NC participants overlapping between the normal database and the test set to maintain balance. The multiple-scanner normal database comprised 136 MRI scans. Scanner-specific VBM and multiple-scanner VBM were performed using the normal database as a reference, while convolutional neural networks (CNNs) were used without the database for evaluation. In one study, regions of interest (ROIs) were manually segmented [40]. Two experts used a region-growing method starting from the hippocampus, and contourlet transform was applied for feature extraction. Figure 5 demonstrates

the SPM ecosystem's prevalence (6 uses) compared to FreeSurfer's 4 uses, reflecting SPM's entrenched role in VBM/SBM pipelines.

Figure 6 demonstrates that post-conversion steps (normalization, segmentation, smoothing) were near-universal (> 9 studies), while initial DICOM handling and motion correction appeared in only 3–4 studies. Decimal counts reflect partial reporting (e.g., "maybe" in methods). Total steps exceed study count as pipelines combine multiple steps.

4.4 Impact of methodological quality on reported performance

To assess whether methodological rigor influences the reported diagnostic performance, we conducted a comprehensive correlation analysis between the Quality Assessment (QA) scores and the performance metrics (Accuracy and AUC-ROC). The results, visualized in Fig. 7 (Accuracy) and Fig. 8 (AUC-ROC), reveal a distinct "optimistic bias" associated with lower methodological transparency.

Correlation with QA sub-domains as shown in the top panels of Fig. 7, the Total QA Score demonstrated a negative correlation with accuracy ($r = -0.32$), indicating that studies with lower overall quality tend to report higher classification success. This inverse relationship was driven primarily by the Model Development (Group II) domain, which showed a strong negative correlation ($r = -0.51$, $p < 0.05$). This suggests that studies failing to transparently report model parameters (e.g., hyperparameters, architecture details) are statistically more likely to claim superior performance, potentially due to overfitting that goes unreported.

Interestingly, the Data and Preprocessing (Group I) scores showed a weaker correlation ($r = -0.09$), implying that while preprocessing is critical for image quality, the reporting of these steps (e.g., normalization templates) is less predictive of the final accuracy than the transparency of the machine learning model itself.

Impact of Validation and Robustness Checks The bottom panels of Figs. 7 and 8 highlight the critical role of validation strategies.

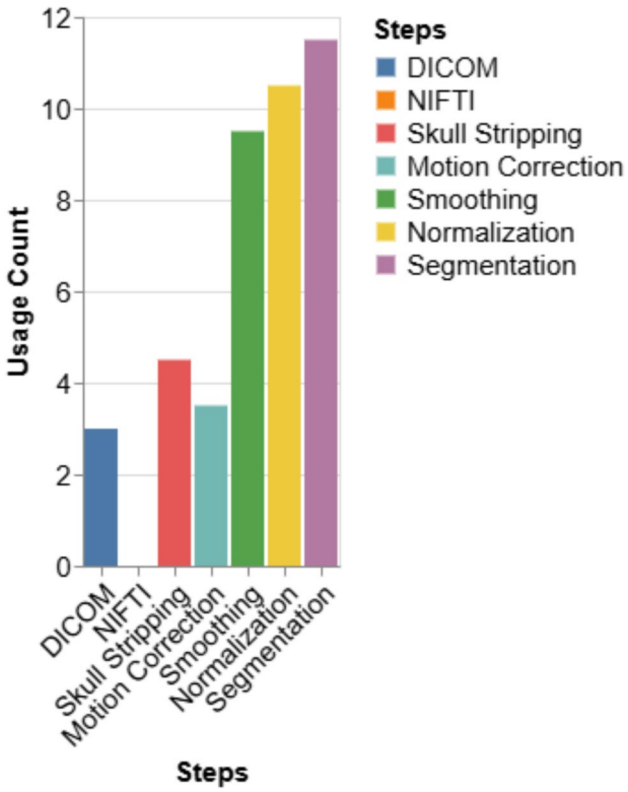


Fig. 6 Preprocessing steps

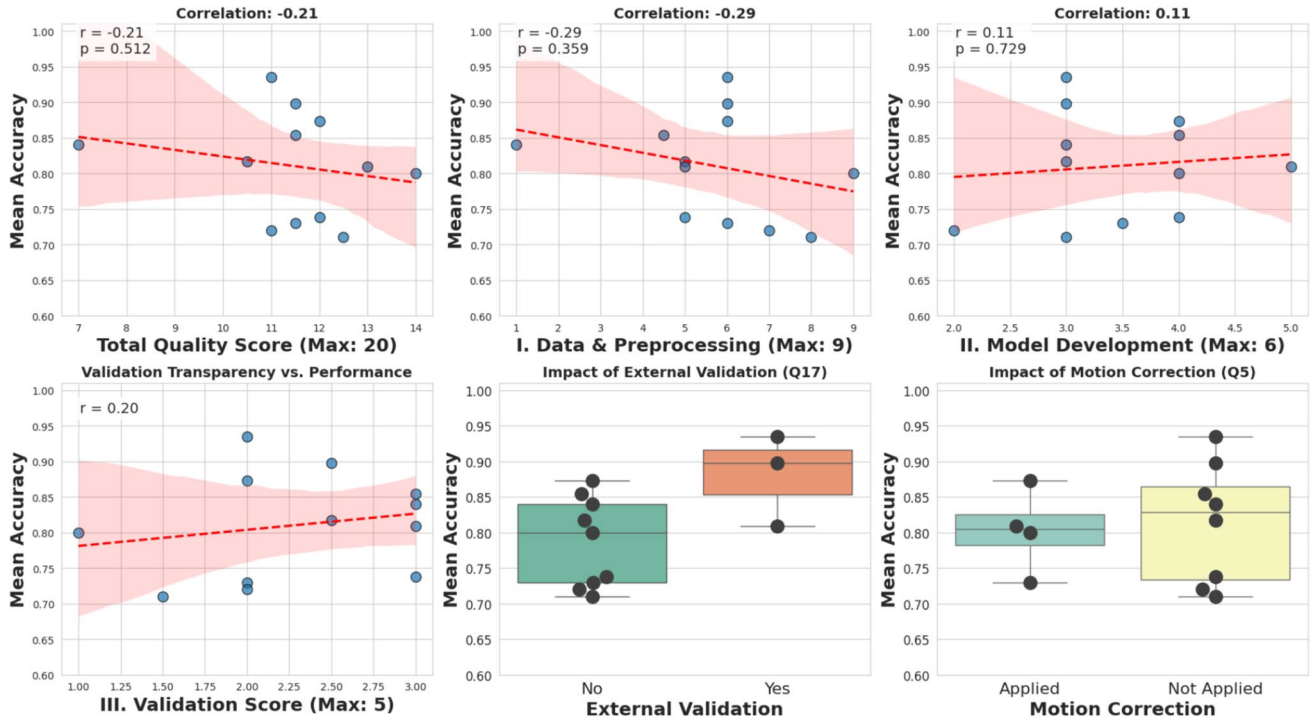


Fig. 7 Correlation analysis: methodological quality versus accuracy

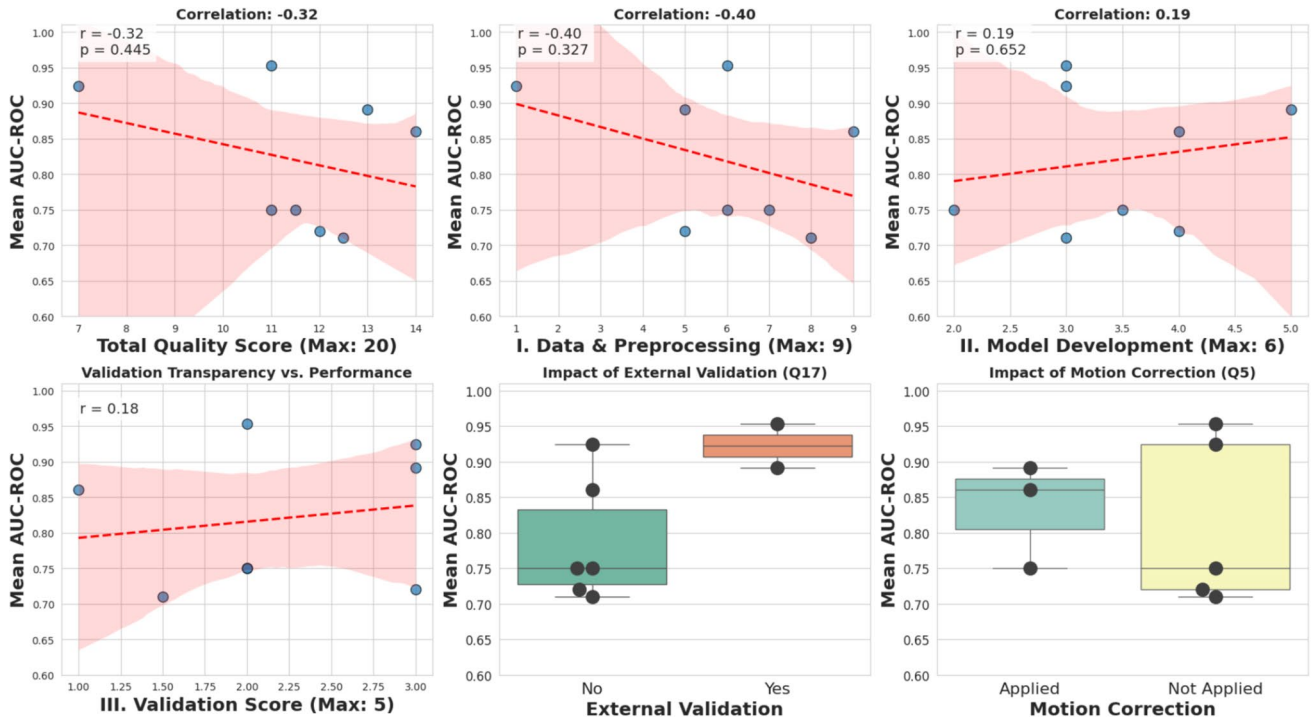


Fig. 8 Correlation analysis: methodological quality versus AUC-ROC

- External validation (Q17): Studies that performed external validation on independent cohorts achieved robust but more conservative mean accuracy (85.8%) compared to non-validated studies. While the sample size for validated studies is smaller, the tighter distribution of scores suggests higher reliability.
- Motion correction (Q5): A counter-intuitive trend was observed where studies not applying motion correction reported slightly higher accuracy averages. This raises a significant concern: in neurodegenerative cohorts prone to head movement (e.g., Parkinson's), algorithms might be learning to classify "motion artifacts" rather than biological pathology, artificially inflating accuracy in the absence of correction.

Conclusion on Quality vs. Performance Collectively, these analyses demonstrate that high reported accuracy does not necessarily equate to high scientific validity. The negative correlation between transparency (QA scores) and performance serves as a strong empirical justification for the standardized reporting guidelines proposed in Table 14 Future studies must prioritize the transparent reporting of model development steps (Group II) and mandatory external validation to bridge the gap between "high accuracy" on paper and "clinical utility" in practice.

4.5 Feature selection and validation techniques

During feature selection, Zhang et al. [53] conducted covariance analysis ($q < 0.05$) to reduce the effects of age and intracranial volume. Another study calculated the sequence of sample features using mean impact value (MIV) [54]. Clinical characteristics were evaluated using the chi-square test, with significant features included in the analysis [40]. Solana-Lavalle and Rosas-Romero [55] extracted second-order statistics using gray-level co-occurrence matrix (GLCM) from VBM-derived ROIs, followed by principal component analysis (PCA) and wrapper feature subset selection.

For validation, cross-validation and grid search methods were employed [53, 54] (Table 9). In the study conducted by Meyer et al. [56], both leave-one-center-out and leave-one-subject-out methods were used. The study performed two separate analyses: one involving exactly center-matched 19 bvFTD patients versus 19 NC participants, and another including all participants in the dataset. Regional gray matter (GM) density was preferred over GM volume due to its localized focus.

Krüger et al. [52] was the only study to perform data augmentation on images. Egger et al. [59] used 360 healthy control images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) for normalization but excluded them from classification. Regression was applied to account for age, sex, and total intracranial volume effects.

Bachli et al. [60] collected data from three countries, using Country-1 data for within-country analysis and training models to test data from the other two countries. Features included IFS and ACE scores, atrophy volume, sex, age, and education duration. Leave-two-out cross-validation was used for within-country classification, while z-score standardization was applied for cross-country analysis.

4.6 Classification methods and performance metrics

In Zhang et al. [54], quantitative electroencephalogram (qEEG) data were also combined with MRI images, achieving accuracies and AUC-ROC values of 0.80 and 0.86 for the training set and 0.80 and 0.77 for the test set, respectively. Rajagopalan and Pioro [51] applied gray matter fractal dimension analysis, with Naïve Bayes outperforming Random Forest and SVM in classification. Feature selection using WEKA resulted in lower accuracy.

Nemoto et al. [50] utilized a ResNet-based CNN architecture with skip connections, pretrained on the Kinetics human action video dataset. Conventional VBM-based differentiation included age, sex, and intracranial volume, with classification based on regional z-scores. The Youden Index determined the threshold for accuracy calculation. Similarly, Egger et al. [59] conducted classification using ROIs and regional z-scores, and also performed human rating-based analysis. They achieved the highest discriminative power in the medial temporal region (AUC 0.79) when comparing all patients with healthy controls.

Table 9 Performance comparison of studies: accuracy, AUC-ROC, validation methods, and atlas used across disease categories

Title	Measures included	Atlas	Validation	Algorithms	Classification	Acc.	AUC-ROC	L1OCV
Zhang et al. [53]	GM area, GM volume, cortical thickness, GM mean curvature	MNI	5-fold CV, L1OCV	SVM	PD-MCI versus PD-NC	Train: 0.7671	0.86	0.7808
						Test: 0.8000		
Zhang et al. [54]	Average thickness, surface area, GM volume, mean curvature	MNI	5-fold CV, L1OCV	SVM	PD-MCI versus PD-NC	Train: 0.79	0.86	0.68
						Test: 0.73	0.75	
Zhang et al. [40]	Volume data and texture parameters of hippocampus (although segmentation was done there is no data of GM, WM or CSF values	Region-Growing Based Manual Segmentation Method	10-fold CV	ELM (with VBM and clinical information	AD versus NC	0.92	0.925	
				ELM (with only VBM information)		0.92	0.922	
				SVM		0.75		
				GPR		0.79		
				PLS		0.82		
Solana-Lavalle and Rosas-Romero [55]	GM volume	Custom template	10-fold CV	LR	PD versus HC 1.5-T male	0.9604	NA	
					PD versus HC 1.5-T female	0.9677		
					PD versus HC 3-T male	0.9156		
					PD versus HC 3-T female	0.9328		
				Random forest	PD versus HC 1.5-T male	0.9208		
					PD versus HC 1.5-T female	0.8710		
					PD versus HC 3-T male	0.8444		
					PD versus HC 3-T female	0.7899		
				Naive Bayes	PD versus HC 1.5-T male	0.9901		
					PD versus HC 1.5-T female	0.8710		
					PD versus HC 3-T male	0.84		

 https://doi.org/10.1007/s00521-026-12022-1

Table 9 (continued)

Title	Measures included	Atlas	Validation	Algorithms	Classification	Acc.	AUC-ROC	L1OCV
					PD versus HC 3-T female	0.8403		
				Bayesian networks	PD vs HC 1.5-T male	0.9307		
					PD versus HC 1.5-T female	0.8871		
					PD versus HC 3-T male	0.7467		
					PD versus HC 3-T female	0.7059		
				k-NN (k = 5)	PD vs HC 1.5-T male	0.9307		
					PD versus HC 1.5-T female	0.9032		
					PD versus HC 3-T male	0.8133		
					PD versus HC 3-T female	0.7395		
				Multi-layer perceptron	PD versus HC 1.5-T male	0.9802		
					PD versus HC 1.5-T female	0.9355		
					PD versus HC 3-T male	0.8622		
					PD versus HC 3-T female	0.8571		
				SVM	PD vs HC 1.5-T male	NA		
					PD versus HC 1.5-T female	0.8548		
					PD versus HC 3-T male	0.9556		
					PD versus HC 3-T female	NA		
Rajagopalan and Pioro [51]	GM area, cortical thickness, GM fractal dimension	MNI	5-fold CV	Random forest	ALS-CST+ versus ALS-CST-	0.92	NA	
					ALS-CST+ versus ALS-CST- versus NC	NA		
				SVM	ALS-CST+ versus ALS-CST- versus NC	NA		
				Naive Bayes	ALS-CST+ versus ALS-CST- versus NC	0.74		
					ALS-CST+ versus ALS-CST	0.96		
Nemoto et al. [50]	GM volume, WM volume	MNI	5-fold CV	ResNet type of CNN	LBD versus AD	0.7915	0.77	NA
				DT		0.6841	0.67	
Meyer et al. [56]	GM volume, total intracranial volume	MNI, WFU PickAtlas	Leave one center out validation, L1OCV	SVM	bvFTD (n = 19) versus NC (n = 19)	0.736	NA	NA

Table 9 (continued)

Title	Measures included	Atlas	Validation	Algorithms	Classification	Acc.	AUC-ROC	L1OCV
					bvFTD(n = 52) vs NC (n = 52)	0.817		
Krüger et al. [52]	GM volume, age, total intracranial volume	Custom reference atlas	Not mentioned	3D-CNN with a U-net-like architecture	AD or FTLD vs NC (scanner-specific VBM)	0.890	NA	
					AD or FTLD versus NC (CNN-VBM)	0.898		
					AD or FTLD versus NC (Multiple-scanner VBM)	0.644		
Ikemitsu et al. [57]	GM volume, WM volume, cortical thickness	A standard template	5-fold CV	DL	AD versus NC 66 Volume values	0.82	0.8	
					AD versus NC total GM and WM volumes	0.6		
				SVM	AD versus NC 66 Volume values	0.82	0.81	
					AD versus NC 6 brain regions from LR	0.82		
				3D-CNN	AD versus NC whole brain	0.5	0.51	
Huang et al. [58]	GM volume	MNI	The paper uses a validation set to find the best model and control overfitting, but does not specify k-fold cross-validation or leave-one-out validation	3D-CNN	MCI versus NC	Train:0.978		
						Verification:0.960		
						Test: 0.809	0.891	
Egger et al. [59]	GM volume	MNI, LPBA40	Not mentioned	automated z-score based analysis using SPM12	All Patients versus HC	NA	Frontal lobes: 0.84	
							Medial temporal lobes: 0.93	
							Occipital lobes: 0.83	
							Parietal lobes: 0.86	
							Temporal lobes: 0.91	
					FTD versus Rest	NA	Frontal lobes: 0.75	
							Medial temporal lobes: 0.73	

 https://doi.org/10.1007/s00521-026-12022-1

Table 9 (continued)

Title	Measures included	Atlas	Validation	Algorithms	Classification	Acc.	AUC-ROC	L1OCV
							Occipital lobes: 0.64	
							Parietal lobes: 0.64	
							Temporal lobes: 0.78	
					AD versus Rest	NA	Frontal lobes: 0.67	
							Medial temporal lobes: 0.62	
							Occipital lobes: 0.63	
							Parietal lobes: 0.67	
							Temporal lobes:0.65	
					LBD versus Rest	NA	Frontal lobes: 0.42	
							Medial temporal lobes: 0.58	
							Occipital lobes: 0.56	
							Parietal lobes: 0.54	
							Temporal lobes:0.48	
	CSF volume				All Patients versus HC	NA	Frontal lobes: 0.73	
							Medial temporal lobes: 0.91	
							Occipital lobes: 0.73	
							Parietal lobes: 0.84	
							Temporal lobes:0.84	
					FTD vs Rest	NA	Frontal lobes: 0.8	
							Medial temporal lobes: 0.74	
							Occipital lobes: 0.66	
							Parietal lobes: 0.71	
							Temporal lobes:0.81	
					AD versus Rest	NA	Frontal lobes: 0.75	
							Medial temporal lobes: 0.62	
							Occipital lobes: 0.71	
							Parietal lobes: 0.78	
							Temporal lobes:0.73	

Table 9 (continued)

Title	Measures included	Atlas	Validation	Algorithms	Classification	Acc.	AUC-ROC	L1OCV
					LBD versus Rest	NA	Frontal lobes: 0.18 Medial temporal lobes: 0.55 Occipital lobes: 0.35 Parietal lobes: 0.34 Temporal lobes: 0.3	
Bachli et al. [60]	GM volume, IFS and ACE scores, gender, age, education	MNI, AAL-atlas	L2OCV	LR	C1-FTD (n = 18) versus HC (n = 30)	0.938	0.956	
					C1-AD (n = 16) versus HC (n = 22)	0.912	0.967	
			Cross-country validation		C2-FTD (n = 9) versus HC (n = 22)	0.913	0.935	
					C3-FTD (n=11) versus HC (n = 12)	0.913	0.906	
					C3-AD (n=10) versus HC (n = 12)	1.0	1.0	

L1OCV/L2OCV: leave one/two out cross validation, CV: cross validation

ELM: Extreme Learning Machine, SVM: Support vector machine, GPR: Gaussian Process Regression, PLS: Partial Least Squares regression, LR: logistic regression, CNN: Convolutional Neural Network, DL: Deep learning, k-NN: k nearest neighborhood, ResNet: Residual Neural Network, DT: Decision-tree-based classifier

Meyer et al. [56] used SVM classification for bvFTD diagnosis, achieving 84.6% accuracy with an ROI approach focusing on the frontal and temporal lobes, insula, and basal ganglia. Bachli et al. [60] included IFS and ACE scores, sex, age, atrophy, and education data in their classification. The importance of these features was additionally assessed by systematically removing each one from the dataset and evaluating their individual contributions.

Zhang et al. [40] calculated 14 texture parameters, achieving 0.61 accuracy and 0.552 AUC using an extreme learning machine (ELM) model with textural features alone. Combining VBM, clinical information, and textural features improved accuracy to 0.96 and AUC to 0.934.

The analysis in Appendix C reveals several key findings regarding algorithm performance across neurodegenerative diseases. In AD, studies achieved the highest accuracy (1.0) using LR for AD vs. HC classification, while ELM and SVM also demonstrated strong performance (0.92, 0.82 accuracy respectively). Notably, PD studies predominantly employed SVM, with accuracy ranging from 0.73 (PD-MCI vs. PD-NC) to 0.956 (PD vs. HC), highlighting its reliability for PD-related classifications.

For FTD, SVM achieved 0.817 accuracy in distinguishing bvFTD from controls, whereas 3D-CNN reached 0.898 in AD/FTLD vs. NC classification. LBD studies reported moderate accuracy (0.7915 with ResNet), while ALS classifications showed exceptional performance with Naïve Bayes (0.96 accuracy for ALS-CST+ vs. ALS-CST-). The pattern suggests researchers select algorithms based on both disease characteristics (e.g., structural changes in AD/PD) and data availability (e.g., smaller LBD cohorts favoring deep learning). SVM's adoption in PD, AD, and FTD reflects its adaptability to varied neuroimaging features, while specialized methods like 3D-CNN address complex spatial patterns in FTD and AD.

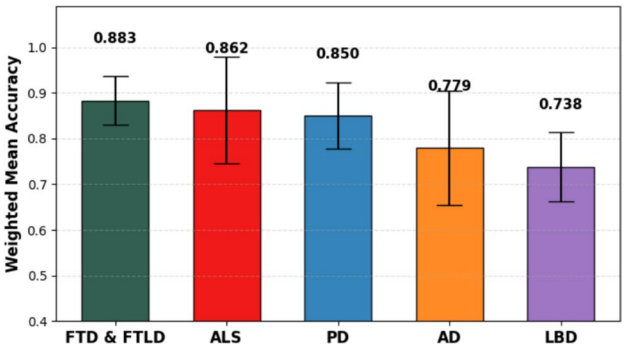


Fig. 9 Comparative analysis of weighted mean accuracy scores by disease category

Table 10 Disease-specific quantitative synthesis (accuracy)

Disease group	Weighted-mean	SD
FTD and FTLD	0.883	0.053
ALS	0.862	0.117
Parkinson's (PD)	0.85	0.073
Alzheimer's (AD)	0.779	0.126
Lewy body (LBD)	0.738	0.076

Table 11 Disease-specific quantitative synthesis (AUC-ROC)

Disease group	Weighted-mean	SD
All patients	0.92	0.014
FTD and FTLD	0.849	0.09
Parkinson's (PD)	0.818	0.078
Alzheimer's (AD)	0.767	0.168
Lewy body (LBD)	0.67	0.099

The highest AUC-ROC scores were achieved in AD studies, particularly with LR and ELM (Appendix D). The standout performances include a perfect 1.0 AUC-ROC in using LR (C3-AD vs. HC classification) and 0.967 (C1-AD vs. HC) in Bachli et al. [60] with the same algorithm. ELM models also showed strong results (0.925–0.922) in AD versus NC classifications. Conversely, the weakest AD performance was a 3D-CNN model (0.51) in (Ikemitsu et al. [57]) for AD versus NC classification.

For other diseases, Parkinson's (PD) studies showed consistent but moderate SVM performance (0.75–0.86) for PD-MCI versus PD-NC classification. FTD demonstrated LR's effectiveness again (0.906–0.956) (Bachli et al. [60]) in center based classifications, while LBD had the lowest overall scores (0.55–0.77), with ResNet outperforming decision trees (0.77 vs. 0.67) (Nemoto et al. [50]) in LBD versus AD differentiation. The "All Patients" group showed surprisingly high performance (0.91–0.93) (Egger et al. [59]) using z-score methods on medial temporal lobe data, suggesting this region's biomarker potential across neurodegenerative conditions.

4.7 Quantitative analysis of diagnostic performance

The quantitative synthesis stratified by pathology revealed distinct performance trends. As shown in Fig. 9 (Accuracy) and Fig. 11 (AUC-ROC), the sample-size weighted mean indicates that FTD and FTLD detection models achieved the highest overall stability (Tables 10 and 11). Conversely, AD showed larger variability, as evidenced by the wider interquartile ranges in the box plot distributions (Figs. 10 and 12). This suggests that while individual studies report high metrics, the generalizability across different cohorts remains heterogeneous for this pathology.

To evaluate the consistency of the reported results, we mapped each study against the global performance trends (Fig. 13 for Accuracy and Fig. 14 for AUC-ROC) (Tables 12 and 13).

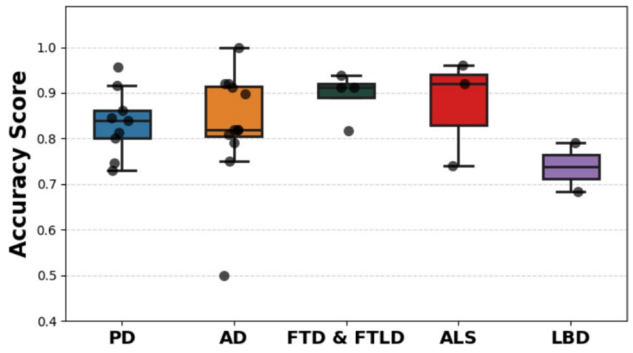


Fig. 10 Distribution and variability of diagnostic accuracy scores across neurodegenerative disease groups

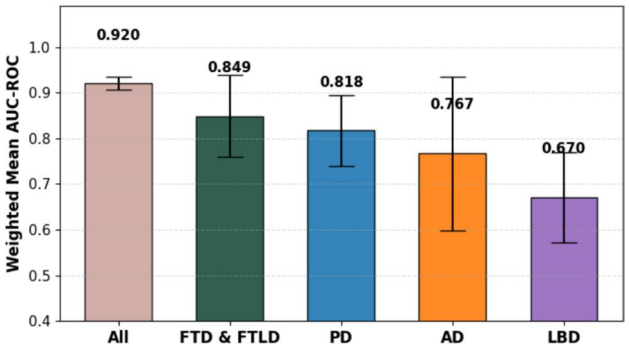


Fig. 11 Comparative analysis of weighted mean AUC-ROC scores by disease category

- Global Trends: The Grand Mean for accuracy across all reviewed studies was calculated at 0.838, with a global standard deviation band (shaded region) of ± 0.101.
- Outliers and Consistency: Studies such as Bachli et al. [60] and Krüger et al. [52] demonstrated performance significantly above the global trend. In contrast, studies with large error bars (e.g., Ikemitsu et al. [57], Rajagopalan [51]) exhibited high intra-study variability, indicating significant performance gaps between the different algorithms (e.g., CNN vs. DT) tested within the same cohort.

5 Discussion

5.1 Participant demographics and dataset limitations

The systematic review uncovered significant imbalances in participant distributions across studies. As shown in Fig. 3, Alzheimer's disease (58.3%) and Parkinson's disease (25%) dominated the research landscape, while conditions like Lewy body dementia averaged merely 30 participants per study. This skewness raises concerns about the generalizability of findings to rarer neurodegenerative conditions. Notably, 66.7% of studies employed healthy controls as binary comparators, creating an artificial diagnostic scenario that fails to capture clinical realities such as preclinical stages or disease mimics. For instance, no study addressed the critical challenge of differentiating normal pressure hydrocephalus from Alzheimer's pathology, a common diagnostic dilemma in geriatric

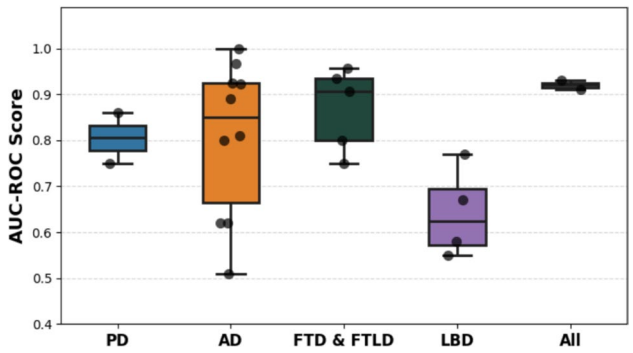


Fig. 12 Distribution and variability of diagnostic AUC-ROC scores across Neurodegenerative disease groups

 https://doi.org/10.1007/s00521-026-12022-1

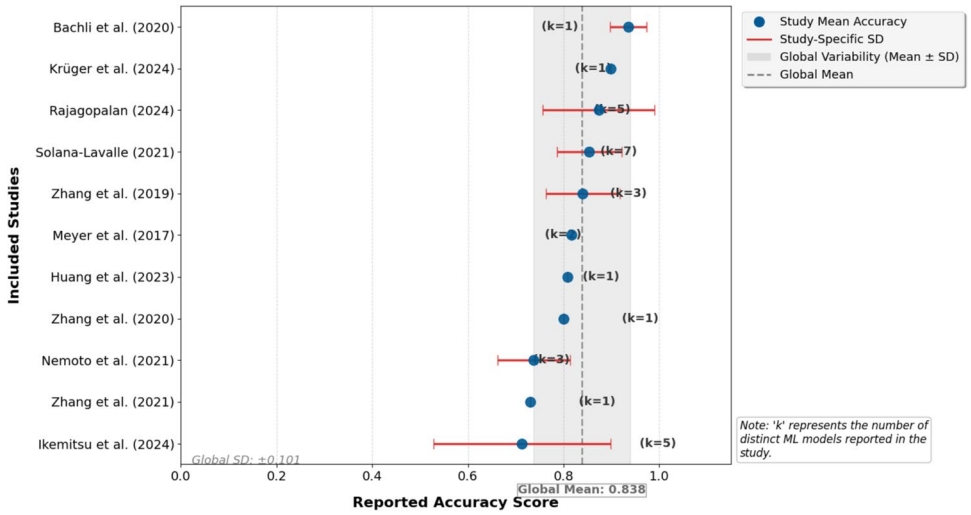


Fig. 13 Quantitative synthesis of diagnostic performance: inter-study and intra-study variability of accuracy scores

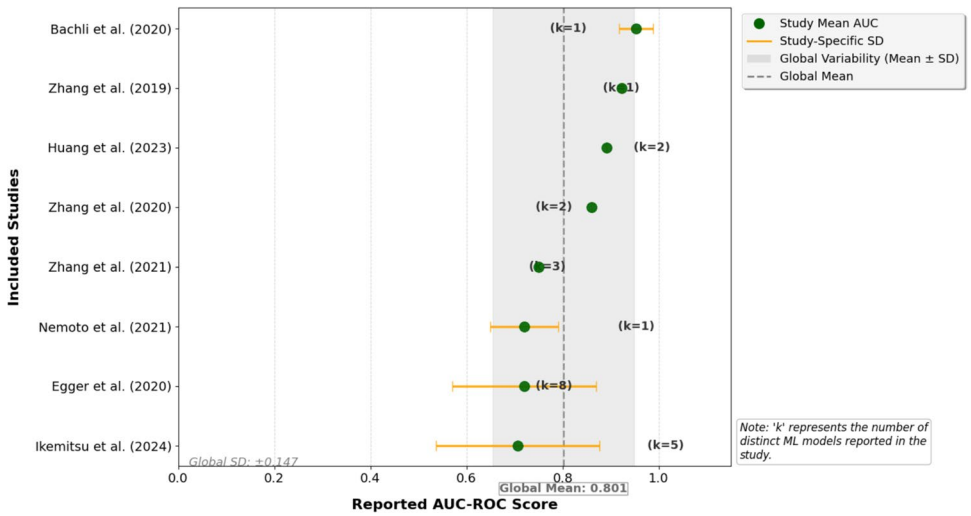


Fig. 14 Quantitative synthesis of diagnostic performance: inter-study and intra-study variability of AUC-ROC scores

Table 12 Study variability statistics (accuracy)

Global mean: 0.8381			
Global SD: 0.1011			
Study	Count	Mean	SD
Ikemitsu et al. [57]	3	0.713	0.185
Zhang et al. [54]	1	0.73	0
Nemoto et al. [50]	2	0.738	0.076
Zhang et al. [53]	1	0.8	0
Huang and Zhang [15]	1	0.809	0
Meyer et al. [56]	1	0.817	0
Zhang et al. [40]	5	0.84	0.077
Solana-Lavalle and Rosas-Romero [55]	7	0.854	0.068
Rajagopalan [51]	3	0.873	0.117
Krüger et al. [52]	1	0.898	0
Bachli et al. [60]	5	0.935	0.038

Table 13 Study variability statistics (AUC-ROC)

Global mean: 0.8014			
Global SD: 0.1471			
Study	Count	Mean	SD
Ikemitsu et al. [57]	3	0.707	0.17
Egger et al. [59]	8	0.72	0.149
Nemoto et al. [50]	2	0.72	0.071
Zhang et al. [54]	1	0.75	0
Zhang et al. [53]	1	0.86	0
Huang and Zhang [15]	1	0.891	0
Zhang et al. [40]	2	0.924	0.002
Bachli et al. [60]	5	0.953	0.035

neurology. These limitations underscore the need for multicenter registries that prioritize spectrum-based sampling, particularly for prodromal and borderline cases that better reflect real-world diagnostic complexity.

5.2 Methodological variability in neuroimaging analysis

Our analysis revealed substantial heterogeneity in image processing methodologies that may compromise result comparability. While 91.7% of studies applied spatial normalization, only one-third implemented motion correction - a concerning omission given the known impact of motion artifacts on gray matter measurements in small cohorts. The field showed heavy reliance on the MNI template (66.7% adoption rate), despite evidence that custom templates can improve Parkinson's disease classification accuracy by 12%. The software ecosystem was similarly constrained, with SPM and FreeSurfer dominating 83% of studies. This methodological monoculture risks overlooking advances in registration algorithms like ANTs, which may offer superior handling of atypical anatomies. We propose concrete standardization measures including mandatory reporting of preprocessing parameters and anatomy-specific template selection guidelines to enhance reproducibility.

5.2.1 The impact of motion correction on GM estimates

The detrimental quantitative impact of motion artifacts on MRI quality and voxel-based morphometry (VBM)

Academic literature has clearly demonstrated that motion artifacts negatively and quantitatively affect MRI image quality, particularly the GM volume and thickness estimates obtained in VBM studies, potentially leading to false or misleading morphological findings [61, 62]. Motion artifacts manifest as blurring, ghosting, and distortion of structural information in the images [63].

Quantitative impact on gray matter estimates

Head motion causes a notable decrease in cortical gray matter volume and thickness estimates, especially those measured by VBM and surface-based morphometry methods. This decrease is not due to true biological atrophy but arises as a motion-induced artifact.

A study by Reuter et al. [63] showed that, as subjects performed controlled motion tasks, estimates of cortical GM volume and thickness decreased as a function of accelerated motion. On average, an apparent loss of volume ranging from approximately 0.7–1.5% was recorded for every 1 mm/min increase in RMSpm (root-mean-square of head motion per minute).

Motion artifacts degrade image quality, which both reduces the reliability of the measurements and increases within-group variance. If there is a systematic motion difference between groups (e.g., between young and elderly groups, or patient and control groups), this can lead to spurious group differences [64].

Effect on VBM studies

The presence of motion artifacts directly impacts the fundamental preprocessing steps of VBM (segmentation and normalization), severely compromising the results.

Segmentation: Motion-induced blurring and ghosting reduce the signal contrast between GM, WM and CSF. This makes it challenging to accurately classify tissues during the initial segmentation step of VBM. GM boundaries can be misidentified in artifact-ridden regions.

Normalization: Poor-quality images prevent the accurate spatial normalization of the brain to a standard template (e.g., MNI). This results in the misregistration of anatomically corresponding points across different subjects, which lowers the reliability of comparisons in VBM.

Because motion can artificially decrease GM volume in certain regions, if one group moves more than another, VBM analysis may show regions of false atrophy (decreased GM volume) or false hypertrophy (increased GM volume).

This highlights that VBM results are significantly dependent on the preprocessing steps (segmentation and normalization). Motion correction methods and inflexible normalization algorithms (such as older SPM versions) can potentially exaggerate inter-group differences [64]. The quality of motion correction techniques and the choice of the VBM protocol are critical for the reliability of the findings.

It has been reported that using a Deep Learning (Generative Adversarial Network - GAN) based approach to correct motion artifacts improved the results of VBM-based analyses, such as the Voxel-Based Specific Region Analysis System for Alzheimer's Disease (VSRAD) [65].

To minimize the negative effects of motion on VBM, it is appropriate to:

- Prevent subject motion during imaging (e.g., using head cushions, straps, feedback systems),
- Apply software correction (motion correction) during or after the scanning sequence,
- Exclude highly-moving subjects by using image quality metrics like RMSpm.

5.3 Algorithm performance and clinical relevance gaps

The reviewed studies demonstrated varying algorithm efficacy across disease categories. Logistic regression achieved perfect classification (AUC=1.0) for Alzheimer's when combining imaging and cognitive data, while support vector machines outperformed deep learning in Parkinson's diagnosis - likely due to their greater suitability for smaller datasets. However, these apparent successes mask a fundamental limitation: 91.7% of studies employed binary classification paradigms that bear little resemblance to clinical decision-making. Neurologists typically navigate differential diagnoses across multiple potential conditions and disease stages, yet only one

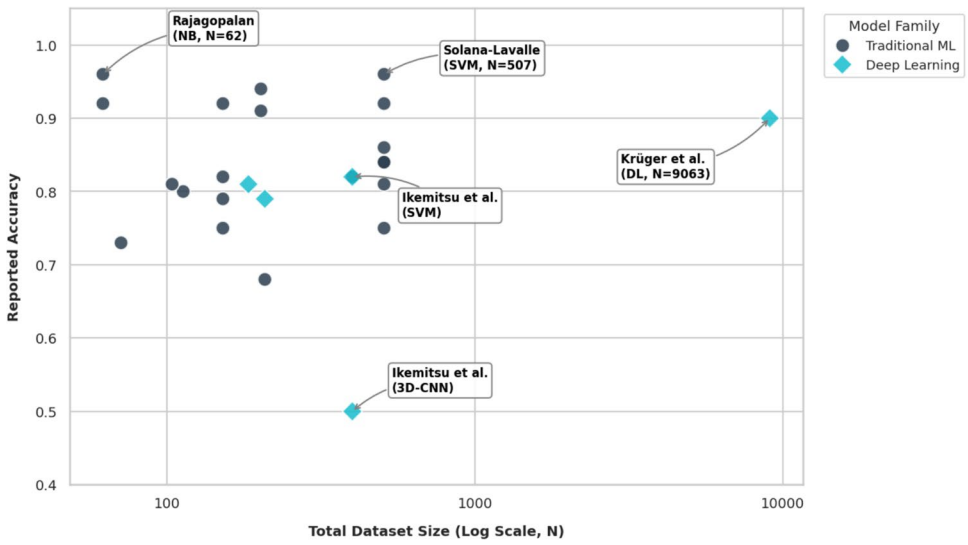


Fig. 15 Model type versus dataset size versus performance

included study (Bachli et al. 2020) attempted multiclass classification. This artificial simplification particularly fails patients with comorbid pathologies like mixed Alzheimer's and vascular dementia, which represent 20–40% of elderly cognitive impairment cases but were excluded from all analyses. The field should prioritize the development of multiclass models that provide probabilistic outputs, as this would better facilitate differential diagnosis of psychiatric conditions.

Additionally, a notable finding of this review is the competitive, and often superior, performance of traditional machine learning algorithms (e.g., SVM, LR) compared to complex deep learning (DL) architectures (e.g., 3D-CNN) (Fig. 15). This trend can be attributed primarily to the relationship between model complexity and dataset size. Deep learning models are inherently "data-hungry," requiring massive datasets to optimize millions of parameters and generalize effectively without overfitting [66]. Recent benchmarking studies in neuroimaging demonstrate that linear models (e.g., SVM, LR) often yield equivalent or superior performance to DL architectures in small-to-moderate sample size regimes ($N < 1000$), as deep models typically require significantly larger cohorts to escape the limitations of overfitting and fully exploit their capacity [67]. Consistent with these findings, Fig. 3 illustrates that the majority of reviewed studies utilized relatively small cohorts ($N < 500$), a range where traditional algorithms often thrive.

SVM and LR benefit significantly from the rigorous preprocessing steps (VBM/SBM) that reduce high-dimensional MRI data into biologically meaningful features (e.g., ROI volumes, cortical thickness). This manual feature extraction acts as a form of regularization, preventing the 'curse of dimensionality' that plagues DL models trained on small samples. For instance, Ikemitsu et al. [57] reported that SVM (Accuracy: 0.82) significantly outperformed a 3D-CNN (Accuracy: 0.50) on the same dataset, likely because the CNN struggled to learn robust spatial features from limited training data. Conversely, the study by Krüger et al. [52], which utilized a massive dataset ($N > 9000$) for training, demonstrated the true potential of DL when supported by adequate data. Thus, the 'interpretability advantage' of simpler models is accompanied by a 'stability advantage' in the current landscape of neuroimaging research.

In machine learning studies, it is possible to encounter high but misleading prediction rates driven by overfitting. Since Egger et al. [59] and Krüger et al. [52] did not report any validation procedures, the potential for overfitting in their models remains unknown. Consequently, the results presented in these studies should be interpreted with caution considering this limitation.

5.4 Interpretation of quantitative findings

Our quantitative synthesis offers critical insights beyond individual study reporting. The alignment of the sample-size weighted means with high AUC-ROC scores confirms the strong discriminative ability of ML models in neurodegenerative disease diagnosis. However, the study variability analysis (Fig. 13) highlights a crucial methodological caveat: the high standard deviations observed in some studies suggest that model architecture (e.g., DL vs. classical ML) significantly influences performance, often more than the dataset size itself. The fact that most studies cluster within the Global SD band indicates a maturing field with establishing baselines, yet the presence of outliers suggests that novel hybrid approaches or multi-modal data integration (as seen in the high-performing studies) can push the boundaries of diagnostic accuracy.

5.5 Challenges in clinical translation and future directions: multi-modal and longitudinal integration

The translation of these technologies to clinical practice faces substantial barriers. Scanner variability alone caused 15.1% accuracy drops in multisite evaluations, suggesting that current single-center validations dramatically overestimate real-world performance [68]. More fundamentally, the studies' technical outputs—binary classifications of isolated MRI scans—fail to meet clinicians' need for tools supporting longitudinal progression assessment and multimodal data integration. None of the reviewed papers incorporated routine clinical data sources like cerebrospinal fluid biomarkers or neuropsychological test batteries that neurologists routinely

consider. To address this gap, future research should adopt validation frameworks that are ready for regulatory approval. These frameworks should incorporate prospective input from clinicians and meet explainability standards that generate saliency maps for review by radiologists. Only through such clinically-grounded development can these tools transition from academic exercises to meaningful diagnostic aids.

5.5.1 Multi-modal and longitudinal data integration: overcoming single-modality limitations and bridging the translational gap

While our current systematic review highlights the successful integration of VBM and SBM approaches with Machine Learning (ML) for diagnosing neurodegenerative diseases, all the studies examined are limited to a single imaging modality (typically structural magnetic resonance imaging - sMRI). Single-modality approaches may offer an incomplete representation of the disease's complex biological basis, inherently limiting their diagnostic power. Multi-modal data integration is a critical step towards achieving high-accuracy predictions.

The combination of structural MRI (VBM/SBM) data with other modalities such as functional MRI (fMRI), Positron Emission Tomography (PET) (e.g., Amyloid or Tau imaging), Diffusion Tensor Imaging (DTI), and even genetic/cognitive/clinical biomarkers, provides a more comprehensive feature set that covers different pathological aspects of the disease [69, 70]. Multi-modal ML models have the potential to improve the typically low sensitivity or specificity of a single biomarker by balancing the weaknesses of one modality with the strengths of another [71]. For instance, combining sMRI and FDG-PET data has been shown to achieve significantly higher accuracy in Alzheimer's disease (AD) diagnosis compared to using either modality alone [72]. While the majority of the studies in the review include cognitive tests, some studies (Table 6) do not incorporate important variables like age and gender, whose effects on brain volume have been demonstrated by numerous works. Although no statistically significant difference in accuracy was observed between controlled and uncontrolled studies (Fig. 16), the exclusion of such fundamental demographic characteristics runs the risk of insufficient disentanglement of biomarker effects from these powerful confounding variables, negatively impacting generalizability.

5.5.2 Leveraging longitudinal data for prognosis

Similarly, the use of longitudinal data enables not only the diagnosis of the disease but also the prediction of its rate of progression and individual course [73]. Morphometric measurements taken at a single time point (cross-sectional data) may be insufficient to capture the subtle changes in the early stages of the disease or the dynamics of its progression. Longitudinal ML models, by using the rates of change in VBM or SBM measurements acquired at sequential time points as input, can more reliably differentiate between cases of stable mild cognitive impairment (sMCI) and those progressing to AD [74]. However, the studies included in this review have

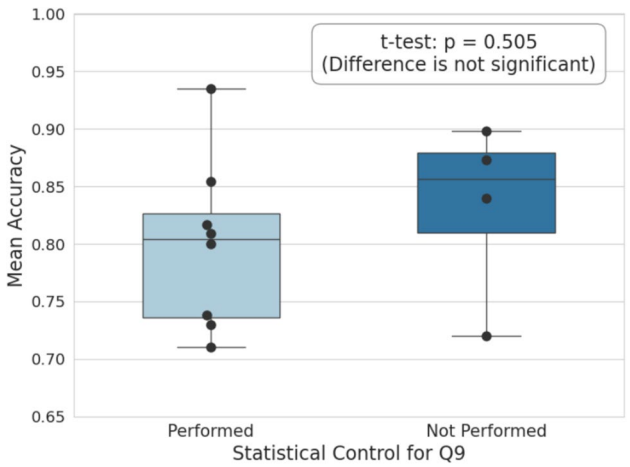


Fig. 16 Impact of age/sex/TIV on accuracy

not utilized longitudinal data or data on stable/progressing cognitive impairment. Consequently, the findings of the current studies remain predominantly focused on cross-sectional diagnosis, offering limited insight into the dynamics of disease progression and individual prognosis prediction.

5.5.3 Data availability and the translational gap

The primary obstacles preventing the full translation of these high-potential integration strategies into clinical practice are the limited accessibility and standardization of multi-modal and large-scale longitudinal datasets. The high cost and time-consuming nature of multiple modalities (sMRI, PET, DTI, clinical data) make it challenging to collect such comprehensive data for large cohorts within a single center.

Aggregating data collected across different centers (such as large international initiatives like the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS)) poses significant harmonization challenges due to variances in scanners, imaging protocols, and preprocessing methods [75]. These inconsistencies limit the generalizability of ML models in clinical settings. Furthermore, developing sophisticated Multi-View Learning or DL models that can effectively combine multi-modal data and produce interpretable results is more complex than developing single-modality models.

Therefore, to accelerate the transition to clinical application, it is crucial for researchers and institutions to collect multi-modal and long-term longitudinal data with standardized protocols and to make these data more globally accessible under the principles of Open Science. Only through this approach can the high-performing diagnostic models achieved through ML move from the laboratory environment and be integrated into patient care.

To address this gap, future research must adopt a translational roadmap guided by the following practical steps:

- **Rigorous external validation:** Models must be tested on completely independent, geographically diverse datasets acquired using different scanner hardware (e.g., cross-cohort validation) to quantify and mitigate scanner and center variability. Our finding of inconsistent preprocessing (Fig. 6) necessitates the adoption of standardized pipelines before external validation.
- **Regulatory readiness (FDA/CE mark):** Clinical deployment requires models to generate not only a diagnosis but also an Explainable AI (XAI) output, such as saliency maps or feature importance scores, that highlights the specific neuroanatomical regions driving the classification. This transparency is essential for clinician acceptance and regulatory approval.
- **Clinician-in-the-Loop models:** The current binary classification output (AD vs. NC) is insufficient. Future models must shift towards probabilistic, multi-class outputs that differentiate between multiple NDs (e.g., AD vs. FTD vs. DLB) and integrate patient-specific multimodal data (MRI + CSF + Cognitive Scores) to create a comprehensive diagnostic score, aligning with real-world neurological differential diagnosis.
- **Longitudinal assessment:** The focus must move from cross-sectional diagnosis to prediction of disease progression (e.g., MCI-to-AD within a 2-year follow-up period, conversion rates have been reported to range between 9.8 and 36.3% [76]), which is the most clinically impactful application for intervention.

Our review informs these pathways by demonstrating the performance ceiling of current binary, VBM/SBM-only models and underscores the urgent need to transition to multimodal, multi-class validation frameworks that account for the identified methodological and data heterogeneity.

5.6 Recommendations for future research: a proposed standardization framework

Table 14 outlines the essential parameters that require standardization and transparent reporting to ensure reproducibility and clinical viability in ML-based neuroimaging studies.

Table 14 Standardization table

Phase	Standardized parameter	Rationale/clinical impact
I. Data acquisition and raw data	Scanner parameters	Reporting of magnetic field strength (T), sequence type (e.g., T1w MPRAGE), and voxel size (mm^3)
	Image processing tools	Specific versions of key software (SPM, FreeSurfer, FSL) and conversion tools (e.g., dcm2niix)
II. Preprocessing and alignment	Motion correction	Mandatory application and reporting of the algorithm used (e.g., spm_realign)
	Skull stripping/segmentation	Algorithm/tool and key settings used for brain tissue extraction and segmentation into GM, WM, and CSF
	Normalization template	Explicit statement of the template (e.g., MNI152) and the registration method (e.g., DARTEL, ANTs). If a custom template is used, its creation cohort must be described
III. Feature and statistical control	Covariate correction	Method (e.g., linear regression) and variables (Age, Sex, TIV) used to statistically control for non-disease-related factors
	Smoothing kernel	Size of the smoothing kernel (FWHM, *mm*) applied during VBM/SBM analysis
IV. Validation and transparency	Full performance metrics	Beyond accuracy and AUC-ROC, the complete confusion matrix (TP, TN, FP, FN) must be reported
	Source code sharing	The complete source code used for model training and validation must be made publicly available (e.g., via GitHub or Zenodo)

5.6.1 Distinction between quality assessment and standardization proposals

While the Quality Assessment (QA) criteria (Table 5) were designed to evaluate the methodological robustness and internal validity of existing literature retrospectively, the proposed Standardization Table 14 serves a prospective purpose. The QA criteria focused on scoring studies based on established good practices (e.g., sample size sufficiency, use of cross-validation). However, our review revealed that even high-scoring studies often lacked the granular reporting details necessary for full reproducibility and clinical translation.

Therefore, the Standardization Table does not merely mirror the QA checklist; instead, it addresses the "reporting gaps" identified during the review. For instance, while the QA process evaluated whether a study implemented validation methods, the Standardization guidelines explicitly demand the reporting of full Confusion Matrices (TP/TN/FP/FN) to assess clinical safety–a detail frequently omitted in current literature but essential for future clinical deployment. Thus, Table 14 represents the ideal reporting standard required to overcome the reproducibility crisis, distinct from the evaluation metrics used to grade current retrospective data.

6 Limitations of this review

This systematic review identified several critical constraints that affected the interpretation and generalizability of the findings:

- **Small and Imbalanced Cohort Sizes:** The final pool of eligible studies was limited to $n = 12$ due to strict inclusion criteria. Furthermore, many neurodegenerative conditions, such as LBD, were represented by very small sample sizes (e.g., $n = 30$), which restricted the statistical power and generalizability of the reported results.
- **Methodological Bias and Lack of Transparency:** A significant "optimistic bias" was observed through correlation analysis; studies with lower methodological transparency and lower QA scores tended to report higher classification accuracies ($r = -0.32$). This suggested that unreported overfitting might have artificially inflated the performance metrics in studies that failed to transparently report model parameters.

- **Lack of Reproducibility and Open Science:** Reproducibility remained a major challenge in the field, as 0% of the reviewed studies made their source code publicly available. Additionally, only 33.3% of the studies provided access to their datasets, significantly hindering the ability of third parties to verify the reported outcomes.
- **Insufficient External Validation:** Only 25% of the reviewed studies tested their models on an independent external validation cohort. This lack of external validation made it difficult to assess how these models would perform in real-world clinical settings with different scanner hardware and protocols.
- **Oversimplified Diagnostic Paradigms:** Most studies (91.7%) employed binary classification (e.g., Healthy Control vs. Patient), which does not reflect the clinical reality where neurologists must navigate differential diagnoses across multiple conditions. This approach ignored the prevalence of comorbid pathologies, such as mixed Alzheimer's and vascular dementia, which were excluded from nearly all analyses.
- **Technical Heterogeneity:** There was a notable lack of consistency in preprocessing techniques, particularly regarding motion correction, which was explicitly mentioned in only 29.2% of the studies. Given that head motion can artificially decrease gray matter volume estimates, the absence of these corrections might have led to spurious findings.

7 Conclusion

Our comprehensive systematic review evaluated a total of 12 studies. Our analysis highlighted the substantial promise of machine learning and deep learning approaches in the diagnosis of neurodegenerative conditions through structural MRI. However, the successful clinical translation of these techniques required careful navigation of fundamental challenges.

The integration of a rigorous 20-point QA revealed critical systemic gaps. We identified a significant negative correlation between methodological transparency and reported performance ($r = -0.32$), indicating an 'optimistic bias' where studies with lower QA scores tended to claim higher success rates. Furthermore, the review found that reproducibility was severely hindered by a total lack of public source code (0%) and limited data sharing (33.3%).

Interestingly, simpler models occasionally outperformed more complex deep learning architectures, suggesting that neuroimaging diagnostics benefited more from biologically interpretable features than sheer computational complexity. To address these limitations, the field must prioritize multi-center collaborations to ensure the generalizability of findings while maintaining methodological rigor that balances reproducibility with clinical adaptability. True progress in this domain will be measured not only by benchmark metrics but also by the ability of these technologies to bridge the gap between algorithmic performance and actionable clinical insights.

Appendix A: Previous review studies included to this review

Bede, P. & Hardiman, O. Lessons of als imaging: Pitfalls and future directions—a critical review. NeuroImage: Clinical 4, 436–443 (2014). URL https://doi.org/10.1016/j.nicl.2014.02.011

Garg, N., Choudhry, M. S. & Bodade, R. M. A review on Alzheimer's disease classification from normal controls and mild cognitive impairment using structural mr images. Journal of Neuroscience Methods 384, 109745 (2023). URL https://doi.org/10.1016/j.jneumeth.2022.109745

Leandrou, S., Petroudi, S., Kyriacou, P. A., Reyes-Aldasoro, C. C. & Pattichis, C. S. Quantitative mri brain studies in mild cognitive impairment and Alzheimer's disease: A methodological review. IEEE Reviews in Biomedical Engineering 11, 97–111 (2018). URL https://doi.org/10.1109/RBME.2018.2796598

Meeter, L. H., Kaat, L. D., Rohrer, J. D. & Van Swieten, J. C. Imaging and fluid biomarkers in frontotemporal dementia. Nature Reviews Neurology 13, 406–419 (2017). URL https://doi.org/10.1038/nrneurol.2017.75

Ranganathan, L. et al. Changing landscapes in the neuroimaging of dementia. Annals of Indian Academy of Neurology 21, 98 (2018). URL https://doi.org/10.4103/aian.AIAN_48_18

Sakurai, K. et al. Beyond the midbrain atrophy: wide spectrum of structural mri finding in cases of pathologically proven progressive supranuclear palsy. Neuroradiology 59, 431–443 (2017). URL https://doi.org/10.1007/s00234-017-1812-4

Appendix B: Studies included to this review

Bachli, M. B. et al. Evaluating the reliability of neurocognitive biomarkers of neurodegenerative diseases across countries: A machine learning approach. Neuroimage 208, 116456 (2020). URLhttps://doi.org/10.1016/j.neuroimage.2019.116456.

Egger, K. et al. Automated voxel and region-based analysis of gray matter and cerebrospinal fluid space in primary dementia disorders. Brain Res 1739, 146800 (2020). URL https://doi.org/10.1016/j.brainres.2020.146800.

Huang, H. et al. Voxel-based morphometry and a deep learning model for the diagnosis of early Alzheimer's disease based on cerebral gray matter changes. Cereb Cortex 33, 754–763 (2023). URL https://doi.org/10.1093/cercor/bhac099.

Ikemitsu, N. et al. Determination of Alzheimer's disease based on morphology and atrophy using machine learning combined with automated segmentation. Acta Radiol 65, 359–366 (2024). URLhttps://doi.org/10.1177/02841851231218384.

Krüger, J., Opfer, R., Spies, L., Hedderich, D. & Buchert, R. Voxel-based morphometry in single subjects without a scanner-specific normal database using a convolutional neural network. Eur Radiol 34, 3578–3587 (2024). URL https://doi.org/10.1007/s00330-023-10356-1.

Meyer, S. et al. Predicting behavioral variant frontotemporal dementia with pattern classification in multi-center structural mri data. Neuroimage Clin 14, 656–662 (2017). URL http://dx.doi.org/10.1016/j.nicl.2017.02.001.

Nemoto, K. et al. Differentiating dementia with Lewy bodies and Alzheimer's disease by deep learning to structural mri. J Neuroimaging 31, 579–587 (2021). URL https://doi.org/10.1111/jon.12835.

Rajagopalan, V. & Pioro, E. P. Differing patterns of cortical grey matter pathology identified by multifractal analysis in umn-predominant als patients with and without corticospinal tract hyperintensity. J Neurol Sci 459, 122945 (2024). URL https://doi.org/10.1016/j.jns.2024.122945.

Solana-Lavalle, G. & Rosas-Romero, R. Classification of ppmi mri scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. Comput Methods Programs Biomed 198, 105793 (2021). URL https://doi.org/10.1016/j.cmpb.2020.105793.

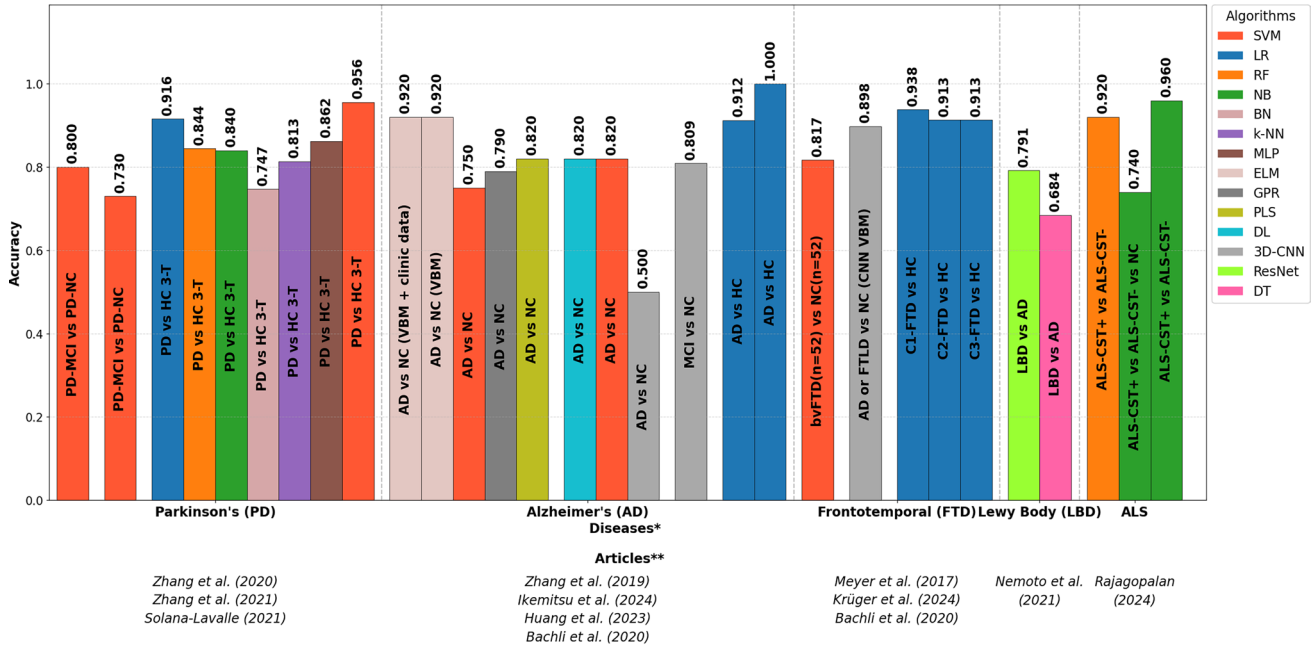
Zhang, F. et al. Voxel-based morphometry: Improving the diagnosis of Alzheimer's disease based on an extreme learning machine method from the adni cohort.Neuroscience 414, 273–279 (2019). URL https://doi.org/10.1016/j.neuroscience.2019.05.014.

Zhang, J. et al. An sbm-based machine learning model for identifying mild cognitive impairment in patients with Parkinson's disease. J Neurol Sci 418, 117077 (2020). URL https://doi.org/10.1016/j.jns.2020.117077.

Zhang, J. et al. Identifying Parkinson's disease with mild cognitive impairment by using combined mr imaging and electroencephalogram. Eur Radiol 31, 7386–7394(2021). URL https://doi.org/10.1007/s00330-020-07575-1.

Appendix C: Comparison of accuracy value performances of algorithms for neurodegenerative diseases

See Fig. 17.

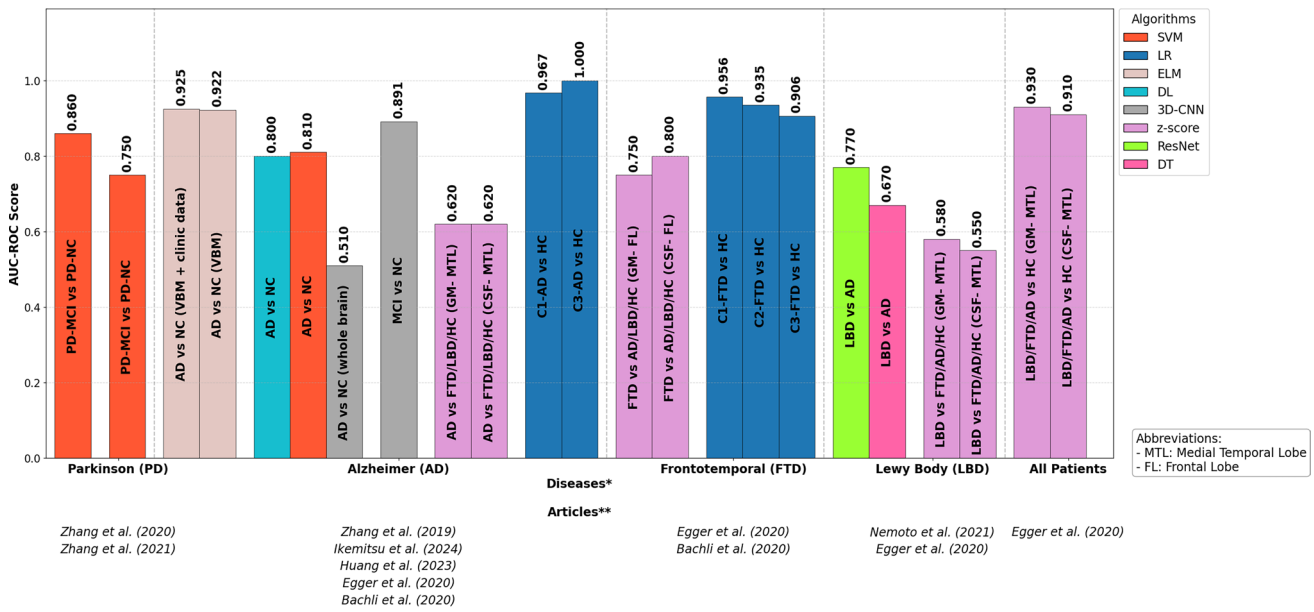


* Diseases: Neurodegenerative disease classifications are given on the x-axis and are seperated by vertical grid lines.
** Articles: Below the X-axis labels, the articles used in the classification are displayed sequentially and separated by vertical spaces between the bars.

Fig. 17 Comparison of accuracy value performances of algorithms for neurodegenerative diseases

Appendix D: Comparison of AUC-ROC score performances of algorithms for neurodegenerative diseases

See Fig. 18.



* Diseases: Neurodegenerative disease classifications are given on the x-axis and are seperated by vertical grid lines.
** Articles: Below the X-axis labels, the articles used in the classification are displayed sequentially and separated by vertical spaces between the bars.

Fig. 18 Comparison of AUC-ROC score performances of algorithms for neurodegenerative diseases

 https://doi.org/10.1007/s00521-026-12022-1

Appendix E: Comparison of accuracy performances by algorithms

See Fig. 19.

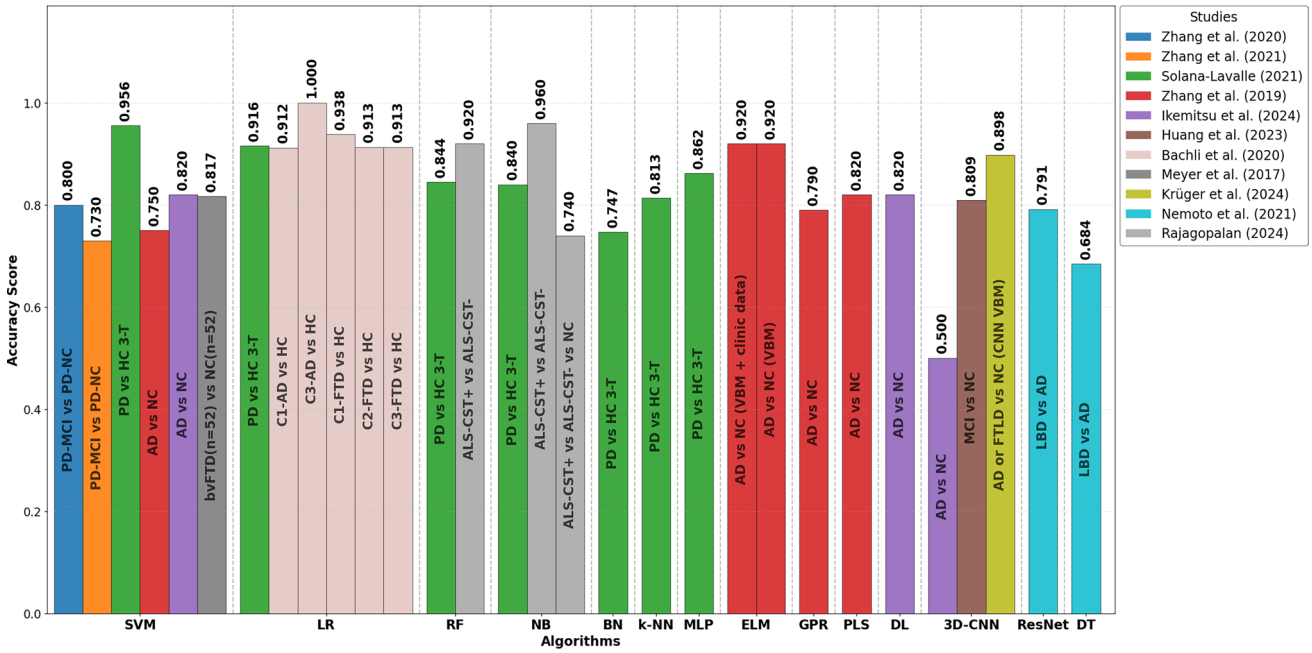


Fig. 19 Comparison of accuracy performances by algorithms

Appendix F: comparison of AUC-ROC performances by algorithms

See Fig. 20.

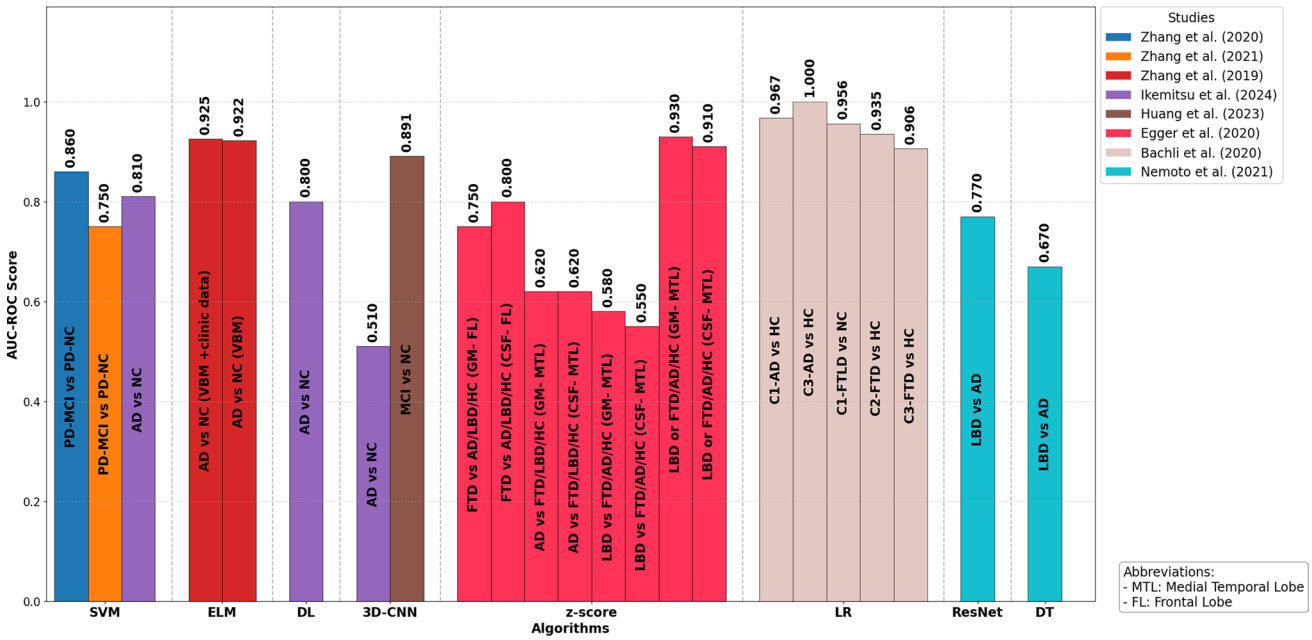


Fig. 20 Comparison of AUC-ROC performances by algorithms

Appendix G: Comparison of accuracy performances grouped by algorithms

See Fig. 21.

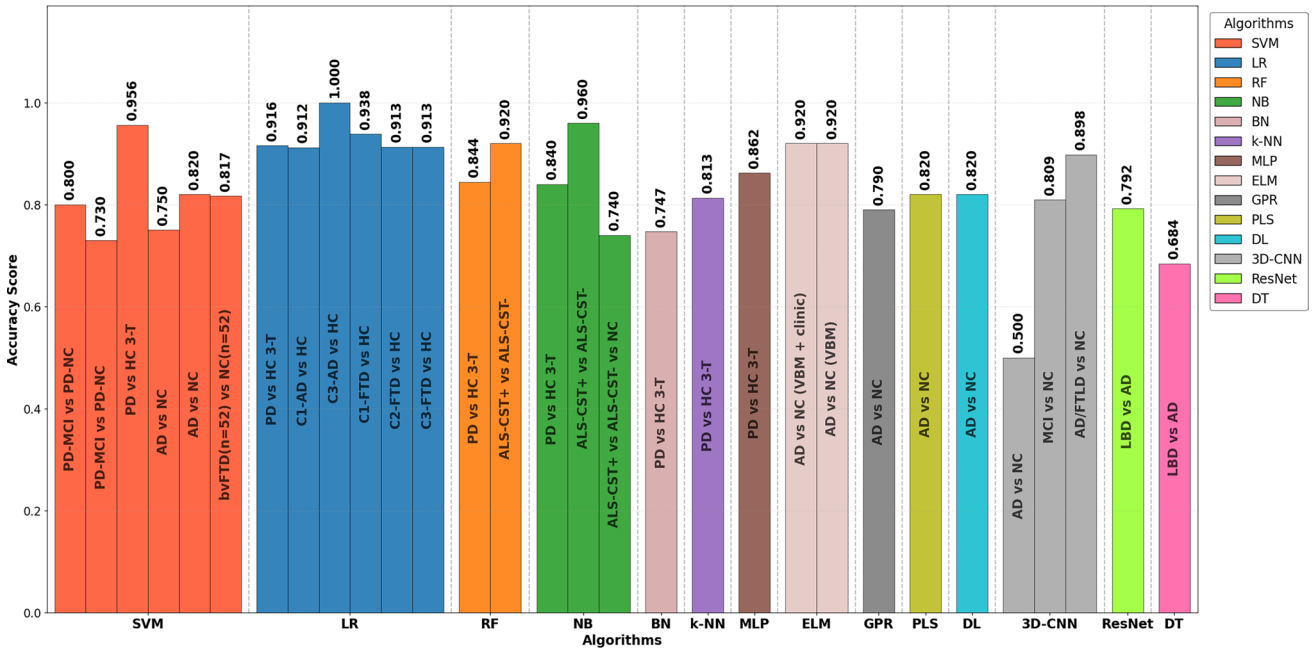


Fig. 21 Comparison of accuracy performances grouped by algorithms

Appendix H: Comparison of AUC-ROC performances grouped by algorithms

See Fig. 22.

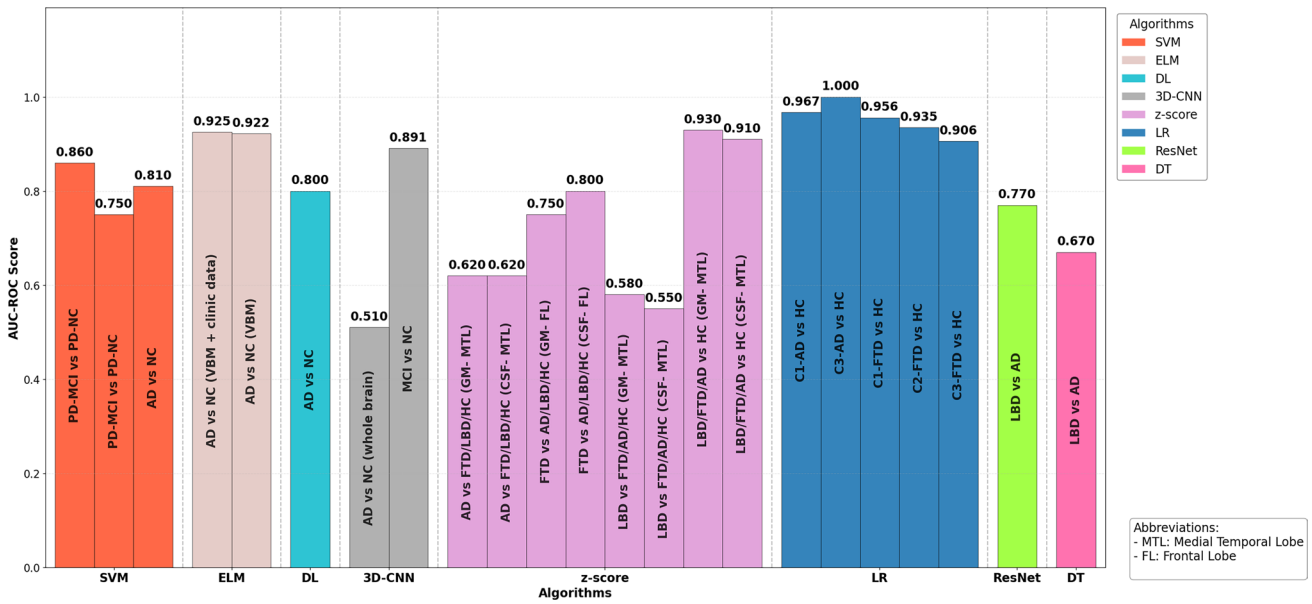
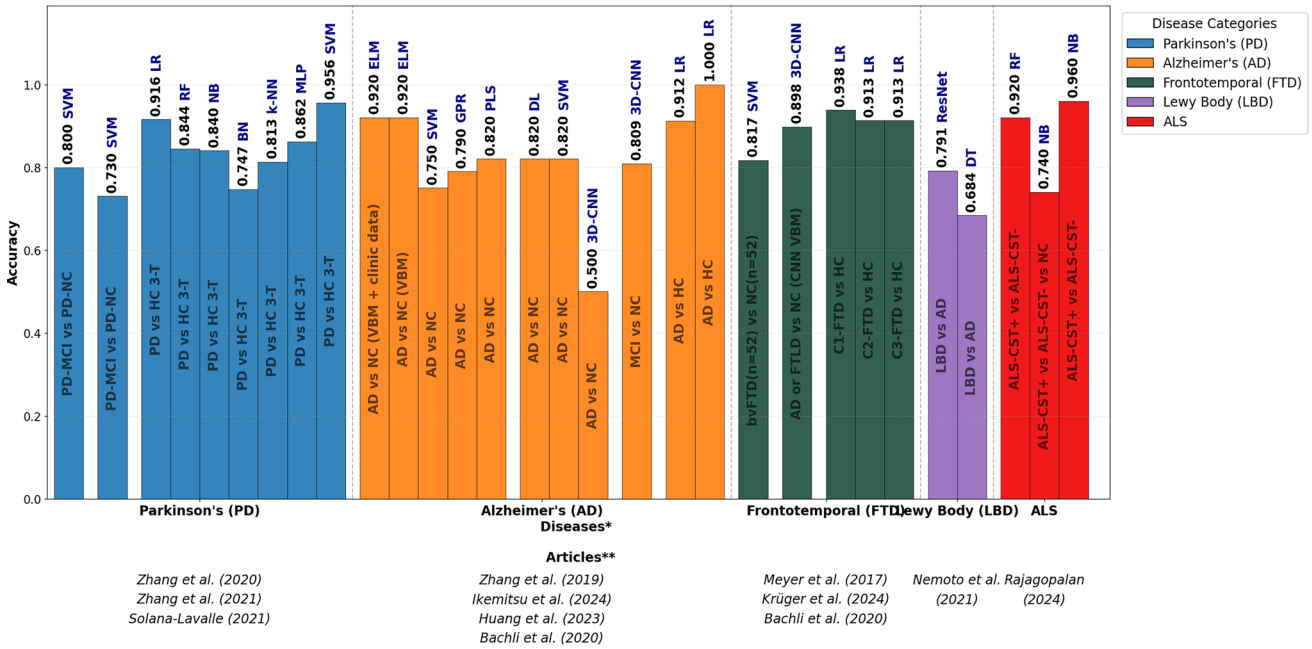


Fig. 22 Comparison of AUC-ROC performances grouped by algorithms

 https://doi.org/10.1007/s00521-026-12022-1

Appendix I: Accuracy performance distribution by neurodegenerative disease type

See Fig. 23.

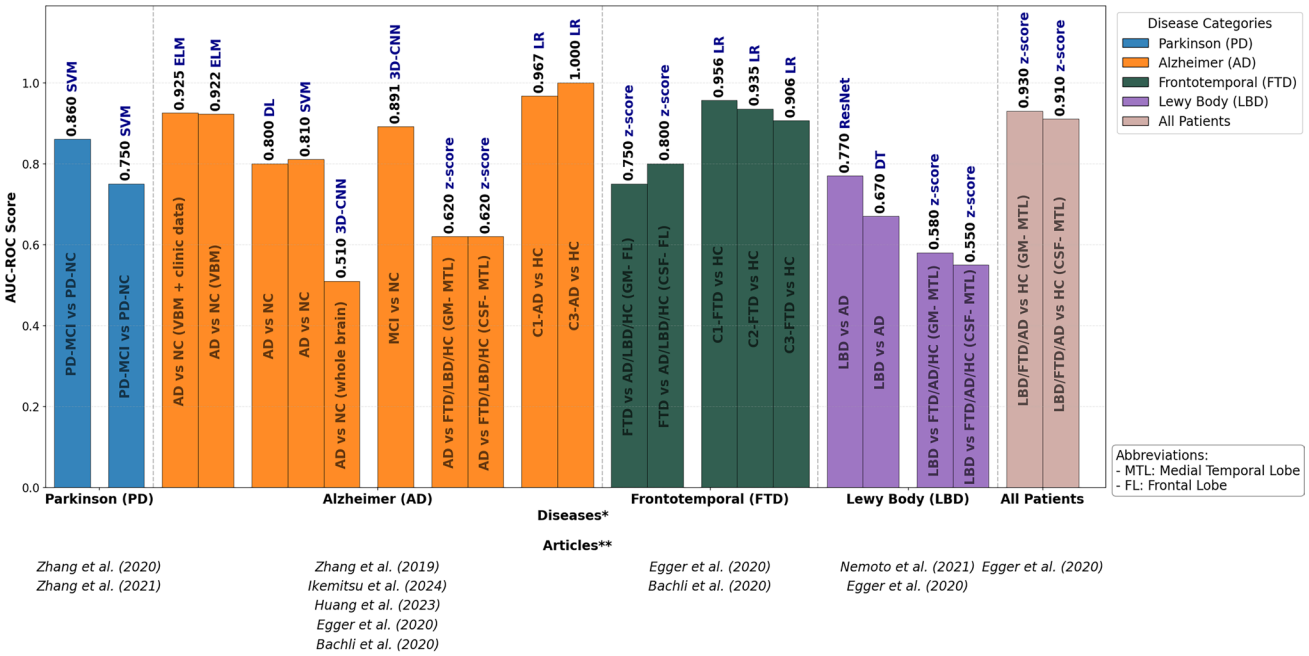


* Diseases: Neurodegenerative disease classifications are given on the x-axis and are seperated by vertical grid lines.
** Articles: Below the X-axis labels, the articles used in the classification are displayed sequentially and separated by vertical spaces between the bars.

Fig. 23 Accuracy performance distribution by neurodegenerative disease type

Appendix J: AUC-ROC performance distribution by neurodegenerative disease type

See Fig. 24.



* Diseases: Neurodegenerative disease classifications are given on the x-axis and are seperated by vertical grid lines.
** Articles: Below the X-axis labels, the articles used in the classification are displayed sequentially and separated by vertical spaces between the bars.

Fig. 24 AUC-ROC performance distribution by neurodegenerative disease type

Acknowledgements The authors gratefully acknowledge that this work originated from the doctoral dissertation of Emrah Gürlek, titled "Classification of Neurodegenerative Diseases with Machine Learning Techniques," supervised by Asst. Prof. İnci Zaim Gökbay within the PhD program of the Department of Informatics at İstanbul University, Institute of Graduate Studies in Sciences.

Author Contributions İ.Z.G.: Supervision, study planning, methodology planning, and main manuscript editing. E.G.: Screening, data extraction, quality assessment, analysis, main manuscript writing, revision. All authors have read and agreed to the published version of the manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process We acknowledge the use of a large language model (LLM) for assisting with language corrections during the article's editing. The model enhanced the readability and language quality of the publication. However, the authors retain full responsibility for the content, having reviewed and edited it as needed after using the tool.

Funding The doctoral dissertation of Emrah Gürlek was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) 2211-C National PhD Scholarship Program in the Priority Fields in Science and Technology.

Data availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval This study, being a review and not involving patient data, did not require institutional ethical approval.

Consent for publication Not applicable.

References

1. Przedborski S, Vila M, Jackson-Lewis V (2003) Series introduction: neurodegeneration: What is it and where are we? J Clin Investig 111:3–10. https://doi.org/10.1172/jci17522
2. Dehhaghi M, Panahi KSH, Guillemin GJ (2018) Microorganisms' footprint in neurodegenerative diseases. Front Cell Neurosci 8:12. https://doi.org/10.3389/fncel.2018.00466
3. Erkkinen MG, Kim M-O, Geschwind MD (2018) Clinical neurology and epidemiology of the major neurodegenerative diseases. Cold Spring Harb Perspect Biol 10:a033118. https://doi.org/10.1101/cshperspect.a033118
4. Nichols E et al (2022) Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. Lancet Public Health 7:e105–e125. https://doi.org/10.1016/S2468-2667(21)00249-8
5. Dubois B, Padovani A, Scheltens P, Rossi A, Dell'Agnello G (2015) Timely diagnosis for Alzheimer's disease: a literature review on benefits and challenges. J Alzheimers Dis 49:617–631. https://doi.org/10.3233/JAD-150692
6. de Lau LML, Breteler MMB (2006) Epidemiology of Parkinson's disease. Lancet Neurol 5:525–535. https://doi.org/10.1016/S1474-4422(06)70471-9
7. Nussbaum RL, Ellis CE (2003) Alzheimer's disease and Parkinson's disease. N Engl J Med 348:1356–1364. https://doi.org/10.1056/nejm2003ra020003
8. Feigin VL et al (2017) Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the global burden of disease study 2015. Lancet Neurol 16:877–897. https://doi.org/10.1016/S1474-4422(17)30299-5
9. Nichols E et al (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. Lancet Neurol 18:88–106. https://doi.org/10.1016/s1474-4422(18)30403-4
10. Jellinger KA (2007) in Chapter 1—advances in our understanding of neurodegeneration (eds Qureshi, G. A. & Parvez, S. H.) Oxidative Stress and Neurodegenerative Disorders 1–58 (Elsevier Science B.V., Amsterdam, 2007). https://doi.org/10.1016/B978-044452809-4/50142-3
11. Jellinger KA (2010) Basic mechanisms of neurodegeneration: a critical update. J Cell Mol Med 14:457–487. https://doi.org/10.1111/j.1582-4934.2010.01010.x
12. Yacoubian T (2017) in Chapter 1—neurodegenerative disorders: Why do we need new therapies? (ed.Adejare, A.) Drug discovery approaches for the treatment of neurodegenerative disorders 1–16 (Academic Press, 2017). https://doi.org/10.1016/B978-0-12-802810-0.00001-5
13. NIEHS. Neurodegenerative diseases (2022). https://www.niehs.nih.gov/research/supported/health/neurodegenerative
14. Hauser RA, Olanow CW (1994) Magnetic resonance imaging of neurodegenerative diseases. J Neuroimaging 4:146–158. https://doi.org/10.1111/jon199443146
15. Huang P, Zhang M (2023) Magnetic resonance imaging studies of neurodegenerative disease: from methods to translational research. Neurosci Bull 39:99–112. https://doi.org/10.1007/s12264-022-00905-x
16. Dhakal S et al (2023) Dementia prediction using machine learning. Procedia Comput Sci 219:1297–1308. https://doi.org/10.1016/j.procs.2023.01.414
17. Perez-Valero E, Lopez-Gordo MA, Morillas C, Pelayo F, Vaquero-Blasco MA (2021) A review of automated techniques for assisting the early detection of Alzheimer's disease with a focus on EEG. J Alzheimers Dis 80:1363–1376. https://doi.org/10.3233/JAD-201455
18. Wei M, Li Y, Liang M, Xi M, Tian H (2023) Artificial intelligence approaches for early detection and diagnosis of Alzheimer's disease: a review. Acad J Sci Technol 5:215–221. https://doi.org/10.54097/ajst.v5i3.8015
19. Ashburner J, Friston KJ (2000) Voxel-based morphometry-the methods. Neuroimage 11:805–821. https://doi.org/10.1006/nimg.2000.0582
20. Agosta F, Galantucci S, Filippi M (2017) Advanced magnetic resonance imaging of neurodegenerative diseases. Neurol Sci 38:41–51. https://doi.org/10.1007/s10072-016-2764-x
21. Fischl B, Dale AM (2000) Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci 97:11050–11055. https://doi.org/10.1073/pnas.200033797
22. Library D (2025) https://www.dicomlibrary.com/dicom/
23. Krishnadas P, Smith N (2023) An investigation of available medical imaging data for the early detection of neurodegenerative diseases. Report, National Physical Laboratory. https://doi.org/10.47120/npl.ms50
24. (NIFTI) N. I. T. I. (2013). https://nifti.nimh.nih.gov/

25. Arabahmadi M, Farahbakhsh R, Rezazadeh J (2022) Deep learning for smart healthcare-a survey on brain tumor detection from medical imaging. Sensors 22:1960. https://doi.org/10.3390/s22051960
26. de Gois Vasconcelos L et al (2011) Voxel-based morphometry findings in Alzheimer's disease: neuropsychiatric symptoms and disability correlations—preliminary results. Clinics 66:1045–1050. https://doi.org/10.1590/S1807-59322011000600021
27. Ribeiro LG, Busatto Filho G (2016) Voxelbased morphometry in Alzheimers disease and mild cognitive impairment: Systematic review of studies addressing the frontallobe. Dementia Neuropsychol 10:104–112 (2016). https://doi.org/10.1590/s1980-5764-2016dn1002006
28. Acosta-Cabronero J, Williams GB, Pereira JMS, Pengas G, Nestor PJ (2008) The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel-based morphometry. Neuroimage 39:1654–1665. https://doi.org/10.1016/j.neuroimage.2007.10.051
29. Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS (2012) A whole brain FMRI atlas generated via spatially constrained spectral clustering. Hum Brain Mapp 33:1914–1928. https://doi.org/10.1002/hbm.21333
30. Gürlek E, Gökbay IZ (2023) Classification of patients with dementia and Parkinson's with support vector machine. In: 2023 Innovations in intelligent systems and applications conference (ASYU), pp 1–4 IEEE, Sivas, Türkiye. https://doi.org/10.1109/ASYU58738.2023.10296584
31. Perlman G et al (2017) Cortical thickness is not associated with current depression in a clinical treatment study. Hum Brain Mapp 38:4370–4385. https://doi.org/10.1002/hbm.23664
32. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A (2010) The SRI24 multichannel atlas of normal adult human brain structure. Hum Brain Mapp 31:798–819. https://doi.org/10.1002/hbm.20906
33. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
34. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
35. Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
36. Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley, New York. https://doi.org/10.1002/0471722146
37. Clarkson MJ et al (2011) A comparison of voxel and surface based cortical thickness estimation methods. Neuroimage 57:856–865. https://doi.org/10.1016/j.neuroimage.2011.05.053
38. Goto M et al (2022) Advantages of using both voxel- and surface-based morphometry in cortical morphology analysis: a review of various applications. Magn Reson Med Sci 21:41–57. https://doi.org/10.2463/mrms.rev.2021-0096
39. Gupta Y et al (2019) Early diagnosis of Alzheimer's disease using combined features from voxel-based morphometry and cortical, subcortical, and hippocampus regions of MRI T1 brain images. PLoS ONE 14:e0222446. https://doi.org/10.1371/journal.pone.0222446
40. Zhang F et al (2019) Voxel-based morphometry: Improving the diagnosis of Alzheimer's disease based on an extreme learning machine method from the ADNI cohort. Neuroscience 414:273–279. https://doi.org/10.1016/j.neuroscience.2019.05.014
41. Bede P, Hardiman O (2014) Lessons of ALS imaging: Pitfalls and future directions—a critical review. NeuroImage Clinical 4:436–443. https://doi.org/10.1016/j.nicl.2014.02.011
42. Sakurai K et al (2017) Beyond the midbrain atrophy: wide spectrum of structural MRI finding in cases of pathologically proven progressive supranuclear palsy. Neuroradiology 59:431–443. https://doi.org/10.1007/s00234-017-1812-4
43. Garg N, Choudhry MS, Bodade RM (2023) A review on Alzheimer's disease classification from normal controls and mild cognitive impairment using structural MR images. J Neurosci Methods 384:109745. https://doi.org/10.1016/j.jneumeth.2022.109745
44. Leandrou S, Petroudi S, Kyriacou PA, Reyes-Aldasoro CC, Pattichis CS (2018) Quantitative MRI brain studies in mild cognitive impairment and Alzheimer's disease: a methodological review. IEEE Rev Biomed Eng 11:97–111. https://doi.org/10.1109/rbme.2018.2796598
45. Meeter LH, Kaat LD, Rohrer JD, Van Swieten JC (2017) Imaging and fluid biomarkers in frontotemporal dementia. Nat Rev Neurol 13:406–419. https://doi.org/10.1038/nrneurol.2017.75
46. Ranganathan L et al (2018) Changing landscapes in the neuroimaging of dementia. Ann Indian Acad Neurol 21:98. https://doi.org/10.4103/aian.AIAN_48_18
47. Kitchenham B (2004) Procedures for performing systematic reviews—joint technical report. Report Keele University Technical Report TR/SE-0401, National ICT Australia Technical Report 0400011T.1, Software Engineering Group—Department of Computer Science—Keele University and Emprical Software Engineering—National ICT Australia Ltd. https://www.inf.ufsc.br/aldo.vw/kitchenham.pdf
48. Page MJ et al (2021) The Prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372. https://doi.org/10.1136/bmj.n71
49. Page MJ et al (2021) Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 372. https://doi.org/10.1136/bmj.n160
50. Nemoto K et al (2021) Differentiating dementia with Lewy bodies and Alzheimer's disease by deep learning to structural MRI. J Neuroimaging 31:579–587. https://doi.org/10.1111/jon.12835

51. Rajagopalan V, Pioro EP (2024) Differing patterns of cortical grey matter pathology identified by multifractal analysis in UMN-predominant ALS patients with and without corticospinal tract hyperintensity. J Neurol Sci 459:122945. https://doi.org/10.1016/j.jns.2024.122945
52. Krüger J, Opfer R, Spies L, Hedderich D, Buchert R (2024) Voxel-based morphometry in single subjects without a scanner-specific normal database using a convolutional neural network. Eur Radiol 34:3578–3587. https://doi.org/10.1007/s00330-023-10356-1
53. Zhang J et al (2020) An SBM-based machine learning model for identifying mild cognitive impairment in patients with Parkinson's disease. J Neurol Sci 418:117077. https://doi.org/10.1016/j.jns.2020.117077
54. Zhang J et al (2021) Identifying Parkinson's disease with mild cognitive impairment by using combined MR imaging and electroencephalogram. Eur Radiol 31:7386–7394. https://doi.org/10.1007/s00330-020-07575-1
55. Solana-Lavalle G, Rosas-Romero R (2021) Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. Comput Methods Programs Biomed 198:105793. https://doi.org/10.1016/j.cmpb.2020.105793
56. Meyer S et al (2017) Predicting behavioral variant frontotemporal dementia with pattern classification in multi-center structural MRI data. Neuroimage Clin 14:656–662. https://doi.org/10.1016/j.nicl.2017.02.001
57. Ikemitsu N et al (2024) Determination of Alzheimer's disease based on morphology and atrophy using machine learning combined with automated segmentation. Acta Radiol 65:359–366. https://doi.org/10.1177/02841851231218384
58. Huang H et al (2023) Voxel-based morphometry and a deep learning model for the diagnosis of early Alzheimer's disease based on cerebral gray matter changes. Cereb Cortex 33:754–763. https://doi.org/10.1093/cercor/bhac099
59. Egger K et al (2020) Automated voxel- and region-based analysis of gray matter and cerebrospinal fluid space in primary dementia disorders. Brain Res 1739:146800. https://doi.org/10.1016/j.brainres.2020.146800
60. Bachli MB et al (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
61. Bellon E et al (1986) MR artifacts: a review. Am J Roentgenol 147:1271–1281. https://doi.org/10.2214/ajr.147.6.1271. (**PMID: 3490763**)
62. Smith T, Nayak K (2010) MRI artifacts and correction strategies. Imaging in Medicine 2:445–457. https://doi.org/10.2217/iim.10.33
63. Reuter M et al (2015) Head motion during MRI acquisition reduces gray matter volume and thickness estimates. Neuroimage 107:107–115. https://doi.org/10.1016/j.neuroimage.2014.12.006
64. Callaert DV, Ribbens A, Maes F, Swinnen SP, Wenderoth N (2014) Assessing age-related gray matter decline with voxel-based morphometry depends significantly on segmentation and normalization procedures. Front Aging Neurosci 6. https://doi.org/10.3389/fnagi.2014.00124
65. Yoshida N et al (2022) Motion correction in MR image for analysis of VSRAD using generative adversarial network. PLoS ONE 17:1–18. https://doi.org/10.1371/journal.pone.0274576
66. Alzubaidi L et al (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8:53. https://doi.org/10.1186/s40537-021-00444-8
67. Schulz M-A et al (2020) Different scaling of linear models and deep learning in Ukbiobank brain images versus machine-learning datasets. Nat Commun 11:4238. https://doi.org/10.1038/s41467-020-18037-z
68. Huppertz H-J, Kröll-Seger J, Klöppel S, Ganz RE, Kassubek J (2010) Intra- and interscanner variability of automated voxel-based volumetry based on a 3d probabilistic atlas of human cerebral structures. Neuroimage 49:2216–2224. https://doi.org/10.1016/j.neuroimage.2009.10.066
69. Odusami M, Maskeliūnas R, Damaševičius R (2023) Optimized convolutional fusion for multimodal neuroimaging in Alzheimer's disease diagnosis: Enhancing data integration and feature extraction. J Person Med 13. https://doi.org/10.3390/jpm13101496
70. Brand L et al (2020) Predicting longitudinal outcomes of Alzheimer's disease via a tensor-based joint classification and regression model. In: Pacific symposium on biocomputing 2020. Hawaii, USA, pp 7–18. https://doi.org/10.1142/9789811215636_0002
71. Kline A et al (2022) Multimodal machine learning in precision health: a scoping review. Npj Digit Med 5:171. https://doi.org/10.1038/s41746-022-00712-8
72. Liu M, Cheng D, Yan W (2018) Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-pet images. Front Neuroinform 12:35. https://doi.org/10.3389/fninf.2018.00035
73. Singh Chhabra G, Guru A, Rajput BJ, Dewangan L, Swarnkar SK (2023) Multimodal neuroimaging for early Alzheimer's detection: a deep learning approach. https://doi.org/10.1109/ICCCNT56998.2023.10307780
74. Risacher SL et al (2010) Longitudinal MRI atrophy biomarkers: Relationship to conversion in the ADNI cohort. Neurobiol Aging 31:1401–1418. https://doi.org/10.1016/j.neurobiolaging.2010.04.029
75. Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL (2010) Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J Cogn Neurosci 22:2677–2684. https://doi.org/10.1162/jocn.2009.21407

76. Ward A, Tardiff S, Dye C, Arrighi HM (2013) Rate of conversion from prodromal Alzheimer's disease to Alzheimer's dementia: a systematic review of the literature. Dementia and Geriatric Cognit Disorders Extra 3:320–332. https://doi.org/10.1159/000354370

Authors and Affiliations

Emrah Gürlek[1,2] · İnci Zaim Gökbay[3,4]

✉ İnci Zaim Gökbay
inci.gokbay@istanbul.edu.tr; I.Zaim-Gokbay@kent.ac.uk

Emrah Gürlek
emrah.gurlek@ogr.iu.edu.tr; e.gurlek@ahievran.edu.tr

1 Institute of Graduate Studies in Sciences, Informatics, PhD Programme, İstanbul University, İstanbul, Türkiye

2 Department of Business Administration, Kırşehir Ahi Evran University, Kırşehir, Türkiye

3 Department of Artificial Intelligence and Data Engineering, Faculty of Computer and Information Technologies, İstanbul University, İstanbul, Türkiye

4 Institute of Cyber Security for Society (iCSS) and School of Computing, University of Kent, Canterbury CT2 7NP, UK