Robust and Fair Parkinson's Disease Detection: A Stacked Ensemble Learning Approach Across Age Groups

Ying-Chang Wu, Yu-Chen Kuo, Sheng-Fu Liang*

Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan

I. Introduction:

Background & Motivation: Voice analysis is a **vital non-invasive biomarker** for Parkinson's Disease (PD), but reliable machine learning tools face challenges due to severe age-dependent bias:

- **Older Subjects (Age >= 60):** Aging voice features mimic PD, causing high **False Positives**.
- **Younger Subjects (Age < 60):** Subtle early symptoms are missed, leading to critical **False Negatives**.

Objective: To propose a **Stacked Ensemble Framework** to mitigate this age bias and significantly improve **diagnostic sensitivity** in early-onset patients.

II.Material and Methodolgy:

The dataset reflects clinical prevalence but introduces severe age bias: The Older Group (age >= 60) is PD dominant (81.6%), while the Younger Group (age < 60) is Control dominant (20.9% PD prevalence). This imbalance causes single-classifier models to exhibit poor sensitivity for younger PD patients.

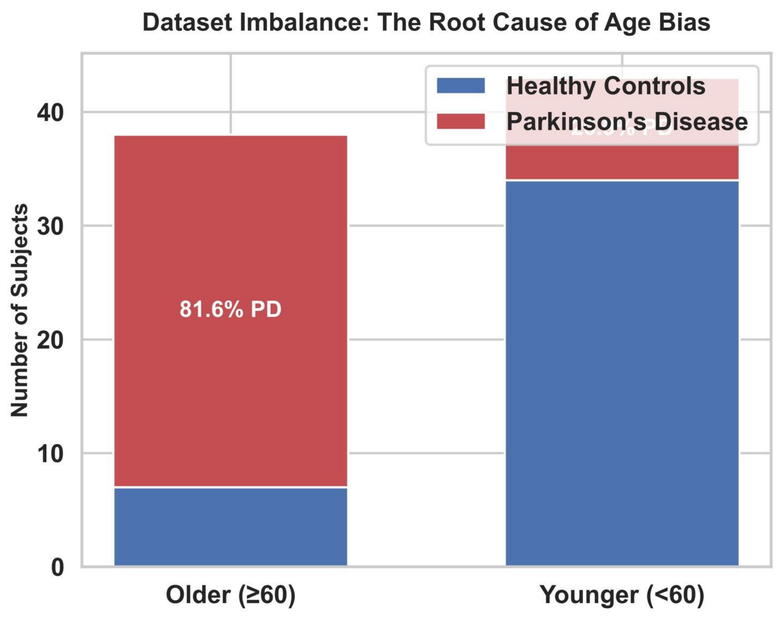

Figure 1: Dataset Imbalance Across Age Groups

Proposed Framework:

We utilize a **Stacked Generalization Architecture** to address age bias:

- **Feature Input:** Multi-modal acoustic features (**MFCCs, LPCs**) capture subtle variations.
- **Architecture:** Consists of **Level-0 heterogeneous base learners** (LR, RF, XGBoost, SVC) integrated by a **Level-1 Logistic Regression Meta-Learner**.
- **Key Advantage:** The meta-learner effectively **corrects systematic diagnostic errors** specific to age subgroups.

III. Experiment and Reaults:

The Stacked Ensemble achieved the highest Overall Accuracy (0.765). Crucially, it significantly boosted sensitivity to 0.778 in the younger group (16.6% relative gain) while maintaining 0.871 in the older group, confirming effective age-bias mitigation.

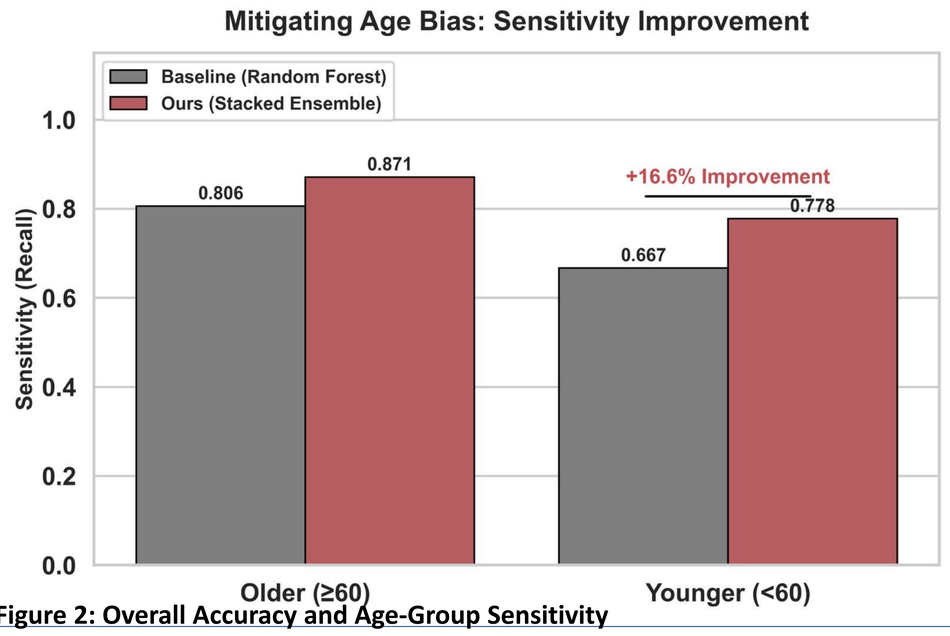

Figure 2: Overall Accuracy and Age-Group Sensitivity

III. Experiment and Reaults:

Minimizing False Negatives (FNs) is the primary clinical priority. Our model significantly reduced the miss rate from 9 FNs (Baseline) to only 6 FNs (Ours). This improvement confirms the model's efficacy in lowering the critical miss rate, making it a safer and more reliable tool for initial PD screening.

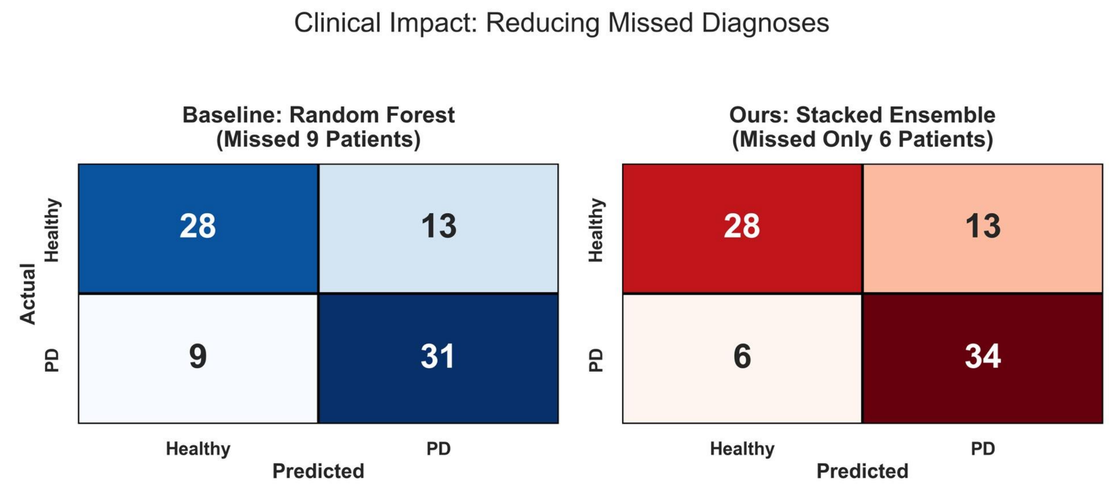

Figure 3: Confusion Matrix and False Negative Reduction

Our model achieves SOTA overall performance, with the highest Accuracy (0.765) and AUC (0.768), significantly outperforming all single, non-ensemble baselines. This robustness is attributed to successfully combining linear and non-linear model strengths across diverse demographics.

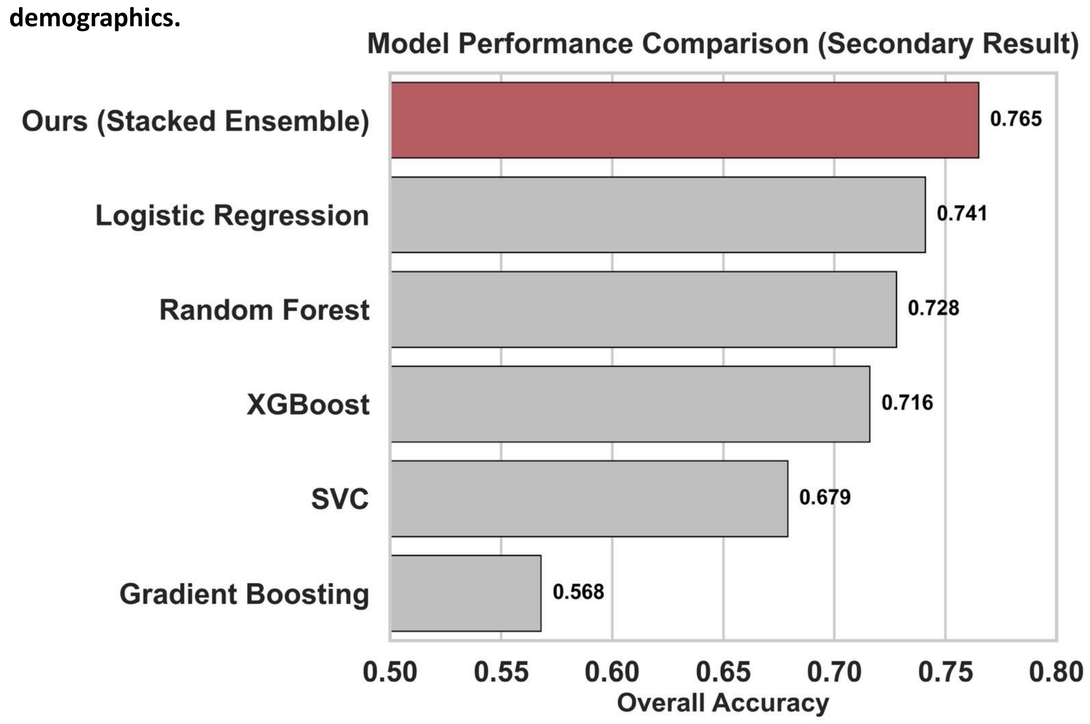

Figure 4: Overall Performance and State-of-the-Art (SOTA) Comparison

IV. Conclusion and Future Work:

Conclusion:

- Traditional machine learning classifiers are susceptible to severe **age-related bias** in PD diagnosis.
- The proposed **Stacked Ensemble Architecture** successfully mitigates this problem, resulting in a diagnostic tool that demonstrates greater **balance, stability, and reliability** across all age groups.
- This robust approach is highly applicable to broader medical AI developments.

Future Work:

- We will incorporate the analysis of **temporal dynamic features of speech** to further enhance model sensitivity to subtle, time-varying acoustic cues, which is crucial for early disease detection and clinical progression monitoring.

Acknowledgment:

The dataset was obtained from the figshare repository at https://figshare.com/articles/dataset/Voice_Samples_for_Patients_with_Parkinson_s_Disease_and_Healthy_Controls/23849127 , and was accessed on 4 August 2023. This work was supported in part by the National Science and Technology Council, Taiwan, under Grant 114-2221-E-006-150 and in part by the Higher Education Sprout Project, Ministry of Education at National Cheng Kung University.