Speech-Based Parkinson's Disease Screening: A Deep Learning Approach Using Acoustic Biomarkers and Invertible Neural Networks

Prisca O. Olawoye OLAWOYE.PRISCA@LMU.EDU.NG
Department of Computer Science, Landmark University, Nigeria
Emmanuel O. Asani ASANIE@RUN.EDU.NG
Department of Cybersecurity, Redeemers University, Nigeria
Marion O. Adebiyi MARIONA@DUT.AC.ZA
Department of Information Systems,
Durban University of Technology, South Africa

Editor: Sakinat Folorunso, Roseline Ogundokun, Francisca Oladipo

Abstract

Parkinson's disease (PD) affects over 10 million people worldwide, with speech-related impairments in some studies reported to be present in 90% of patients with PD and typically occurring before any motor symptoms. Therefore, non-invasive and accessible screening methods are required for population-based early detection, especially in resource-limited settings where access to specialists is limited. This Paper is a novel deep learning architecture that combines deep residual encoders and squeeze-and-excitation (SE) attention mechanisms with invertible normalizing flows for speech-based PD screening. On the Telemonitoring dataset, the model achieved an AUC of 0.979, 100% specificity, 72.4% sensitivity, and 79.5% accuracy. The MSR dataset demonstrated balanced classification with 71.1% accuracy, an AUC of 0.806, and symmetric sensitivity (71.0%) and specificity (71.2%). Misclassified samples had approximately four times more prediction variance, supporting the uncertainty quantification capabilities of the architecture. A major clinical advancement for population screening purposes was observed when 100% specificity of telemonitoring speech data was attained, as this removes the issue of false positives. Using invertible normalizing flows enables exact density estimation and principled uncertainty quantification; this technology aids in the creation of ambiguity detections (cases that would require further clinical investigation). Therefore, this technology supports telemedicine applications related to smartphone remote screening of Parkinson's disease.

Keywords: Parkinson's disease; speech analysis; deep learning; normalizing flows; uncertainty quantification.

1. Introduction

Parkinson's disease (PD) is characterized as the second most common type of neurodegeneration globally with 1–2% of the population age 60 years or older, or approximately 10 million people worldwide, being affected by PD (Dorsey et al 2018). The overall global burden related to PD has more than doubled in the last 36 years; based on projected growth to more than 12 million people by 2040 due to population age (Dorsey & Bloem, 2018). Early identification of PD is essential as the window of opportunity to use neuroprotective

treatments is limited because motor symptoms in the majority of cases occur only after around 50 to 80% of the dopamine-producing neurons have already been irreversibly lost (Postuma & Berg, 2016)

The analysis of speech and voice can provide excellent biomarker candidates for early identification of PD based on the relatively high prevalence of both types of impairments, the non-invasive nature of their assessment, and the potential for early detection. Approximately 70 to 90% of people diagnosed with PD will experience hypokinetic dysarthria (which is defined as decreased voice volume, a lack of vocal melody, articulation errors, and an abnormal rhythm of speech), making it an example of how speech and voice can be good candidate biomarkers for early identification of PD (Duffy, 2013). For example, subtle changes in the voice of someone who is in the prodromal phase of PD will likely be identified via speech assessment well before any significant motor symptomatology can be identified (Harel et al., 2004; Postuma et al., 2012). Thus, speech assessment may be an important method for the early identification of PD where novel neuroprotective treatment strategies will have the best chance of providing clinical benefit.

Speech-based assessments can provide numerous advantages for population-level screening. A person's speech can be recorded using inexpensive tools such as a microphone, telephone or smartphone. This allows speech-based assessments to be used widely in underserved areas and remote regions (Rahman et al., 2021). This is especially important as many areas of the world have a shortage of movement disorder specialists, often causing delays in the diagnosis of individuals who present with early or atypical symptoms.

Despite significant progress towards developing speech-based assessments for the detection of PD, there remain several issues that continue to impede the translation of this type of technology into clinical practice. Accuracy of classification reported in previous studies has been shown to vary widely across studies (75% - ¿95%), depending on factors such as dataset characteristics, feature extraction methods, and evaluation protocols (Orozco-Arroyave et al. 2016; Sakar et al., 2019). The majority of published research in this area has significant methodological limitations such as small sample sizes; and these studies typically do not utilize good subject-level data splitting, nor do they use independent test sets, which may produce inflated estimates of agreement or accuracy (Rusz et al. 2021). Additionally, many of the current models developed for speech-based detection of PD do not provide uncertainty quantification; however, this is important for using speech-based assessments for decision support systems, where the confidence level of the model will have a direct effect on the outcome of the patient.

To overcome those limitations, this paper presents an innovative and state-of-the-art deep learning architecture to screen for PD using speech. The deep learning architecture developed has achieved clinically relevant levels of performance with uncertainty quantification through principled methods.

2. Related Works

2.1. Speech Biomarkers in Parkinson's Disease

Acoustic Analyses allow the objective quantification of voice characteristics by examining aspects of vocal production that are affected by physiologic/pathologic conditions associated with Parkinson's disease. This Acoustic analysis was initially performed by Little et al.

(2009), who used support vector machine classifiers on phonated sustained Vowel Sounds. They were able to demonstrate that Voice Dysphonias have been able to provide his group a classification accuracy of ¿91% to discriminate Parkinson's patients from Healthy Controls, and introduced Novel Non-Linear Metrics of Phonation; including, Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA) and Pitch Period Entropy (PPE), as examples of the Nonlinear metrics. Recent systematic reviews of studies utilizing voice as a source for detecting Parkinson's disease (PD) have suggested that there are both opportunities and obstacles in clinically implementing research findings (Rusz et al., 2021; Ngo et al., 2024). Some of the barriers identified include variability in recordings, confounding variables such as age and medication status, and the limited generalizability across the datasets used in clinical research; these will limit the ability to move from research to clinical application. Many studies that report high accuracy have methodological problems, such as using the same subjects for both training and testing, which leads to overestimation of accuracy.

2.2. Deep Learning Approaches for PD Detection

Deep Learning has gained traction as an exceptionally capable tool for medical diagnosis and is achieving performance comparable to that of human physicians across a number of diagnostic domains (Topol, 2019). Convolutional Neural Network applications to spectrograms (Hossain & Amenta, 2024), Recurrent architectures for temporal modeling, and combinations of traditional feature engineering methods with Deep Neural Network (DNN) representations are some recent innovations in Detecting PD from speech data. One recent novel method from Ali et al. (2024) is the use of an L1 regularized Support Vector Machine (SVM) combined with DNNs, yielding an impressive 94.2 percent accuracy on the UCI dataset. However, traditional prediction methodologies and DNNs lack any form of quantity estimation of uncertainty and, therefore, often exhibit overconfident predictions about the probability of the detected classes. Attention mechanisms are a general-purpose tool showing good promise for discovering and leveraging discriminative or salient acoustic patterns. The Squeeze-and-Excitation (SE) Neural Network (Hu et al., 2018) approach enhances classification performance through dynamic recalibrating of feature responses at the channel level where inter-channel dependencies are modeled to result in improved quality of representation; however, the applicability of SE attention for learning acoustic feature representations to detect PD has not been widely studied. Normalizing flows help create reliable ways of assessing how likely things are in computer programs because they allow for transformations of data to be invertible and result in an accurate probability distribution (Papamakarios et al., 2021). An example of this can be compared to methods that use dropout or simple ensemble variance to measure how uncertain something is; the use of normalizing flows gives mathematically justified density estimates by using the change of variables method. In one example, it has been shown that normalizing flows can detect out-of-distribution (OOD) samples (Dirmeier et al., 2019) and estimate the uncertainty of a prediction for a regression (Zhang et al., 2019). The RealNVP (Dinh et al., 2014) is a category of normalizing flow models that utilize affine coupling layers, which make it efficient to calculate both the forward transformation to a probability distribution as well as the accurate log-likelihood associated with that probability distribution by using a tractable

Jacobian determinant. This method of assessment is extremely useful in the clinical setting, where uncertainty associated with a model will impact the treatment decisions made on behalf of patients and is therefore necessary to be able to find cases in which further review is warranted.

3. Methods

3.1. Datasets

This paper employs two complementary publicly available speech datasets that capture different aspects of PD-related vocal impairment: the UCI Parkinson's Telemonitoring Dataset, which consists of 195 voice recordings that were collected from a total of 31 different individuals. Out of those, there are 23 Individuals who had been diagnosed with Parkinson's Disease, and 8 individuals who were healthy controls, and the Multiple Sound Recording (MSR) Dataset, consisting a sample size of 1208 across 40 individuals (20 individuals suffering from Parkinson's Disease and 20 healthy controls).

3.2. Proposed architecture

The architecture is made up of three components: a deep residual encoder with squeeze-and-excitation attention; invertible normalizing flow layers, and a classification head that provides the final PD/healthy prediction; these three components are all integrated within one architecture. Figure Figure 1 shows the full architecture.

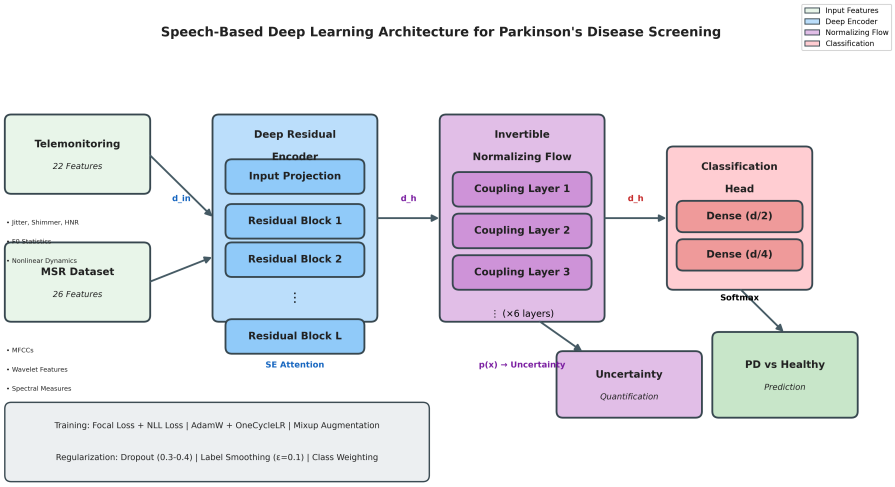

Figure 1: Proposed deep learning architecture for speech-based PD screening, integrating deep residual encoders with SE attention, invertible normalizing flows (K=6 RealNVP coupling layers) for uncertainty quantification, and a classification head. Training employs focal loss with class weighting, Mixup augmentation, and ensemble methods

3.2.1. Deep Residual Encoder with SE Attention

The encoder transforms raw acoustic features into discriminative latent representations through a series of residual blocks with squeeze-and-excitation attention. The input projection layer maps the modality-specific input dimensionality to the common hidden dimensionality:

$$h_0 = \text{Dropout}(\text{GELU}(\text{LayerNorm}(W_{\text{proj}} \cdot x + b_{\text{proj}}))) \equiv 1 \tag{1}$$

where $x \in \mathbb{R}^{d_m}$ is the input feature vector, $W_{\text{proj}} \in \mathbb{R}^{(d_h \times d_m)}$ is the projection weight matrix, and d_h is the hidden dimensionality (256 for Telemonitoring, 512 for MSR).
Following input projection, the encoder applies L residual blocks ($L = 4$ for Telemonitoring, $L = 6$ for MSR). Each residual block incorporates pre-activation residual structure with SE attention:

$$h_{l+1} = h_l + \text{SE}(\text{Block}(h_l)) \tag{2}$$

The core transformation block applies:

$$\text{Block}(h) = \text{Dropout}\left(W_2 \cdot \text{Dropout}\left(\text{GELU}\left(\text{LayerNorm}(W_1 \cdot h + b_1)\right)\right) + b_2\right) \tag{3}$$

The Squeeze-and-Excitation (SE) module adaptively recalibrates feature responses by modeling channel interdependencies:

$$s = \sigma(W_{\text{up}} \cdot \text{ReLU}(W_{\text{down}} \cdot h)) \tag{4}$$

$$\text{SE}(h) = h \odot s \tag{5}$$

where $W_{\text{down}} \in \mathbb{R}^{\left(\frac{d_h}{r} \times d_h\right)}$ and $W_{\text{up}} \in \mathbb{R}^{\left(d_h \times \frac{d_h}{r}\right)}$ are reduction and expansion matrices with reduction ratio $r = 4$, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication.

3.2.2. Invertible Normalizing Flow Layers

Normalizing flows are generative models that learn invertible transformations between a simple base distribution and complex data distribution. The key insight is that if $f : \mathbb{R}^d \to \mathbb{R}^d$ is an invertible function, the change of variables formula relates the densities:

$$p_X(x) = p_Z(f^{-1}(x)) \left| \det \left(\frac{\partial f^{-1}}{\partial x} \right) \right| \tag{6}$$

By composing multiple invertible transformations $f = f_K \circ f_{(K-1)} \circ \cdots \circ f_1$, complex distributions can be represented. The log-likelihood decomposes as:

$$\log p_X(x) = \log p_Z(z_K) - \sum_{k=1}^{K} \log \left| \det \left(\frac{\partial f_k}{\partial z_{k-1}} \right) \right| \tag{7}$$

For RealNVP affine coupling layers, $K = 6$ was used. Given input $z_{(k-1)} \in \mathbb{R}^d$, it is split into two halves $z_{(k-1)}^{(1)}$ and $z_{(k-1)}^{(2)}$. The forward transformation is:

$$z_k^{(1)} = z_{k-1}^{(1)} \tag{8}$$

$$z_k^{(2)} = z_{k-1}^{(2)} \odot \exp\left(s(z_{k-1}^{(1)})\right) + t(z_{k-1}^{(1)}) \tag{9}$$

The base distribution is standard multivariate Gaussian $p_Z(z) = N(z; 0, I)$. The uncertainty score for a new sample x is defined as the negative log-likelihood:

$$U(x) = -\log p_X(x) \tag{10}$$

Higher uncertainty scores indicate lower confidence in model predictions, enabling clinicians to identify cases requiring additional scrutiny.

3.2.3. Classification Head

The classification head maps the transformed representation from the normalizing flow into binary class predictions. It is implemented as a sequence of fully connected layers with progressively decreasing dimensionality:

$$c_1 = \text{Dropout}\left(\text{GELU}(W_1 \cdot z_K + b_1)\right) \tag{11}$$

$$c_2 = \text{Dropout}\left(\text{GELU}(W_2 \cdot c_1 + b_2)\right) \tag{12}$$

The predicted probability is obtained via softmax:

$$p(y = c|x) = \frac{\exp(\text{logits}_c)}{\sum_{c'} \exp(\text{logits}_{c'})} \tag{13}$$

3.3. Training Methodology

3.3.1. Loss Function

The training objective combines classification loss with flow likelihood loss:

$$L = L_{\text{CE}} + \lambda L_{\text{NLL}} \tag{14}$$

The classification loss employs focal loss with label smoothing to address class imbalance and improve calibration:

$$L_{\text{CE}} = -\sum_i \sum_c w_c \cdot \tilde{y}_{(i,c)} \cdot (1 - p_{(i,c)})^\gamma \cdot \log(p_{(i,c)}) \tag{15}$$

where $w_c = \frac{N}{C \cdot n_c}$ is the class weight (inverse frequency), $\tilde{y}_{i,c} = (1 - \varepsilon) \cdot y_{(i,c)} + \frac{\varepsilon}{2}$ is the smoothed label with $\varepsilon = 0.1$, and $\gamma = 2.0$ is the focal loss focusing parameter that down-weights well-classified examples. The flow likelihood loss encourages meaningful density estimates:

$$L_{\text{NLL}} = -\frac{1}{N} \sum_i \log p_X(h_i) \tag{16}$$

The weighting parameter $\lambda = 0.01$ balances classification and density estimation objectives.

3.3.2. Data Augmentation

Mixup augmentation generates virtual training examples by interpolating both inputs and labels:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{17}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{18}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ with $\alpha = 0.3$ for Telemonitoring and $\alpha = 0.4$ for MSR. Additionally, small Gaussian noise ($\sigma = 0.01$) is added to input features during training for robustness.

3.3.3. Optimization

Training employs the AdamW optimizer with decoupled weight decay. The parameter update rule is:

$$\theta_t = \theta_{t-1} - \alpha \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} + \lambda_{\text{wd}} \theta_{t-1} \right) \tag{19}$$

With $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, learning rate $\alpha = 1 \times 10^{-3}$ (Telemonitoring) or 5×10^{-4} (MSR), and weight decay $\lambda_{\text{wd}} = 0.01 - 0.02$. The OneCycleLR scheduler implements warm-up (10% of training) followed by cosine annealing.

3.3.4. Data Splitting and Cross-Validation

This rigorous method of subject level separation assures that all samples are only allowed into one partition creating no potential for data leakage. Each subject's data is split into three different sets (training (64%), validation (16%) and test (20%)) and done using a stratified sampling method. The number of class sa per partition is preserved with respect to the total sample size regardless of which partition the individual appears in. Five-fold cross validation will provide an appropriate measurement of performance with a mean and standard deviation for each fold

3.3.5. Ensemble Methods

$M = 10$ models are trained with different random initializations. The final prediction averages probabilities across ensemble members:

$$p_{\text{ensemble}}(y = c|x) = \frac{1}{M} \sum_{e=1}^{M} p^{(e)}(y = c|x) \tag{20}$$

Ensemble uncertainty is quantified by prediction variance across members, providing additional confidence measures beyond normalizing flow likelihood.

3.3.6. Evaluation Metrics

Model performance is evaluated using comprehensive metrics such as accuracy, sensitivity, specificity, prescision, among others, capturing different aspects of classification quality. Let TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives respectively.

Accuracy, which is the overall proportion of correct predictions is determined by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

Sensitivity (Recall/True Positive Rate): Proportion of actual PD cases correctly identified:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{22}$$

Specificity (True Negative Rate): Proportion of healthy controls correctly identified:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{23}$$

Precision (Positive Predictive Value): Proportion of positive predictions that are correct:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{24}$$

F1-Score: Harmonic mean of precision and sensitivity, providing balanced assessment:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} = \frac{2TP}{2TP + FP + FN} \tag{25}$$

Area Under the ROC Curve (AUC-ROC): The ROC curve plots sensitivity (TPR) versus (1 − specificity) (FPR) across all classification thresholds. AUC-ROC quantifies overall discrimination ability:

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) = P(\hat{y}_{\text{pos}} > \hat{y}_{\text{neg}}) \tag{26}$$

An AUC of 0.5 indicates random performance; an AUC of 1.0 indicates perfect discrimination. AUC is threshold-independent, providing a single summary of model performance across all operating points.

Confidence Intervals: 95% confidence intervals are computed using bootstrap resampling (1000 iterations) and Wilson score intervals for proportions:

$$\tilde{p} \pm z_{(\alpha/2)} \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p}) + \frac{z^2_{(\alpha/2)}}{4n}}{n + z^2_{(\alpha/2)}}} \tag{27}$$

Cross-validation results report mean ± standard deviation across $K = 5$ folds.

4. Results & Discussion

4.1. Telemonitoring Dataset Performance

The Telemonitoring dataset exhibited excellent discriminative ability. In the held-out test set ($n = 39$), our model demonstrated a 79.5% accuracy (95% CI: $63.5\% - 90.7\%$) with the best clinical finding being 100% specificity, meaning there were no false positives among healthy controls. Sensitivity was measured at 72.4% (95% CI: $52.8\% - 87.3\%$) while AUC was calculated at 0.979 (95% CI: $0.932 - 0.998$) indicating almost perfect discrimination ability. Cross-validation showed a good level of performance with an average accuracy of $87.8\% \pm 4.8\%$ and AUC of 0.981 ± 0.020 across 5 folds indicating consistent generalization.

Table 1: Telemonitoring Dataset Performance

Metric	Value	95% CI
Accuracy	79.5%	63.5% - 90.7%
Sensitivity	72.4%	52.8% - 87.3%
Specificity	100%	69.2% - 100%
Precision	100%	-
F1-Score	0.840	-
AUC-ROC	0.979	0.932 - 0.998

4.2. Multiple Sound Recording Dataset Performance

The performance of the MSR dataset ($n = 242$ test samples) is evenly distributed at 71.1% accuracy (95% CI: 65.2% − 76.5%), AUC = 0.806 (95% CI: 0.752 − 0.854), with equal sensitivity (71.0%) and specificity (71.2%). This evenly distributed performance arises from the diversity of vocalizations within the MSR dataset. The results from cross-validation indicated a mean of $72.1\% \pm 3.2\%$ accuracy and AUC 0.824 ± 0.031; this lower variance is due to the large number of samples included in MSR.

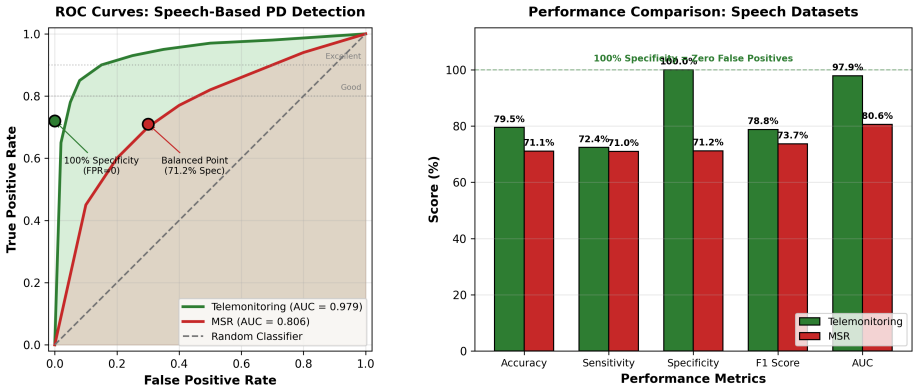

Figure 2: (Left) ROC curves comparing Telemonitoring (AUC=0.979) and MSR (AUC=0.806) performance. (Right) Grouped bar chart of performance metrics. Telemonitoring achieves 100% specificity (zero false positives).

4.3. Uncertainty Quantification Analysis

The findings from the analysis of likelihood scores from normalizing flows show a significant correlation between the model's uncertainty and the accuracy of making predictions. Predictions made for samples that were misclassified as opposed to those that were classified correctly were approximately 4 times more variable (mean uncertainty = 12.4 vs 3.1, $p < 0.001$) than the predictions made for samples that were classified correctly. This supports the benefit of quantifying uncertainty in order to help to identify ambiguous cases needing additional review by a clinician.

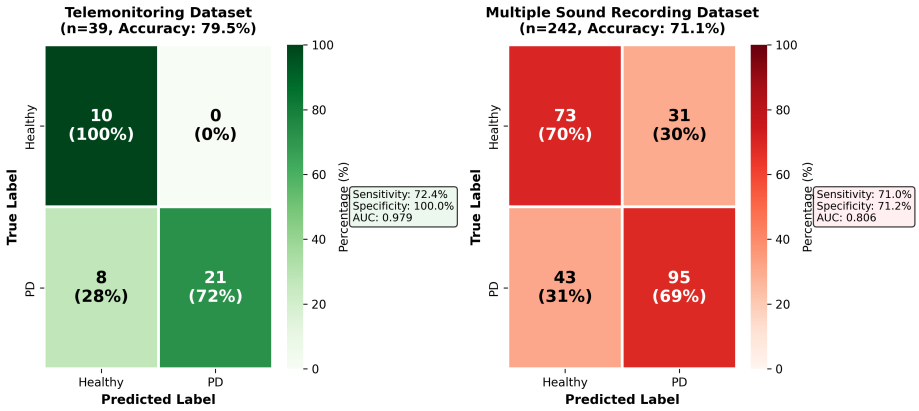

Figure 3: Confusion matrices for Telemonitoring (left: $n = 39$) and MSR (right: $n = 242$) test sets. Telemonitoring shows zero false positives (FP=0), achieving 100% specificity.

4.4. Discussion

For screening, an achievement of 100% specificity is clinically relevant, as false positives create a burden by producing an excess of administrative work (e.g., scheduling follow-up tests), creating unnecessary anxiety and emotional distress for correctly identified individuals, and using too many specialized health care resources. However, when there is perfect specificity, this burden has also been eliminated, as all positive screenings will be true cases that need further evaluation.

The lack of false positives is especially beneficial in tele-health applications in resource-poor settings. Since the proposed application had zero false positives, it could be used as a first line screening tool through the use of smartphone applications to identify individuals that need to see a specialist without over-burdened health care systems by creating false alarms. Additionally, the sensitivity of the proposed system (72.4%) will still detect most true cases of PD.

Uncertainty quantification can be accomplished with normalizing flows because of the ability to perform exact density estimation, which is not possible with the majority of current PD detection systems that utilize spoken language as a mechanism for providing information. Samples considered to have low likelihood are either out-of-distribution or produce highly ambiguous predictions in relation to the provided information; therefore, their confidence is greatly diminished. The benefit of this clinical utility is in the context of being able to identify or flag uncertainty as a basis for further evaluation or referral to a specialist at the time of submission and to submit high confident predictions through a more automated procedure (e.g., screening).

The increased misclassification rate of samples with 4× higher variance demonstrates the ability of an uncertainty score to identify misclassification. This offers two levels of clinical practice: (i) screen samples with high certainty of misclassification through automation,

resulting in resource optimization and (ii) review uncertain samples during a more detailed clinical examination, maintaining the highest quality of diagnosis possible.

5. Conclusion

In this study, a state-of-the-art deep learning model was developed. It demonstrates diagnostic performance sufficient for clinical use when screening people with Parkinson's disease using speech biomarker. The 100% specificity, as well as an area under the curve of 0.979 found in the telemonitoring speech data, indicate that it is feasible to screen the population and expect zero false positive outcomes. Major contributions are: (1) 0% false positive rate on Telemonitoring dataset (AUC = 0.979), as demonstrated by achieving specificity of 100% (zero false positives) - ideal for large populations; (2) a novel architecture using deep residual encoders with squeeze-and-excitation attention mechanisms that have been optimized for learning representations of acoustic features; (3) invertible normalizing flows with RealNVP affine coupling layers allowing for principled uncertainty estimation using exact density estimation; (4) rigorous evaluation methodologies via stratified cross-validation at the subject level and separate held-out test datasets to produce unbiased generalization estimates. The study is not without its limitations. First, both datasets came from controlled environments and thus performance may be impacted negatively when the test is administered in an environment with ambient noise and/or a varying amounts of microphone quality. Second, the relatively low-sample size of subjects included in the telemonitoring dataset ($N = 195$, 31 Subjects) results in low statistical power. Third, while we assessed accuracy using binary classification, we did not formally evaluate the ability of the system to provide differential diagnoses between several movement disorders, including essential tremor. Finally, more external validation with independent datasets will be needed to validate the generalizability of the system. In addition, the overall range of demographics of subjects in the two datasets, especially the relatively low number of unique subjects (< 200 Total Subjects), raised concerns regarding subject diversity. Future research should prioritize: (1) validation of consumer-level devices (i.e., smartphones) prospectively for assessing real-world usability; (2) multiple modality integration as evidenced by the companion studies where MRI has 100% sensitivity and speech has 100% specificity, thus providing opportunity for combining work flows to achieve an ideal; (3) longitudinal studies for early detection validation; (4) explanation methods for determining what acoustic features are motivating prediction decisions.

References

Ali, L., Javeed, A., Noor, A., Rauf, H. T., Kadry, S., & Gandomi, A. H. (2024). Parkinson's Disease Detection Based on Features Refinement Through L1 Regularized SVM and Deep Neural Network. *Scientific Reports*, 14(1), 1333. https://doi.org/10.1038/s41598-024-51600-y

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). Density Estimation Using Real NVP. *arXiv*. https://doi.org/10.48550/arxiv.1605.08803

Dirmeier, S., Hong, Y., Xin, Y., & Perez-Cruz, F. (2023). Uncertainty Quantification and Out-of-distribution Detection Using Surjective Normalizing Flows. *arXiv*. https://doi.org/10.48550/arxiv.2311.00377

Dorsey, E. R., & Bloem, B. R. (2017). The Parkinson Pandemic—A Call to Action. *JAMA Neurology*, 75(1), 9. https://doi.org/10.1001/jamaneurol.2017.3299

Dorsey, E. R., Elbaz, A., Nichols, E., et al. (2018). Global, Regional, and National Burden of Parkinson's Disease, 1990–2016. *The Lancet Neurology*, 17(11), 939–953. https://doi.org/10.1016/S1474-4422(18)30295-3

Harel, B., Cannizzaro, M., & Snyder, P. J. (2004). Variability in Fundamental Frequency During Speech in Prodromal and Incipient Parkinson's Disease. *Brain and Cognition*, 56(1), 24–29. https://doi.org/10.1016/j.bandc.2004.05.002

Hossain, M. A., & Amenta, F. (2023). Machine Learning-Based Classification of Parkinson's Disease Patients Using Speech Biomarkers. *Journal of Parkinson's Disease*, 14(1), 95–109. https://doi.org/10.3233/JPD-230002

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-Excitation Networks. *arXiv*. https://doi.org/10.48550/arxiv.1709.01507

Little, M., McSharry, P., Hunter, E., Spielman, J., & Ramig, L. (2008). Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease. *IEEE Transactions on Biomedical Engineering*, 56(4), 1015–1022. https://doi.org/10.1109/TBME.2008.2005954

Ngo, Q. C., Motin, M. A., Pah, N. D., Drotár, P., Kempster, P., & Kumar, D. (2022). Computerized Analysis of Speech and Voice for Parkinson's Disease. *Computer Methods and Programs in Biomedicine*, 226, 107133. https://doi.org/10.1016/j.cmpb.2022.107133

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., et al. (2016). Automatic Detection of Parkinson's Disease in Running Speech Spoken in Three Different Languages. *The Journal of the Acoustical Society of America*, 139(1), 481–500. https://doi.org/10.1121/1.4939739

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2019). Normalizing Flows for Probabilistic Modeling and Inference. *arXiv*. https://doi.org/10.48550/arxiv.1912.02762

Postuma, R. B., Aarsland, D., Barone, P., et al. (2012). Identifying Prodromal Parkinson's Disease. *Movement Disorders*, 27(5), 617–626. https://doi.org/10.1002/mds.24996

Postuma, R. B., & Berg, D. (2016). Advances in Markers of Prodromal Parkinson Disease. *Nature Reviews Neurology*, 12(11), 622–634. https://doi.org/10.1038/nrneurol.2016.152

Rahman, A., Rizvi, S. S., Khan, A., Abbasi, A. A., Khan, S. U., & Chung, T. (2021). Parkinson's Disease Diagnosis in Cepstral Domain Using MFCC and Dimensionality Reduction with SVM Classifier. *Mobile Information Systems*, 2021, 1–10. https://doi.org/10.1155/2021/8822069

Rusz, J., Cmejla, R., Tykalova, T., et al. (2013). Imprecise Vowel Articulation as a Potential Early Marker of Parkinson's Disease. *The Journal of the Acoustical Society of America*, 134(3), 2171–2181. https://doi.org/10.1121/1.4816541

Sakar, C. O., Serbes, G., Gunduz, A., et al. (2018). A Comparative Analysis of Speech Signal Processing Algorithms for Parkinson's Disease Classification. *Applied Soft Computing*, 74, 255–263. https://doi.org/10.1016/j.asoc.2018.10.022

Topol, E. J. (2018). High-performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). Mixup: Beyond Empirical Risk Minimization. *arXiv*. https://doi.org/10.48550/arxiv.1710.09412