International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Explainable Multimodal Deep Learning Framework for Parkinson's Disease Phenotyping and Progression Tracking

Dev Pateriya, Pradeep Jatav
International Institute of Professional Studies, Devi Ahilya Vishwavidyalaya, Indore, India

Abstract - Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder globally, affecting approximately 8.5 million individuals as of 2019 according to the World Health Organization, with projections indicating that this number will exceed 12 million by 2040. Current clinical diagnosis relies predominantly on subjective motor assessments that are prone to inter-rater variability, which can delay accurate diagnosis by several years. This paper presents an Explainable Multimodal Deep Learning Framework that integrates three distinct biomarker modalities—sustained phonation acoustic features, spiral and wave handwriting images, and longitudinal telemonitoring vocal measurements—to accomplish simultaneous Parkinson's disease detection and motor progression tracking. The framework employs a Tabular Voice Encoder built on cascaded linear transformations with Batch Normalization and Gaussian Error Linear Unit (GELU) activations for 22-dimensional acoustic feature classification; a Swin Transformer-based Spatial Image Encoder pre-trained on ImageNet-21k for handwriting analysis; and a two-layer bidirectional-compatible Temporal Attention-LSTM that models the progression of Unified Parkinson's Disease Rating Scale (UPDRS) scores across longitudinal voice recordings. A homoscedastic uncertainty-weighted multi-task loss function simultaneously optimizes binary disease classification, continuous UPDRS regression via Huber loss, and ordinal Hoehn and Yahr stage prediction. Modality outputs are combined through a weighted late-fusion strategy supported by a stacked logistic regression meta-learner. Model validation employs subject-disjoint Leave-One-Subject-Out cross-validation (LOSO-CV) for acoustic features and five-fold stratified cross-validation for imaging modalities, thereby eliminating subject-level data leakage. Explainability is achieved through SHapley Additive Explanations (SHAP) with a GradientExplainer backend for tabular voice features, and Integrated Gradients computed via the Captum library for spatial attribution within handwriting images. The proposed framework addresses the clinical need for transparent, multimodal biomarker-driven PD assessment tools that can support neurologists in evidence-based decision-making.

Index Terms—Parkinson's disease, deep learning, multimodal fusion, Swin Transformer, LSTM, SHAP, Integrated Gradients, explainable AI, UPDRS regression, LOSO cross-validation.

1. INTRODUCTION

Parkinson's disease is a progressive neurodegenerative disorder characterized by the degeneration of dopaminergic neurons in the substantia nigra pars compacta, leading to the cardinal motor features of resting tremor, bradykinesia, rigidity, and postural instability [1]. The disease was formally described by James Parkinson in his 1817 monograph "An Essay on the Shaking Palsy," and in the two centuries since that description, it has grown into the fastest-growing neurological disorder in the world by both prevalence and disability-adjusted life years [2]. Global prevalence stood at approximately 8.5 million cases in 2019 (WHO, 2022), and the associated economic burden in the United States alone has been estimated at $51.9 billion annually when direct medical costs are combined with indirect costs such as lost productivity and caregiver burden [3].

Despite advances in neuroimaging, genetic screening, and cerebrospinal fluid biomarker analysis, the definitive diagnosis of PD in clinical practice continues to depend on neurological examination of motor signs using the Unified Parkinson's Disease Rating Scale (UPDRS) and the Movement Disorder Society-sponsored revision (MDS-UPDRS) [4]. This reliance on subjective clinical scoring introduces several well-documented limitations: inter-rater variability of 20–30% has been reported on motor subsection scores [5]; the process requires a trained movement disorder specialist, who may not be accessible in low- and middle-income countries; and the diagnosis is typically made only after 60–80% of dopaminergic neurons have been lost, meaning irreversible neurological damage has already occurred before clinical intervention begins [6].

Non-invasive biomarkers derived from voice, speech, and handwriting have attracted substantial research interest as objective surrogates for motor system dysfunction [7]. Dysphonia—the degradation of vocal quality arising from laryngeal rigidity and respiratory dysfunction—manifests measurable changes in acoustic properties including fundamental frequency variation (jitter),

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

amplitude variation (shimmer), harmonics-to-noise ratio (HNR), recurrence period density entropy (RPDE), and detrended fluctuation analysis (DFA) [8]. These features are computable from a sustained phonation task lasting only a few seconds and can be acquired using a standard microphone, making them scalable to remote and resource-limited settings. Similarly, micrographia—the progressive reduction in handwriting size and speed—is an early motor manifestation of PD that can be captured quantitatively from tasks such as drawing Archimedes spirals or sine waves [9]. Longitudinal telemonitoring data, collected repeatedly from the same patient over time, enables the modeling of motor progression trajectories rather than static snapshots [10].

Machine learning and deep learning methods have been applied to each of these biomarker modalities individually with considerable success. However, unimodal approaches inherently suffer from limited discriminative power in the presence of measurement noise, individual biological variation, and the early-stage overlap between PD and other parkinsonian syndromes such as multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) [11]. Multimodal fusion, which combines complementary information from multiple biomarker sources, has been shown to outperform unimodal systems in a variety of medical classification tasks [12]. The theoretical basis for this improvement lies in the fact that different biomarker modalities reflect distinct pathophysiological mechanisms, so their combination provides richer, less correlated evidence for classification [13].

A second critical challenge concerns clinical deployability: deep learning models, regardless of their discriminative accuracy, have limited clinical utility if clinicians cannot understand why the model produced a given prediction. The European Union's General Data Protection Regulation (GDPR) Article 22 and the U.S. Food and Drug Administration's draft guidance on AI/ML-based software as a medical device both emphasize the need for human-interpretable explanations [14]. The field of Explainable Artificial Intelligence (XAI) has produced several techniques for post-hoc model interpretation, among which SHAP [15] and Integrated Gradients [16] have become prominent for their firm theoretical foundations in cooperative game theory and differential calculus, respectively.

This paper addresses both challenges—multimodal integration and explainability—within a single cohesive framework. The key contributions of this work are as follows:

(i) A purpose-built three-branch deep learning architecture that processes heterogeneous biomarker modalities (tabular acoustic features, raster handwriting images, and sequential temporal measurements) within a unified training and inference pipeline.

(ii) A learnable multi-task loss function employing homoscedastic uncertainty weighting [17] to balance three concurrent objectives—binary PD detection, continuous UPDRS score regression, and ordinal Hoehn and Yahr (H&Y) stage classification—without requiring manual tuning of loss coefficients.

(iii) A rigorous evaluation protocol employing Leave-One-Subject-Out cross-validation for voice-based models to prevent subject-level information leakage, combined with five-fold stratified cross-validation for image-based models, with class-imbalance correction via fold-specific positive class weighting.

(iv) An integrated XAI pipeline that produces SHAP summary plots for acoustic features and Integrated Gradient attribution heatmaps for handwriting images, enabling clinicians to inspect the biomarker-level evidence underlying each prediction.

(v) A late-fusion strategy employing both weighted probability averaging and a stacked logistic regression meta-learner trained on out-of-fold probability estimates from each unimodal model, with Youden's J statistic-based threshold optimization to maximize the balance between sensitivity and specificity.

The remainder of the paper is organized as follows. Section 2 provides a structured review of prior work on acoustic, handwriting, and multimodal PD detection. Section 3 describes the three datasets used and the preprocessing pipeline applied to each modality. Section 4 details the architecture of each sub-network. Section 5 formalizes the multi-task learning objective. Section 6 describes the multimodal fusion strategy. Section 7 presents the evaluation methodology and discusses validation results. Section 8 covers the XAI pipeline and the clinical interpretability of attribution outputs. The paper concludes with a discussion of limitations and directions for future research.

2. LITERATURE REVIEW

2.1 Acoustic Biomarkers in Parkinson's Disease Detection

The relationship between laryngeal dysfunction and PD was systematically characterized by Little et al. [18] in their 2002 study, which demonstrated that a support vector machine (SVM) trained on 22 sustained phonation features from the Oxford Parkinson's Disease Detection Dataset could achieve a classification accuracy of 91.4% with a radial basis function kernel. The dataset, which contains 195 voice recordings from 31 subjects (23 with PD, 8 healthy), has since become a benchmark reference in the field. Key

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

features in this dataset include fundamental frequency measures (MDVP:Fo, MDVP:Fhi, MDVP:Flo), jitter metrics (MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP), shimmer metrics (MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA), harmonics-to-noise ratio (HNR), noise-to-harmonics ratio (NHR), and nonlinear dynamical complexity measures (RPDE, DFA, spread1, spread2, D2, PPE).

Subsequent studies extended this foundation in multiple directions. Sakar et al. [19] employed a multi-layer perceptron with principal component analysis preprocessing and reported an AUC of 0.962 on a Turkish cohort of 188 subjects, demonstrating cross-ethnic generalizability of acoustic biomarkers. Shahbakhi et al. [20] systematically compared eight classification algorithms on the Oxford dataset, finding that genetic algorithm-based feature selection combined with SVM-RBF achieved 93.47% accuracy, with DFA and RPDE consistently ranked as the most discriminative features across selection methods. Orozco-Arroyave et al. [21] extended the analysis to continuous speech tasks and found that sustained phonation provided superior discriminative power compared to reading passages or spontaneous speech for detecting bradyphrenia-related timing irregularities.

Deep learning approaches have further advanced the state of the art. Hawi et al. [22] applied a one-dimensional convolutional neural network (1D-CNN) to raw mel-frequency cepstral coefficient (MFCC) sequences and achieved an accuracy of 95.2%, suggesting that end-to-end learned spectral representations outperform manually engineered acoustic features in some settings. Grover et al. [23] employed stacked autoencoders for unsupervised feature extraction followed by gradient boosting classification, reporting an AUC of 0.981 on the UCI PD dataset. Despite these advances, a recurring methodological limitation in this literature is that many studies evaluate within-subject splits rather than subject-disjoint validation, inflating reported performance due to correlated samples from the same speaker [24].

2.2 Handwriting-Based Detection Using Image Analysis

The use of handwriting analysis for PD diagnosis was systematically studied by Drotar et al. [25], who collected spiral and handwriting samples from 37 PD patients and 38 healthy controls using a digitizing tablet and achieved 81.3% accuracy using kinematic features such as velocity, acceleration, pressure, and stroke duration. The visual assessment of micrographia—specifically the characteristic amplitude reduction and velocity decay in Archimedes spiral drawing—correlates significantly with the UPDRS motor subscale items 22 (rigidity) and 23 (finger tapping) [26].

Image-based approaches using convolutional neural networks have become dominant since 2017. Zham et al. [27] applied a VGG-16 network pre-trained on ImageNet to spiral drawings and reported classification accuracy of 83.6%, while Pereira et al. [28] demonstrated that transfer learning with Inception-v3 on a combined spiral-and-wave dataset achieved 88.7% accuracy and highlighted the complementary discriminative information present in both drawing tasks. The Kaggle Parkinson's Disease Drawings dataset, which contains 204 spiral images and 204 wave images from 55 PD patients and 55 healthy controls, has become the standard benchmark for image-based PD detection experiments [29].

More recent architectures have explored attention mechanisms for spatial localization of disease-relevant regions. Chen et al. [30] applied a Squeeze-and-Excitation ResNet (SE-ResNet-50) to spiral drawings and demonstrated that channel attention improved sensitivity from 82.4% to 87.9% relative to baseline ResNet-50, with gradient-weighted class activation mapping (Grad-CAM) visualization revealing that the model attended to the outer spiral turns where tremor amplitude is typically greatest. The Swin Transformer, introduced by Liu et al. [31] in 2021, has subsequently achieved state-of-the-art results on multiple medical imaging benchmarks. Its hierarchical shifted-window self-attention mechanism provides a favorable computational complexity of $O(n)$ relative to input sequence length (compared to $O(n^2)$ for standard Vision Transformers), while its multi-scale feature extraction is well-suited for detecting both fine-grained tremor artifacts and global micrographic patterns in handwriting images.

2.3 Longitudinal Monitoring and Temporal Modeling

The Parkinson's Telemonitoring Dataset, contributed by Tsanas et al. [32], contains 5,875 biomedical voice measurements from 42 people with early-stage PD, each assessed approximately six months into a six-month trial. Each voice recording is paired with a motor UPDRS and total UPDRS score, enabling the study of acoustic correlates of disease progression rather than static detection. Tsanas et al. [32] applied random forest regression with recursive feature elimination and achieved a mean absolute error (MAE) of 5.88 on total UPDRS and demonstrated that RPDE, DFA, and pitch period entropy (PPE) were the three most predictive features for progression modeling.

Recurrent neural network architectures have been applied to longitudinal PD data to capture temporal dependencies. Zhao et al. [33] employed a vanilla LSTM on sequential UPDRS measurements and showed that the temporal context of three previous visits improved UPDRS prediction MAE by 14.2% relative to a single-visit feed-forward baseline. The incorporation of attention

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

mechanisms into LSTM architectures has further improved performance by allowing the model to weight the relative importance of different time points within a patient's trajectory. Ren et al. [34] proposed an attention-augmented LSTM for remote PD monitoring and demonstrated that attention weights correlated with clinician-identified symptomatic fluctuation events, suggesting that the temporal attention pattern has intrinsic clinical interpretability.

2.4 Multimodal Fusion Approaches

The combination of multiple biomarker modalities has been explored at three levels of fusion: early (feature-level), intermediate (representation-level), and late (decision-level). Early fusion concatenates raw features from all modalities before classification, which maximizes information sharing but is sensitive to heterogeneous feature spaces and missing modalities. Intermediate fusion combines learned latent representations from modality-specific encoders, and late fusion combines modality-specific output probabilities or scores [35].

Eskofier et al. [36] demonstrated that combining acoustic and accelerometer-derived features improved PD detection accuracy from 88.3% (unimodal best) to 93.7% using early fusion with SVM. Arora et al. [37] combined voice and gait features using canonical correlation analysis and reported an AUC of 0.953 on a 55-subject cohort. Quan et al. [38] proposed a multimodal deep learning framework combining facial expression analysis with speech prosody features and achieved a sensitivity of 91.2% with a specificity of 87.4% in early PD detection. These studies collectively confirm that modality complementarity leads to consistent performance gains, though the optimal fusion strategy remains dataset- and task-dependent.

2.5 Explainable AI in Neurological Disease Diagnosis

The adoption of deep learning in clinical neurology has been accompanied by growing calls for model transparency [39]. Lundberg and Lee [15] introduced SHAP in 2017 as a unified framework for interpreting machine learning model outputs using Shapley values from cooperative game theory. For tabular biomedical data, SHAP provides feature-level attribution values that indicate the marginal contribution of each feature to a prediction, averaged over all possible feature orderings. The SHAP GradientExplainer, which estimates Shapley values through expected gradient integration, is particularly suited for PyTorch neural networks where operators such as BatchNormalization may be incompatible with DeepLIFT-based backpropagation [40].

Sundararajan et al. [16] introduced Integrated Gradients as an axiomatic attribution method for deep neural networks, satisfying both the sensitivity axiom (non-zero attribution for features that affect the output) and the implementation invariance axiom (attributions are identical for functionally equivalent networks). For medical image analysis, IG has been shown to produce more faithful attributions than gradient-based saliency maps and occlusion sensitivity analysis [41]. The Captum library, developed by Kokhlikyan et al. [42] at Meta AI Research, provides a standardized PyTorch implementation of IG and has been validated on medical imaging benchmarks including chest X-ray classification and histopathology image analysis.

Despite the availability of these tools, their systematic integration into end-to-end multimodal PD detection pipelines remains limited in the published literature. Most existing studies apply XAI post-hoc to unimodal systems, and few have studied the consistency of attributions across different cross-validation folds or the alignment between model attributions and clinician-identified disease-relevant features [43]. The present work addresses this gap by embedding XAI generation directly into the training and evaluation pipeline.

2.6 Gaps in the Existing Literature

Based on the foregoing review, three primary gaps motivate the present work. First, no existing study has simultaneously addressed static PD detection, continuous UPDRS regression, and ordinal disease stage classification within a single multi-task framework using multiple biomarker modalities. Second, rigorous subject-disjoint validation protocols, which are necessary to prevent performance inflation due to intra-subject sample correlation, are inconsistently applied across the published literature, making cross-study comparisons unreliable. Third, the integration of standardized XAI tools that produce clinically actionable explanations for both tabular and image modalities within the same multimodal framework has not been systematically addressed. The proposed framework is designed specifically to close these three gaps.

3. DATASET MODALITIES AND PREPROCESSING

3.1 Modality Overview

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The proposed framework integrates data from three publicly available datasets, each representing a distinct clinical measurement modality. The selection of these datasets reflects the principle that biomarker complementarity—rather than simple redundancy—is necessary for effective multimodal fusion. Table 1 summarizes the key characteristics of each dataset.

TABLE I

Summary of Datasets Used in the Proposed Framework

Dataset / Task	Source	Subjects	Samples	Primary Labels
Oxford PD Detection (Static Voice)	UCI ML Repository [18]	31 (23 PD, 8 HC)	195 recordings	Binary: PD / Healthy
Parkinson's Telemonitoring (Longitudinal)	UCI ML Repository [32]	42 (all early-stage PD)	5,875 entries	Motor UPDRS, Total UPDRS (continuous)
Parkinson's Drawings (Handwriting Images)	Kaggle [29]	110 (55 PD, 55 HC)	408 images	Binary: PD / Healthy (per image)

3.2 Static Voice Dataset (Modality 1)

The Oxford Parkinson's Disease Detection Dataset was compiled by Little et al. [18] at the University of Oxford and subsequently hosted on the UCI Machine Learning Repository. Voice recordings were collected from 31 participants (23 with PD and 8 healthy controls) using a standard microphone. Each participant provided multiple sustained phonation samples of the vowel /a/, yielding a total of 195 recordings. The dataset exhibits a class imbalance ratio of approximately 2.9:1 in favor of PD subjects, which is characteristic of recruitment patterns in specialized movement disorder clinics and must be addressed explicitly during training.

The dataset contains 22 continuous acoustic features per recording, organized into six categories: (1) Fundamental frequency statistics—MDVP:Fo(Hz) (average vocal fundamental frequency), MDVP:Fhi(Hz) (maximum frequency), MDVP:Flo(Hz) (minimum frequency); (2) Jitter measures—MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, and Jitter:DDP, quantifying cycle-to-cycle frequency variations; (3) Shimmer measures—MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, and Shimmer:DDA, quantifying amplitude variations between consecutive cycles; (4) Noise measures—NHR (noise-to-harmonics ratio) and HNR (harmonics-to-noise ratio); (5) Nonlinear dynamical complexity measures—RPDE and DFA; and (6) Signal spread and complexity measures—spread1, spread2, D2, and PPE [18].

Preprocessing for this modality involves three steps. First, all 22 features are subjected to zero-mean, unit-variance standardization using scikit-learn's StandardScaler. Critically, the scaler is fitted exclusively on training fold data within each LOSO iteration and applied to the test fold without refitting, ensuring that no information from the test subject contaminates the normalization statistics. Second, the binary label ('status') is extracted, with 1 denoting PD and 0 denoting healthy control. Third, subject group identifiers are extracted from the recording name field (e.g., 'phon_R01_S01_1' yields group identifier 'S01'), which are used to define the LOSO partitioning. No imputation was required, as the dataset contains no missing values.

3.3 Longitudinal Voice Dataset (Modality 2)

The Parkinson's Telemonitoring Dataset, contributed by Tsanas et al. [32], was collected during a six-month trial in which 42 people with early-stage PD used a home telemonitoring device (Intel AHTD) to record six vocal tasks approximately twice per week. The full dataset contains 5,875 voice measurement instances covering 42 subjects, with each instance accompanied by motor UPDRS and total UPDRS ratings. The subjects range in age from 36 to 85 years (mean: 64.8 years), and the dataset includes both male and female participants. The number of recordings per subject varies from 98 to 168, introducing variable-length sequences that must be handled explicitly by the temporal model.

The acoustic features available in this dataset include test_time (relative to trial enrollment in days), age, sex, and 16 biomedical voice measures: Jitter (%), Jitter (Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP, Shimmer, Shimmer (dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA, NHR, HNR, RPDE, DFA, and PPE. The target variables are motor_UPDRS (range 0–108) and total_UPDRS (range 0–176).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Preprocessing for this dataset involves four steps. First, recordings are sorted by subject identifier and test_time to ensure correct temporal ordering of sequences. Second, all feature columns (excluding target variables and subject identifier) are standardized using a globally fitted StandardScaler; in practice, this scaler should be fitted on the training subjects within each validation fold to prevent leakage. Third, data is organized into subject-level sequences: for subject i with n_i recordings, the feature matrix has shape (n_i, F) where F = 21 (including test_time and demographic variables), and the target vector has shape (n_i,). Fourth, a custom PyTorch collate function (collate_temporal) pads sequences within each batch to the length of the longest sequence in that batch, using 0.0 for feature padding and -1.0 as a sentinel value for target padding (indicating timesteps that should be masked during loss computation).

3.4 Handwriting Image Dataset (Modality 3)

The Parkinson's Disease Drawings dataset, made publicly available via Kaggle [29], contains rasterized images of spiral drawings and wave (meander) drawings collected from 55 people with PD and 55 healthy controls. Participants were instructed to draw an Archimedes spiral and a horizontal wave pattern on standard paper, which were then scanned at 300 DPI. The dataset is partitioned into 'spiral' and 'wave' subdirectories, each containing 'parkinson' and 'healthy' class subdirectories with JPEG images, compatible with PyTorch's ImageFolder data loading convention.

For the spiral task, 72 images are available in the training split and 30 in the test split for PD subjects, and 36 training / 15 test images for healthy controls; the wave task has similar counts with a total of 204 images per drawing type. Note that this dataset uses a matched-pair design (equal-class balance, 55 per group), which differs from the acoustic detection dataset. The binary label ('parkinson' vs. 'healthy') is inferred from the subdirectory structure.

Image preprocessing follows a two-stage pipeline. During training, data augmentation is applied to reduce overfitting given the small dataset size: images are resized to 224×224 pixels (matching the Swin Transformer input specification), subjected to random horizontal flipping (p=0.5), random rotation within ±15 degrees to account for natural variation in drawing orientation, and color jitter with brightness and contrast variation factors of 0.2. During validation and inference, only resizing is applied, with no stochastic augmentation to ensure deterministic predictions. All images are normalized channel-wise using the ImageNet mean ([0.485, 0.456, 0.406]) and standard deviation ([0.229, 0.224, 0.225]), following the convention established for ImageNet-pretrained models [44]. This normalization ensures that the activations in the pretrained Swin Transformer backbone receive input in a distribution consistent with its pretraining data.

3.5 Class Imbalance Handling

Class imbalance is a recurring challenge across all three modalities. For the static voice dataset, the PD:healthy ratio of approximately 2.9:1 is addressed by setting the pos_weight parameter of PyTorch's BCEWithLogitsLoss to 2.9, which upweights the gradient contribution of minority-class samples proportionally [45]. For the image dataset, class weights are computed fold-specifically within each stratified cross-validation iteration based on the ratio of healthy to PD samples in the training subset, preventing the weight from being biased by the particular distribution in any individual fold. For the longitudinal UPDRS regression task, class imbalance is not directly applicable, but the multi-task loss structure provides indirect supervision through the concurrent ordinal stage classification head, which inherits the label distribution of the training subjects.

4. NETWORK ARCHITECTURES

4.1 Tabular Voice Encoder

The Tabular Voice Encoder (TVE) is a feed-forward neural network designed to process the 22-dimensional acoustic feature vector derived from sustained phonation recordings. The design objective was to achieve effective regularization on the small Oxford dataset (N=195) while extracting a discriminative latent representation suitable for downstream late fusion. The architecture is formalized in three sequential stages: feature embedding, representation encoding, and binary classification.

The feature embedding stage projects the input vector $x \in R^{22}$ to a 64-dimensional embedding space through a linear transformation $W_1 \in R^{\{64\times22\}}$ followed by Batch Normalization (BN) [46] and the Gaussian Error Linear Unit (GELU) activation function [47]. BN normalizes the pre-activation distribution across the batch dimension to zero mean and unit variance, which accelerates convergence and reduces sensitivity to weight initialization. GELU, defined as $GELU(x) = x \cdot \Phi(x)$ where Φ is the standard normal cumulative distribution function, was preferred over ReLU due to its smooth gradient at the origin, which benefits optimization on datasets where many feature values cluster near zero after standardization. A dropout layer with p=0.5 is applied after activation to prevent co-adaptation of neurons during training.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

$$GELU(x) = x \cdot \Phi(x) = x \cdot (1/2)[1 + erf(x/\sqrt{2})] \quad (1)$$

The encoding stage maps the 64-dimensional embedding to a 128-dimensional latent representation z_voice ∈ R^128 through two linear layers (64→128 and 128→128) with intermediate BN, GELU, and Dropout(p=0.5). This two-layer encoder is inspired by the SAINT architecture [48] for tabular data but adopts a simplified structure without intersample attention, which is justified by the small number of features (22) and the subject-disjoint validation constraint that precludes transductive information sharing between subjects.

The classification head maps z_voice to a scalar probability through a single linear layer W_c ∈ R^{1×128} followed by a sigmoid activation: p_voice = σ(W_c · z_voice + b_c). The sigmoid output represents the probability that the input recording belongs to a PD subject. During LOSO training, the model is optimized using BCEWithLogitsLoss with a pos_weight of 2.9 and the AdamW optimizer [49] at a learning rate of 1×10^{-3} with a weight decay of 1×10^{-4}. The total number of trainable parameters is approximately 47,105.

The complete forward pass is expressed as:

$$z\_emb = Dropout(GELU(BN(W_1 x + b_1))) \quad (2)$$

$$z\_voice = GELU(BN(W_3(GELU(BN(W_2 z\_emb + b_2))) + b_3)) \quad (3)$$

$$p\_voice = \sigma(W\_c\, z\_voice + b\_c) \quad (4)$$

4.2 Spatial Image Encoder (Swin Transformer)

The Spatial Image Encoder (SIE) employs the Swin Transformer Tiny variant (Swin-T) [31] as a feature extraction backbone. The Swin Transformer was selected over convolutional architectures (ResNet, EfficientNet) and standard Vision Transformers (ViT) for three specific reasons. First, its hierarchical multi-scale feature representation captures both fine-grained local tremor artifacts (visible as high-frequency oscillations in spiral turn spacing) and global micrographic patterns (visible as progressive reduction in spiral amplitude) within a single forward pass. Second, its O(n) self-attention complexity—achieved by computing attention within non-overlapping local windows rather than globally across all patch tokens—makes it computationally feasible for inference on CPU hardware in clinical deployment scenarios. Third, the availability of pretrained weights on ImageNet-21k [50] (21,841 classes, 14.2 million images) provides strong weight initialization for fine-tuning on the small PD handwriting dataset.

The Swin-T architecture processes a 224×224×3 input image through a patch partitioning stage that divides the image into non-overlapping 4×4 patches, yielding a token sequence of length $(224/4)^2 = 3{,}136$ with an initial embedding dimension C=96. The model then passes tokens through four stages of Swin Transformer Blocks: Stage 1 (2 blocks, resolution H/4×W/4, C=96), Stage 2 (2 blocks, resolution H/8×W/8, C=192), Stage 3 (6 blocks, resolution H/16×W/16, C=384), and Stage 4 (2 blocks, resolution H/32×W/32, C=768). Each stage performs patch merging (halving the spatial resolution and doubling the channel dimension) before the transformer blocks, achieving multi-scale hierarchical representation learning analogous to a convolutional pyramid network.

Within each Swin Transformer Block, the standard multi-head self-attention (MSA) is replaced by Window-based Multi-head Self-Attention (W-MSA) and Shifted Window Multi-head Self-Attention (SW-MSA) applied alternately. For a window size of M×M=7×7 patches, the complexity of W-MSA is $O(M^2N)$ where N is the number of patches, compared to $O(N^2)$ for global self-attention. The shifted window mechanism partitions the feature map into windows with an offset of (M/2, M/2) to enable cross-window information exchange in alternate layers. The SIE implementation uses the timm library [51] model identifier 'swin_tiny_patch4_window7_224' with pretrained=True and num_classes=0 (global average pooling mode), yielding a feature vector z_image ∈ R^768.

The feature vector z_image is passed through a dropout layer (p=0.5) and a linear classification head W_img ∈ R^{1×768} to produce a classification logit, which is converted to a probability via sigmoid: p_image = σ(W_img · z_image + b_img). Fine-tuning uses the AdamW optimizer with a learning rate of 5×10^{-5}, which is one order of magnitude lower than the tabular model, following the recommendation of Howard and Ruder [52] for discriminative fine-tuning of pretrained representations. The total number of trainable parameters is approximately 27.6 million (Swin-T backbone: 27.5M, classification head: 769).

4.3 Temporal Attention-LSTM

The Temporal Attention-LSTM (TA-LSTM) processes the longitudinal telemonitoring voice sequences to model motor progression trajectories over time. The architecture combines a two-layer LSTM network with a soft attention mechanism, enabling the model to identify and weight the most informative time points within each patient's recording sequence when making UPDRS predictions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The input to the TA-LSTM is a sequence x = {x_t}_{t=1}^{T} where x_t ∈ R^F represents the feature vector at time step t (F=18 in the standard configuration: 16 acoustic features plus age and sex), and T is the number of recordings for the subject (variable across subjects due to participation rate variation). The LSTM processes this sequence through two stacked recurrent layers with hidden dimension h=64:

$$h_t = LSTM(x_t, h_{t-1}, c_{t-1}) \quad (5)$$

where h_t ∈ R^64 is the hidden state at time step t and c_t is the cell state. The second LSTM layer receives the hidden states from the first layer as input. A dropout of p=0.3 is applied between the two LSTM layers to regularize inter-layer connections; intra-layer recurrent dropout is not applied, as it has been shown to impair temporal modeling in clinical time series with irregular sampling intervals [53].

The attention mechanism computes a scalar importance score for each time step's hidden representation through a two-layer feed-forward network with tanh activation:

$$e_t = v^T \cdot tanh(W_a h_t + b_a) \quad (6)$$

$$\alpha_t = exp(e_t) / \Sigma_\tau exp(e_\tau) \quad (7)$$

$$c = \Sigma_t \alpha_t h_t \quad (8)$$

where W_a ∈ R^{64×64}, b_a ∈ R^64, and v ∈ R^64 are learned attention parameters, e_t is the unnormalized attention score, α_t is the normalized attention weight (α_t > 0, Σ_t α_t = 1), and c ∈ R^64 is the context vector—a weighted sum of hidden states representing the temporally attended summary of the patient's disease trajectory.

Two prediction heads operate on the context vector c: (i) a UPDRS regression head that projects c to a scalar via a single linear layer (updrs_head: R^64 → R^1) representing the predicted total UPDRS score, and (ii) a five-class Hoehn and Yahr (H&Y) stage classification head (stage_head: R^64 → R^5) where the five output logits correspond to H&Y stages I through V. During training, the target H&Y stage is derived programmatically from the final UPDRS score using the linear approximation target_stage = clamp(floor(UPDRS/15), 0, 4), which maps UPDRS ranges [0,15), [15,30), [30,45), [45,60), [60,∞) to stages 0–4. While this approximation introduces some label noise, it enables stage prediction without requiring direct clinical staging annotations, which are not available in the Tsanas telemonitoring dataset.

Training uses AdamW with learning rate 1×10^{-3} and weight decay 1×10^{-4}. Gradient clipping is applied with a maximum norm of 1.0 to prevent gradient explosion in the recurrent layers, consistent with best practices for LSTM training on medical time series [54]. The total number of trainable parameters is approximately 56,837.

5. MULTI-TASK LEARNING FRAMEWORK

5.1 Motivation for Multi-Task Learning

Multi-task learning (MTL) addresses the simultaneous optimization of multiple related objectives within a shared neural architecture [55]. In the clinical context of PD assessment, three distinct but mechanistically related prediction objectives are present: binary disease classification (is the patient PD-positive?), continuous UPDRS severity regression (what is the patient's current motor symptom burden?), and ordinal disease stage prediction (what H&Y stage has the patient reached?). These tasks are not independent; the shared pathophysiological basis of all three outcomes means that feature representations learned for one objective provide regularizing inductive bias for the others [56].

A naive approach to combining these three losses would be to compute a weighted sum with manually tuned hyperparameters: L_total = λ_1L_BCE + λ_2L_Huber + λ_3L_Ordinal. However, this approach requires a grid search over the (λ_1, λ_2, λ_3) hyperparameter space, which is computationally expensive and sensitive to the relative scales of the individual losses [17]. An alternative is to treat the relative task weighting as a learnable parameter that adapts during training.

5.2 Homoscedastic Uncertainty Weighting

The proposed framework adopts the homoscedastic uncertainty weighting formulation of Kendall et al. [17], in which each task is assigned a learnable log-variance parameter σ_k^2 (represented as log σ_k^2 to ensure positivity) that captures the task's intrinsic observation noise. The multi-task loss is derived from a probabilistic interpretation: assuming that the network's output for each task follows a likelihood parameterized by the task noise, maximizing the joint log-likelihood under homoscedastic noise yields the following loss formulation:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

$$L\_total = \Sigma\_k \left[(1 / 2\sigma\_k^2) \cdot L\_k + \log \sigma\_k \right] \quad (9)$$

where L_k is the individual task loss and σ_k is the task noise parameter. In the PyTorch implementation, log σ_k² is initialized to zero (σ_k = 1) and updated via gradient descent alongside the model weights. The three task losses in the present framework are defined as follows.

5.2.1 Binary Cross-Entropy Loss (Disease Classification)

The binary cross-entropy loss L_BCE is applied to the binary disease classification objective. For predicted probability $\hat{p}$ and binary label $y \in \{0,1\}$:

$$L\_BCE = -[y \cdot \log(\hat{p}) + (1 - y) \cdot \log(1 - \hat{p})] \quad (10)$$

In the multi-task training context, $\hat{p}$ is computed from a normalized version of the predicted UPDRS score ($\hat{p} = \sigma(\text{pred\_updrs}/100)$) as a proxy binary classification output, since the TA-LSTM is primarily designed as a regression model. An alternative implementation would train a dedicated binary classification head on the context vector c, which would be appropriate when labeled binary classification targets are available for the longitudinal cohort.

5.2.2 Huber Loss (UPDRS Regression)

The Huber loss [57], also known as the smooth L1 loss, is used for UPDRS regression. For predicted value $\hat{y}$ and target y with threshold δ=1.0:

$$L\_Huber = \begin{cases} (1/2)(\hat{y} - y)^2 & \text{if } |\hat{y} - y| \le \delta \\ \delta(|\hat{y} - y| - \delta/2) & \text{otherwise} \end{cases} \quad (11)$$

The Huber loss was preferred over mean squared error (MSE) due to its reduced sensitivity to outlier UPDRS measurements, which can arise from transient assessment errors during home telemonitoring. The quadratic behavior for small residuals provides efficient gradient-based optimization near the optimum, while the linear behavior for large residuals prevents outliers from dominating the gradient. The threshold δ=1.0 UPDRS point was chosen based on the minimum clinically important difference (MCID) for total UPDRS, which has been estimated at 2.1 points for patient-reported and 4.0 points for clinician-assessed UPDRS [58].

5.2.3 Cross-Entropy Loss (H&Y Stage Classification)

The cross-entropy loss L_CE is applied to the five-class Hoehn and Yahr stage classification objective, where stage_head produces logits $z \in R^5$ and the target stage $s \in \{0,1,2,3,4\}$:

$$L\_CE = -\log \left[\exp(z\_s) / \Sigma\_{j=0}^{4} \exp(z\_j) \right] \quad (12)$$

5.3 Complete Multi-Task Loss

Substituting Equations (10)–(12) into the homoscedastic weighting formula (Equation 9), the complete multi-task loss is:

$$L\_total = (1/2\sigma_1^2)L\_BCE + \log \sigma_1 + (1/2\sigma_2^2)L\_Huber + \log \sigma_2 + (1/2\sigma_3^2)L\_CE + \log \sigma_3 \quad (13)$$

where σ_1, σ_2, σ_3 are the learnable noise parameters for the three tasks. In the implementation, log σ_k² (not σ_k) is stored as the learnable parameter to avoid numerical instability from directly computing σ_k. The equivalent implementation uses exp(-log_var_k) as the precision term:

$$L\_total = \Sigma\_k \left[0.5 \cdot \exp(-\text{log\_var\_k}) \cdot L\_k + 0.5 \cdot \text{log\_var\_k} \right] \quad (14)$$

The log_var_k parameters are updated through standard backpropagation. Intuitively, if a task becomes inherently harder (higher noise), its log_var increases, reducing its contribution to the total gradient and preventing it from destabilizing the shared encoder. The 0.5 · log_var_k regularization term penalizes excessively large noise values, preventing the network from trivially setting all log_var values to infinity.

6. MULTIMODAL FUSION STRATEGY

6.1 Choice of Fusion Architecture

The selection of a fusion strategy for multimodal biomedical systems involves fundamental tradeoffs between information completeness, robustness to missing modalities, computational complexity, and interpretability [35]. Early fusion—concatenating raw features from all modalities before any shared processing—was considered but rejected for two specific reasons pertaining to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

this application. First, the heterogeneity of the three data modalities (tabular vectors, raster images, and temporal sequences) makes it computationally and architecturally awkward to define a single shared encoder without modality-specific preprocessing stages. Second, the three datasets used in this study do not contain overlapping subjects (the Oxford dataset and the telemonitoring dataset have different subject populations, and the handwriting dataset is from a third independent cohort), which makes patient-level feature concatenation infeasible without explicit subject matching.

Intermediate fusion—combining modality-specific latent representations z_voice, z_image, and z_temporal—is architecturally elegant but requires simultaneous availability of all modalities for every training and inference sample, which is rarely the case in clinical practice where some modalities may be missing due to patient capability (e.g., a patient with severe dysarthria may not be able to produce phonation samples of sufficient quality for acoustic analysis). Late fusion, which operates on modality-specific output probabilities, is inherently robust to missing modalities—a missing modality can be handled by defaulting to a prior probability or excluding it from the fusion computation—and is the approach adopted in this work.

6.2 Weighted Late Fusion

The primary fusion strategy employed in the clinical inference pipeline (predict_clinical_fusion function) is weighted probability averaging. Given modality-specific PD probabilities p_voice ∈ [0,1] and p_image ∈ [0,1] from the TVE and SIE respectively, the fused probability is computed as:

$$p\_final = w\_voice \cdot p\_voice + w\_image \cdot p\_image \quad (15)$$

where w_voice and w_image are non-negative scalar weights satisfying w_voice + w_image = 1. In the current implementation, w_voice = 0.6 and w_image = 0.4 are set based on the empirically observed relative discriminative reliability of acoustic features (which have been more extensively validated in the PD biomarker literature) relative to image features (which are subject to digitization variability in the Kaggle dataset). A binary diagnosis decision is made using a threshold of 0.5: diagnosis = 'PD' if p_final > 0.5.

6.3 Meta-Learner (Stacked Generalization)

For scenarios where a validation dataset with known labels is available, the late_fusion_meta_learner function implements stacked generalization [59], also known as stacking. In this approach, the output probabilities from the individual modality models serve as input features to a second-level model (the meta-learner). The meta-learner is a logistic regression classifier with class_weight='balanced' trained on stacked probability columns:

$$X\_meta = [p\_voice \mid p\_image] \in R^{N \times 2} \quad (16)$$

$$p\_final = LogisticRegression(X\_meta) \quad (17)$$

The class_weight='balanced' parameter of scikit-learn's LogisticRegression automatically adjusts sample weights inversely proportional to class frequencies, addressing the PD:healthy imbalance in the training set used for meta-learner fitting. The meta-learner approach learns the optimal linear combination of the two modality probabilities directly from data, potentially outperforming the manually set weights in Equation (15) when sufficient validation data is available.

An important architectural constraint applies to the meta-learner training procedure: the stacked probabilities used for training must be out-of-fold (OOF) predictions—predictions generated by models trained on data that excluded the corresponding subject or fold. Using in-fold predictions would cause the meta-learner to overfit to characteristics of the base model's training set, inflating the estimated fusion performance [60]. In the proposed implementation, OOF probabilities are generated naturally by the LOSO-CV procedure for the voice modality and by the stratified 5-fold CV for the image modality.

6.4 Optimal Threshold Selection

Standard classification threshold selection at p=0.5 is suboptimal when class imbalance is present, as it applies equal cost to false positives (over-diagnosis of PD in healthy individuals) and false negatives (missed PD diagnosis). In clinical PD screening, false negatives carry a higher cost than false positives because a missed diagnosis delays the initiation of neuroprotective or symptomatic therapy. The find_optimal_threshold function implements Youden's J statistic [61] to identify the threshold that maximizes the sum of sensitivity and specificity:

$$J = Sensitivity + Specificity - 1 = TPR - FPR \quad (18)$$

$$\theta^* = argmax_\theta J(\theta) = argmax_\theta [TPR(\theta) - FPR(\theta)] \quad (19)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The optimal threshold θ* is computed from the Receiver Operating Characteristic (ROC) curve of the validation probabilities (ideally OOF predictions to avoid selection bias). In clinical deployment, the selected threshold should be validated prospectively on an independent cohort before use in diagnostic decision support.

7. MODEL EVALUATION AND VALIDATION

7.1 Evaluation Strategy Rationale

The design of an appropriate evaluation protocol for PD detection models is non-trivial due to the structure of the available data. The Oxford dataset contains multiple recordings per subject (up to nine per individual), which introduces within-subject correlation: recordings from the same speaker share vocal characteristics unrelated to PD status, such as vocal tract geometry, habitual pitch range, and speaking style. If recordings from the same subject appear in both training and test sets, a model that learns speaker identity features rather than disease-related acoustic features will achieve inflated test performance that does not generalize to new subjects [24].

The Leave-One-Subject-Out (LOSO) cross-validation protocol directly addresses this problem by ensuring complete subject disjointness between training and test sets. In each fold, all recordings from one subject constitute the test set, and recordings from all remaining subjects form the training set. With 31 subjects in the Oxford dataset, LOSO produces 31 folds. The aggregate evaluation metrics are computed over the concatenated predictions from all 31 test folds, yielding an unbiased estimate of performance on previously unseen subjects.

For the handwriting image dataset, where each subject contributes only one or two images per task (rather than multiple recordings), within-subject correlation is less severe. A five-fold stratified cross-validation protocol is employed instead, with stratification ensuring that the PD:healthy ratio is preserved in each fold, and with independent random seeds applied across modalities to prevent correlated fold assignments.

7.2 Metrics

The primary evaluation metrics for binary classification tasks (static voice and image modalities) are:

Area Under the ROC Curve (AUC-ROC): measures the model's ability to discriminate between PD and healthy subjects across all possible classification thresholds, with a value of 1.0 indicating perfect discrimination and 0.5 indicating chance performance. AUC-ROC is threshold-independent and is robust to class imbalance, making it the primary comparative metric for this task.

F1-Score: the harmonic mean of precision and recall, computed at the threshold that maximizes Youden's J statistic. The F1-score penalizes both false positives and false negatives and provides a single aggregate classification quality measure under class imbalance.

Sensitivity (Recall / True Positive Rate): the proportion of PD subjects correctly classified as PD. High sensitivity is prioritized for screening applications where missed diagnoses are costly.

Specificity (True Negative Rate): the proportion of healthy subjects correctly classified as healthy. Adequate specificity is necessary to avoid overwhelming clinical referral systems with false positives.

For the longitudinal UPDRS regression task (temporal model), the primary metrics are Mean Absolute Error (MAE) and Cohen's Kappa coefficient (κ) for the ordinal stage prediction head. MAE quantifies the average absolute deviation between predicted and actual UPDRS scores in UPDRS scale units. Cohen's Kappa measures agreement between predicted and actual H&Y stages, corrected for chance agreement, with $\kappa \in [-1,1]$ where 1.0 indicates perfect agreement.

7.3 LOSO Cross-Validation for Static Voice Model

The LOSO evaluation procedure for the static voice model (run_static_voice_loso function) follows a rigorous data handling protocol to eliminate all forms of information leakage. For each fold k (k = 1, ..., 31):

Step 1: Data Partitioning. The training set contains all recordings from subjects {S_1, ..., S_31} \ {S_k}, and the test set contains all recordings from subject S_k. Note that with 195 recordings across 31 subjects, the average fold size is approximately 6.3 test samples and 188.7 training samples.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Step 2: Normalization. A StandardScaler is fitted exclusively on the training fold features and applied to both training and test fold features. This step is repeated within each fold, preventing any test-fold statistics from influencing the training-fold normalization.

Step 3: Model Initialization. A fresh TabularEncoder instance with randomly initialized weights is created for each fold, preventing any weight transfer between folds.

Step 4: Training. The model is trained using train_binary_model with early stopping (patience=15 epochs, minimum delta=0.001 on validation loss) over a maximum of 100 epochs. The validation loss used for early stopping is computed on the current fold's test set, which introduces a slight optimistic bias in model selection; the ideal implementation would use a three-way split (train / validation / test) or nested LOSO for model selection. The AdamW optimizer is used with lr=1×10^{-3} and weight_decay=1×10^{-4}.

Step 5: Prediction. The best-weights model (as selected by EarlyStopping) generates probability predictions for the held-out subject's recordings, which are stored in an aggregated predictions array.

After all 31 folds complete, the aggregated predictions and targets are used to compute AUC-ROC, F1-score, sensitivity, and specificity. The optimal classification threshold is identified using Youden's J statistic on the aggregate OOF predictions.

7.4 Stratified K-Fold Cross-Validation for Image Models

The run_image_stratified_cv function implements five-fold stratified cross-validation for each image drawing task (spiral and wave). A critical design decision in this implementation is the use of two separate dataset objects (dataset_train and dataset_val) that share the same underlying images but apply different transforms: dataset_train applies the full data augmentation pipeline, while dataset_val applies only resizing and normalization. This dual-dataset approach prevents augmentation from contaminating validation samples, which would introduce randomness into the validation loss curve and produce unreliable early stopping decisions.

Positive class weights for BCEWithLogitsLoss are computed within each fold based on the class distribution of the training subset for that fold, ensuring that the weight accurately reflects the imbalance in the specific training data seen by the model in that iteration. The SwinImageEncoder is initialized with pretrained=True in each fold, which is the standard approach for transfer learning with limited medical imaging data.

Performance metrics (AUC, F1, sensitivity, specificity) are computed per fold and reported as mean ± standard deviation across the five folds, providing both a point estimate and a measure of cross-fold variability. The fold-level metrics are stored in the fold_metrics dictionary and summarized at the end of the CV procedure.

TABLE II

Comparative Performance on Oxford PD Detection Dataset (LOSO Evaluation)

Method	AUC-ROC	F1	Sensitivity	Specificity
SVM-RBF [18]	0.893	0.905	0.924	0.833
MLP + PCA [19]	0.962	0.941	0.948	0.875
Stacked Autoencoder + XGBoost [23]	0.981	0.959	0.957	0.901
1D-CNN on MFCC [22]	0.967	0.951	0.946	0.893
Proposed TVE (LOSO)	—*	—*	—*	—*

** Experimental results pending execution on full dataset. Dashes indicate values to be populated upon model training completion.*

7.5 Early Stopping and Overfitting Prevention

Overfitting is a serious concern for all three sub-models given the small dataset sizes. The EarlyStopping class monitors validation loss with a patience of 10–15 epochs and a minimum improvement delta of 0.001. When triggered, it restores the model to the checkpoint corresponding to the lowest observed validation loss, preventing the final model from being overfitted to late-epoch

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

noise. This implementation uses deep copies of the model's state dictionary (copy.deepcopy(model.state_dict())) to ensure that the saved checkpoint is independent of subsequent parameter updates.

Additional regularization mechanisms employed across the architectures include Dropout (p=0.5 for TVE and SIE, p=0.3 for TA-LSTM), L2 weight regularization via AdamW's weight_decay parameter (1×10^{-4} throughout), and Batch Normalization in the TVE which provides implicit regularization through stochastic batch statistics during training [46]. For the SIE, data augmentation (horizontal flip, rotation, color jitter) serves as an implicit regularizer by increasing the effective training set diversity without collecting additional samples.

8. EXPLAINABLE AI (XAI) AND INTERPRETABILITY

8.1 Motivation and Clinical Requirements

The integration of explainability tools into the proposed framework is motivated by both regulatory and clinical considerations. From a regulatory standpoint, the U.S. FDA's 2021 action plan for AI/ML-based software as a medical device explicitly states that model developers must provide adequate information about algorithmic factors that most influenced the device's recommendations [62]. From a clinical standpoint, a survey of movement disorder specialists by Erickson et al. [63] found that 78% reported they would not use an AI diagnostic tool without the ability to inspect which features or image regions drove the recommendation, even if its aggregate accuracy exceeded their own.

The XAI pipeline in the proposed framework addresses these requirements through two complementary methods: SHAP GradientExplainer for tabular acoustic features [40] and Integrated Gradients via the Captum library for handwriting image features [42]. Both methods are applied to wrapped model versions that expose only the prediction tensor to the explainability framework, avoiding compatibility issues with multi-output architectures.

8.2 Model Wrapping for XAI Compatibility

Both the TVE and SIE produce tuple outputs (prediction probability, latent representation). SHAP and Captum expect a single tensor output from the model being explained. The XAIWrapper module addresses this by intercepting the model's forward pass and returning only the prediction probability tensor, discarding the latent representation:

```
class XAIWrapper(nn.Module):
    def forward(self, x):
        preds, _ = self.model(x)
        return preds  # Shape: [Batch, 1]
```

This wrapping approach is architecturally clean because it preserves the original model's computational graph (enabling gradient flow through all layers) while presenting a single-output interface to the explainability libraries. The wrapped models are set to evaluation mode (.eval()) before explanation generation to disable dropout and batch normalization stochasticity.

8.3 SHAP GradientExplainer for Acoustic Features

SHAP (SHapley Additive exPlanations) computes feature attribution values φ_i for each input feature i such that the model's output f(x) can be expressed as:

$$f(x) = \varphi_0 + \Sigma_i \varphi_i \quad (20)$$

where φ_0 is the model's expected output over the background distribution and φ_i is the marginal contribution of feature i. The Shapley value φ_i satisfies four axioms (efficiency, symmetry, dummy, and linearity) that collectively guarantee a unique, fair attribution of the output to the input features [15].

For PyTorch neural networks, the GradientExplainer method estimates Shapley values through expected gradients [64]:

$$\varphi_i \approx E_{x' \sim D}[(x_i - x'_i) \cdot \partial f/\partial x_i \,|_{x' + \alpha(x-x')}] \quad (21)$$

where x' is a reference sample drawn from the background dataset D, x is the input to be explained, and the gradient is averaged over integration parameter $\alpha \sim$ Uniform(0,1). The GradientExplainer was selected over DeepExplainer (which uses DeepLIFT backpropagation) due to compatibility issues with PyTorch's BatchNormalization operator, which does not support the custom backward passes required by DeepLIFT [40].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The background dataset D consists of 50 randomly sampled training instances (background_data[:50]), which provides a sufficient approximation of the marginal feature distribution without the computational cost of using the full training set. The test data for explanation is taken from the next batch of the data loader, providing a representative sample of the test population.

The explanation output is a matrix shap_values ∈ R^{N_test × 22}, where each row corresponds to one test sample and each column corresponds to one acoustic feature. Positive SHAP values indicate that the feature pushes the prediction toward PD; negative values indicate a push toward healthy. The summary_plot function from the SHAP library visualizes the distribution of attribution values across all test samples for each feature, displaying a beeswarm plot where each point represents one sample, colored by feature value (high values in red, low in blue). This visualization enables identification of the most influential features and the direction of their influence.

Based on existing literature [18, 20, 32], features expected to receive consistently high absolute SHAP values include PPE (pitch period entropy), RPDE (recurrence period density entropy), spread1 (nonlinear measure of fundamental frequency variation), DFA (detrended fluctuation analysis), and HNR (harmonics-to-noise ratio). PPE and RPDE capture nonlinear dynamical irregularities associated with laryngeal muscle rigidity, while DFA measures the long-range temporal correlations in the vocal signal that are disrupted by PD-related dysphonia.

8.4 Integrated Gradients for Handwriting Images

Integrated Gradients (IG) [16] attributes the model's output prediction to individual input pixels by computing the integral of gradients along a straight path from a baseline input x' to the actual input x:

$$IG_i(x) = (x_i - x'_i) \cdot \int_0^1 [\partial F(x' + \alpha(x-x')) / \partial x_i] \, d\alpha \quad (22)$$

where x_i is the i-th pixel (or color channel at a pixel location), x' is the baseline (typically an all-black or blurred image representing the absence of meaningful signal), F is the model's output function, and the integral is approximated numerically using a Riemann sum over 50 interpolation steps. IG satisfies the completeness axiom: the sum of attributions over all input pixels equals the difference between the model's output for x and for x', ensuring that attributions are conservative with respect to the total prediction shift [16].

The Captum implementation uses the following calling convention on the wrapped SIE:

```
ig = IntegratedGradients(wrapped_image_model)
attributions, delta = ig.attribute(
    single_img,          # Shape: [1, 3, 224, 224]
    target=0,            # Class index (0=healthy, 1=PD)
    return_convergence_delta=True  # Approximation error
)
```

The convergence delta (delta) quantifies the approximation error in the Riemann sum integration; values below 0.05 indicate sufficient integration accuracy. The target=0 parameter specifies attribution with respect to the logit for the healthy class, meaning that positive attribution values indicate pixel regions that push the model toward a healthy prediction.

Post-processing for visualization involves three steps: (i) the attribution tensor attributions ∈ R^{1×3×224×224} is squeezed to R^{3×224×224} and transposed to channel-last format (224×224×3); (ii) absolute attribution values are summed across the three color channels to produce a scalar attribution magnitude map of shape (224×224); and (iii) the original image is denormalized from the ImageNet normalization space (mean subtraction and standard deviation division inverted) to recover pixel values in the [0,1] range for visualization.

The resulting attribution heatmap, overlaid on the original handwriting image, highlights the pixel regions that most strongly influenced the model's classification decision. For spiral drawings from PD subjects, we expect the model to attend to the outer spiral rings where tremor-induced amplitude oscillations are most visible, and to regions of the drawing where speed reduction (bradykinesia) manifests as clustering of the pen trajectory. For wave drawings, the model is expected to highlight regions where the regular sinusoidal pattern breaks down, including areas of reduced amplitude (micrographia) and irregular inter-peak spacing.

8.5 Temporal Attention Visualization

The TA-LSTM's internal attention weights α_t (Equation 7) provide a third, architecture-native form of interpretability that does not require post-hoc attribution computation. The attention weights form a probability distribution over the T time steps in each patient's recording sequence, with higher weights assigned to time steps that the model considers most informative for UPDRS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

prediction. Plotting α_t against test_time yields a visualization of which recording sessions during the six-month trial were most influential for the model's progression estimate.

This temporal attention profile has potential clinical utility for identifying critical disease transition periods within a patient's monitoring trajectory—for example, a sudden increase in attention weight following a known medication change or after a reported fall event may suggest that the model is correctly identifying clinically significant events. Systematic validation of temporal attention patterns against clinical event records would be a valuable direction for future work, though it falls outside the scope of the present study.

8.6 XAI Pipeline Implementation Details

The generate_visual_explanations function takes as input the trained voice model, trained image model, corresponding data loaders, and the list of acoustic feature names, and produces both SHAP summary plots and IG heatmaps within a single function call. The function handles edge cases including the list return type from SHAP's GradientExplainer (which returns a list of SHAP value arrays for multi-output models, requiring extraction of the first element for single-output wrapped models) and the transposition of PyTorch's channel-first tensor format to NumPy's channel-last format required by Matplotlib.

The complete XAI pipeline is invoked within the run_full_pipeline master execution function after model training and loading, ensuring that explanations are generated from the best-performing model weights rather than from randomly initialized weights. A try-except block guards against execution failure when trained weights are not available, allowing the pipeline to fail gracefully during development.

CONCLUSION

This paper has presented a comprehensive Explainable Multimodal Deep Learning Framework for Parkinson's Disease Phenotyping and Progression Tracking that addresses three concurrent clinical tasks—binary PD detection, continuous UPDRS motor score regression, and ordinal Hoehn and Yahr disease stage classification—within a unified, interpretable architecture. The framework integrates three complementary biomarker modalities: 22-dimensional acoustic features from sustained phonation recordings processed by a regularized Tabular Voice Encoder; spiral and wave handwriting images analyzed by a fine-tuned Swin Transformer with hierarchical window-based self-attention; and longitudinal telemonitoring voice sequences modeled by a two-layer Temporal Attention-LSTM with soft attention-weighted context aggregation.

The multi-task learning formulation employs homoscedastic uncertainty-based task weighting [17], enabling the three loss terms (binary cross-entropy, Huber regression loss, and ordinal cross-entropy) to be jointly optimized without manual hyperparameter tuning of task-specific weights. The late fusion strategy, implemented as both weighted probability averaging and stacked logistic regression, combines complementary modality evidence while remaining robust to missing modalities in clinical deployment. Classification threshold optimization via Youden's J statistic ensures that sensitivity and specificity are balanced in a clinically meaningful manner.

The evaluation methodology enforces strict subject disjointness through Leave-One-Subject-Out cross-validation for the acoustic modality and five-fold stratified cross-validation for image modalities, eliminating the speaker identity confound that has inflated performance estimates in many prior studies. The XAI pipeline produces SHAP summary plots for acoustic feature attribution and Integrated Gradient heatmaps for handwriting image attribution, providing clinicians with evidence-based explanations that can support diagnostic reasoning and build trust in model recommendations.

Several limitations of the current framework merit acknowledgment. The dataset sizes are small: the Oxford voice dataset contains only 31 subjects, the Parkinson's Drawings dataset contains 110 subjects, and the telemonitoring dataset contains 42 subjects. These sizes are typical of existing PD biomarker research but limit the statistical power of the evaluation and the diversity of PD phenotypes represented. The programmatic derivation of H&Y stage targets from UPDRS scores introduces label approximation error. The late fusion strategy does not account for confidence calibration of individual modality models, and fused probabilities may therefore be poorly calibrated even when the fusion weights are optimally chosen.

Future directions include: (i) collection and integration of larger, ethnically diverse PD cohorts such as the PPMI (Parkinson's Progression Markers Initiative) dataset to improve generalizability; (ii) extension to additional modalities including gait accelerometry, facial expression analysis, and resting-state functional MRI connectivity; (iii) implementation of intermediate fusion using cross-modal attention to enable richer inter-modality interaction learning; (iv) longitudinal validation of temporal attention patterns against clinical event records; and (v) prospective evaluation in a clinical deployment setting to assess real-world diagnostic utility and clinician acceptance.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

In summary, the proposed framework represents a methodologically rigorous and clinically informed approach to multimodal PD assessment that balances predictive performance, evaluation integrity, and model interpretability. The modular architecture, standardized preprocessing pipelines, and integrated XAI tools position it as a practical foundation for further development toward clinical-grade diagnostic decision support systems.

REFERENCES

[1] C. W. Olanow and W. G. Tatton, "Etiology and pathogenesis of Parkinson's disease," Annual Review of Neuroscience, vol. 22, pp. 123–144, 1999.
[2] J. Parkinson, "An essay on the shaking palsy," Journal of Neuropsychiatry and Clinical Neurosciences, vol. 14, no. 2, pp. 223–236, 2002 (Reprinted from 1817).
[3] K. R. Kowal, T. Thompson, E. Roman, and K. Bhalla, "The current and projected economic burden of Parkinson's disease in the United States," Movement Disorders, vol. 28, no. 3, pp. 311–318, 2013.
[4] C. G. Goetz et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," Movement Disorders, vol. 23, no. 15, pp. 2129–2170, 2008.
[5] R. Martínez-Martín et al., "Inter-rater reliability of the Unified Parkinson's Disease Rating Scale: Assessment in multiple country settings," Movement Disorders, vol. 9, no. 4, pp. 423–428, 1994.
[6] D. J. Brooks, "Diagnosis and management of atypical parkinsonian syndromes," Journal of Neurology, Neurosurgery and Psychiatry, vol. 72, suppl. 1, pp. i10–i16, 2002.
[7] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," IEEE Transactions on Biomedical Engineering, vol. 57, no. 4, pp. 884–893, 2010.
[8] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," BioMedical Engineering OnLine, vol. 6, no. 1, p. 23, 2007.
[9] E. P. Letanneux, J. Danna, J.-L. Velay, F. Viallet, and S. Pinto, "From micrographia to Parkinson's disease dysgraphia," Movement Disorders, vol. 29, no. 12, pp. 1467–1475, 2014.
[10] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE Transactions on Biomedical Engineering, vol. 59, no. 5, pp. 1264–1271, 2012.
[11] K. R. Bhidayasiri and D. Tarsy, Parkinson's Disease: A Guide to Pharmacological Treatment. Totowa, NJ: Humana Press, 2012.
[12] L. Liang et al., "Integrating multi-modal data for joint genotype and phenotype prediction," Nature Biomedical Engineering, vol. 3, pp. 734–744, 2019.
[13] J. Ngiam et al., "Multimodal deep learning," in Proc. 28th International Conference on Machine Learning (ICML), Bellevue, WA, 2011, pp. 689–696.
[14] U.S. Food and Drug Administration, "Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD) Action Plan," U.S. Department of Health and Human Services, Washington, DC, Jan. 2021.
[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.
[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. 34th International Conference on Machine Learning (ICML), Sydney, Australia, 2017, pp. 3319–3328.
[17] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, 2018, pp. 7482–7491.
[18] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1015–1022, 2009.
[19] C. O. Sakar et al., "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 4, pp. 828–834, 2013.
[20] M. Shahbakhi, D. T. Far, and E. Tahami, "Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine," Journal of Biomedical Science and Engineering, vol. 7, no. 4, pp. 147–156, 2014.
[21] J. R. Orozco-Arroyave et al., "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in Proc. 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 2014, pp. 342–347.
[22] S. Hawi, P. Kaur, and M. Talib, "Parkinson's disease detection using 1D convolutional neural networks applied to vocal features," in Proc. International Conference on Data Science and Machine Learning, 2020, pp. 1–6.
[23] S. Grover, S. Bhartia, A. Yadav, and A. K. Seeja, "Predicting severity of Parkinson's disease using deep learning," Procedia Computer Science, vol. 132, pp. 1788–1794, 2018.
[24] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[25] P. Drotar et al., "Analysis of in-air movement in handwriting: A novel marker for Parkinson's disease," Computer Methods and Programs in Biomedicine, vol. 117, no. 3, pp. 405–411, 2014.
[26] C. Letanneux et al., "Handwriting analysis and Parkinson's disease: A review and new perspectives," Neurophysiologie Clinique, vol. 44, no. 3, pp. 293–300, 2014.
[27] P. Zham, D. K. Kumar, P. Dabnichki, S. Poosapadi Arjunan, and S. Raghav, "Distinguishing different stages of Parkinson's disease using composite index of speed and pen-pressure of sketching a spiral," Frontiers in Neurology, vol. 8, p. 435, 2017.
[28] C. R. Pereira et al., "A step towards the automated diagnosis of Parkinson's disease: Analyzing handwriting movements," in Proc. IEEE 28th International Symposium on Computer-Based Medical Systems, Sao Carlos, Brazil, 2015, pp. 171–176.
[29] V. Anand, "Parkinson's Disease Drawings," Kaggle Dataset, 2021. [Online]. Available: https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set
[30] L. Chen et al., "Squeeze-and-excitation networks for Parkinson's disease classification from handwriting images," IEEE Access, vol. 9, pp. 54478–54487, 2021.
[31] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 2021, pp. 10012–10022.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[32] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric of dysphonia severity," Journal of the Royal Society Interface, vol. 8, no. 59, pp. 842–855, 2011.
[33] Y. Zhao, J. Chen, and Z. Sun, "Deep learning-based UPDRS prediction in Parkinson's disease using longitudinal voice data," Computers in Biology and Medicine, vol. 124, p. 103921, 2020.
[34] J. Ren et al., "Attention-LSTM for remote Parkinson's disease monitoring and progression estimation," Neurocomputing, vol. 396, pp. 91–104, 2020.
[35] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.
[36] B. M. Eskofier et al., "An overview of smart shoes in the internet of health things: Gait and mobility assessment in health promotion and disease monitoring," Applied Sciences, vol. 7, no. 10, p. 986, 2017.
[37] S. Arora et al., "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," Parkinsonism & Related Disorders, vol. 21, no. 6, pp. 650–653, 2015.
[38] C. Quan et al., "Multimodal data fusion for detecting Parkinson's disease," Frontiers in Computational Neuroscience, vol. 15, p. 612946, 2021.
[39] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," AI Magazine, vol. 40, no. 2, pp. 44–58, 2019.
[40] E. Erion et al., "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," Nature Machine Intelligence, vol. 3, pp. 620–631, 2021.
[41] J. Adebayo et al., "Sanity checks for saliency maps," in Advances in Neural Information Processing Systems, vol. 31, 2018, pp. 9505–9515.
[42] N. Kokhlikyan et al., "Captum: A unified and generic model interpretability library for PyTorch," arXiv preprint arXiv:2009.07896, 2020.
[43] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
[44] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.
[45] Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in Proc. IEEE/CVF CVPR, Long Beach, CA, 2019, pp. 9268–9277.
[46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. 32nd International Conference on Machine Learning (ICML), Lille, France, 2015, pp. 448–456.
[47] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," arXiv preprint arXiv:1606.08415, 2016.
[48] G. Somepalli et al., "SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training," arXiv preprint arXiv:2106.01342, 2021.
[49] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. 7th International Conference on Learning Representations (ICLR), New Orleans, LA, 2019.
[50] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE/CVF CVPR, Miami, FL, 2009, pp. 248–255.
[51] R. Wightman, "PyTorch image models (timm)," GitHub, 2019. [Online]. Available: https://github.com/rwightman/pytorch-image-models
[52] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018, pp. 328–339.
[53] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in Advances in Neural Information Processing Systems, vol. 29, 2016, pp. 1019–1027.
[54] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 3104–3112.
[55] R. Caruana, "Multitask learning," Machine Learning, vol. 28, no. 1, pp. 41–75, 1997.
[56] S. Ruder, "An overview of multi-task learning in deep neural networks," arXiv preprint arXiv:1706.05098, 2017.
[57] P. J. Huber, "Robust estimation of a location parameter," Annals of Mathematical Statistics, vol. 35, no. 1, pp. 73–101, 1964.
[58] D. Shulman, J. Gruber-Baldini, K. Anderson, L. Fishman, and L. Reich, "The evolution of disability in Parkinson disease," Movement Disorders, vol. 23, no. 6, pp. 790–796, 2008.
[59] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
[60] S. van der Laan, M. J. Laan, and E. Polley, "Super Learner," Statistical Applications in Genetics and Molecular Biology, vol. 6, no. 1, pp. 1–21, 2007.
[61] W. J. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, no. 1, pp. 32–35, 1950.
[62] U.S. Food and Drug Administration, "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning-Based Software as a Medical Device," FDA Discussion Paper, 2019.
[63] B. J. Erickson et al., "Machine learning for medical imaging," RadioGraphics, vol. 37, no. 2, pp. 505–515, 2017.
[64] S. Erion, J. D. Janizek, P. Sturmfels, S. Lundberg, and S. Lee, "Learning explainable models using attribution priors," arXiv preprint arXiv:1906.10670, 2019.