Compact ResNet18 with Test-Time Adaptation: Balancing Accuracy and Latency for Parkinson's Disease Screening

Maria V. Leyba-Mesa, *IEEE member,* and Buket D. Barkana*, IEEE member*

***Abstract*— Parkinson's disease (PD) is a progressive neurodegenerative disorder where early and reliable diagnosis is critical for timely intervention and disease management. Handwriting analysis is recognized as a promising digital biomarker; however, deep learning models still face challenges with inter-patient variability and practical deployment constraints. In this work, we propose a Test-Time Adaptation framework, called TTA-ResNet18, built on a ResNet18 backbone for patient-specific PD detection across spiral and meander handwriting tasks. The proposed TTA ResNet18 model improves accuracy from 86.5% to 92.2% (spiral) and from 84.9% to 90.7% (meander), with corresponding gains in macro F1 and ROC AUC. Mean ROC AUCs of 0.925 ± 0.044 (spiral) and 0.914 ± 0.049 (meander) across a 5-fold cross-validation. Computational analysis shows that TTA introduces only modest runtime overhead while preserving the underlying architecture's efficiency. These results indicate that test time adaptation is a practical mechanism for improving PD handwriting classification performance with minimal additional computational cost**

I. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting approximately 1% of individuals over age 60, with prevalence expected to double by 2040 due to aging populations [1]. Early and accurate diagnosis remains crucial for timely intervention, disease management, and improved patient outcomes [2]. Traditional diagnostic methods rely heavily on clinical assessment and neurological examinations by specialists [3], which are inherently subjective, time-intensive, and often fail to detect subtle early-stage symptoms like mild motor fluctuations [4].

Handwriting analysis has emerged as a promising, non-invasive biomarker for PD detection [5]. Motor impairments characteristic of PD manifest in distinct patterns, including micrographia (progressive reduction in letter size), tremor-induced irregularities such as spiraling strokes, and decreased writing speed or pressure variability [6]. These kinematic and spatial features, captured from digitized handwriting samples via tablets or digitizing pens, provide objective, quantifiable data that correlate strongly with disease severity [7]. These features were typically fed into classical machine learning models, including support vector machines and decision trees, achieving moderate diagnostic performance but limited generalizability due to strong dependence on feature engineering.

With the rise of deep learning, convolutional neural networks (CNNs) have become the dominant paradigm for handwriting-based PD classification. Architectures such as continuous convolutional networks and EfficientNet variants have demonstrated improved performance by automatically learning hierarchical spatial features from raw images. For instance, Li et al. [7] proposed a CNN-based framework for offline handwriting analysis that improved the accuracy of early PD detection, while more recent work using EfficientNetV2 architectures [8] has shown further gains in classification metrics across the spiral and wave datasets.

Despite these advances, prevailing deep learning models for handwriting-based PD detection face significant hurdles. They often struggle with inter-patient variability, where individual differences in baseline handwriting styles, age-related tremors unrelated to PD, or comorbidities confound generalization [9]. Moreover, these models are typically computationally burdensome: large architectures require substantial GPU resources during both training and inference, limiting deployment on edge devices or in resource-constrained clinical settings such as primary care offices [10]. This inefficiency hampers real-world scalability, especially for screening large at-risk populations where low-latency processing is essential.

Test-time adaptation (TTA) offers a principled way to address these challenges. Broadly, TTA refers to a family of techniques in which a model updates a small subset of its parameters or normalization statistics at inference time, using only the unlabeled test data it receives [11]. Instead of treating the model as fixed after training, TTA exploits distributional information from each new input, or batch of inputs, to reduce train–test mismatch and improve robustness without requiring additional labeled examples [12]. In medical applications, this enables the model to compensate for patient-specific shifts, such as variations in writing styles, acquisition devices, or noise characteristics, while keeping the core architecture lightweight and efficient [13].

Despite the progress in both handwriting-based PD detection and test-time adaptation, a clear gap remains at their intersection. Existing handwriting classification models do not account for patient-specific variability at inference time,

M. V. Leyba Mesa is with the Biomedical Engineering Department, The University of Akron, Akron, OH, 44325 USA (phone: 216-250-6850; e-mail: mvl16@uakron.edu).

B. D. Barkana is with the Biomedical Engineering Department, The University of Akron, Akron, OH, 44325 USA (e-mail: bbarkana@uakron.edu).

while current TTA approaches have not been tailored to lightweight, interpretable adaptation in this domain.

This paper addresses this gap by introducing a compact and patient-specific test-time adaptation framework built on a ResNet18 backbone. Unlike prior work, the proposed method:

- avoids full-network adaptation by introducing a low-dimensional feature shift vector,
- enables per-patient personalization without labeled data, and
- maintains low computational overhead suitable for clinical deployment.

In this way, the proposed approach bridges the gap between high-performance deep learning models and practical, adaptive systems for real-world PD screening.

We evaluate our approach on the HandPD dataset and demonstrate that patient-specific adaptation improves classification performance while maintaining computational efficiency suitable for clinical deployment.

II. Methods

A. Dataset and Preprocessing

We evaluate on the HandPD dataset [14] comprising hand-drawn spirals and meanders from 92 participants: 18 controls and 74 PD patients. Each patient contributes 4 drawings, yielding 736 images in total, with 368 images in each of the spiral and meander subsets. Images capture characteristic PD tremors through spiral and meander irregularities, as illustrated in Fig. 1. The dataset exhibits significant class imbalance (72 control images vs 296 PD images per subset), reflecting real-world screening scenarios in which controls are less prevalent.

Images were preprocessed with the following pipeline: (1) Resizing to 256×256 pixels followed by random cropping to 224×224, (2) Data augmentation: horizontal flip (p=0.5), rotation (±15°), color jitter, (3) ImageNet normalization and random erasing (p=0.25). Patient-level stratified 5-fold cross-validation ensured that all samples from a given patient appeared exclusively in either training or testing sets, preventing data leakage and providing realistic performance estimates.

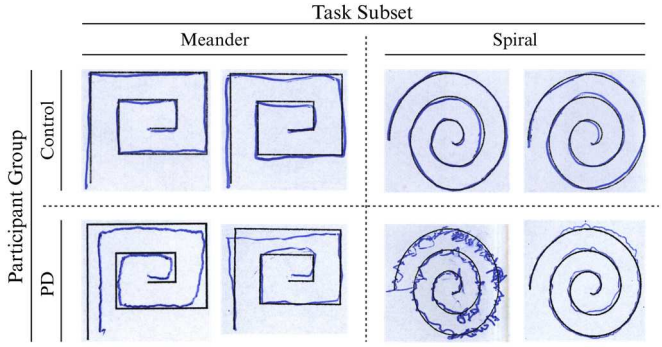

Figure 1. Representative handwriting samples from the HandPD dataset. The figure displays typical spiral and meander task examples from both control (top row) and PD patient (bottom row) groups.

B. Model Architecture

Our architecture builds upon the ResNet18 pre-trained on ImageNet and is modified for patient-specific adaptation. The model consists of, (1) Feature Extractor: ResNet18 backbone (early layers frozen for stability) (2) Global Average Pooling: Reduces spatial dimensions to 512-D patient embedding, (3) Batch Normalization: 512-dimensional feature normalization, (4) Adaptation Layer: Learnable delta parameter (Δ) for patient-specific shifts (5) Dropout (p=0.5) and fully connected classifier (512→2). The proposed model architecture, TTA-ResNet18, is illustrated in Fig. 2.

To enable patient-specific adaptation, a learnable feature shift vector $\Delta \in \mathbb{R}^{512}$ was introduced at the feature level. This vector is added to the pooled feature representation during test-time adaptation, allowing controlled, low-dimensional personalization without modifying the core network weights. This lightweight formulation ensures both computational efficiency and clinical interpretability, as Δ directly encodes patient-specific deviations amenable to visualization and clinical correlation

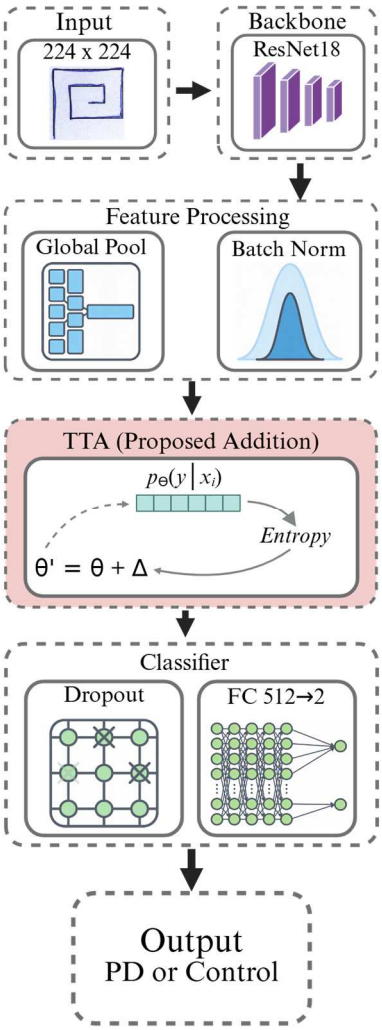

Figure 2. Proposed TTA-ResNet18 architecture for patient-specific PD handwriting classification.

C. Test-Time Adaptation

To address inter-subject variability during inference, we employ an unsupervised TTA strategy based on entropy minimization [11]. For each test patient, all available handwriting samples are jointly adapted, assuming predictions should be consistent across samples from the same subject.

Let $p_\Theta(y \mid x_i)$ denote the model's probability assigned to class y for sample x_i and let N be the number of samples for a patient. The adaptation objective minimizes the average predictive entropy (1), encouraging confident, low-entropy predictions. It is selectively applied to uncertain samples, defined as those with a maximum class probability below 0.9. By focusing on ambiguous predictions, the adaptation process avoids reinforcing already confident outputs.

$$\mathcal{H}_{TTA} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{y\in\{0,1\}} p_\theta(y|x_i+\Delta)\log p_\theta(y|x_i+\Delta) \quad (1)$$

This objective encourages confident, low-entropy predictions and is selectively applied to uncertain samples, defined as those with maximum class probability below 0.9. By focusing on ambiguous predictions, the adaptation process avoids reinforcing already confident outputs. Rather than updating the full network parameters, we introduce a constrained, patient-specific offset (2) where the offset Δ is optimized over 30 iterations using the Adam optimizer with a learning rate of 10^{-3}. Bounding Δ limits the magnitude of adaptation and mitigates overfitting, ensuring that patient-level personalization does not degrade the globally learned representation.

$$\theta' = \theta + \Delta, \Delta \in [-0.5, 0.5] \quad (2)$$

This approach exploits the low-density separation principle, encouraging decision boundaries to move away from high-density regions of patient-specific data while preserving the original classifier structure. As a result, the model adapts to individual handwriting characteristics without requiring labeled test data.

D. Training Strategy

The model was trained using the Adam optimizer with weight decay of 1×10^{-4} and a cosine-annealed learning rate schedule of 3×10^{-4}. Cross-entropy loss with label smoothing was employed over 10 epochs to improve robustness under class imbalance and limited sample sizes. Gradient clipping was applied to stabilize training. All training was performed using mini-batches sampled from patient-disjoint training sets.

Experiments were conducted on a high-performance workstation equipped with an NVIDIA RTX 2000 Ada Generation for accelerated training and inference. The system also included an Intel(R) Core(TM) Ultra 9 285 (2.50 GHz) processor and 32GB of system memory. The experiments were implemented in Python using PyTorch as the deep learning framework, with CUDA and cuDNN support enabled for GPU computation. Model training and testing were carried out on this setup to ensure consistent performance across all trials.

E. Evaluation Metrics

Given the strong class imbalance, performance was primarily assessed using balanced accuracy (3), sensitivity (4), precision (5), macro-averaged F1-score (6), calculated from the F1-Score per class (7), and ROC-AUC. Confusion matrices and ROC curves were generated for each fold, along with mean ROC curves across folds. These metrics provide a robust assessment of diagnostic performance under skewed class distributions.

$$\text{Balanced Acc} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (3)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$F1_{macro} = \frac{1}{2}(F1_{PD} + F1_{control}) \quad (6)$$

$$F1_{class} = 2 \cdot \frac{Precision_{class} \cdot Sensitivity_{class}}{Precision_{class} + Sensitivity_{class}} \quad (7)$$

ROC-AUC was calculated as the area under the receiver operating characteristic curve, integrating true positive rate (TPR = Sensitivity) against false positive rate (FPR = FP/(FP+TN)) across decision thresholds.

To assess practical feasibility, we conducted a detailed computational cost analysis. Training time, test-time adaptation time, inference time, GPU memory usage, and parameter counts were recorded for each fold. All computational statistics were aggregated and reported alongside classification performance.

III. Results

A. Classification Performance

Table 1 summarizes the classification performance of the proposed TTA-enhanced ResNet18 compared to the static baseline across the spiral and meander handwriting tasks. For the spiral task, the baseline model achieved an accuracy of 86.5%, sensitivity of 0.919, precision of 0.938, macro-F1 of 0.831, and ROC-AUC of 0.901, while the TTA model obtained 92.2% accuracy, sensitivity of 0.942, precision of 0.977, macro-F1 of 0.876, and ROC-AUC of 0.928. Similarly, on the meander task, the baseline reached 84.9% accuracy, sensitivity of 0.907, precision of 0.928, macro-F1 of 0.809, and an ROC-AUC of 0.893, whereas the TTA variant achieved 90.7% accuracy, sensitivity of 0.936, Precision of 0.974, macro-F1 of 0.864, and ROC-AUC of 0.917.

Figure 3 presents the ROC analysis for the TTA-enhanced ResNet18 across both spiral and meander handwriting tasks from the HandPD dataset, evaluated using 5-fold cross-validation.

TABLE I. CLASSIFICATION PERFORMANCE

Metric	Task Subset			
	Spiral Proposed	***Spiral ResNet18***	***Meander Proposed***	***Meander ResNet18***
Balanced Acc.	92.2%	86.5%	90.7%	84.9%
Sensitivity	0.942	0.919	0.936	0.907
Precision	0.977	0.938	0.974	0.928
Macro F1	0.876	0.831	0.864	0.809
ROC AUC	0.925	0.901	0.917	0.893

The ROC curves for individual folds (left panels) demonstrate consistent discriminative performance, with all folds achieving AUC values between 0.87–0.98 for spiral and 0.86–0.99 for meander tasks. The mean ROC curve with shaded standard deviation (right panels) further confirms robust generalization, yielding mean AUCs of 0.925 ± 0.044 (spiral) and 0.914 ± 0.049 (meander).

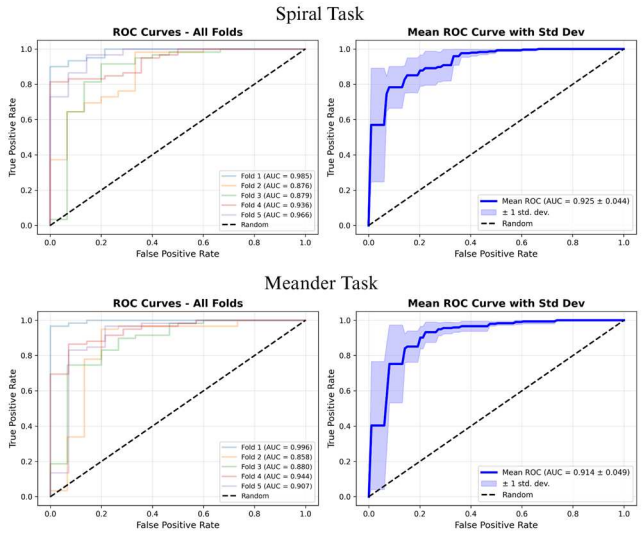

Figure 3. ROC curves for all folds of both the spiral and meander subsets using the proposed TTA-ResNet18 model. The X-axis shows the True Positive Rate, and the Y-axis shows the False Positive Rate.

B. Computational Efficiency

Table 2 reports the computational cost for the baseline and TTA models in terms of total trainable parameters, inference time per sample, peak memory usage, and total runtime per task. Despite introducing an additional adaptation step, the TTA configuration increases the parameter count by only 1.2% compared to the baseline and incurs 22.4% overhead in FLOPs for the spiral task and 10.9% for meander. For the spiral task, total runtime increases from 154.3 s to 188.8 s (+22.4%), while for the meander task, it changes from 137.8 s to 152.8 s (+10.9%). Peak memory consumption remains at 169 MB across all variants, indicating that the method is compatible with commodity GPU hardware and suitable for edge deployment.

TABLE II. COMPUTATIONAL EFFICIENCY REPORT

Metric	Task Subset			
	Spiral Proposed	***Spiral ResNet18***	***Meander Proposed***	***Meander ResNet18***
Total Parameters	11,179,074	11,178,562	11,179,074	11,178,562
Training Time	155.9 s	152.4 s	123.2 s	136.0 s
Inference Time	0.8 s	1.9 s	0.7 s	1.8 s
TTA Time	32.1 s	-	28.9 s	-
Peak Memory	169 MB	169 MB	169 MB	169 MB
Total Runtime	188.8 s	154.3 s	152.8 s	137.8 s

Within this framework, we also quantified the computational overhead introduced by test-time adaptation. Compared to the static baseline, the proposed TTA-ResNet18 model added an average of 78 ms of processing time per patient during inference, corresponding to the additional optimization steps required to update the patient-specific shift vector Δ.

IV. DISCUSSION

The experimental results demonstrate that test-time adaptation can enhance handwriting-based PD classification. Across both spiral and meander tasks, the proposed TTA-enhanced ResNet18 yields consistent improvements in PD detection performance in accuracy, sensitivity, precision, macro F1, and ROC AUC relative to the static baseline. The increase of up to 6% in accuracy and macro-F1 on both tasks suggests that the adaptation step effectively compensates for patient-specific differences in handwriting style and disease manifestation rather than merely overfitting to folds (Table 1). The 5-fold ROC analysis further supports this interpretation: the narrow spread of AUC values across folds and the high mean ROC-AUCs indicate that the approach generalizes well despite the intrinsic heterogeneity of the HandPD dataset (Fig. 3).

The inclusion of macro F1-score alongside accuracy provides a more balanced and informative assessment of performance in the presence of potential class imbalance between PD and control subjects. Unlike accuracy, which can obscure poor minority-class performance, macro F1 equally weights precision and recall across classes, offering a clearer view of how well the model performs for both PD and healthy controls. The improvements observed in macro F1 therefore indicate that test-time adaptation enhances discriminative performance across classes, rather than disproportionately benefiting the majority class or exploiting dataset priors.

In addition, the consistent gains in ROC AUC suggest that TTA improves the overall separability of prediction score distributions. This property is particularly relevant in clinical settings, where decision thresholds may be adjusted to emphasize sensitivity for early screening or specificity for confirmatory diagnosis. Improved score separability enables more flexible threshold selection without substantial performance degradation, thereby increasing the model's practical utility across diverse clinical use cases.

From a deployment standpoint, the computational profile is equally compelling. The 78 ms per-patient overhead represents a small addition for clinical workflows, equivalent to ~0.5 seconds for a typical 6-image screening session. This latency may be dwarfed by human assessment times, enabling real-time adaptation during patient intake.

Importantly, the proposed approach requires neither labeled test data nor patient-specific retraining. This eliminates the need for manual annotation, repeated offline optimization, or storage of patient data beyond the inference window, factors that often constitute significant barriers to clinical deployment. By avoiding explicit fine-tuning, the method evades common privacy, regulatory, and logistical challenges associated with adapting models in healthcare settings, where data governance constraints and limited access to expert labels are the norm rather than the exception.

As a result, test-time adaptation provides a practical and scalable mechanism for mitigating covariate shift across acquisition devices, clinical sites, and patient populations. This capability is particularly valuable in real-world deployments, where handwriting characteristics may vary substantially due to differences in hardware, task instructions, or patient motor variability, and where static models often degrade in performance when exposed to such shifts.

Several limitations of the current study warrant consideration. First, the evaluation is limited to a single backbone architecture (ResNet18) and a single dataset (HandPD). Second, the analysis primarily focuses on aggregate performance metrics. A more granular evaluation of per-patient performance, model calibration, and robustness under controlled domain shifts (e.g., variations in writing conditions or acquisition hardware) would provide deeper insight into clinical reliability. Finally, although the computational overhead of TTA is small, additional optimizations could further reduce inference latency while preserving adaptation benefits.

Future work can explore several directions. On the modeling side, combining TTA with model compression techniques, such as pruning and quantization, could further reduce latency and energy consumption while preserving, or even enhancing, the observed performance gains. On the evaluation side, extending the analysis to include calibration metrics, per-patient trajectories, and out-of-distribution scenarios would help quantify the robustness of the adapted model under deployment-like conditions. Finally, integrating the proposed framework into a broader multimodal PD assessment pipeline, such as combining handwriting with speech, gait, or imaging biomarkers, could yield a more comprehensive and reliable digital diagnostic and monitoring tool.

V. Conclusion

This study introduced a test-time adaptation framework, called TTA-ResNet18, for patient-specific PD detection from spiral and meander handwriting tasks, using a compact ResNet18 backbone. By optimizing a lightweight adaptation module during inference, the proposed method significantly improves accuracy, sensitivity, precision, macro F1, and ROC AUC over a static baseline while maintaining a modest computational footprint. The ROC analyses across multiple cross-validation folds highlight the robustness of the approach and its ability to generalize across heterogeneous handwriting patterns. Although TTA introduces a small increase in total runtime, peak memory remains low and training cost unchanged, making the method suitable for deployment in resource-constrained clinical environments.

Taken together, the results show that test-time adaptation is an effective and practical strategy to address interpatient variability in handwriting-based PD detection. The framework offers a promising direction for building adaptive, efficient, and clinically meaningful decision support tools that can better accommodate the diversity of real-world patient populations.

References

[1] E.R. Dorsey, T. Sherer, M. S. Okun, and B.R. Bloem. “The emerging evidence of the Parkinson pandemic”. Journal of Parkinson’s disease, vol.8, no. s1, pp. S3-8, 2018.

[2] P. Mahlknecht, K. Marini, M. Werkmann, W. Poewe, and K. Seppi. “Prodromal Parkinson's disease: hype or hope for disease-modification trials?”. Translational neurodegeneration, vol. 11, no. 1, pp. 11, 2022.

[3] L. di Biase, P. M. Pecoraro, and V. Di Lazzaro. “Validating the accuracy of Parkinson's disease clinical diagnosis: a UK Brain bank case–control study”. Annals of Neurology, vol. 97, no. 6, pp. 1110-1121, 2025.

[4] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino. “Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis”. Neurology, vol. 86, no. 6, pp. 566-576, 2016.

[5] G. Marano, S. Rossi, E. M. Marzo, A. Ronsisvalle, L. Artuso, G. Traversi, A. Pallotti, F. Bove, C. Piano, A. R. Bentivoglio, and G. Sani. “Writing the Future: Artificial Intelligence, Handwriting, and Early Biomarkers for Parkinson’s Disease Diagnosis and Monitoring”. Biomedicines, vol. 13, no. 7, pp. 1764, 2025.

[6] F. Asci, G. Saurio, G. Pinola, M. Falletti, A. Zampogna, M. Patera, F. Fattapposta, S. Scardapane, and A. Suppa. “Micrographia in Parkinson's disease: automatic recognition through artificial intelligence”. Movement Disorders Clinical Practice, vol. 12, no. 12, pp. 2187-96, 2025.

[7] Z. Li, J. Yang, Y. Wang, M. Cai, X. Liu, and K. Lu. “Early diagnosis of Parkinson's disease using Continuous Convolution Network: Handwriting recognition based on off-line hand drawing without template”. Journal of biomedical informatics, vol. 130, p. 104085, 2022.

[8] A. Razaq, S. Ramzan, S. Jabbar, M. M. Iqbal, M. A. Habib, and U. Raza. “A Deep Learning Framework for Early Parkinson’s Disease Detection: Leveraging Spiral and Wave Handwriting Tasks with EfficientNetV2-S”. Diagnostics, vol. 15, no. 21, pp. 2795, 2025.

[9] A. Sar, P. S. Puri, H. Naz, S. Aich, T. Choudhury, and L. A. Gabralla. “Multi-modal deep learning framework for early detection of Parkinson’s disease using neurological and physiological data for high-fidelity diagnosis”. Scientific Reports, vol. 15, no. 1, pp. 34835, 2025.

[10] M. A. Islam, M. Z. Majumder, M. A. Hussein, K. M. Hossain, and M. S. Miah. “A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets”. Heliyon, vol. 10, no. 3, 2024.

[11] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. “Tent: Fully test-time adaptation by entropy minimization”. arXiv preprint, arXiv:2006.10726, 2020.

[12] J. Liang, R. He, and T. Tan. “A comprehensive survey on test-time adaptation under distribution shifts”. International Journal of Computer Vision, vol. 133, no. 1, pp. 31-34, 2025.

[13] W. Ma, C. Chen, S. Zheng, J. Qin, H. Zhang, and Q. Dou. “Test-time adaptation with calibration of medical image classification nets for label distribution shift”. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022, pp. 313-323.

[14] C. R. Pereira, D. R. Pereira, F. A. Silva, J. P. Masieiro, S. A. Weber, C. Hook, and J. P. Papa. “A new computer vision-based approach to aid the diagnosis of Parkinson's disease”. Computer methods and programs in biomedicine, vol. 136, pp. 79-88, 2016.