Accessible, At-Home Detection of Parkinson's Disease via Multi-Task Video Analysis

Md Saiful Islam[1]; Tariq Adnan[1]; Jan Freyberg[2]; Sangwu Lee[1]; Abdelrahman Abdelkader[1]; Meghan Pawlik[3]; Cathe Schwartz[4]; Karen Jaffe[4]; Ruth B. Schneider[3]; Ray Dorsey[3]; Ehsan Hoque[1]

[1]University of Rochester, [2]Google Research, [3]University of Rochester Medical Center, [4]InMotion

Summary

Challenge: Limited access to neurological care leads to missed diagnosis of Parkinson's Disease (PD), the fastest-growing neurological disorder[1].

Proposed Solution: Introduced the largest multi-task video dataset (**finger tapping**, **facial expression**, **speech**) from **845** participants (**272 PD**) and a multimodal fusion network (**UFNet)** for comprehensive PD assessment.

Performance: Achieved **87.3%** predictive **accuracy and 92.8% AUROC**. Built-in uncertainty measures enhance reliability by **withholding predictions** in cases of low model confidence.

Global Impact: The proposed framework promotes **health equity** by enabling **accessible, home-based PD screening** using just a **webcam and microphone.**

Why Multimodal Learning

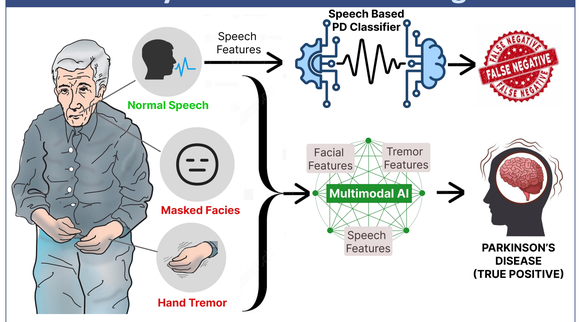
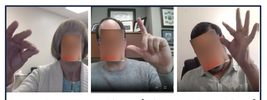


Figure 1. *Multi-faceted disease PD needs multimodal AI for comprehensive assessment as a unimodal model fails if symptoms are absent in that modality. Part of the figure was obtained from Lees et al.[2]*

Dataset

- **Diverse** participants recruited via a brain health study registry, social media, a PD wellness center, and clinician referrals.
- Approved by the University of Rochester **IRB.**

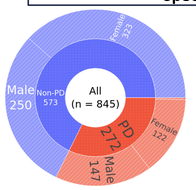
Facial Expression (Smile)



Motor Function (Finger-tapping)

The quick brown fox jumps over the lazy dog. The dog wakes up and follows the fox into the forest. But again, the quick brown fox jumps over the lazy dog.

Speech (Pangram Utterance)



Global participants
Age: 18 – 93

Figure 2. *Dataset overview*

Feature Extraction

Smile and Finger Tapping Tasks

Hand-crafted features[3,4] outperformed deep video models (ViViT[5], TimeSformer[6], VideoMAE[7], and Uniformer[8)] in terms of accuracy and resource-utilization.

Speech Task

WavLM[9] outperformed other feature choices[10].

Finger-tapping Features

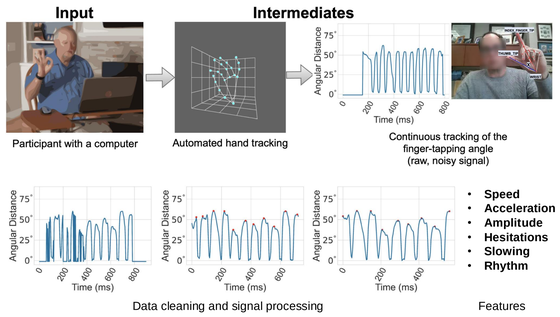


Figure 3. *Overview of feature-extraction from the finger-tapping task[3]*

Smile Features

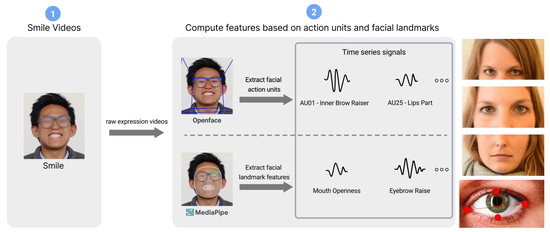


Figure 4. *Overview of feature-extraction from the smile task[4]*

Speech Features

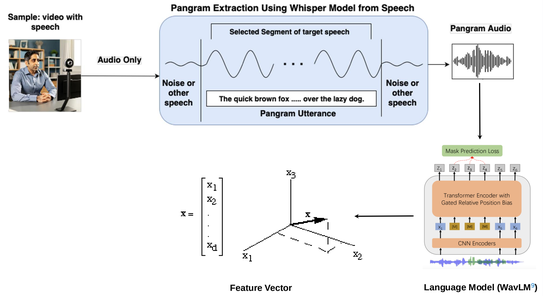


Figure 5. *Overview of feature-extraction from the speech task[10]*

Task-specific Modeling

Model Choice

Shallow neural network: Light architecture considering the dataset size and structured features.

Uncertainty Modeling

Monte-Carlo (MC) Dropout was applied to obtain multiple random predictions at inference time. **Standard deviation of logits** was used to estimate task-specific uncertainty.

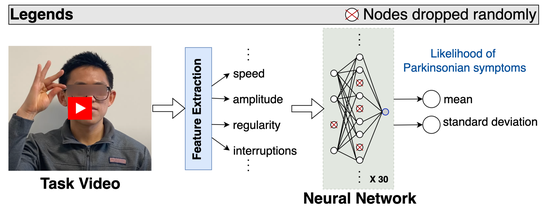


Figure 6. *Task-specific modeling with Monte Carlo dropout*

Multi-Task Fusion

Projection

Task-specific features are first projected into the same shared dimension using task-specific projection layer.

Calibrated Self-Attention

- **Customized self-attention** prioritizes informative tasks while accounting for **task-specific uncertainty.**
- **Down-weights** contributions from tasks with higher prediction uncertainty to improve reliability.

$$A = softmax(\frac{Q.K^T}{\sqrt{d}} - \eta\Sigma)$$

- $\Sigma = [\sigma_1, \sigma_2, \sigma_3] \rightarrow$ standard deviations of task-specific logits
- η (hyper-param) $\rightarrow$ controls the weight of the forced calibration

Final Predictor

- **Linear layer** for **binary** classification (PD/Non-PD)
- Uncertain predictions are withheld using MC Dropout

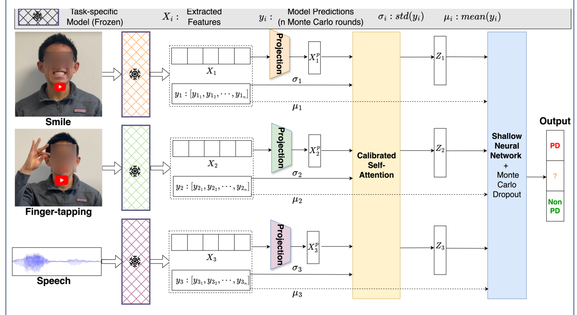


Figure 7. *Overview of the UFNet architecture*

Results

Effect of Multi-Task Modeling

Task Combination	Accuracy	F_1 score	AUROC
All three tasks	$\mathbf{87.3 \pm 0.4}$	$\mathbf{81.0 \pm 0.6}$	$\mathbf{92.8 \pm 0.2}$
Tapping + Smile	78.0 ± 0.8	65.6 ± 1.7	84.8 ± 0.5
Tapping + Speech	84.1 ± 0.3	77.3 ± 0.4	91.4 ± 0.2
Smile + Speech	85.2 ± 0.3	75.0 ± 0.4	91.2 ± 0.1
Tapping	73.1 ± 0.7	61.7 ± 0.9	74.9 ± 0.7
Smile	77.6 ± 0.2	67.5 ± 0.3	83.6 ± 0.1
Speech	85.1 ± 0.2	72.1 ± 0.6	87.8 ± 0.1

Table 1. *Multi-task combinations perform significantly better than corresponding single tasks.*

Analysis of Bias and Model Calibration

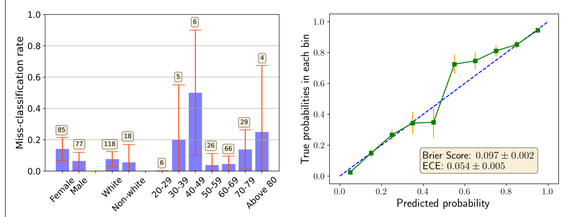


Figure 8. *(Left) Misclassification rate of the best UFNet model across demographic subgroups; (Right) Calibration curve showing the alignment between predicted probability and true observations.*

Comparison against Baselines and Ablation Studies

Model	Accuracy	AUROC	F_1 score	Precision	Recall
Baseline models					
Majority Voting	85.3	89.6	78.2	80.0	76.5
Neural Late Fusion	84.1 ± 0.4	91.7 ± 2.2	73.2 ± 8.3	73.5 ± 7.5	76.3 ± 9.4
Early Fusion Baseline	83.6 ± 0.6	91.0 ± 0.2	76.7 ± 0.7	75.4 ± 1.1	78.1 ± 0.9
Hybrid Fusion Baseline	84.1 ± 0.3	91.4 ± 0.2	77.3 ± 0.4	76.2 ± 0.7	$\mathbf{78.6 \pm 0.6}$
Attention variants					
Dot product self-attention[11]	85.5 ± 0.4	$\mathbf{92.9 \pm 0.2}$	78.3 ± 0.6	80.7 ± 0.6	76.1 ± 1.1
LRFormer[12]	$\underline{86.2 \pm 0.5}$	92.6 ± 0.3	79.5 ± 0.7	81.7 ± 0.9	77.6 ± 1.0
UFNet (ours)	$\underline{\mathbf{87.3 \pm 0.4}}$	92.8 ± 0.2	$\mathbf{81.0 \pm 0.6}$	$\underline{\mathbf{83.8 \pm 0.5}}$	78.4 ± 1.0
Early vs. hybrid fusion					
Early Fusion	$\underline{86.7 \pm 0.5}$	92.7 ± 0.3	79.9 ± 0.8	$\underline{83.3 \pm 0.7}$	76.9 ± 1.4
Hybrid Fusion	$\underline{\mathbf{87.3 \pm 0.4}}$	92.8 ± 0.2	$\mathbf{81.0 \pm 0.6}$	$\underline{\mathbf{83.8 \pm 0.5}}$	78.4 ± 1.0

Table 2. *UFNet performed significantly better than traditional fusion approaches. Ablation shows the efficacy of the proposed attention module.*

Discussion

Ethics: Mispredictions in PD detection can cause harm — false positives may lead to stress and financial burden, while false negatives delay essential care.

Bias: Our model performance is consistent across sex and ethnic subgroups, but accuracy drops for ages below 50 and above 80.

Future work: Expand the model for non-English speakers and tailor decision thresholds based on individual preferences and healthcare settings.

Dataset access: We release extracted features and code for extending the dataset (QR code below), but raw video data cannot be shared due to HIPAA compliance.

Live demo: Scan the QR code below to try it out.

Contact

mislam6@ur.rochester.edu
tadnan@ur.rochester.edu
mehoque@cs.rochester.edu

Extended Paper (AAAI 2025)

Code & Data

Live Demo

References

1. Dorsey, E. Ray, et al. "The emerging evidence of the Parkinson pandemic." *Journal of Parkinson's disease* 8.s1 (2018): S3-S8.
2. Lees, Andrew J., John Hardy, and Tamas Revesz. "Parkinson's disease." The Lancet 373.9680 (2009): 2055-2066.
3. Islam, Md Saiful, et al. "Using AI to measure Parkinson's disease severity at home." *npj Digital Medicine* 6.1 (2023): 156.
4. Adnan, Tariq, et al. "Unmasking Parkinson's Disease with Smile: An AI-enabled Screening Framework." *arXiv preprint arXiv:2308.02588* (2023).
5. Arnab, Anurag, et al. "Vivit: A video vision transformer." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
6. Bertasius, G., Wang, H., & Torresani, L. (2021, July). Is space-time attention all you need for video understanding?. In *ICML* (Vol. 2, No. 3, p. 4).
7. Tong, Zhan, et al. "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training." *Advances in neural information processing systems* 35 (2022): 10078-10093.
8. Li, Kunchang, et al. "Uniformer: Unifying convolution and self-attention for visual recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10 (2023): 12581-12600.
9. Chen, Sanyuan, et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1505-1518.
10. Adnan, Tariq, et al. "A Novel Fusion Architecture for PD Detection Using Semi-Supervised Speech Embeddings." arXiv preprint arXiv:2405.17206 (2024).
11. Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).
12. Ye, Wenqian, et al. "Mitigating transformer overconfidence via Lipschitz regularization." *Uncertainty in Artificial Intelligence*. PMLR, 2023.