scientific reports

Check for updates

OPEN

Web based AI-driven framework combining multi-modal data with CNN and LLM for Parkinson's disease diagnosis

Priyadharshini S[1]✉, Ramkumar K[2]✉, Narasimhan K[3], V. B. Surya Prasath[4] & Venkatesh S[3]

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by a wide spectrum of motor and non-motor symptoms, often leading to delayed or inaccurate diagnosis. Conventional diagnostic methods frequently suffer from limited sensitivity, scalability, and interpretability, thereby restricting their utility in clinical settings. To address these limitations, this study presents a novel AI-driven diagnostic framework that integrates multimodal data fusion, deep learning-based classification, and generative language modeling to improve diagnostic accuracy and enable personalized reporting. The proposed framework leverages the Parkinson's Progression Marker Initiative (PPMI) dataset, incorporating structural Magnetic resonance imaging (MRI), Single-Photon Emission Computed Tomography (SPECT) imaging, cerebrospinal fluid (CSF) biomarkers, and clinical assessments. Statistical analysis was employed to select 14 key biomarkers–including dopamine transporter SBR values and CSF protein levels–from a total of 21 features identified as clinically relevant. A 1D Convolutional Neural Network (1D-CNN) was developed and trained using 121 engineered features, comprising radiomic descriptors and biologically derived metrics. Preprocessing and extensive feature engineering were conducted prior to a 70:30 train-test split, with data augmentation applied to the training set to enhance model generalization. The classifier achieved an accuracy of 93.7%, surpassing baseline approaches and emphasizing the value of domain-informed feature design. To improve interpretability and clinician usability, a Mini ChatGPT-4.0 Large Language Model (LLM) was fine-tuned using approximately 1,000 domain-specific prompt-response pairs generated from literature, classifier-derived eXplainable AI (XAI) feature scores, and expert annotations. The generated responses were evaluated using a custom scoring metric (0.0-5.0) based on their semantic alignment with ground truth completions. This LLM module produces patient-specific diagnostic summaries and treatment suggestions. Additionally, a cloud-based interface was developed to facilitate real-time MRI uploads, automated inference, and chatbot-driven consultations. Overall, the framework demonstrates high diagnostic performance, transparency, and user accessibility, offering significant potential for real-world clinical deployment in PD diagnosis and decision support.

Keywords PD diagnosis, Clinical decision support systems, Multimodal data analysis, Fine-tuned LLM, Personalized medical report, Cloud-Based Diagnostic Platform

Parkinson's disease (PD) is a progressive neurological disorder that predominantly affects older adults but can also occur in younger individuals. The condition severely impacts motor coordination, cognitive function, and overall quality of life. Clinical diagnosis of PD often relies on symptom presentation, neuroimaging findings, laboratory evaluations, and response to dopaminergic therapies[1]. However, the variability in symptom onset and progression complicates timely and accurate identification. Although a definitive cure remains elusive, early-stage detection is critical to initiating targeted interventions and mitigating symptom escalation such as cognitive deterioration and motor dysfunction[2]. In this context, artificial intelligence (AI) and machine learning (ML)

[1]Department of Electrical and Electronics Engineering, Saveetha Engineering College, Chennai, Tamil Nadu, India. [2]School of Computing, SASTRA Deemed University, Thanjavur, India. [3]School of Electrical and Electronics Engineering, SASTRA Deemed University, Thanjavur, India. [4]Department of Biomedical Informatics, College of Medicine, University of Cincinnati, Cincinnati, OH 45267, USA. ✉email: spriyadharshini@saveetha.ac.in; ramkumar@eie.sastra.edu

technologies are gaining prominence for their potential to extract subtle diagnostic patterns from heterogeneous patient data, thereby improving diagnostic sensitivity and reliability.

PD presents a growing global health challenge, affecting approximately 1% of individuals over the age of 50 and rising to 2.5% among those over 70. The lifetime risk is estimated at 2.0% for men and 1.3% for women, with nearly 10% of cases occurring before age 50[3]..In the United States alone, around 60,000 new diagnoses are reported annually. The 2019 Global Burden of Disease study highlighted a sharp rise in global PD cases–from 2.5 million in 1990 to 6.2 million by 2017. This number is projected to reach nearly 9.8 million by 2025[4].In India, the current prevalence is estimated at 0.61 million, with a steep upward trajectory[5]. These rising numbers, combined with evidence of regional variations in biomarker expression and disease progression, underscore the need for scalable, population-specific diagnostic solutions. Moreover, a recent survey revealed that 26% of individuals received an incorrect diagnosis prior to a confirmed PD identification, indicating substantial gaps in current clinical screening methodologies[6]. The lack of attention to unique genetic and environmental risk factors in underrepresented populations further complicates early detection efforts.

Recent advances in AI, particularly in natural language processing (NLP), have led to the emergence of powerful tools in healthcare. LLMs like ChatGPT, trained via reinforcement learning from human feedback (RLHF), are capable of generating human-like responses and adapting to varied clinical scenarios. These models are increasingly applied in diagnostics, decision support, and patient interaction, showing potential to synthesize medical knowledge and assist clinicians in real time[7,8].However, their integration into clinical workflows remains limited. Most LLMs operate independently of domain-specific diagnostic systems and lack access to structured, multimodal patient data. This gap restricts their ability to deliver personalized, interpretable outputs–an essential requirement for managing complex, heterogeneous diseases like PD.

To address these limitations, this study introduces a novel cloud-based diagnostic framework that combines a 1D-CNN with a fine-tuned lightweight LLM to improve PD diagnosis and personalized disease management. The system integrates heterogeneous data modalities, including MRI, SPECT, CSF biomarkers, and clinical assessments enabling comprehensive patient profiling. To ensure transparency and trust, the framework incorporates XAI techniques that highlight the most influential features contributing to classification decisions. The core component, a deep learning-based 1D-CNN, processes radiomics features extracted from MRI scans and fuses them with multimodal clinical and biological data to classify PD cases with high accuracy. The salient features identified during classification are passed to a fine-tuned Mini ChatGPT-4.0 model, which generates individualized diagnostic summaries and actionable clinical recommendations. The entire pipeline is deployed via a user-friendly cloud interface that supports real-time MRI uploads, rapid inference, and interactive chatbot consultations–making it scalable, accessible, and suitable for diverse healthcare settings.

Research questions

This study is guided by the following key research questions, aimed at enhancing early detection, interpretability, and user engagement in Parkinson's disease diagnosis:

RQ1: Which combinations of multimodal inputs—such as MRI-based radiomics, CSF biomarkers, and clinical assessment scores—contribute most significantly to accurate PD classification when processed through a 1D-CNN architecture?

RQ2: Can a fine-tuned lightweight LLM, guided by XAI outputs such as SHAP and LIME, effectively generate patient-specific diagnostic narratives and respond meaningfully to queries from clinicians and patients?

RQ3: What is the practical clinical value of deploying this diagnostic framework via a cloud-based platform? Specifically, how does it enable real-time data upload, accelerate diagnostic inference, and provide an interactive, user-friendly experience across diverse healthcare environments?

Contribution

To address the limitations of traditional PD diagnostic approaches, this study presents a novel AI-driven framework that integrates Deep learning (DL), explainable AI, and generative language models. The key contributions of this work are summarized as follows:

1. A 1D-CNN is developed and integrated with explainability techniques such as SHAP and LIME. This architecture enhances both diagnostic accuracy and interpretability, offering clinicians insights into the most influential features involved in classification.
2. The proposed framework fuses heterogeneous inputs, including clinical scores (e.g., UPDRS, MoCA), neuroimaging features (MRI and DaTscan-derived SBR values), and CSF proteins biomarkers. This fusion improves the robustness of the diagnosis and facilitates finer differentiation between PD, prodromal stages, and healthy control subjects.
3. A lightweight LLM (ChatGPT-4.0 Mini) is fine-tuned using structured inputs derived from the classification model and XAI feature scores. The model generates patient-specific diagnostic narratives and answers contextual queries, enabling a human-in-the-loop interaction paradigm in clinical settings.
4. A cloud-accessible platform is implemented, supporting real-time data upload, model inference, and interactive chatbot-based consultations. This enhances usability for both clinicians and patients, particularly in resource-constrained or remote environments.
5. The framework highlights the diagnostic utility of ratio-based features and multimodal correlations, improving early-stage detection and subtype differentiation of PD.

This paper is organized into five sections. Section Related works reviews related works, providing context and insights into existing approaches. Section Proposed Method outlines the proposed methodology in detail. Section Results presents the results and discusses their implications. Finally, Section Discussion concludes the study, highlighting key findings and potential directions for future research.

Related works

Recent advancements in AI have significantly advanced the diagnosis of neurological disorders by integrating neuroimaging, biomarkers, and clinical assessments. Despite this progress, the early detection of PD remains a major challenge due to its heterogeneous symptoms and highly variable disease progression. Traditional diagnostic workflows primarily rely on clinical judgment, observable symptoms, and imaging-based biomarkers. While these approaches are effective in later disease stages, they often fall short in identifying subtle prodromal signs that are critical for early intervention. This limitation underscores the need for data-driven diagnostic frameworks that offer greater sensitivity and reliability, especially during the early, less obvious phases of the disease.

A wide range of ML and DL models have been explored for diagnosing PD, particularly in distinguishing PD patients from healthy controls. Early studies primarily utilized traditional ML classifiers such as Support Vector Machines[9], multi-layer perceptron[10], logistic regression[11] and k-nearest neighbors[12], all of which depended on handcrafted feature extraction. These approaches, however, often suffered from limited generalizability and overlooked subtle diagnostic cues. DL methods addressed this by automatically learning hierarchical features from raw data, leading to improved performance. For instance[13], used a CNN based on AlexNet to classify PD and prodromal cases from MRI scans, achieving 88.9% accuracy. In[14],, showed that 3D-CNN models trained on multi-source MRI data significantly outperformed both 2D CNNs and traditional ML models, highlighting the value of volumetric feature learning. Transformer-based architectures have further advanced medical image analysis by modeling long-range dependencies and contextual relationships. In Alzheimer's Disease research[15], proposed a Regularized Transformer with an adaptive token fusion strategy to aggregate multi-slice MR images. This reduced token redundancy and improved spatial coherence. Additionally, L2-SP regularization was used to retain useful pretrained representations and reduce overfitting especially important for small medical datasets. These techniques are highly relevant to PD, where spatially localized changes and limited data pose significant challenges.

Motivated by this need for robust volumetric analysis[16], presented a hybrid deep learning architecture that combined a 3D-CNN with an enhanced 3D-ResNet. The model was further optimized through Canonical Correlation Analysis (CCA)-based feature fusion and bio-inspired feature selection techniques, ultimately achieving an impressive accuracy of 97.2%. This reinforces the value of integrating anatomical and functional information across modalities for robust diagnostic performance. In[17], a review on DL pipelines for colorectal cancer emphasized factors such as dataset quality, annotation consistency, and interpretability–challenges that are equally pertinent in PD diagnosis. Similarly[18], showcased interpretable CNN models for malaria detection using XAI techniques, reinforcing the growing demand for models that are not only accurate but also transparent and clinician-trustworthy. Despite the successes of single-modality models, their ability to represent the full complexity of PD symptoms remains limited. As emphasized in recent literature, multimodal integration–combining imaging, biomarkers, and clinical data–is essential for achieving comprehensive and accurate classification of PD.

Recent efforts have expanded the boundaries of multimodal learning by aligning visual and textual domains to improve clinical interpretability. In[19]a dual-branch network was introduced that employed large adaptive filters alongside an Aligning Normalized Network (ANNet) to facilitate multi-level alignment between chest X-ray images and associated radiology reports. By leveraging textual priors to guide visual features, the model achieved improved cross-modal representation and interpretability. This approach is particularly valuable in neurodegenerative diagnostics, where integrating neuroimaging with clinical reports or cognitive assessments may enhance decision-making. However, diagnostic models based on a single modality often face limitations due to incomplete representations of disease characteristics. Recognizing this[20], emphasized the importance of multimodal data fusion, demonstrating that combining multiple input streams–such as structural MRI, CSF biomarkers, and clinical scores–can significantly improve the robustness and generalizability of PD classification models. Yet, a persistent barrier to clinical translation remains: the black-box nature of many deep models.

This has led to increased interest in XAI methods, which help demystify model predictions by providing transparency[21]. Techniques such as LIME and SHAP, as applied in the referenced[22–24], quantify the influence of input features on model outputs and generate interpretable visualizations. These tools empower clinicians to understand and validate AI decisions–crucial for complex, high-stakes diagnoses like PD, where trust and accountability are paramount.

Reinforcing the value of interpretability[25],, proposed a hybrid ensemble approach combining deep networks with Extreme Learning Machines (ELMs), augmented with XAI-based visualization modules. This ensemble was applied to gastrointestinal disease detection, but the methodological framework holds promise for PD and other multifaceted neurological conditions. In a related advancement[26], introduced CTBViT, a compact Vision Transformer architecture optimized for tuberculosis classification. Despite its lightweight structure–featuring modular blocks and randomized classifier heads–the model achieved competitive accuracy on constrained datasets, making it a compelling option for deployment in low-resource or edge environments. Such architectures are directly relevant to real-time PD diagnostics, where balancing performance with efficiency is critical. Taken together, these contributions highlight the emerging consensus: successful clinical AI systems must combine multimodal learning with interpretability and deployment readiness. Yet, a fully integrated pipeline that unifies these strengths–capable of learning from diverse medical data while remaining transparent and lightweight–remains an unmet need in the field.

Transformer-based architectures, especially LLMs, have emerged as pivotal tools in medical AI due to their strengths in contextual reasoning, multimodal fusion, and generative capabilities. Models such as GPT-3.5 and GPT-4 have shown remarkable proficiency in diverse tasks–ranging from clinical summarization to differential diagnosis and interactive patient engagement[27]. A recent mini-review on ChatGPT[28] emphasized its potential to synthesize complex clinical information, especially in neurodegenerative contexts, reinforcing its suitability for early detection and decision support. Furthermore, domain-adapted LLM variants like ChatGPT-4o Mini[29] offer reduced computational overhead and faster inference, positioning them as ideal candidates for embedded or edge-based healthcare solutions.

In biomedical question answering and clinical decision-making, fine-tuned LLMs have demonstrated high precision when navigating structured and semi-structured datasets[30]. The GPT-4 Technical Report[31], further highlights the model's multimodal reasoning capacity–capable of processing textual and visual inputs with near-human comprehension. These features position LLMs as ideal components in diagnostic systems that demand explainability, natural language output, and contextual intelligence.

Bridging the gap between vision and language[19], presented a cross-modal dual-branch network that aligned chest X-ray images with radiology reports using large adaptive filters and normalized embeddings. This design showcased the feasibility of tightly coupling clinical language with imaging features for robust performance. Such hybrid strategies–blending LLMs with visual cues–offer significant promise for neurodegenerative disorders like PD, where multimodal data integration is critical.

On the system level, cloud-based AI infrastructures are increasingly embraced to support scalable and distributed diagnostics. The Cloud-MRI framework[32] for example, integrates 6G communication, edge computing, and blockchain technologies to facilitate secure and real-time MRI data sharing. However, many existing cloud-based platforms still lack personalized diagnostic reasoning and natural interaction capabilities, which are essential for chronic, complex conditions such as PD. A comparative summary of prior research is provided in Table 1, illustrating key methodologies, performance insights, and research gaps. This overview consolidates developments across machine learning, deep learning, XAI, and LLM-based methods, and justifies the need for a unified, multimodal diagnostic pipeline as proposed in this study.

To address these gaps, the present study proposes a comprehensive, cloud-deployable diagnostic ecosystem that synergistically combines deep CNN-based imaging classification, multimodal data fusion, XAI outputs, and LLM-driven report generation. Unlike previous approaches that compartmentalize these modules, our architecture ensures seamless end-to-end integration. Specifically, it supports: (i) ingestion of heterogeneous inputs including radiomics, biomarkers, and clinical scores; (ii) interpretable predictions via LIME/SHAP visualizations; (iii) real-time narrative generation through a fine-tuned LLM interface; and (iv) clinician interaction and feedback through a secure web-based portal. This unified pipeline marks a significant advancement toward scalable, transparent, and context-aware AI for PD diagnosis and beyond.

Proposed method

The proposed framework integrates multimodal data from neuroimaging (MRI, SPECT), CSF biomarkers, and clinical assessments to enhance PD diagnosis. A 1D-CNN model is employed for classification. A fine-tuned GPT-4o Mini model facilitates medical query analysis, leveraging explainable AI techniques such as LIME and SHAP to ensure clinical interpretability. Additionally, a cloud-based diagnostic system enhances real-time accessibility, integrating secure AI-driven analytics for personalized patient insights. This approach bridges the gap between deep learning, explainability, and interactive AI for scalable and reliable PD diagnostics. Each of the modules is discussed in detail in this section.

Ref No	Year	Method/Approach	Advantages	Limitations & Motivation for This Study
[9–12]	2019–2023	Traditional ML (SVM, MLP, LR, k-NN)	Simple, fast, interpretable	Poor generalization; manual feature engineering → Move to DL to capture complex patterns automatically
[13]	2020	CNN on MRI (AlexNet)	Learns features from raw data	Lower accuracy (88.9%), no multimodal fusion → Use deeper/multimodal models for better performance
[15]	2025	Regularized Transformer (Alzheimer's)	Adaptive token fusion, L2-SP for regularization	Needs domain-specific adaptation for PD → Inspired multimodal fusion strategy for PD
[16]	2024	3D-CNN + 3D-ResNet + CCA + WOA	Very high accuracy (97.2%), optimized features	No cloud deployment or LLM → Build interpretable cloud system integrating LLM
[18,22,23]	2022–2025	XAI-based CNN (SHAP, LIME)	Improves clinician trust and interpretability	Limited to image features, no language → Fuse XAI scores into report generation via LLM
[26]	2025	CTBViT (lightweight ViT for TB)	Compact, fast, high accuracy	Not specialized for neurodegenerative conditions → Shows feasibility of lightweight models for deployment
[27,28,30,31]	2023–2025	GPT-3.5/4, LLMs for medical tasks	Contextual reasoning, patient-specific output	Lack structured data fusion, no task-specified diagnostics → Fine-tuned LLM on structured PD data for real-world use
Proposed Work	–	1D-CNN + XAI feature score + fine-tuned LLM in cloud platform	High accuracy (93.7%), interpretability, real-time diagnosis, multimodal inputs	Needs broader validation → Unified, interpretable, scalable PD diagnostic tool

Table 1. Comparative summary of existing PD diagnosis methods highlighting key approaches, strengths, limitations, and the motivation for developing the proposed unified AI framework.

Data collection

This study used data from the PPMI, a longitudinal, multinational research initiative focused on identifying optimal biomarkers for the early diagnosis of PD. The data were accessed through the PPMI website Dataset link. The dataset consists of T1-weighted MRI scans from 150 participants comprising 55 PD patients, 45 prodromal subjects, and 50 healthy controls who had undergone both MRI and SPECT imaging during their most recent visits. Additionally, clinical data, specific binding ratio (SBR) features of four striatal regions, and CSF protein markers were collected from control subjects, individuals with PD, and prodromal cases. 'The pre-processing steps include brain-extracted, registered, and intensity-normalized for MRI data, while clinical scores, SBR, and biomarker values were z-score standardized. Data augmentation addressed class imbalance via transformations and resampling. A 70:30 train-test split was used, with 5-fold cross-validation to ensure robustness.

Delineation of multimodal data features

This study uses multimodal data features from diverse sources to enhance the accurate classification and analysis of PD and its related subtypes.

Neuroimaging data

The neuroimaging data comprised MRI and DaT/SPECT scans. MRI scans were acquired using a SIEMENS Prisma 3.0T scanner with both sagittal and axial acquisition. The imaging protocol included a T1-weighted 3D gradient-echo sequence with parameters: TE = 3.0 ms, TI = 900.0 ms, TR = 2300.0 ms, and a flip angle of 9°. The resulting images featured a 1 mm slice thickness, pixel spacing of 1.0 mm in both X and Y dimensions, and a matrix size of $256\times256\times192$. SPECT imaging was performed using a SIEMENS NM detector with a step-and-shoot acquisition method (3° angular steps over a 180° scan arc) and a parallel collimator. A DAT radiopharmaceutical dose of 185 MBq was administered for the imaging procedure. All neuroimaging data were initially provided in DICOM format and subsequently converted to NIfTI format to facilitate 3D analyses across axial, sagittal, and coronal planes. This conversion ensured compatibility with advanced neuroimaging tools and methodologies, supporting comprehensive evaluations in this study.

SBR Values from SPECT

Dopamine Transporter (DaTscan) imaging was conducted using I-123 Ioflupane to quantify SBR values in key striatal regions. These regions, which are central to the pathology of PD, include the right caudate (RC), left caudate (LC), right putamen (RP), and left putamen (LP). SBR values provide critical insights into dopaminergic activity, serving as a vital biomarker for evaluating motor symptoms and tracking disease progression in PD[33].

Biological features (CSF proteins biomarkers)

CSF biomarkers were incorporated into this study to capture the underlying neurochemical changes in PD. Four important key markers were used they are α-synuclein (α-syn), which is associated with PD pathology and neurodegeneration; Amyloid-β_{1-42} (Aβ_{1-42}) is a marker of amyloid plaque formation; total Tau (tTau) indicates neuronal damage; and Phosphorylated Tau (pTau$_{181}$), is indicative of tau pathology. These biomarkers provide critical information for differentiating PD from related neurodegenerative conditions and understanding disease progression[34].

Clinical data

The clinical data in this study include detailed assessments of motor and non-motor symptoms using the Unified Parkinson's Disease Rating Scale (UPDRS) and cognitive evaluations through the Montreal Cognitive Assessment (MoCA). The UPDRS-1 evaluates non-motor experiences of daily living, including mood, behavior, and therapy-related complications, with a range of 0–52. The UPDRS-2 assesses motor experiences of daily living, such as speech, swallowing, and handwriting, also ranging from 0–52. UPDRS-3 focuses on motor examinations, including rigidity, tremors, bradykinesia, posture, and gait, with a range of 0–132, while UPDRS-4 examines motor complications of therapy, with a range of 0–24. Cognitive function was assessed using the MoCA, a standardized test evaluating memory, attention, language, and executive functions. These clinical scores provide a comprehensive understanding of both motor and non-motor symptoms, aiding in the evaluation of disease progression and severity[35]. Table 2 presents the range of values for each multimodal feature, including clinical, neuroimaging, and biomarker data, across control, PD, and prodromal groups, along with demographic details such as age, sex, weight, and height. The range of protein biomarker values is also included, highlighting that α-syn and Aβ_{1-42} exhibit significantly larger values compared to pTau$_{181}$ and tTau, underscoring their importance in disease characterization.

1D-CNN architecture

The 1D CNN classifier has been utilized to distinguish between three classes: PD, Control, and prodromal variants. This model, a type of deep neural network, is specifically designed for processing one-dimensional data, such as time series or sequential datasets. In this study, protein biomarkers, SBR, clinical data, and neuroimaging features are used as input for the model. As illustrated in Fig. 1, the 1D CNN architecture comprises an input layer, followed by three consecutive 1D convolutional layers, and each convolutional layer employs a kernel size of 3 and applies the Leaky ReLU activation function, defined as (1), to introduce non-linearity while avoiding vanishing gradients problems. Mathematically, a 1D convolution operation for the i-th filter can be expressed as in Eq. (2).

$$f(x) = \max(0.01x, x) \tag{1}$$

Variable	PD Subjects	Prodromal Subjects	Control Subjects	p-value
Clinical Data				
Gender (F/M)	25/25	21/29	23/27	>0.05[b]
Age	67.3 ± 13.09	72.39 ± 14.13	74.2 ± 13.85	>0.05[a]
History of PD (Yes/No)	5/45	3/47	0/50	>0.05[b]
Weight	77.8 ± 21.3	75.46 ± 28.22	72.15 ± 18.77	>0.05[a]
Height	168.1 ± 10.1	169.56 ± 10.02	170.6 ± 11.5	>0.05[a]
Heart rate (stand)	71.98 ± 15.011	64.7 ± 12.81	61.56 ± 5.88	>0.05[a]
Heart rate (sit)	80.14 ± 16.91	74 ± 15.504	68.48 ± 5.4	>0.05[a]
UPDRS-1	28.94 ± 15.09	16.14 ± 3.07	-	<0.05[a]
UPDRS-2	28.7 ± 15.92	18 ± 4.29	-	<0.05[a]
UPDRS-3	65.64 ± 36.37	31.4 ± 7.82	-	<0.05[a]
UPDRS-4	14.6 ± 6.16	8.28 ± 2.29	-	<0.05[a]
Total UPDRS score	118.28 ± 40.62	65.54 ± 11.10	-	<0.05[a]
MOCA score	15.04 ± 2.18	21.28 ± 2.20	28.28 ± 1.249	<0.05[a]
DaTscan (SBR Features)				
Right caudate	1.1 ± 0.39	1.61 ± 0.50	2.10 ± 0.83	<0.05[a]
Left caudate	1.26 ± 0.48	1.56 ± 0.47	2.00 ± 0.97	<0.05[a]
Right putamen	0.50 ± 0.21	0.74 ± 0.31	0.92 ± 0.79	>0.05[a]
Left putamen	0.55 ± 0.24	0.71 ± 0.30	0.91 ± 0.74	>0.05[a]
Protein Biomarkers				
α-synuclein (pg/ml)	1927.56 ± 51.47	1989.79 ± 48.8	1874.76 ± 42.7	<0.05[a]
$A\beta_{1-42}$ (pg/ml)	884.5 ± 313.84	1070.7 ± 421.37	964.9 ± 422.6	<0.05[a]
t-Tau (pg/ml)	17.6 ± 66.25	19.80 ± 66.08	18.7 ± 71.96	<0.05[a]
p-Tau_{181} (pg/ml)	14.36 ± 6.31	15.89 ± 5.02	17.27 ± 8.18	<0.05[a]

Table 2. Multimodal biomarker profiling across PD, prodromal, and control groups with statistical significance analysis. **Note:**[a] = ANOVA test, [b] = Chi-square test. (UPDRS: Unified Parkinson's Disease Rating Scale, MoCA: Montreal Cognitive Assessment, SBR: Specific Binding Ratio, DaTscan: Dopamine Transporter Scan).

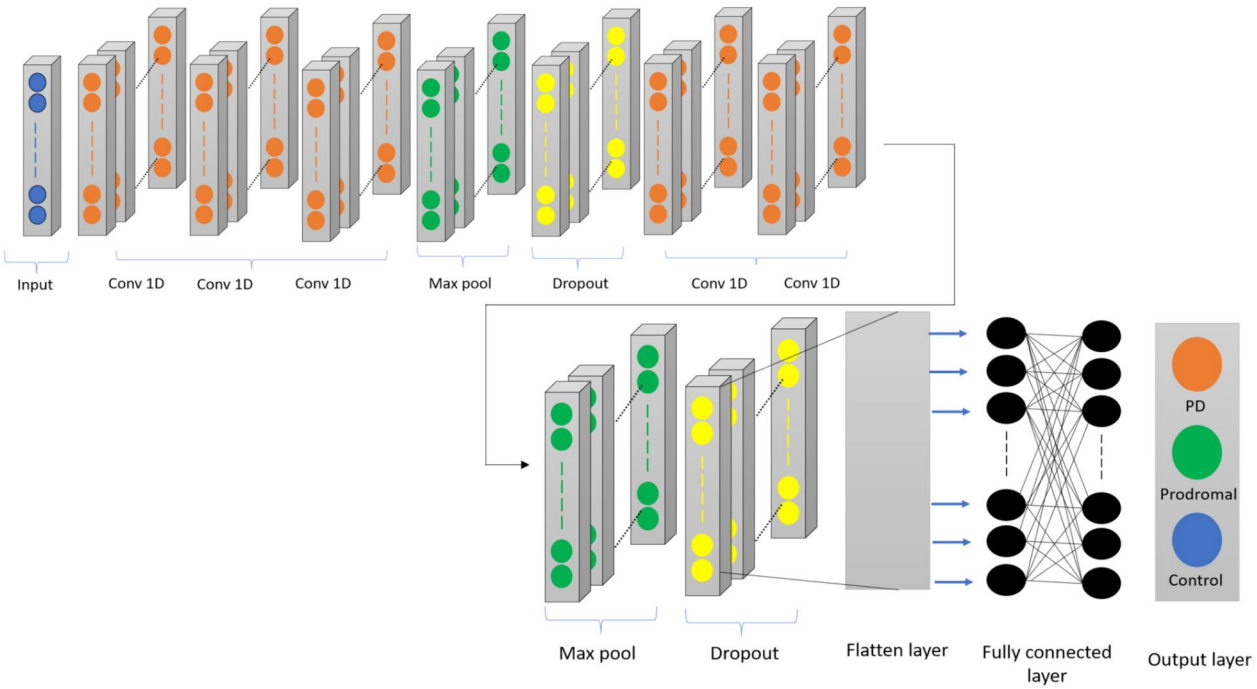

Fig. 1. Schematic representation of the 1D-CNN architecture for multimodal classification of PD stages.

$$y^{(i)}[t] = \sum_{k=1}^{K} w_k^{(i)} \cdot x[t+k-1] + b^{(i)} \tag{2}$$

where $y^{(i)}[t]$ represents the output at position t, $w_k^{(i)}$ are the filter weights, $x[t]$ is the input signal, and $b^{(i)}$ is the bias term. The combination of Leaky ReLU and convolutional operations enhances feature extraction and improves the model's ability to capture complex patterns in sequential data. Max-pooling layers with a pool size of 2 and a stride of 2 follow the convolutional layers to downsample the dimensionality of the feature maps while retaining the most salient features. Dropout layers with a learning rate of 0.5 are used to reduce overfitting. Two additional 1D convolutional layers further refine feature extraction, followed by another pooling layer, dropout layer, flattening operation, and a fully connected layer. The output layer, designed for three-class classification, employs the SoftMax activation function, and the model is optimized using the categorical cross-entropy loss function, defined in Eq. (3).

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}) \tag{3}$$

where N is the number of samples, C is the number of classes, y_{ij} is the true label, and $\hat{y}_{ij}$ is the predicted probability. 'To ensure reproducibility of our 1D-CNN implementation, the following hyperparameters were used: the model includes five 1D convolutional layers with filter sizes of [32, 64, 128, 64, 32], each using a kernel size of 3 and Leaky ReLU activation with $\alpha = 0.01$. Max pooling was applied after each convolutional block using a pool size of 2. Dropout layers with a dropout rate of 0.5 were introduced after the convolutional stack to reduce overfitting. A dense fully connected layer with 128 neurons was used prior to the final output layer. The model was trained using the Adam optimizer with a learning rate of 0.0005, a batch size of 16, and a maximum of 100 epochs. Early stopping was enabled with a patience of 10 epochs. The loss function used was categorical cross-entropy, and SoftMax was used for multi-class prediction in the output layer. All experiments were conducted in a Python 3.10 environment using PyTorch 1.13.1 with CUDA 11.7. The system configuration included a 64-bit Intel(R) Xeon W-2255 CPU @ 3.70 GHz with 128 GB RAM, and these environment settings have been detailed for reproducibility.

Fine-tuned large language model for PD analysis

LLM are advanced AI systems designed to understand and generate human-like text, enabling automation of tasks such as language comprehension, content creation, and domain-specific problem-solving in fields like healthcare. ChatGPT-4o Mini is a compact version of GPT-4, optimized for handling text and image inputs. This model leverages the Transformer architecture with self-attention mechanisms and layered Transformer blocks to extract meaningful patterns from sequential data. By employing techniques like parameter pruning and quantization, this model achieves high efficiency, making it well-suited for environments with limited computational resources, such as edge devices. In this study, ChatGPT-4o Mini has been fine-tuned to enhance its contextual understanding of tasks related to PD diagnosis. Fine-tuning involves adapting a pre-trained model using task-specific datasets to improve its performance, as described by Ouyang[36]. This process often incorporates supervised learning and RLHF, which helps align the model's responses with user expectations while boosting task-specific accuracy. For PD diagnosis, the fine-tuned LLM was trained on multimodal data that included patient-specific features, XAI feature scores derived from classification models, and relevant PD research. This integration equips the model to extract meaningful insights and generate precise interpretations tailored to the needs of PD diagnosis. The fine-tuned LLM creates standardized patient reports that combine diagnostic data, extracted features, and their clinical interpretations. It also enables personalized user interactions by addressing queries with responses customized to individual health records and XAI insights. As illustrated in Fig. 2, this framework aids clinicians in decision-making by leveraging multimodal data analysis and providing personalized recommendations for PD diagnosis.

Cloud-based interactive health inquiry system

Cloud-based applications for medical imaging offer more convenience. They incorporate advanced DL tools for early diagnosis and provide explainable results, effectively bridging the gap between AI and clinical practice. In[37], authors demonstrated that cloud-based frameworks have proven highly effective in deploying AI models for diagnosing diseases from medical images. Building on this, the proposed cloud-based platform allows patients to effortlessly upload MRI scans for detecting PD. This system not only simplifies tracking disease progression but also supports proactive management of the condition. By securely storing diagnostic outcomes and treatment histories, it ensures personalized care and enhances accessibility. Real-time data collection enhances the system's accuracy, continuously refining its diagnostic capabilities through insights from real-world cases. This dynamic approach ensures precise PD mapping and fosters personalized, convenient care for patients.

Results

Statistical analysis

Statistical analysis was performed using SPSS software to evaluate the multimodal data summarized in Table 2. Descriptive statistics, such as mean and standard deviation, were calculated for all variables. Categorical data were analyzed using chi-square tests, while hypothesis tests like ANOVA were applied to numerical variables to assess their significance. A p-value threshold of 0.05 was used to determine statistical significance for each biomarker

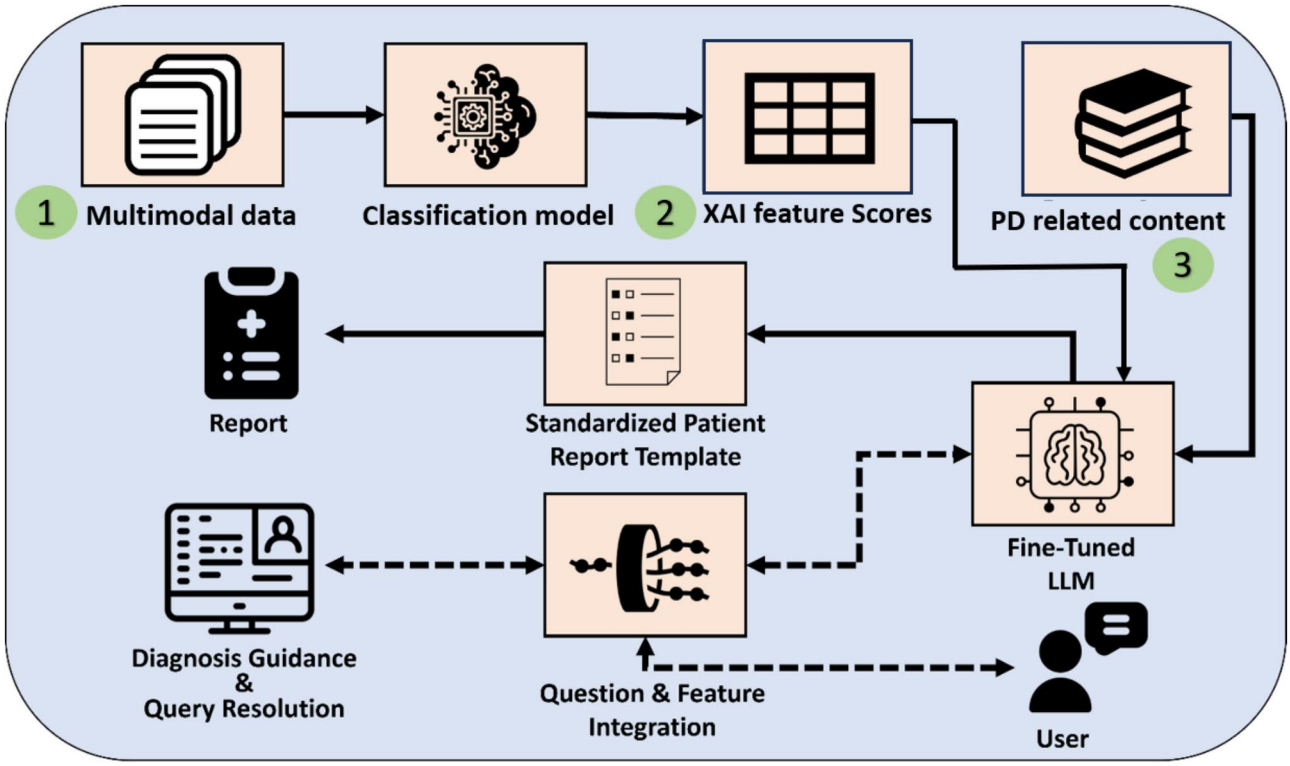

Fig. 2. Integration framework of fine-tuned LLM for PD diagnosis and clinical decision support.

across the cohorts (PD, prodromal, and control). Features with p-values greater than 0.05 were considered statistically insignificant and excluded from further analysis. Out of the 21 features analyzed, 14 were found to be statistically significant and selected for the next phase of the study. These included clinical data (e.g., UPDRS scores, MoCA scores), DaTscan-derived SBR features, and protein biomarkers (α-synuclein, Aβ_{1-42}, t-tau, and p-tau). For example, UPDRS scores and most protein biomarkers showed significant differences between cohorts ($p < 0.05$), whereas features like height and heart rate did not meet the significance threshold. This focused approach ensured that only the most relevant features were used for subsequent PD classification and analysis. 'The statistical significance testing in this section supports the feature selection process shown in Table 2. Additionally, the model performance comparisons in Table 3 were validated using 5-fold cross-validation, yielding low variance across folds, thereby reducing the need for additional pairwise significance testing.

Radiomics features from MRI scan

The MRI scans underwent several pre-processing steps, as shown in Fig. 3. A detailed explanation of the process, along with the feature extraction, is documented in the previous work[38]. Radiomics features in neuroimaging provide detailed quantitative data from medical images, capturing subtle changes in brain structure and function. These features were extracted from the subcortical regions for all three classes: PD, control, and prodromal. A total of 107 radiomics features were collected for each of the 150 participants.

Performance evaluation of 1D-CNN classifier

In this study, 121 multimodal features were utilized for classification, including 14 features from SPECT, CSF proteins, and clinical data, as well as 107 radiomics features extracted from MRI scans of 150 patients. These features underwent pre-processing steps, including min-max normalization, to address variations in feature ranges. This normalization technique scaled all feature values to a range of 0 to 1, improving the model's ability to identify relationships and enhancing both accuracy and reliability. The formula for min-max normalization is given in Eq. (4):

$$X_i' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{4}$$

Feature engineering was applied to create five ratio-based features:

- P-tau$_{181}$/Total-tau
- Total-tau/Aβ_{1-42}
- P-tau$_{181}$/Aβ_{1-42}
- Right caudate/Left caudate
- Right putamen/Left putamen

S.no	Features	Split	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
1	SPECT + CSF proteins + Clinical data	1	88.2	90.6	87.8	91.6
		2	89.5	89.8	88.4	90.8
		3	90.6	90.2	89.2	89.7
		4	89.8	92.4	87.3	91.8
		5	90.9	89.3	89.4	90.5
		Avg.	**89.8**	**90.4**	**88.4**	**90.8**
2	MRI data	1	90.2	90.5	90.8	92.3
		2	88.4	91.8	92.5	93.2
		3	89.9	90.6	91.4	93.8
		4	91.8	90.9	89.9	91.7
		5	90.7	90.3	92.3	90.6
		Avg.	**90.2**	**90.8**	**91.3**	**92.3**
3	SPECT + CSF proteins + Clinical data + MRI data	1	92.4	92.2	94.5	93.6
		2	93.5	91.6	92.8	92.2
		3	92.3	90.8	93.9	94.8
		4	91.8	89.7	91.8	95.7
		5	93.8	92.8	92.5	90.8
		Avg.	**92.7**	**91.4**	**93.1**	**93.4**
4	All + Ratio based features	1	92.9	94.8	94.3	95.1
		2	93.2	94.7	95.8	94.4
		3	94.4	94.5	96.7	93.6
		4	93.6	95.4	95.2	94.8
		5	94.8	93.2	95.1	95.9
		Avg.	**93.7**	**94.5**	**95.4**	**94.7**

Table 3. Comparative Performance of Multimodal Feature Combinations for Multiclass Classification of PD.

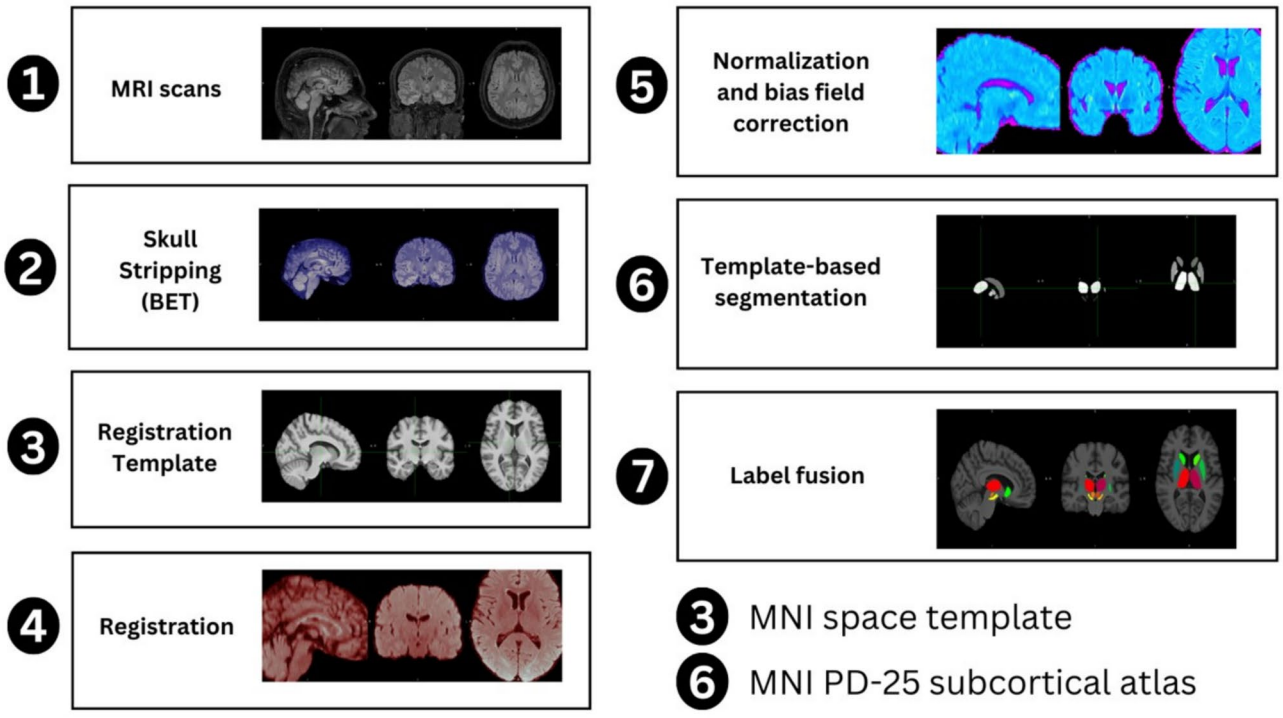

Fig. 3. MRI Preprocessing and Subcortical Structure Segmentation Pipeline for PD Analysis.

These ratios were derived by comparing high- and low-value features. These derived features highlight significant relationships between biomarkers and structural measures, aiding in the early detection of PD and prodromal conditions[39,40]. To address data imbalance, data augmentation was applied to the minority class within the training dataset, ensuring balanced representation and improving classification performance. 'Specifically, a 70:30 stratified train-test split was performed first, and Synthetic Minority Oversampling Technique (SMOTE) was then applied only to the training data. This ensured that the test set remained completely untouched and unbiased, preserving the reliability of the model's performance evaluation. Table 2 reflects the original class distribution before augmentation. The multimodal data, after pre-processing and feature engineering, was fed into the 1D-CNN classifier. The dataset was split into 70% training data and 30% testing data, with augmentation performed only on the training data to generate synthetic samples for the minority class. The architecture of the 1D-CNN classifier, as detailed in Section 1.3, processes the multimodal inputs for multiclass classification, distinguishing between PD, prodromal, and control groups. The complete workflow of the classification process is illustrated in Fig. 4.

Table 3 highlights the performance comparison of various feature combinations for distinguishing between control, PD, and prodromal stages in a multiclass classification. The evaluation metrics, including accuracy, precision, recall (sensitivity), and F1-Score, are employed to assess the effectiveness of the classification model. These metrics provide a comprehensive understanding of the model's ability to classify instances across all classes for accurate diagnosis of PD progression. 'Combining all features (SPECT, CSF proteins, clinical data, and MRI) improved accuracy to 92.7%. The addition of ratio-based features further boosted the performance, achieving the highest accuracy of 93.7%, along with enhanced recall, precision, and F1-score. The mathematical expressions of the evaluation metrics are discussed below in Eq. (5) to Eq. (8). Here, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

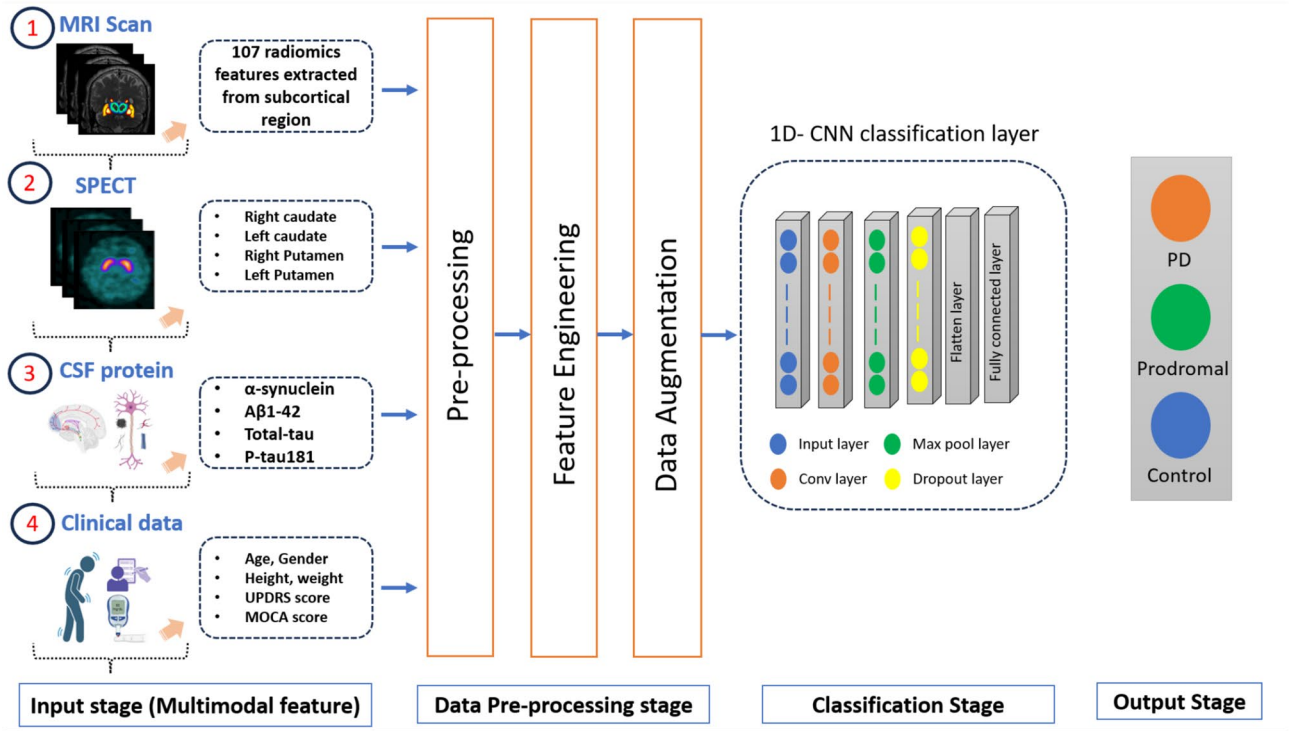

Fig. 4. Overview workflow of the 1D-CNN classifier model for PD classification using multimodal data.

eXplainable AI (XAI) feature score

The XAI feature scores extracted from the 1D-CNN model quantify the importance of individual features across various instances and cohorts. Techniques like SHAP and LIME, provide a transparent view of model's decision-making process. This analysis highlights the most influential features driving predictions, improving interpretability, and validating the model's outcomes. By identifying critical features, XAI facilitates fine-tuning and integration into higher-level layers of the LLM. This approach creates a seamless connection between feature-level insights and contextual predictions, enhancing the reliability and trustworthiness of the model's outputs. XAI strengthens the interpretability of AI-driven diagnostic systems, supporting informed clinical decisions. 'A detailed SHAP and LIME-based interpretability analysis was already conducted in our prior study[38], where top multimodal features were visualized and discussed extensively. To avoid redundancy, we have not repeated similar figures in this manuscript. Instead, we utilize the XAI-derived scores as direct inputs to the fine-tuned LLM, enabling it to generate personalized diagnostic narratives grounded in interpretable model outputs.

LLM-based diagnosis guidance and query resolution

This study utilized a fine-tuned version of OpenAI's ChatGPT-4.0 Mini, customized with datasets specific to PD, to analyze key features for medical diagnosis prediction. The model aims to empower both patients and clinicians by offering informed guidance for initiating or managing treatment options. To enhance the model's performance for PD-specific analysis, a fine-tuning process was undertaken. This involved a comprehensive review of medical literature to identify critical diagnostic features, which were curated into a specialized dataset encompassing diverse medical cases and data types. The fine-tuning process required configuring an advanced environment, including the installation of Python dependencies such as openai, pandas, and matplotlib, and enabling GPU acceleration to reduce training time significantly. Multimodal data, PD-related research papers and XAI feature scores were utilized. Since feeding raw documents and textbooks into the model was not feasible, a prompt-completion method was adopted to structure the data effectively. The ScaleXI tool was used to streamline dataset extraction and organization, ensuring efficient and accurate data preparation. 'To generate prompt-completion pairs, PD-related content was rewritten into clinical Q&A form, while key numerical features (e.g., MoCA, UPDRS, SHAP scores) were embedded as contextual inputs. Each pair was stored in JSONL format following OpenAI standards. These steps ensured structured, reproducible fine-tuning of the LLM. Ground truth data were obtained from the "Expanded Library for Parkinson's Disease Prompts" a rich resource offering validated tools for monitoring PD, educational materials for patients and caregivers, and insights from leading specialists on symptom management and treatment innovations[41,42]. This library also includes diverse media formats, such as books, queries, and videos, aimed at improving the quality of life for individuals with PD.

By integrating the fine-tuned LLM model with meticulously prepared datasets and automated acquisition via ScaleXI, the study achieved improved diagnostic accuracy and supported informed healthcare decision-making. The model helps clinicians and patients understand the severity of PD and provides actionable guidance for treatment. It supports multimodal inputs, including text, images, and audio, enabling it to address queries comprehensively. Personalized responses are generated by combining user inputs with extracted feature values, ensuring tailored and insightful interactions. As demonstrated in Fig. 5, the fine-tuned model effectively processes patient scans and segmented MRI images, delivering detailed findings and reasoning related to PD-specific features, such as subcortical region intensity and structural anomalies. These personalized responses highlight the model's capability to analyze imaging data and provide valuable insights to aid diagnosis and treatment planning. Figure 6 illustrates the model's ability to integrate radiomic data with DaTscan imaging, significantly enhancing diagnostic precision. It identifies patterns of dopaminergic dysfunction and structural brain changes, ensuring actionable and context-specific responses to user queries. Additionally, the model maintains relevance by filtering unrelated questions; for example, if a user poses a question not related to PD, the model appropriately redirects by confirming its focus on PD-related topics. The integration of radiomics data, clinical insights, and advanced AI techniques underscores the model's transformative potential in PD diagnosis and care, offering clarity, precision, and actionable recommendations across various scenarios.

LLM-driven personalized medical report generation

Figure 7 displays the AI-generated personalized medical report. The website, from the initial page, gathers basic patient details and medical history. Within the user interface of our proposed cloud-based platform, the segmented subcortical region is obtained. Subsequently, a description of the image is generated by the fine-tuned model. Additionally, an LLM model utilizes neuroimaging data and suggests lab tests based on individual health conditions. With the integration of clinical data and neuroimaging analysis, the report displays protein prediction analysis. Clinical impressions are provided based on predictive biomarkers, clinical data, and neuroimaging data. treatment recommendations, including medications and physical activities, are suggested by the LLM model.

Integrated cloud-based comprehensive record management

This study introduces a cloud-based comprehensive record management platform designed to archive patients' records, including medication history and progress over time, enabling efficient tracking and facilitating quality follow-up recommendations. This data is also utilized to enhance the prediction model's accuracy with real-case scenarios periodically. The user interface of the proposed platform is illustrated in Fig. 8. In Fig. 8 (a), a protein prediction-based disease classification method is depicted, providing users with quick access to their PD status, even without neuro-imaging data like MRI scans. Upon acquiring an MRI scan, users can utilize the neuroimaging data analysis tab, as shown in Fig. 8 (b), to upload their scans and obtain status updates. Subsequently, users can utilize the Parkinson-GPT tab, as shown in Fig. 8 (c), leveraging a fine-tuned LLM

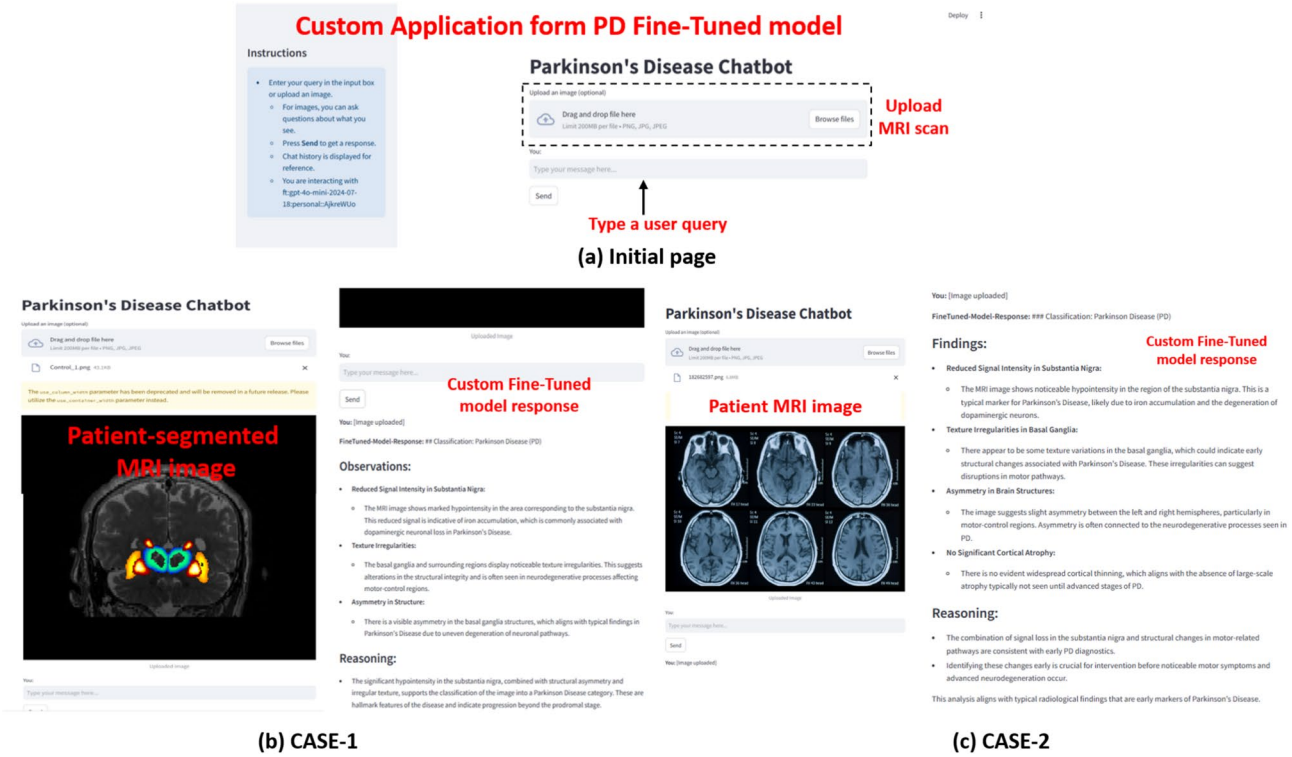

Fig. 5. Interactive Interface and Case-Based Outputs from the Fine-Tuned PD Language Model.

model to address the users' queries by integrating MRI scan features and user-provided clinical data. The account management tab, as shown in Fig. 8 (d), allows patients to monitor their history and recommendations.

Figure 9 illustrates the outcomes of preprocessing, registration, and image correction processes displayed within the user interface during MRI scan uploads in the framework. The framework accepts 3D MRI scan data in Neuroimaging Informatics Technology Initiative (NIfTI) file format to execute the classification model and derive results.

Discussion

Comparison with models based on multi-modal data

This study integrates 107 radiomics features from MRI, 21 key clinical features, SBR imaging-derived measurements, protein biomarkers, and five ratio-based features generated through feature engineering. These diverse features were processed and fed into a 1D-CNN architecture for classification. Table 4 presents a comparative analysis of our proposed framework with recent state-of-the-art multimodal diagnostic models. While most existing studies primarily report accuracy as the key performance metric, accuracy alone may not sufficiently reflect the diagnostic capability of a model, particularly under class imbalance conditions. To address this, we employed SMOTE-based augmentation to balance the training data and evaluated our model using a broader set of metrics. The proposed model achieves an accuracy of 93.7%, with a precision of 97.2%, recall of 94.4%, and F1-score of 96.5%, all of which outperform most prior works and are more clinically informative than accuracy alone. Moreover, unlike prior studies that often focus on binary classification, our framework is designed to handle a more clinically realistic multi-class problem making the task more challenging but aligned with real-world diagnostic needs. The model also benefits from a richer and more heterogeneous feature set, including radiomics, CSF biomarkers, SBR values, clinical scores, and ratio-based features, which enhances its generalizability across diverse cohorts. In contrast, many of the compared models rely on unimodal or limited feature sets, which may inflate performance in controlled settings but lack robustness in practical deployment scenarios. 'Although our framework achieves strong performance, it is important to acknowledge that the limited sample size (150 subjects) may affect generalizability, and larger-scale validation is planned for future work.

Datatypes for finetuning and testing

A total of 1,000 prompt-completion pairs were curated for fine-tuning. This included 250 manually collected pairs, 250 from the "Expanded Library for Parkinson's Disease Prompts", and 500 generated Q&A pairs using this scalexi package are structured prompt-completion datasets created from contextual inputs. Each prompt is carefully crafted based on predefined question types (e.g., open-ended, yes/no, demographic, classification) to ensure diverse and meaningful responses. The completions are generated using OpenAI's API, ensuring high-quality, contextually relevant answers. The dataset follows a standardized JSON/CSV format, making it suitable for AI model fine-tuning and evaluation. The dataset was structured into 915 pairs for fine-tuning and

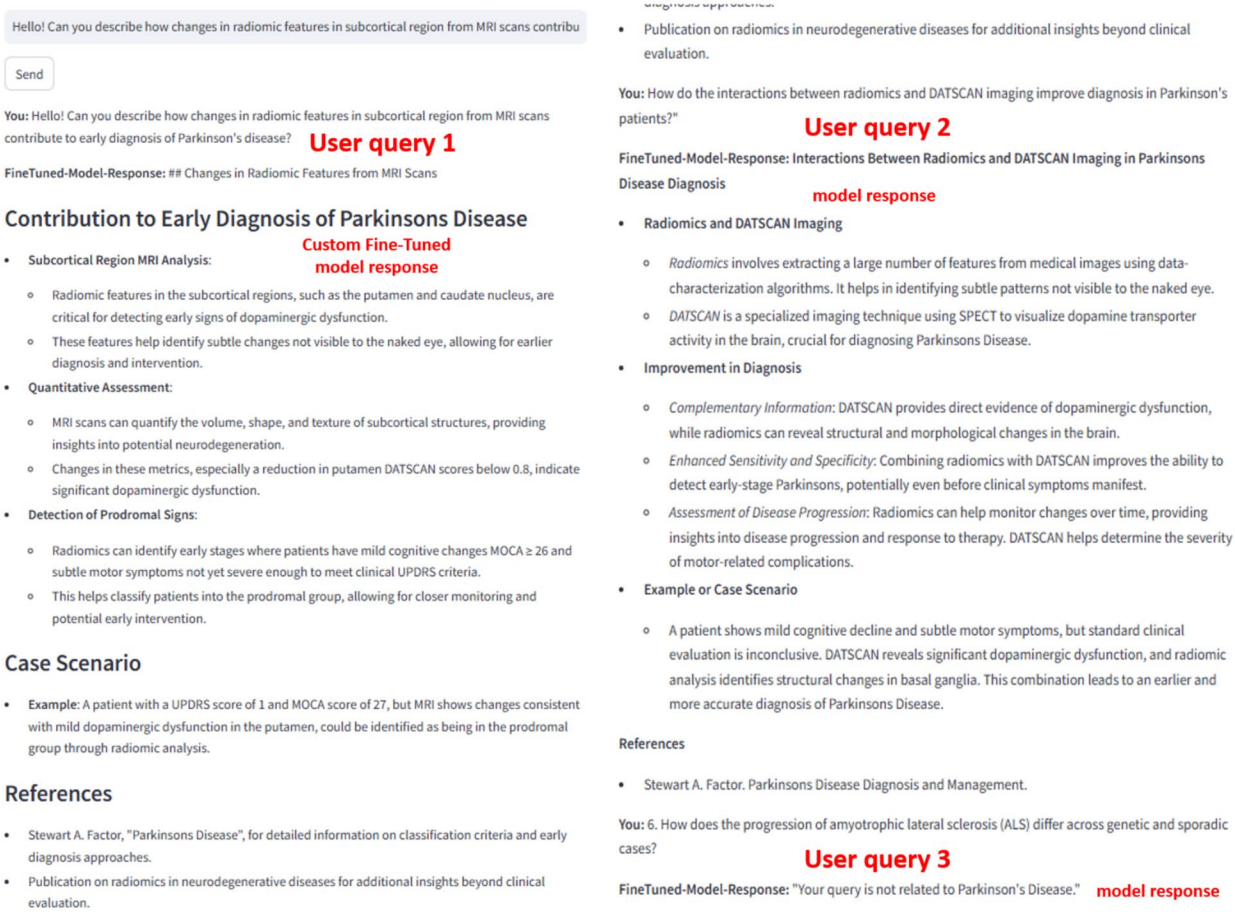

Fig. 6. Text-Based User Interactions and Corresponding Responses from the Fine-Tuned PD LLM.

85 pairs for testing, covering diverse data sources to ensure comprehensive model training and evaluation. Table 5 provides a detailed breakdown of the dataset distribution across different data types.

Parameter fine-tuning LLM

The fine-tuning of LLMs was conducted to enhance response accuracy and minimize hallucinations in patient queries, anchoring the model in curated medical data specific to PD diagnosis. The objective was to refine domain specificity, reliability, and precision, ensuring that the model produces trustworthy and clinically relevant outputs. For consistency, GPT-4 (1 trillion parameters) was fine-tuned using a standardized custom system prompt that instructed the model to function as a concise and accurate medical chatbot for PD-related queries. To optimize learning efficiency, three training epochs were applied. Fine-tuning parameters were carefully selected to balance computational efficiency and model performance. The input and output lengths were capped at 256 tokens each, maintaining a total sequence length of 512 tokens. The model was optimized using the token-averaged cross-entropy loss function and Adam optimizer with a learning rate of 0.0005, ensuring stable adjustments. A batch size of 16 was used to balance memory usage and training efficiency. 'Additionally, temperature (0.3) and repetition penalty (1.2) were configured to enhance response diversity while maintaining factual accuracy. The Scalexi Python library was employed to automate dataset preparation, cost estimation, and model evaluation, significantly simplifying the fine-tuning workflow. This streamlined process enabled efficient optimization, faster iterations, and improved model performance for PD-related tasks. The fine-tuning API was used to handle dataset uploads, initiate fine-tuning jobs, monitor progress, and deploy the fine-tuned model for real-world testing. These optimizations resulted in a specialized LLM capable of delivering precise, clinically relevant responses for PD diagnosis and patient support.

Evaluation method for finetuned LLM response

After fine-tuning the GPT-4o-mini model on a curated dataset of PD-related medical queries, it was crucial to assess its performance in generating accurate and reliable responses. This evaluation was conducted using an automated framework that followed a structured three-step process, leveraging GPT-4 as an evaluator (LLM-based judge).

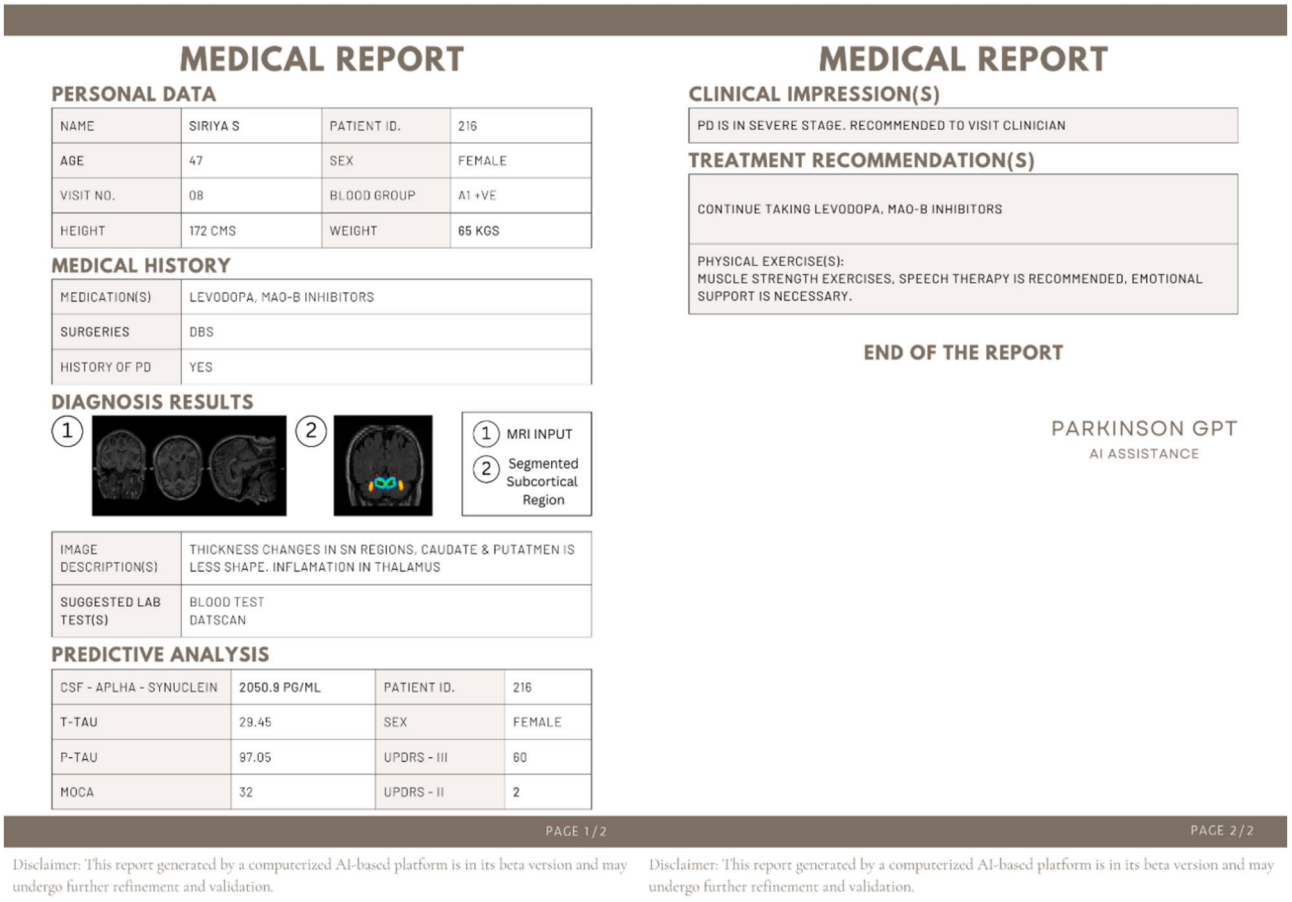

Fig. 7. AI-Generated PD Report Using the Fine-Tuned LLM in a Standardized Clinical Template.

Structured prompt-completion pairs for evaluation

The evaluation dataset was formatted in a structured JSONL format, where each entry contained a medical prompt, a ground truth response, and the model-generated response. This format enabled systematic assessment of the semantic and factual correctness of the generated outputs, as shown in Table 6. The "`prompt`" represents the medical user query, the "`ground_truth_completion`" is the data collected from the PD library and verified as the correct response, and the "`model_completion`" is the response generated by the fine-tuned LLM. The accuracy and reliability of the model were determined by comparing the "`model_completion`" against the "`ground_truth_completion`" using an automated scoring system. The accuracy and reliability of the fine-tuned model were evaluated by assessing the model-generated responses in comparison to the ground truth responses, using a scoring mechanism implemented in GPT-4.

Evaluating using GPT-4 as a judge

To ensure objective assessment, GPT-4 was employed as an LLM-based evaluator, assigning a quantitative score (0.0 - 5.0) to each response, as shown in Table 7. The scoring system evaluated factual correctness, coherence, specificity, and relevance. Each evaluation instance followed a structured scoring prompt template, where GPT-4 was instructed to act as an impartial medical evaluator:

> *"You are a friendly and brilliant medical chatbot, designed to provide concise and accurate answers with regards to all PD-related queries. Given a user prompt, a correct ground truth response, and a generated response, assign a score (0.0 to 5.0) based on factual accuracy, coherence, and specificity. Additionally, provide a concise justification ($\leq$ 50 words) explaining the score. The output must be in CSV format: {"score": value, "score_reason": "justification"}."*

Three-step automated evaluation process

The evaluation framework followed a structured three-step methodology to assess the accuracy and reliability of the fine-tuned model systematically. **Step 1:** The fine-tuned GPT-4o-mini model was tested on the test data, with key generation parameters such as temperature (0.3) and repetition penalty (1.2) optimized for reliability. **Step 2:** The model-generated responses were systematically compared against ground truth answers, ensuring an objective assessment of factual correctness and medical relevance. **Step 3:** GPT-4 was employed as an evaluator, analyzing each generated response and assigning a quantitative score (0.0 - 5.0) based on semantic similarity, factual correctness, and coherence. The final evaluation results were stored in a CSV format, containing five

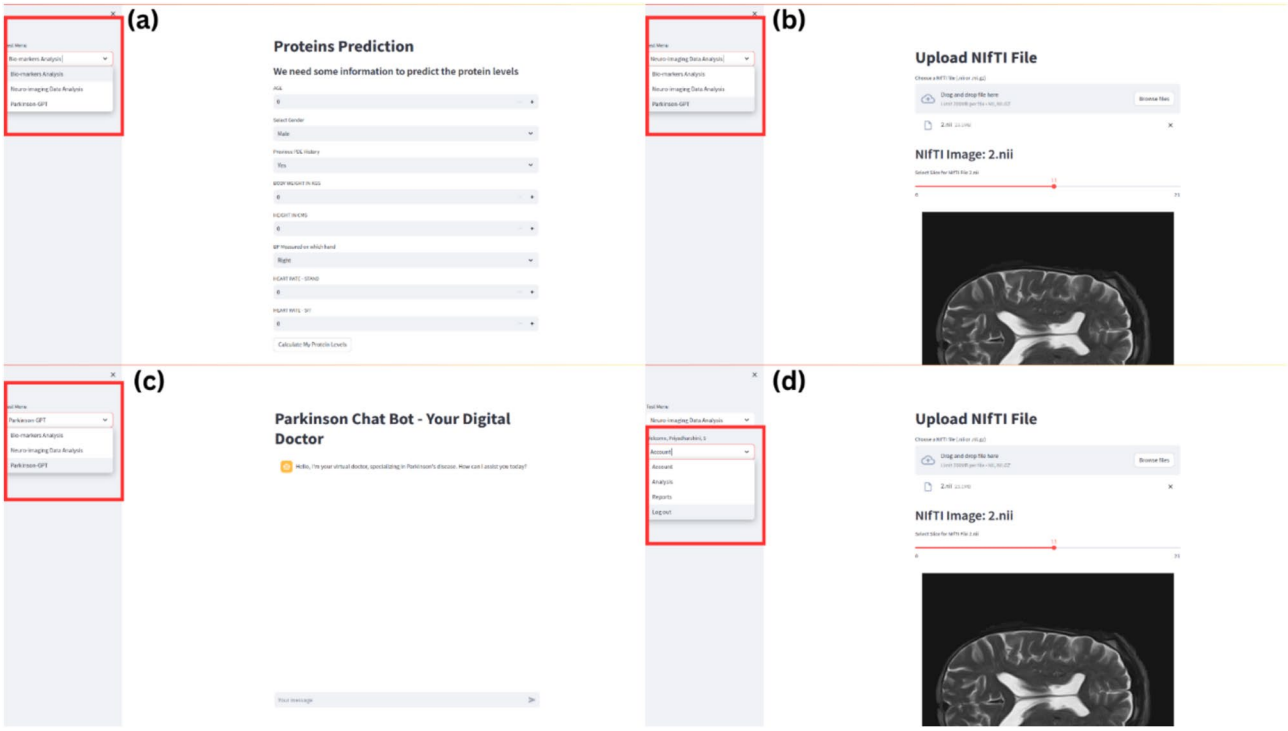

Fig. 8. Graphical User Interface (GUI) of the Proposed Cloud-Based AI Platform for PD Management. The platform integrates multiple modules to streamline clinical and diagnostic workflows: (**a**) protein level prediction using patient-specific clinical inputs, (**b**) neuroimaging analysis through NiFTI file uploads for automated radiomic evaluation, (**c**) an AI-powered chatbot trained for PD-specific interactions and guidance, and (**d**) a centralized navigation panel enabling access to all analytical and support tools within the system.

essential columns: 1. User prompt 2. Ground truth response 3. Model-generated response 4. Assigned score (0.0 - 5.0) 5. Brief justification ($\leq$50 words). This format ensures a compact yet comprehensive evaluation of model performance.

Limitations and managerial implication

While the proposed AI-driven framework demonstrates promising results in integrating multimodal data for PD diagnosis, several limitations warrant consideration. First, the dataset size, although curated with care from validated sources like PPMI, remains modest, which may impact generalizability across broader clinical populations. External validation on larger, more diverse cohorts is necessary to confirm robustness. Additionally, while the 1D-CNN and LLM components were optimized for performance, future comparisons with newer transformer-based or lightweight edge-deployable models could provide further insights. From a deployment standpoint, integration with hospital information systems poses challenges related to data interoperability, privacy regulations (e.g., HIPAA/GDPR), and clinical workflow alignment. Real-time inference via web-based interfaces also introduces latency and reliability concerns in resource-constrained settings. Moreover, regulatory clearance and ethical approval will be crucial before clinical adoption. Despite these limitations, the framework offers substantial value in augmenting diagnostic decisions and providing interpretable, patient-specific insights. It lays the groundwork for future AI-assisted platforms in neurology and personalized medicine.

Conclusion

This study presents a comprehensive AI-driven diagnostic framework for PD that integrates multimodal data fusion, interpretable deep learning, and personalized LLM-based assistance. The proposed methodology combines clinical scores, SPECT-derived SBR values, CSF protein biomarkers, and 107 radiomics features extracted from T1-weighted MRI scans. These imaging data were pre-processed through brain extraction, registration, and intensity normalisation to ensure consistency. Initially, 21 features spanning clinical, imaging, and biological domains were subjected to statistical evaluation, from which 14 significant features were retained. These were combined with the 107 radiomics features and 5 ratio-based engineered features, yielding a total of 126 multimodal inputs. These were used to train a 1D-CNN for multiclass classification. To ensure robustness and fairness, a 70:30 stratified train-test split and 5-fold cross-validation were employed. Class imbalance was mitigated using SMOTE-based data augmentation. The model achieved 93.7% accuracy, along with high precision, recall, and F1-score–surpassing several state-of-the-art baselines.

To enhance interpretability, explainable AI techniques such as SHAP and LIME were used to extract feature importance scores. These scores were then integrated into downstream LLM-based narratives, bridging the

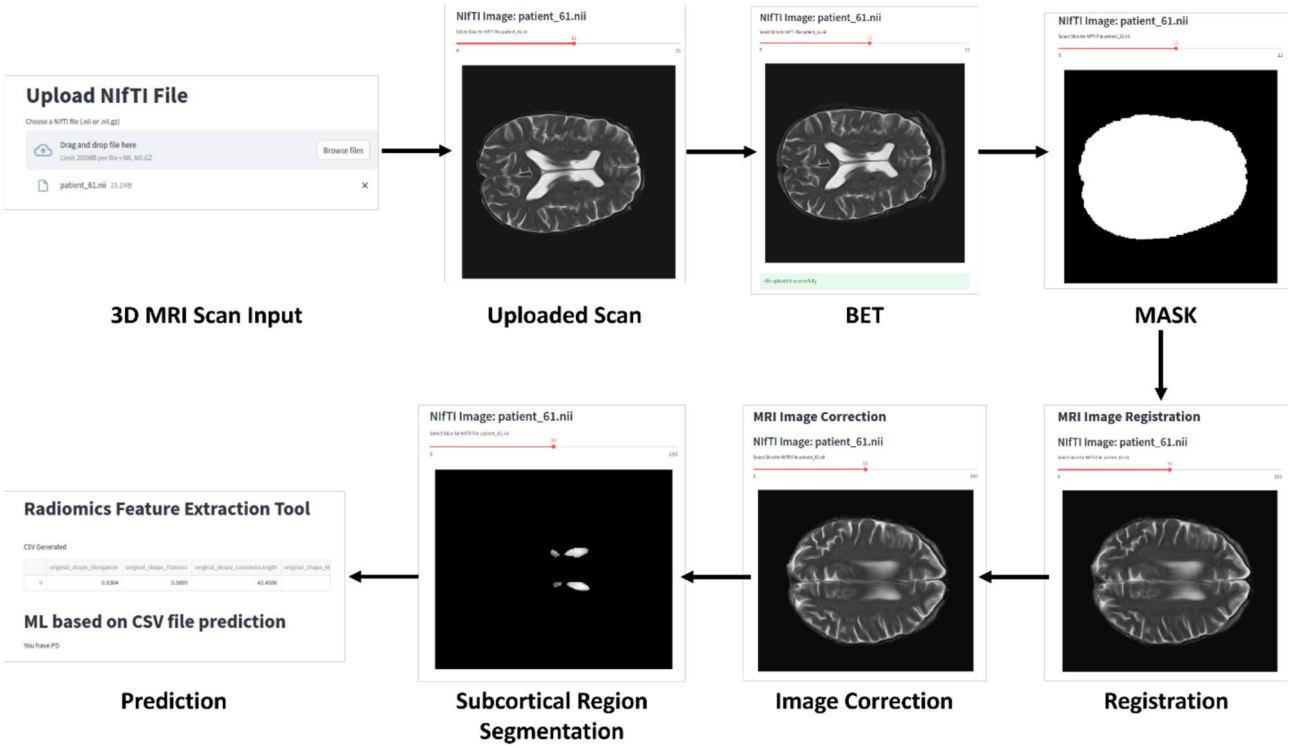

Fig. 9. MRI processing and radiomics-based prediction workflow in the cloud-based PD platform. The interface demonstrates the end-to-end processing pipeline beginning with (i) NiFTI file upload, followed by (ii) image visualization, (iii) segmentation, (iv) brain mask generation, (v) image correction, and (vi) MRI registration. Subsequently, (vii) radiomic features are extracted and exported as a CSV file, which is then used by the integrated machine learning model to predict the presence of PD based on learned feature patterns.

Reference	Year	Classification	Features	Inferences
[43]	2024	PD vs MCI vs Control	SBR + MRI features	Accuracy: 86.79%, Sensitivity: 85.5%, Specificity: 89.9%, AUC: 92.3
[34]	2024	PD vs Healthy vs SWEDD	SBR and biological features	Sensitivity: 96.55%
[33]	2022	PD vs Healthy	MRI + SBR + CSF features	Accuracy: 90.48%, Specificity: 86.05%, F1-score: 92.06%, Geometric Mean: 89.72%, Sensitivity: 93.55%
[35]	2024	PD vs Healthy	SBR + CSF features	AUC: 98%, CI: 0.97–0.99, Sensitivity: 95%, Specificity: 92%
[44]	2024	PD vs Healthy vs SWEDD	Ratio-based features + SBR	AUC: 99.3%
Proposed Method	–	PD vs Healthy vs Prodromal	Ratio-based features + MRI (radiomics) + SBR + CSF + Clinical	Accuracy: 93.7%, Recall: 94.4%, Precision: 97.2%, F1-score: 96.5%

Table 4. Comparative performance analysis of the proposed multimodal diagnostic framework against existing state-of-the-art studies.

Data Type	Q/A Pairs (Fine-tuned Model)	Q/A Pairs (Testing)
Clinical data	194	10
Protein data	174	18
MRI data	151	11
SPECT	196	9
XAI feature score	78	13
PD related content	122	24
Total	**915**	**85**

Table 5. Distribution of multimodal data sources used for multiclass classification and evaluation in PD.

Key	Value
"prompt"	"What are the early symptoms of Parkinson's disease?"
"ground_truth_completion"	"The early symptoms of Parkinson's include tremors, bradykinesia, rigidity, and postural instability."
"model_completion"	"Symptoms of Parkinson's can include hand tremors, stiffness, and slow movement."

Table 6. Sample JSON structure used for fine-tuning.

Score (0.0 - 5.0)	Evaluation Criteria
5.0	Perfect response - fully aligned with the ground truth.
4.0 - 4.9	Mostly correct - minor missing details.
3.0 - 3.9	Partially correct - some factual errors or missing elements.
2.0 - 2.9	Limited relevance - significant omissions or inaccuracies.
1.0 - 1.9	Incorrect but related - contains relevant concepts but is factually incorrect.
0.0 - 0.9	Completely off-topic or misleading response.

Table 7. Scoring rubric for evaluating the quality of model-generated responses.

gap between model prediction and clinical understanding. In parallel, a lightweight ChatGPT-4.0 Mini model was fine-tuned using 1,000 structured prompt-completion pairs derived from curated PD literature, clinical metrics, and XAI-derived insights. The ScaleXI library was used to automate dataset preparation and fine-tuning management. The LLM supports text, image, and audio modalities, enabling personalized diagnostic summaries and clinician-patient interactions. The entire system is deployed on a cloud-based platform offering four key modules: protein-level prediction, neuroimaging upload and analysis, PD-focused chatbot interaction, and centralized patient record management. This setup enables real-time inference, contextual feedback, and long-term monitoring–particularly beneficial for remote or resource-limited clinical environments.

While the framework demonstrates strong diagnostic performance, its current evaluation is limited to a modest dataset of 150 subjects, which may affect generalizability across broader populations. Additionally, the LLM module could benefit from more diverse real-world interaction scenarios to further refine its clinical utility. These results underscore the framework's potential for clinical translation. Future work will explore validation on larger, more diverse cohorts, incorporation of Retrieval-Augmented Generation (RAG) for improved contextual intelligence, and development of agent-based AI modules for adaptive patient engagement and clinical decision support. By integrating explainability, scalability, and personalization, this work sets the foundation for next-generation AI-assisted neurodiagnostics.

Data availability

The training and testing data used in this study is available from the Parkinson's Progression Markers Initiative (PPMI) which is an open-access database.

Received: 14 March 2025; Accepted: 29 September 2025
Published online: 04 November 2025

References

1. Keus, S., Bloem, B., Hendriks, E., Bredero-Cohen, A. & Munneke, M. Evidence-based analysis of physical therapy in parkinson's disease with recommendations for practice and research. *Mov. Disord.* **22**, 451–460. https://doi.org/10.1002/mds.21244 (2007).
2. Clarke, C. Parkinson's disease. *Br. Med. J.* **335**, 441–445. https://doi.org/10.1136/bmj.39289.437454.AD (2007).
3. Surathi, P., Jhunjhunwala, K., Yadav, R. & Pal, P. Research in parkinson's disease in india: A review. *Annals Indian Acad. Neurol.* **19**, 9–20. https://doi.org/10.4103/0972-2327.167713 (2016).
4. Initiative, B. & Disorders, N. The burden of neurological disorders across the states of india: the global burden of disease study 1990–2019. *The Lancet Glob. Heal.* **9**, e1129–e1144. https://doi.org/10.1016/S2214-109X(21)00164-9 (2021).
5. Willis, A. et al. Incidence of parkinson disease in north america. *NPJ Park. Dis.* **8**, Article 10, https://doi.org/10.1038/s41531-022-00410-y (2022).
6. Owolabi, L. Progressive supranuclear palsy misdiagnosed as parkinson's disease: A case report and review of literature. *Annals Med. Heal. Sci. Res.* **3**, 44. https://doi.org/10.4103/2141-9248.121221 (2013).
7. Bhatia, P. Chatgpt for academic writing: A game changer or a disruptive tool?. *J. Anaesthesiol. Clin. Pharmacol.* **39**, 1–2. https://doi.org/10.4103/joacp.joacp_84_23 (2023).
8. Khan, R., Jawaid, M., Khan, A. & Sajjad, M. Chatgpt-reshaping medical education and clinical management. *Pak. J. Med. Sci.* **39**, 605–607. https://doi.org/10.12669/pjms.39.2.7653 (2023).
9. O. Access. Svm-based approach for parkinson's detection using vocal data. *Int. Res. J. Mod. Eng. Technol. Sci.* 253–258, https://doi.org/10.56726/irjmets37926 (2023).
10. Govindu, A. & Palwe, S. Early detection of parkinson's disease using machine learning. *Procedia Comput. Sci.* **218**, 249–261. https://doi.org/10.1016/j.procs.2023.01.007 (2022).
11. Shalaby, M., Belal, N. & Omar, Y. Classification of parkinson's disease. In *Proceedings of [Conference/Workshop name if available]*, 1587–1593 (2021).
12. K.N.N. Algorithm. Prediction of parkinson disease using. *[Journal/Conference name if available]* **6**, 132–135 (2019).

13. Sivaranjini, S. & Sujatha, C. Deep learning based diagnosis of parkinson's disease using convolutional neural network. *Multimed. Tools Appl.* **79**, 15467–15479. https://doi.org/10.1007/s11042-019-7469-8 (2020).
14. Hu, M., Sim, K., Zhou, J., Jiang, X. & Guan, C. Brain mri-based 3d convolutional neural networks for classification of schizophrenia and controls. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 1742–1745, https://doi.org/10.1109/EMBC44109.2020.9176610 (2020).
15. Lu, S., Zhang, Y. & Yao, Y. A regularized transformer with adaptive token fusion for alzheimer's disease diagnosis in brain magnetic resonance images. *Eng. Appl. Artif. Intell.* **155**, https://doi.org/10.1016/j.engappai.2025.111058 (2025).
16. Priyadharshini, S. et al. Bio-inspired feature selection for early diagnosis of parkinson's disease through optimization of deep 3d nested learning. *Sci. Reports* **14**, 1–17. https://doi.org/10.1038/s41598-024-74405-5 (2024).
17. Islam, M., Ahamed, M., Islam, M., Nahiduzzaman, M. & Ahsan, M. Detection, localization, segmentation, and classification in colorectal cancer screening using deep learning: A systematic review. *Biomed. Signal Process. Control.* **110**, https://doi.org/10.1016/j.bspc.2025.108202 (2025).
18. Ahamed, M. et al. Improving malaria diagnosis through interpretable customized cnns architectures. *Sci. Reports* **15**, 1–35. https://doi.org/10.1038/s41598-025-90851-1 (2025).
19. Lu, S., Zhu, Z., Zhang, Y. & Yao, Y. Tuberculosis and pneumonia diagnosis in chest x-rays by large adaptive filter and aligning normalized network with report-guided multi-level alignment. *Eng. Appl. Artif. Intell.* **158**, https://doi.org/10.1016/j.engappai.2025.111575 (2025).
20. Ranson, J. et al. Predictors of dementia misclassification when using brief cognitive assessments. *Neurol. Clin. Pract.* **9**, 109–117. https://doi.org/10.1212/CPJ.0000000000000566 (2019).
21. Skaramagkas, V., Pentari, A., Kefalopoulou, Z. & Tsiknakis, M. Multi-modal deep learning diagnosis of parkinson's disease - a systematic review. *IEEE Transactions on Neural Syst. Rehabil. Eng.* **31**, 2399–2423. https://doi.org/10.1109/TNSRE.2023.3277749 (2023).
22. Achar, C. P., Anvesh, M., Kodipalli, A. & Rao, T. Exploring computational models for parkinson's disease diagnosis: Unveiling insights with lime and shap explainability techniques. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, vol. 1, 1–6, https://doi.org/10.1109/ICKECS61492.2024.10616831 (2024).
23. Nguyen, H. & Byeon, H. Prediction of parkinson's disease depression using lime-based stacking ensemble model. *Mathematics* **11**, https://doi.org/10.3390/math11030708 (2023).
24. Liu, Y., Liu, Z., Luo, X. & Zhao, H. Diagnosis of parkinson's disease based on shap value feature selection. *Biocybern. Biomed. Eng.* **42**, 856–869. https://doi.org/10.1016/j.bbe.2022.06.007 (2022).
25. Ahamed, M. et al. Detection of various gastrointestinal tract diseases through a deep learning method with ensemble elm and explainable ai. *Expert. Syst. with Appl.* https://doi.org/10.1016/j.eswa.2024.124908 (2024).
26. Lu, S., Zhu, Z., Tang, Y., Zhang, X. & Liu, X. Ctbvit: A novel vit for tuberculosis classification with efficient block and randomized classifier. *Biomed. Signal Process. Control.* **100**, https://doi.org/10.1016/j.bspc.2024.106981 (2025).
27. Vaswani, A. et al. Attention is all you need. *In Advances in Neural Information Processing Systems (NeurIPS)* **30**, 5999–6009 (2017).
28. Briganti, G. How chatgpt works: a mini review. *Eur. Arch. Oto-Rhino-Laryngology* **281**, 1565–1569. https://doi.org/10.1007/s00405-023-08337-7 (2024).
29. Thelwall, M. Chatgpt for complex text evaluation tasks. *J. Assoc. for Inf. Sci. Technol.* 1–4, https://doi.org/10.1002/asi.24966 (2024).
30. Tao, K. et al. Fine-tuned large language models for answering questions about full-text biomedical research studies. *medRxiv* https://doi.org/10.1101/2024.10.28.24316263 (2024).
31. OpenAI et al. Gpt-4 technical report. *Tech. Rep.* 2303.08774, arXiv (2023).
32. Zhou, Y. et al. Cloud-magnetic resonance imaging system: In the era of 6g and artificial intelligence. *Magn. Reson. Lett.* 200138, https://doi.org/10.1016/j.mrl.2024.200138 (2024).
33. Pahuja, G. & Prasad, B. Deep learning architectures for parkinson's disease detection by using multi-modal features. *Comput. Biol. Medicine* **146**, https://doi.org/10.1016/j.compbiomed.2022.105610 (2022).
34. Aggarwal, N., Saini, B. & Gupta, S. A deep 1-d cnn learning approach with data augmentation for classification of parkinson's disease and scans without evidence of dopamine deficit (swedd). *Biomed. Signal Process. Control.* **91**, https://doi.org/10.1016/j.bspc.2024.106008 (2024).
35. Xiang, Y. et al. Estimating the sequence of biomarker changes in parkinson's disease. *Park. Relat. Disord.* **118**, https://doi.org/10.1016/j.parkreldis.2023.105939 (2024).
36. Ouyang, L. et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35 (2022).
37. Hroub, N., Alsannaa, A., Alowaifeer, M., Alfarraj, M. & Okafor, E. Explainable deep learning diagnostic system for prediction of lung disease from medical images. *Comput. Biol. Medicine* **170**, https://doi.org/10.1016/j.compbiomed.2024.108012 (2024).
38. Priyadharshini, S. et al. A comprehensive framework for parkinson's disease diagnosis using explainable artificial intelligence empowered machine learning techniques. *Alex. Eng. J.* **107**, 568–582. https://doi.org/10.1016/j.aej.2024.07.106 (2024).
39. Prashanth, R., Roy, S. D., Mandal, P. & Ghosh, S. High-accuracy detection of early parkinson's disease through multimodal features and machine learning. *Int. J. Med. Informatics* **90**, 13–21. https://doi.org/10.1016/j.ijmedinf.2016.03.001 (2016).
40. Hall, S. et al. Csf biomarkers and clinical progression of parkinson disease. *Neurology* **84**, 57–63. https://doi.org/10.1212/WNL.0000000000001098 (2015).
41. Health, D. et al. Datacc by dime expands library of digital measurement products to include parkinson's disease (2024). White paper / report.
42. Gorman, A. Living your best life. *Analysis (United Kingdom)* **81**, 568–576. https://doi.org/10.1093/analys/anab053 (2021).
43. Cheng, J. et al. Alzheimer's disease prediction algorithm based on de-correlation constraint and multi-modal feature interaction. *Comput. Biol. Medicine* **170**, https://doi.org/10.1016/j.compbiomed.2024.108000 (2024).
44. Aggarwal, N., Saini, B. & Gupta, S. Feature engineering-based analysis of datscan-spect imaging-derived features in the detection of swedd and parkinson's disease. *Comput. Electr. Eng.* **117**, https://doi.org/10.1016/j.compeleceng.2024.109241 (2024).

Acknowledgements

The authors are grateful to Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India (IN) for their financial support (CRG/2021/006697) for the project titled 'Design and development of flexible and context-specific AI tool for the early detection and grading of PD' and SASTRA Deemed to be University, India, for providing the necessary resources and support for the successful completion of this research work.

Author contributions

P.S: Writing - review & editing, Writing - original draft, Visualization, Methodology, Investigation, Conceptualization. R.K: Writing - review & editing, Methodology, Formal analysis, Validation, Supervision. N.K: Writing - review & editing, Methodology, Investigation, Formal analysis, Validation, Supervision. V.B.S: Data curation, Investigation, Validation, Visualization. V.S: Formal analysis, Resources.

Declarations

Competing interests

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.

Additional information

Correspondence and requests for materials should be addressed to P.S. or R.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025