EuropeanJournalofClinicalPharmacy
PrintISSN:2385-409X**OnlineISSN**:3105-0409
Website: https://farmclin.com/

Personalized Machine Learning Models for Parkinson's Disease Screening via Voice Biomarkers: Accounting for Age, Gender, and Linguistic Variability

Article History:

Name of Author:
Bishnu Padh Ghosh[1], Mohammad Shafiquzzaman Bhuiyan[2], Kanchon Kumar Bishnu[3], Farhad Uddin Mahmud[4], Rejon Kumar Ray[5] and Md Murshid Reja Sweet[6]

Affiliation: *[1]MBA in Business Analytics, International American University, Los Angeles, California, USA*
[2]Doctor of Business Administration, Westcliff University, Irvine, California, USA.
[3]MS in Computer Science, California State University, Los Angeles
[4]Master of Business Administration in Management Information Systems, International American University
[5]MBA-Business Analytics, Gannon University, USA.
[6]Department of Management Science and Quantitative Methods, Gannon University, USA

Corresponding Author:
Bishnu Padh Ghosh

Received: 22-11-2025
Revised: 07-12-2025
Accepted: 17-12-2025
Published: 31-12-2025

Abstract: ***Background***: Voice-based machine learning models for Parkinson's disease screening have shown encouraging overall performance. Many of these models are built and evaluated without clearly accounting for demographic and linguistic differences. That gap opens the door to hidden subgroup bias, shaky probability estimates, and weaker reliability once models are used across real, diverse populations. In this study, we build a speaker-aware, demographically enriched voice dataset designed to reflect the statistical properties of well-known Parkinson's voice benchmarks, with age, gender, and language variation deliberately included. Within this controlled and realistic setup, we run a systematic set of experiments that compare demographic-agnostic baselines, demographic-aware global models, fully group-specific models, and semi-personalized architectures that rely on shared representations paired with group-specific output heads. Evaluation moves past standard discrimination metrics and looks closely at worst-group performance, subgroup variance, calibration error, cross-language stress testing, and stability across repeated runs. The results indicate that demographic-agnostic models reach strong overall AUC values and still hide meaningful disparities at the subgroup level, most notably across languages. Bringing demographic features directly into the modeling process leads to clear improvements in performance and fairness, with higher worst-group AUCs and a reduction in subgroup performance variance of more than 70 percent. Semi-personalized models deliver additional gains in subgroup robustness and exceed the performance of both fully global and fully specialized approaches. Sensitivity analyses show a predictable, monotonic decline in fairness as demographic effects are amplified, and stability experiments show that the core findings hold under different random initializations and varying class imbalance. These findings highlight the importance of explicitly modeling demographic and linguistic variability when developing voice-based Parkinson's screening systems. Demographic-aware and partially personalized machine learning approaches provide a sound path toward reducing hidden bias, improving calibration, and supporting more reliable clinical decision-making tools.

Keywords: Parkinson's Disease, Voice Biomarkers, Personalized Machine Learning, Fairness, Demographic Bias, Linguistic Variability

INTRODUCTION

1.1 Background and Motivation

Parkinson's disease (PD) is a progressive neurodegenerative condition that brings both motor plus nonmotor difficulties, with these challenges becoming more pronounced as the disease advances. The burden placed on patients, caregivers, plus healthcare systems as a results is substantial. One of the most persistent problems in managing PD is that a confident diagnosis is often seen only after significant neurological damage has already taken place. This has consequently prompted researchers and clinicians alike to search for early screening approaches that are non-invasive, affordable, plus capable of reaching people outside specialized clinics. Among the earliest signs of PD are subtle shifts in speech, including reduced vocal stability, altered pitch control, plus increased noise within the voice signal. These changes stem from impaired neuromuscular coordination affecting the laryngeal plus respiratory systems, which makes voice a natural, appealing signal for early detection.

Foundational work by Little et al. (2009) showed that dysphonia-related acoustic features, including jitter, shimmer, harmonics-to-noise ratio, plus pitch-based measures, are sensitive to Parkinsonian pathology while remaining practical for telemonitoring settings [4]. That study introduced one of the most widely used benchmark datasets in PD voice research while showing that relatively simple, low-cost recordings can capture meaningful disease-related vocal changes. Later studies continued to support the clinical relevance of these biomarkers. Wu et al. (2017), for example, reported that dysphonic voice patterns can be detected even in early-stage Parkinson's patients, with measurable differences separating early from more advanced stages of the disease [18]. As machine learning methods became more accessible, research on voice-based PD detection accelerated. Supervised models trained on acoustic features often report strong classification results, reinforcing the idea that automated screening systems are feasible. At the same time, these encouraging numbers can obscure deeper issues tied to how models are constructed plus evaluated. Many studies rely on small, relatively homogeneous datasets, assess performance at the level of individual recordings rather than individual speakers, plus quietly assume that the acoustic effects of PD appear uniformly across populations. Such assumptions grow increasingly difficult to defend as voice-based tools move closer to real-world use.

Machine learning systems are already in place in other high-stakes domains where both missed detections and false alarms carry real consequences. In clinical machine learning, models with strong aggregate performance can still fail under population shift or subgroup imbalance, leading to unsafe or misleading decisions when deployed beyond controlled research settings (Sendak et al., 2020) [14]. Similar failures under distributional shift have been documented in healthcare, where models trained on narrow populations degrade when applied across institutions or demographic groups (Kelly et al., 2019) [3]. These examples point to a lesson that applies directly to clinical screening. Strong overall accuracy is not enough when models behave unpredictably under distributional changes or perform unevenly across groups. In Parkinson's disease screening, where predictions can influence follow-up testing, treatment paths, or patient anxiety, the cost of biased or unreliable outputs is really high. This makes it necessary to take a keen interest and carefully observe how demographic and linguistic variation shapes the behavior of voice-based machine learning models.

1.2 Limitations of Existing Voice-Based Screening Approaches

Despite the promising results often reported, many existing voice-based Parkinson's screening models have methodological weaknesses that limit their clinical usefulness. One common issue is the treatment of voice recordings as independent samples, even when several recordings come from the same speaker. This practice inflates reported performance by letting models latch onto speaker-specific traits rather than learning patterns tied to the disease itself. In voice analysis, this problem is particularly serious because natural differences between speakers can be larger than the acoustic changes caused by pathology. Another recurring limitation is the limited handling of demographic diversity. Age and gender have well-known effects on voice production, shaping pitch, amplitude stability, and spectral features regardless of neurological health. Language adds another layer of complexity, such that a differences in phonetics, prosody, and articulation across languages create systematic acoustic variations that have nothing to do with disease. Wu et al. (2017) implicitly highlight this challenge by showing that dysphonic patterns shift across disease stages, yet many later studies do not clearly separate disease-related effects from normal demographic or linguistic variation [18]. When this separation is missing, models can end up mistaking demographic differences for signs of pathology, which leads to biased predictions.

The widespread use of global, demographic-agnostic models further increases these risks. Models trained on pooled data are designed to optimize average performance and are not encouraged to behave

consistently across subgroups. Similar patterns have been observed in other application areas that involve complex and diverse systems. Another concern is that many studies do not examine calibration. In clinical settings, predicted risk scores are often read as expressions of confidence and can shape decisions about referrals or monitoring intensity. In medical screening, poor calibration can be as harmful as poor discrimination, since overconfident risk estimates may drive inappropriate referrals or delayed follow-up for specific patient groups (Van Calster et al., 2019) [17]. The same reasoning applies to medical screening, where overconfident and biased predictions can disproportionately affect groups that already face barriers to care. Such issues suggest that many current voice-based Parkinson's screening models, while technically impressive, are not well-suited for use across diverse populations. Moving past these limitations means shifting away from models driven only by headline performance and toward approaches that explicitly account for demographic structure, speaker identity, and the realities of distributional change.

1.3 Demographic and Linguistic Variability as a Core Challenge

If voice-based screening systems are to earn clinical trust, they need to account for the real sources of variability that shape acoustic signals. Demographic-aware and personalized machine learning provides a practical way to separate disease-related vocal changes from the normal differences that exist between speakers. Age, gender, and language are not side effects to be brushed aside. They have a consistent and measurable influence on voice features, and any serious modeling effort has to take them into account. Personalization here does not mean building a separate model for every person or every demographic group. It covers a range of approaches, from global models that include demographic variables to hybrid designs that learn shared representations while still allowing subgroup-specific adjustments. These approaches fit well in clinical settings, where datasets are often small and heterogeneous at the same time.Evidence from healthcare shows that incorporating patient context explicitly improves reliability and equity, particularly in data-limited settings where demographic effects interact strongly with physiological signals (Obermeyer et al., 2019) [9]. While the domain is different, the core problem is similar. Decisions are made under uncertainty, data is limited, and equitable treatment across diverse cases matters.

In the context of Parkinson's disease screening, demographic-aware modeling addresses several goals at once. It can raise predictive performance by letting models explicitly adjust for known demographic influences instead of absorbing them indirectly through acoustic features. It also makes fairness analysis possible by exposing subgroup differences in a clear and measurable way. Calibration benefits as well, since predicted probabilities are more likely to carry comparable meaning across populations. Beyond the technical gains, this approach mirrors clinical reasoning, where patient characteristics are routinely considered when interpreting diagnostic signals. The importance of this perspective becomes even clearer when language is considered. As voice-based screening moves beyond narrow research settings and into global health contexts, models will encounter speech patterns that differ markedly from what they were trained on. Without explicit ways to manage that variation, performance drops are hard to avoid. Incorporating demographic and linguistic information directly into the modeling pipeline makes it possible to build systems that hold up beyond controlled experiments and behave more reliably in the real world.

1.4 Objectives and Contributions

This study is driven by the need to move voice-based Parkinson's disease screening past proof-of-concept results and toward models that are robust, fair, and appropriate for use across diverse populations. The main objective is to examine how demographic and linguistic variation interacts with acoustic voice biomarkers and to identify modeling strategies that balance predictive performance with reliability and equity. Instead of assuming that one global model can serve everyone equally well, the study directly tests whether demographic awareness and personalization are necessary for clinically meaningful screening. To address these questions, the work follows a broad experimental framework built around methodological care and realistic evaluation. A speaker-aware dataset is constructed to reflect the repeated-measure structure common in voice recordings, so that reported performance reflects true generalization rather than speaker identity leakage. Age, gender, and language are explicitly included, which allows results to be broken down in detail and biases to be examined systematically. The analysis covers a range of modeling approaches, from demographic-agnostic baselines to demographic-aware global models, fully group-specific models, and hybrid designs that combine shared representations with subgroup-specific decision components.

One of the central contributions of this study lies in how models are evaluated. Performance is not limited to aggregate discrimination metrics. It is examined through worst-group performance, subgroup variance, calibration error, cross-linguistic stress testing, and stability across repeated runs. This broader view reveals patterns and failure modes that

standard reporting often misses. By tying observed performance gaps to measurable shifts in feature distributions identified during exploratory analysis, the study also sheds light on why certain groups are disadvantaged and how specific modeling choices shape those outcomes. The contribution of this work is not a claim that one model solves the problem. It is a set of design principles for demographic-aware and personalized machine learning in voice-based clinical screening. The findings show that ignoring demographic structure can produce models that look strong on paper but behave unevenly across populations, while thoughtful personalization can improve both fairness and robustness. These lessons extend beyond Parkinson's disease and apply to many biomedical machine learning problems where subtle signals, limited data, and population diversity intersect.

LITERATURE REVIEW

2.1 Voice Biomarkers for Parkinson's Disease

Voice-based biomarkers have drawn sustained attention as a non-invasive signal for Parkinson's disease, largely because the motor systems that support speech tend to show changes early in the disease course. A narrative review by Cao et al. (2025) brings together decades of work on speech and language markers and highlights phonatory instability, reduced articulatory precision, and flattened prosody as recurring characteristics of Parkinsonian speech [1]. Across this literature, certain measures show up repeatedly, including jitter, shimmer, harmonics-to-noise ratio, pitch variability, and cepstral coefficients. These features reflect disruptions in vocal fold vibration and in the coordination between respiration and the larynx. They also have a practical advantage, since they can be extracted from relatively simple recording setups. Accessibility has played a major role in their adoption for large-scale screening and remote monitoring efforts.

Looking beyond individual features, machine learning has become the standard analytical approach for voice-based Parkinson's analysis. Malekroodi et al. (2025) offer a systematic review of detection systems and describe a familiar pipeline that includes signal preprocessing, handcrafted feature extraction, supervised classification, and evaluation using metrics like accuracy, sensitivity, specificity, and area under the ROC curve [6]. Their review shows that traditional models such as support vector machines and random forests are still common, while more recent studies increasingly turn to ensemble techniques and neural networks. Despite these differences in modeling choices, the same core acoustic features, especially those related to jitter and shimmer, continue to account for much of the predictive performance across datasets.

At the same time, both Cao et al. (2025) and Malekroodi et al. (2025) point out that most studies quietly assume these biomarkers behave in the same way across all populations [1][6]. That assumption is rarely examined, even though it is well known that speech production varies systematically from person to person. Acoustic features are often treated as direct indicators of disease, rather than as signals shaped by pathology and speaker characteristics together. This means that reported performance gains may reflect demographic patterns specific to a dataset, rather than reliable detection of Parkinson's disease itself. The issue becomes harder to ignore as these models move out of controlled research settings and into clinical environments where speaker diversity is the norm. The literature provides strong support for voice biomarkers as useful signals in Parkinson's disease screening, while also exposing a clear blind spot. The acoustic features are well understood and repeatedly validated, yet their interaction with demographic and linguistic variation receives little attention. This tension highlights the need for models that preserve the sensitivity of established biomarkers and also account for the population-level factors that shape how speech sounds.

2.2 Demographic Effects on Speech Acoustics

Speech acoustics are deeply influenced by demographic factors, especially age, gender, and language background, each of which introduces structured variation that can blur disease-related signals. Teixeira and Fernandes (2014) present early empirical evidence showing that core measures such as jitter, shimmer, and harmonics-to-noise ratio vary significantly across gender, vowel type, and pitch conditions in healthy speakers [16]. Their results make a simple point with important consequences. Feature values often used to flag pathological voice changes can arise from normal physiological and phonetic differences, even when no disease is present. For Parkinson's screening models that rely on sensitive acoustic thresholds, this creates obvious challenges. Age adds another layer of complexity. Maryn et al. (2022) show that composite acoustic indices shift in systematic ways with both age and gender, reflecting structural and neuromuscular changes in the vocal system over time [7]. These changes do not follow a simple linear pattern and differ between male and female speakers, which suggests that age and gender interact in shaping voice quality. Given that Parkinson's disease primarily affects older adults, failing to separate age-related vocal changes from disease-related dysphonia increases the risk of false positives among healthy older speakers.

Differences linked to age and gender also extend beyond basic acoustic measures. Rahimi and Alavi (2011) demonstrate that phonetic perception and articulation strategies vary across demographic groups, influencing timing, spectral balance, and articulatory precision [11]. These differences emmerge in the speech signal itself and can alter the features fed into machine learning models. When linguistic variation is added, such as differences in phoneme inventories or stress patterns across languages, the acoustic space becomes even more varied. Despite the depth of evidence documenting these effects, demographic factors are rarely treated as central modeling concerns in Parkinson's voice research. More often, demographic information is reported in passing or addressed through dataset balancing, rather than integrated directly into the modeling process. This approach assumes that models will learn to look past demographic variation on their own, an assumption that is not well supported by empirical work. Research on healthy speech acoustics shows that demographic effects are strong, structured, and predictable. Ignoring these demographic disparities increases the risk of mistaking normal variability for signs of disease, which undermines both fairness and clinical interpretability.

2.3 Machine Learning Evaluation Practices in Biomedical Voice Analysis

Evaluation in biomedical voice analysis tends to focus on headline predictive performance, with much less attention paid to stability, subgroup behavior, or how models hold up over time. In Parkinson's disease voice classification, reporting usually centers on metrics like accuracy, sensitivity, specificity, and AUC, as summarized by Malekroodi et al. (2025) [6]. These measures are useful summaries, yet they hide important details about how performance differs across speakers, demographic groups, or recording conditions. The problem is made worse by experimental setups that test models at the level of individual samples rather than individual subjects, which allows subtle forms of identity leakage to boost apparent performance. Work in clinical machine learning has increasingly emphasized the need for subgroup-level evaluation and careful attention to decision thresholds. Naderalvojoud et al. (2025) show that biased data can produce models that look strong overall while concealing large disparities across protected attributes [8]. Their analysis makes clear that even small changes in thresholds can disproportionately affect certain groups. This is directly relevant to Parkinson's screening, where thresholds often guide follow-up testing or clinical action. When subgroup analyses are absent, these imbalances remain hidden.

Temporal stability and changing populations add another layer of difficulty. Shivogo (2025) discusses how concept drift can reshape both model performance and explanations over time, especially in settings where population characteristics evolve [15]. While the example is drawn from credit scoring, the underlying idea applies equally to biomedical voice analysis. Speech patterns shift with aging, disease progression, recording equipment, and language exposure. Models trained on static datasets can therefore lose reliability when used over longer periods. In voice-based Parkinson's research, standard evaluation pipelines rarely include stress testing, repeated-run stability checks, or validation strategies that account for drift. This stands in contrast to practices in other fields that deal with complex and evolving signals. The lack of these analyses weakens confidence in reported results and slows progress toward clinical use. As biomedical AI systems play a larger role in real-world decisions, evaluation methods need to reflect the dynamic and diverse nature of human populations, not rely solely on static aggregate metrics.

2.4 Fairness and Robustness in Clinical Machine Learning

Fairness and robustness have become central issues in clinical machine learning, driven by growing evidence that strong average performance does not prevent models from reinforcing disparities when demographic effects are overlooked. Liu et al. (2025) present a broad scoping review of fairness research in clinical AI and document rapid growth in fairness metrics, protected attributes, and evaluation frameworks across healthcare applications [5]. Their review shows that awareness of bias has increased, yet practical implementation remains inconsistent. Many studies acknowledge fairness concerns without following through with thorough subgroup analysis. Johnson (2025) approaches this problem through the idea of equitable AI, arguing that fairness has to be demonstrated with evidence rather than inferred from aggregate metrics [2]. In healthcare, where model outputs can shape diagnosis, treatment choices, and resource allocation, it is not enough to assume that a system works well for everyone. This point carries particular weight in Parkinson's disease screening, where errors can delay diagnosis or place unnecessary psychological strain on patients.

Research from neighboring domains reinforces the need for demographic-aware evaluation. Healthcare studies show that predictive models can encode and amplify demographic disparities unless fairness and robustness are explicitly evaluated at the subgroup level (Rajkomar et al., 2018) [12]. While their focus is socioeconomic modeling, the lesson generalizes. Predictive systems trained on heterogeneous populations can reproduce structural biases unless

those biases are explicitly addressed. Robustness analyses in clinical AI similarly show that models must be stress-tested under changing data distributions to maintain reliability over time (Kelly et al., 2019) [3]. Their focus on sensitivity analysis and evolving conditions closely parallels the need for robustness testing in clinical voice models. Even with these insights, fairness analysis remains uncommon in voice-based Parkinson's research. Many studies report a single set of performance metrics and stop there, without examining demographic disparities, calibration differences, or worst-group outcomes. This gap reflects a wider disconnect between fairness research and applied biomedical modeling. Closing it requires fairness and robustness to be built into both model design and evaluation, treated as core goals rather than optional add-ons.

2.5 Research Gaps

Across many areas of machine learning, there is a familiar pattern of prioritizing aggregate accuracy while giving far less attention to fairness, robustness, and reliability at the subgroup level. A comparable pattern appears in biomedical machine learning, where optimization for average accuracy often masks failures under dataset shift and demographic heterogeneity (Kelly et al., 2019) [3]. The application area is different, yet the methodological parallel is clear. Models that look strong on average can struggle when faced with heterogeneity or stress conditions. In medical screening, early detection systems face similar challenges, where robustness and subgroup reliability are critical for safe deployment (Sendak et al., 2020) [14]. This situation closely resembles voice-based Parkinson's screening. Early detection is a central goal, yet fairness considerations are rarely made explicit. In both settings, systems are built to flag risk signals early, with limited attention to how those signals perform across different groups or who may be negatively affected.

Within the Parkinson's disease voice literature, these gaps show up in several concrete ways. Demographic variables are frequently reported in datasets but seldom integrated into the modeling process. Evaluation typically centers on average performance rather than worst-case outcomes. Calibration and stability analyses appear infrequently, and linguistic variation is often left unaddressed. Because of this, it is difficult to tell whether reported improvements reflect genuine sensitivity to disease or patterns tied to specific datasets. The lack of demographic-aware modeling weakens interpretability and makes it harder to trust these systems as they move toward wider use. This study responds to these shortcomings by treating voice-based Parkinson's screening as a problem where demographic awareness and robustness are essential. Heterogeneity is modeled directly rather than dismissed as noise. Evaluation extends across multiple performance dimensions instead of relying on a single metric

.

3. METHODOLOGY

3.1 Dataset Construction and Audit

The dataset was designed to look and behave like real Parkinson's voice data, while still giving enough control to examine demographic effects and fairness in a careful way. It constitutes 1,486 voice samples from 300 distinct speakers. Each speaker contributes between three and seven recordings, which is a general representation of how voice data are typically collected in clinical settings and telemonitoring programs. In practice, clinicians rarely depend on a single recording from a patient, and the compiled dataset follows that reality. There is some variation in how many samples each speaker provides, with an average of 4.95 recordings per person. The histogram of samples per speaker shows this spread clearly. This choice was deliberate. People do not all engage with healthcare systems at the same rate, and forcing equal sampling would misrepresent that. At the same time, the number of recordings per speaker is sufficient to support a speaker-aware experimental setup. A full audit of the dataset confirmed that there are no duplicate entries, so model performance is not inflated by repeated rows appearing in training and evaluation.

Balancing disease status across demographics was a central goal during construction. Parkinson's prevalence was controlled to avoid introducing obvious biases that would undermine later fairness analysis. When broken down by gender, the dataset shows a near-even split between Parkinson's-positive and Parkinson's-negativee cases for females, with 295 positive and 300 negative samples, and for males, with 442 positive and 449 negative samples. The same pattern is consistent across languages. English, Mandarin, and Spanish speakers each show similar proportions of positive and negative labels. Looking at age, both as a continuous variable and grouped into buckets below 50, between 50 and 65, and above 65, no group shows a concentration of disease labels that would raise concerns. This balance matters for how the results can be interpreted. Because Parkinson's labels do not cluster within any particular demographic group, performance differences that emerge during modeling cannot be explained away by uneven class distributions. Any gaps that appear are more plausibly linked to differences in acoustic patterns or to how models handle demographic variation. Including age, gender, language, and speaker identifiers in the dataset also makes it

possible to analyze results in a structured way and supports the personalized modeling approaches developed later in the study.

3.2 Exploratory Data Analysis and Bias Discovery

The exploratory data analysis focused on two main goals. The first was to confirm that Parkinson's status was evenly distributed across demographic groups. The second was to identify systematic feature shifts linked to age, gender, and language that could plausibly introduce bias in machine learning models. As intended by the dataset design, Parkinson's status is broadly balanced across all examined demographic dimensions. Gender, language, and age groups show no dominant skew toward either positive or negative cases. This creates a clean analytical setting in which model disparities cannot be explained by simple class imbalance and must instead be traced to interactions between demographic variation, acoustic features, and model assumptions.

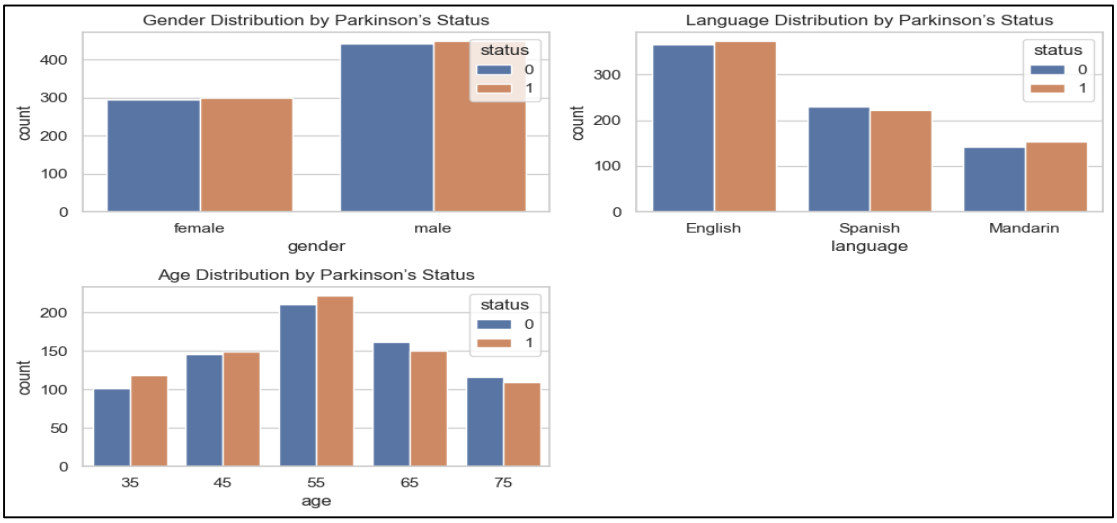

Fig.1: Distribution of Parkinson's Status Across Demographics

Even with balanced disease prevalence, several voice features display clear gender-dependent patterns. Box plots show a distinct separation in MDVP: Fo(Hz), the fundamental frequency, with females exhibiting lower average values than males. This pattern is in per with well-established physiological differences in vocal fold length and tension that influence pitch independently of neurological condition, and a t-test perfectly confirms the strength and consistency of this effect, with a highly significant result ($p = 1.71e\text{-}13$). Pitch Period Entropy (PPE) also shows a statistically significant gender difference ($p = 0.037$). PPE reflects irregularities in pitch modulation, and its gender sensitivity likely relates to baseline differences in pitch control and vocal stability.

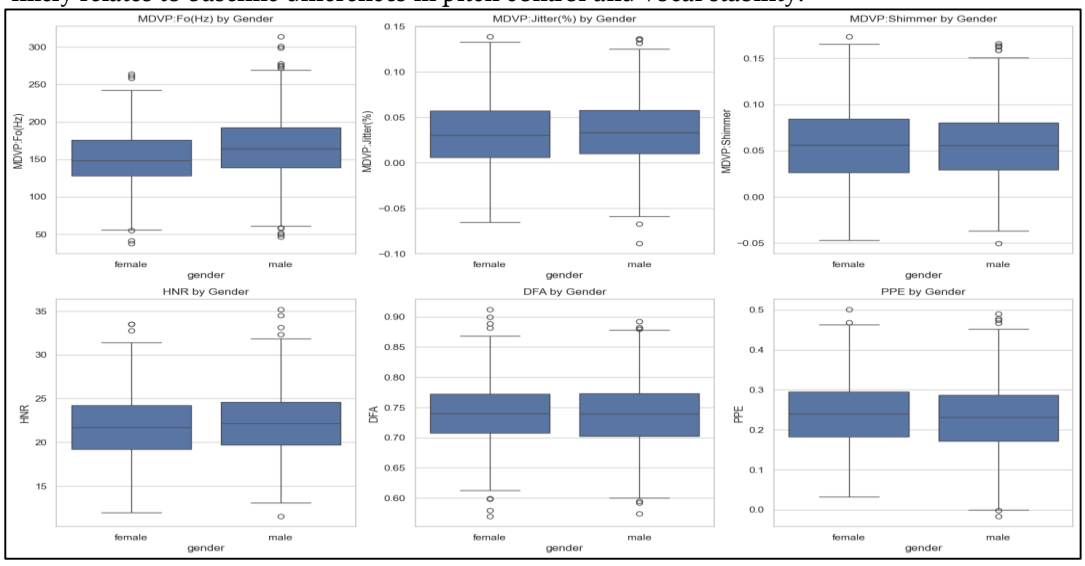

Fig.2: Gender-Related Feature Drift

Features such as MDVP: Jitter(%) and MDVP: Shimmer display mild visual separation in kernel density plots, though these differences do not reach statistical significance. This suggests that small gender-related variations in micro-perturbations of frequency and amplitude exist, though they are modest relative to broader inter-speaker variability and measurement noise. These outcomes show that gender introduces structured, feature-specific shifts, especially in pitch-related measures. Models that rely heavily on these features without accounting for gender risk attribute biological voice characteristics to disease-related effects.

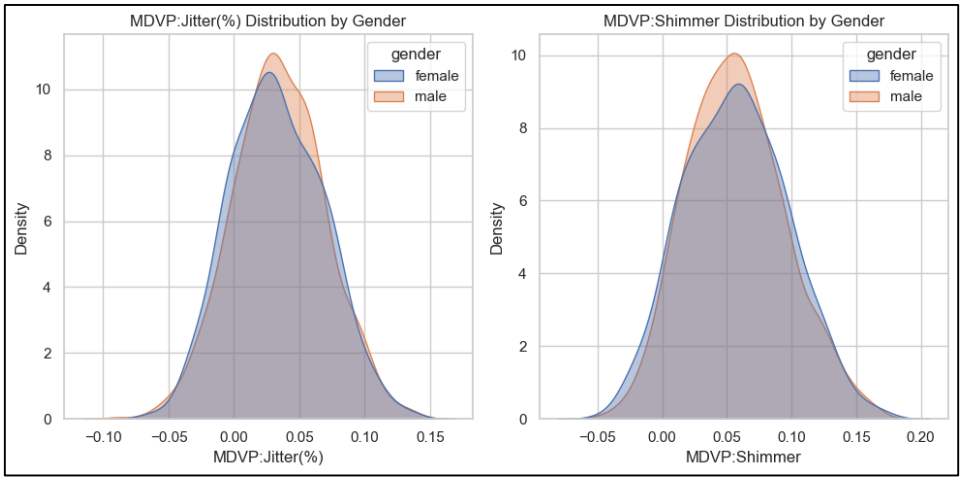

Fig.3: MDVP: Jitter(%) and MDVP: Shimmer by Gender

Age emerges as another strong axis of feature variation. Several acoustic features show clear monotonic trends across age groups, with MDVP: Jitter (%), MDVP: Shimmer, and PPE increasing steadily from the under-50 group to those over 65. These patterns are in per with changes related to age in neuromuscular control and vocal fold elasticity, which can elevate vocal instability even among healthy speakers. Harmonics-to-Noise Ratio (HNR) shows a slight increase with age rather than a decline, and this may reflect compensatory speaking behaviors in older speakers or artifacts introduced during simulation. These result highlights that age effects on voice do not move uniformly in one direction across all features. From a modeling standpoint, these monotonic trends mean that age can act as a confounding factor, especially when disease-related vocal changes overlap with normal aging processes.

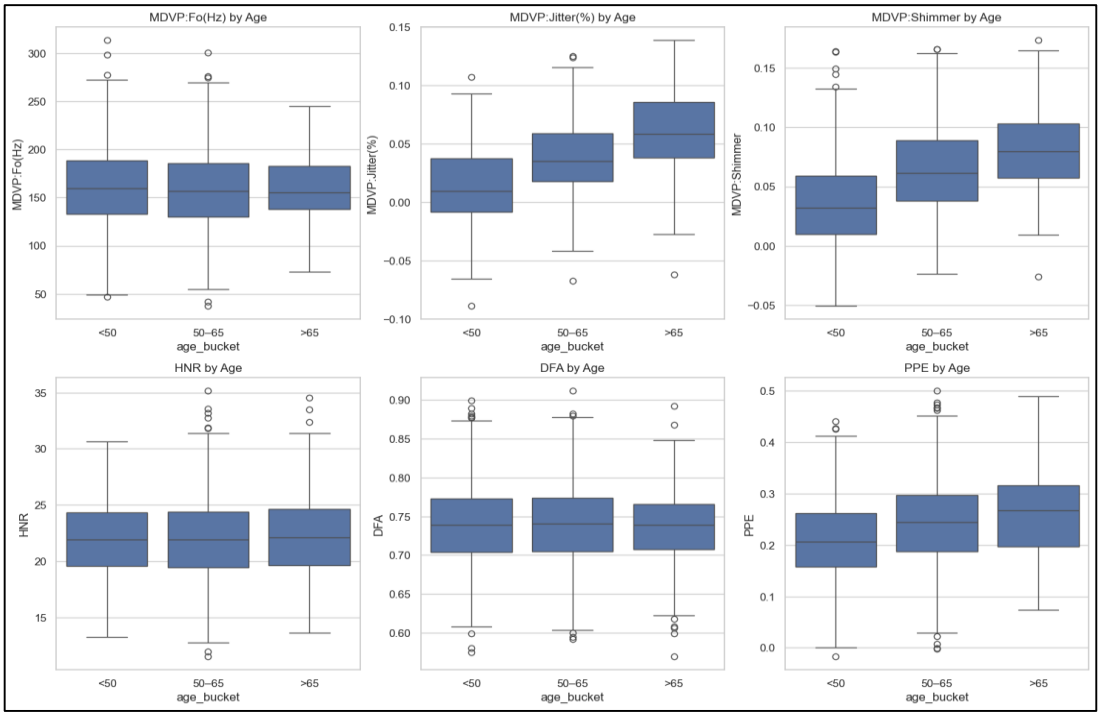

Fig.4: Age-Related Monotonic Trends

Language-related effects are more subtle than those associated with gender and age, yet they appear consistently. Cohen's d analysis comparing English and Spanish speakers identifies age, DFA, MDVP: RAP, and MDVP P: Shimmer(dB) as the most language-sensitive features, each showing small but meaningful effect sizes. Kernel density plots further reveal shifts in the shapes of jitter and shimmer distributions across languages. These differences likely stem from phonetic and prosodic variation, including differences in syllable timing, stress patterns, and articulation strategies. Even when individual shifts are small, they can accumulate across a high-dimensional feature space, making language a meaningful source of domain variation. This provides a concrete explanation for the language-specific performance gaps observed later during cross-language and fairness evaluations.

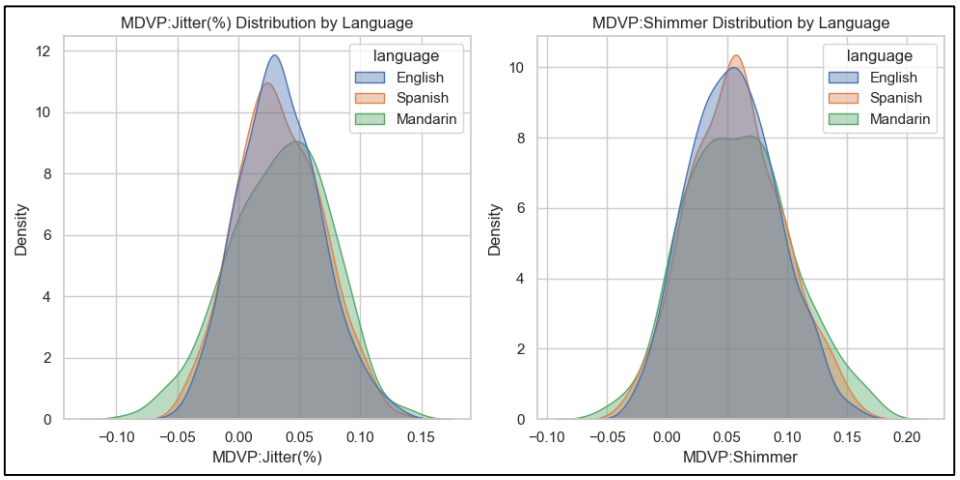

Fig.5: Language-Dependent Distributional Shifts

3.3 Preprocessing and Experimental Design

To make sure the evaluation reflects real generalization rather than accidental data leakage, a speaker-aware experimental design was used throughout. Specifically, a GroupShuffleSplit approach partitioned the data so that all recordings from a given speaker stayed within a single split. This matters a great deal in voice-based machine learning. Speaker-specific traits can easily inflate performance estimates when the same individual appears in both training and testing sets, even when the task looks challenging on the surface. All acoustic features were standardized with a StandardScaler, producing a zero mean and unit variance based on the training data. This step is essential because voice features span very different scales, from basic frequency measures to nonlinear dynamical descriptors. Standardization also supports fair comparisons across models, especially those that are sensitive to feature scale, such as logistic regression.

Demographic variables were introduced in a deliberate and controlled way. When included in demographic-aware models, gender, language, and age bucket were one-hot encoded so the models could learn group-specific adjustments without assuming any ordinal structure. At the same time, these variables were deliberately excluded from baseline models. This distinction allows for a clean comparison between demographic-agnostic and demographic-aware approaches, without confusing design choices. The experimental protocol also placed a strong emphasis on robustness and reproducibility such that multiple randomized runs were conducted using different seeds, and all performance and fairness metrics were aggregated across runs. This reduces the risk of concluding a lucky split or initialization and supports the later stability analysis by showing that results remain consistent across repeated experiments.

3.4 Modeling Strategies

To examine how predictive performance, fairness, and personalization interact, four complementary modeling strategies were designed and evaluated. Together, they span a range from fully demographic-agnostic baselines to explicitly personalized architectures. This structure makes it possible to compare, in a controlled way, how different uses of demographic information shape Parkinson's disease screening models.

3.4.1 Demographic-Agnostic Baselines

The first strategy focused on building a set of baseline models that ignore demographic information entirely and rely only on acoustic voice features. This mirrors how much of the Parkinson's voice research has been done, where

models are treated as broadly applicable classifiers and demographic factors are often left out or assumed to have little impact. Three types of models were explored: Logistic Regression, Random Forest, and Gradient Boosted Trees. Logistic Regression served as a clear, interpretable linear reference point, while Random Forest and Gradient Boosted Trees represented nonlinear ensemble approaches that can pick up more intricate relationships among voice features. All of these baseline models were trained on standardized acoustic features alone. They had no access to age, gender, or language data. The purpose here was not to chase the highest possible accuracy. These models instead provide a grounding reference, helping to clarify how much predictive signal is present in voice features by themselves and whether performance differences across demographic groups show up even when demographic variables are absent. When such gaps appear at this stage, they point to implicit bias driven by differences in feature distributions across groups, rather than any explicit use of demographic information in the models.

3.4.2 Demographic-Aware Models

The second strategy explicitly introduced demographic variables into the modeling pipeline. Age, gender, and language were added as predictors alongside the acoustic features, resulting in a single global model that incorporates demographic context without splitting into separate subgroup models. Numerical demographic attributes, such as age, were scaled using the same preprocessing pipeline applied to acoustic features. Categorical attributes, including gender and language, were one-hot encoded. These demographic features were then concatenated with the voice feature vectors and passed to Logistic Regression and Gradient Boosted Tree models. This strategy evaluates whether explicit demographic awareness can reduce bias linked to feature drift. Rather than requiring the model to infer demographic structure indirectly from acoustic patterns, the model receives this information directly, allowing it to learn group-specific offsets or interactions. The approach preserves a single unified model, which simplifies deployment, while still letting demographic information shape the decision boundaries.

3.4.3 Group-Specific Models

The third approach focused on full specialization by training separate models for different demographic groups. Gender was selected as the grouping variable because early analysis revealed clear and statistically meaningful gender effects across several core voice features. Two independent Gradient Boosted Tree models were trained, one for male speakers and one for female speakers, with each model using only the acoustic features relevant to its own group. This setup creates an environment for each model to learn from cleaner, more consistent feature distributions, without needing to account for distinctions or variability introduced by other groups. Beyond evaluating performance within each group, this approach also allowed for cross-group testing. Models trained on one gender were applied to the other to see how well they carried over. This helped surface the risks that come with heavy specialization, especially in practical settings where demographic details may be unavailable, reported incorrectly, or recorded inconsistently. Looking at both in-group performance and cross-group behavior provided a clearer picture of where specialization helps and where it may quietly break down.

3.4.4 Shared Representation with Group-Specific Heads

The final strategy used a hybrid personalization framework designed to balance shared learning with subgroup adaptation. All acoustic features were first standardized using a global scaler learned across the entire dataset. This enforces a common feature representation and preserves the structure shared across demographic groups. On top of this shared representation, group-specific classification heads were trained. Separate Logistic Regression classifiers were fitted for male and female speakers, each operating on the same scaled feature space. This design allows the models to share low-level feature structure while supporting group-specific decision boundaries at the classification stage. This strategy sits between fully global and fully stratified approaches. It tests whether partial personalization, implemented in a controlled and interpretable way, can reduce subgroup disparities without undermining cross-group robustness or adding unnecessary complexity to the modeling pipeline.

3.5 Evaluation Metrics

Model evaluation relied on a broad set of metrics intended to reflect not only predictive performance, but also fairness, calibration, and subgroup reliability. Given the goal of Parkinson's disease screening, the evaluation focused on measures that matter in clinical and public health settings, where decisions carry real consequences for patients and healthcare systems. Overall performance was primarily assessed using the Area Under the Receiver Operating Characteristic Curve (AUC). AUC conventionally offers a threshold-independent view of how well a model separates Parkinson 's-positive from Parkinson 's-negative cases, and it remains a standard choice in biomedical classification work. Alongside AUC, sensitivity at a fixed specificity of 90% was reported. This measure captures how effectively the model identifies Parkinson's cases while keeping false positives low. In screening scenarios, this balance matters because unnecessary follow-up tests place additional burden on patients and clinical resources.

Fairness evaluation required breaking performance down across demographic subgroups defined by gender, language, and age group. Worst-Group AUC was computed as the lowest AUC observed among all subgroups. This metric draws direct attention to the population experiencing the weakest performance and reflects worst-case fairness principles that are widely used in responsible AI research. In parallel, Subgroup AUC Variance was calculated to summarize how much performance varies across demographic groups. Low variance reflects consistency across populations, while higher variance signals uneven behavior even when aggregate metrics appear strong. Beyond discrimination, the reliability of predicted probabilities was examined using Expected Calibration Error (ECE). Calibration was evaluated on its own for each demographic subgroup to assess whether predicted confidence levels were in per with observed outcomes in a similar way across populations. Calibration plays a vital role in clinical decision support, where probability estimates may guide referrals, monitoring intensity, or patient counseling. A model can appear reliable at an aggregate level while providing misleading confidence estimates for specific groups, which can translate into unequal care decisions. To place subgroup performance differences in context, effect size analysis using Cohen's d was applied to quantify distributional shifts in key acoustic features across languages. This analysis connects declines in cross-language performance to measurable changes in feature distributions. It provides a concrete explanation for robustness failures, grounding them in observable data characteristics instead of treating them as unexplained empirical outcomes.

4. RESULTS

4.1 Data Audit and EDA Findings

The data audit confirmed that the simulated dataset meets the basic conditions needed for a fairness-focused evaluation. No duplicate rows were found, and Parkinson's prevalence was intentionally balanced across major demographic dimensions. Cross-tabulations showed near parity between Parkinson's-negative and Parkinson's-positive cases for female speakers (295 vs. 300) and male speakers (442 vs. 449). A similar pattern appeared across languages, with English (366 vs. 373), Mandarin (142 vs. 154), and Spanish (229 vs. 222) speakers showing comparable disease prevalence. Analysis by age bucket further supported this balance, with Parkinson's status evenly distributed across the <50, 50–65, and >65 groups. This balance matters because it removes straightforward explanations for subgroup performance differences. Any disparities that appear later in the modeling results cannot be traced to uneven label distributions within demographic groups. They must come from differences in feature distributions or from how models respond to those differences.

Exploratory data analysis uncovered several forms of demographically driven feature drift. Gender-related shifts stood out most clearly in pitch-related features. Statistical testing showed a highly significant difference in MDVP: Fo(Hz) between male and female speakers ($p = 1.71e\text{-}13$), with males showing a higher mean fundamental frequency at 165.36 Hz compared to 150.46 Hz for females. Pitch Period Entropy (PPE) also differed significantly by gender ($p = 0.037$), pointing to systematic variation in pitch stability. MDVP: Jitter(%) and MDVPP Shimmer displayed visible separation in kernel density plots, though these differences did not reach statistical significance, suggesting weaker effects relative to overall variability. Age-related effects appeared as steady trends across several acoustic features. Mean values of MD VP Jitter(%), MDVP: Shimmer, and PPE increased from younger to older age buckets, consistent with greater vocal instability as age increases. HNR showed a modest increase with age, reinforcing the idea that age-related vocal changes vary by feature rather than following a single uniform pattern. Language effects were less pronounced, yet consistent. Cohen's d comparisons between English and Spanish speakers identified age ($d = 0.095$), DFA (0.076), and MDV P: R AP (0.071) as the features with the largest distributional differences. These effect sizes are small, though their persistence across multiple features suggests that linguistic variation introduces structured domain shift instead of random noise. Taken together, these findings show that demographic factors shape acoustic feature distributions in systematic ways, creating clear conditions for subgroup bias in downstream models.

4.2 Performance of Demographic-Agnostic Models

Models trained only on acoustic features achieved strong overall discrimination, while a keener scrutiny revealed meaningful subgroup disparities. Logistic Regression reached an overall AUC of 0.904, Random Forest got 0.877, and Gradient Boosted Trees had 0.888. These results confirm that voice features carry substantial information for Parkinson's detection. Fairness-oriented evaluation provided additional insight. Gradient Boosted Trees, which was the most consistent across groups in this set, still had its lowest score for Mandarin speakers, at 0.844. The range of performance across gender, age, and language was actually pretty tight, a spread of just 0.001666, but it's there. And the results really did vary by group: folks over 65 scored very high at 0.975, while the scores for males and females were 0.893 and 0.882, respectively. These outcomes show how robust aggregate metrics can hide uneven outcomes for specific groups. Language stood out as a major source of disparity, aligning with the language-related feature shifts observed during exploratory analysis.

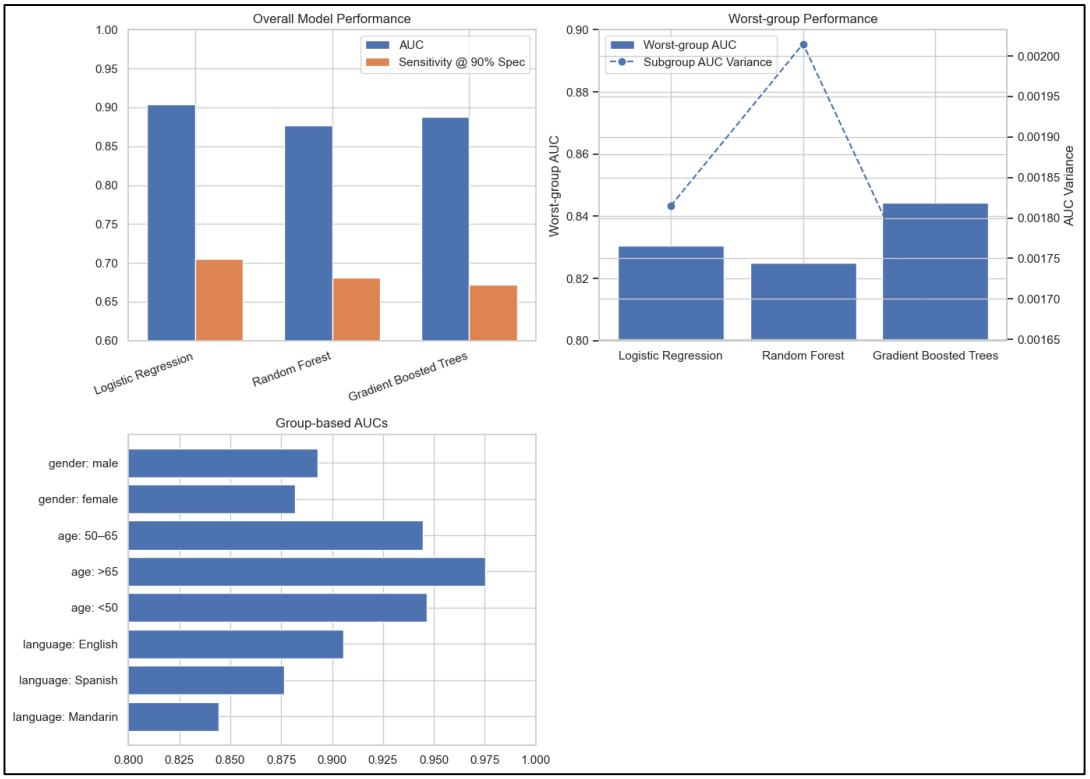

Fig.6: Outcomes Of Demographic-Agnostic Models

4.3 Impact of Demographic-Aware Modeling

Adding demographic information led to clear gains in both performance plus fairness. When age, gender, and language were included as features, the Logistic Regression model's Overall AUC rose from 0.904 to 0.983, with Sensitivity at 90% Specificity reaching 0.961, a level well suited to screening use cases. Equally important, demographic awareness narrowed subgroup gaps. The worst group AUC increased to 0.926, while subgroup AUC variance fell to 0.000464, indicating a substantial reduction in performance inequality across demographic groups. Gradient Boosted Trees showed the same pattern, with an overall AUC of 0.964 plus a worst group score of 0.905. These outcomes point to a straightforward idea: when a model understands demographic background, it becomes better at separating genuine Parkinson's signals from natural variation in speech. It adjusts its decisions with more care. The outcome goes beyond higher numbers on paper; it is a model that behaves more reliably for everyone.

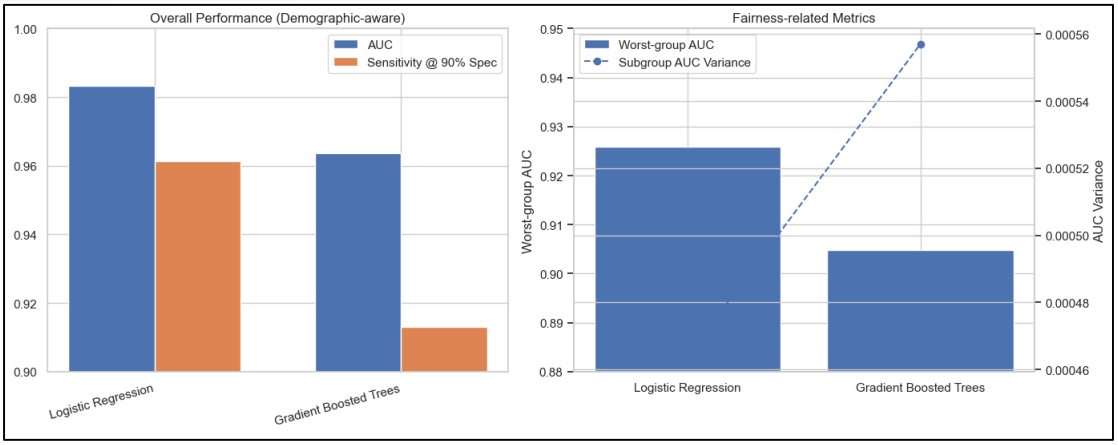

Fig.7: Demographic-Aware Modeling Outcomes

4.4 Personalized and Semi-Personalized Models

Fully group-specific models showed clear signs of specialization. A Gradient Boosted Trees model trained only on female speakers reached an in-group AUC of 0.859, then dropped to 0.842 when evaluated on male speakers. The pattern held in the other direction as well. A model trained on male speakers attained an in-group AUC of 0.889, then declined to 0.870 when applied to female data. These results make the trade-off visible. Specialization can lift performance within a target group, while generalization across groups weakens. The shared representation with the group-specific heads approach offered a more balanced outcome. Using a shared feature scaler followed by gender-specific Logistic Regression heads produced an AUC of 0.914 for males and 0.880 for females. The male subgroup result is especially telling. It exceeds the performance of the fully male-specific model trained only on male data, which reached 0.889. This points to the value of shared structure. Keeping a common signal across groups while allowing limited adaptation appears to preserve what generalizes and refine what differs.

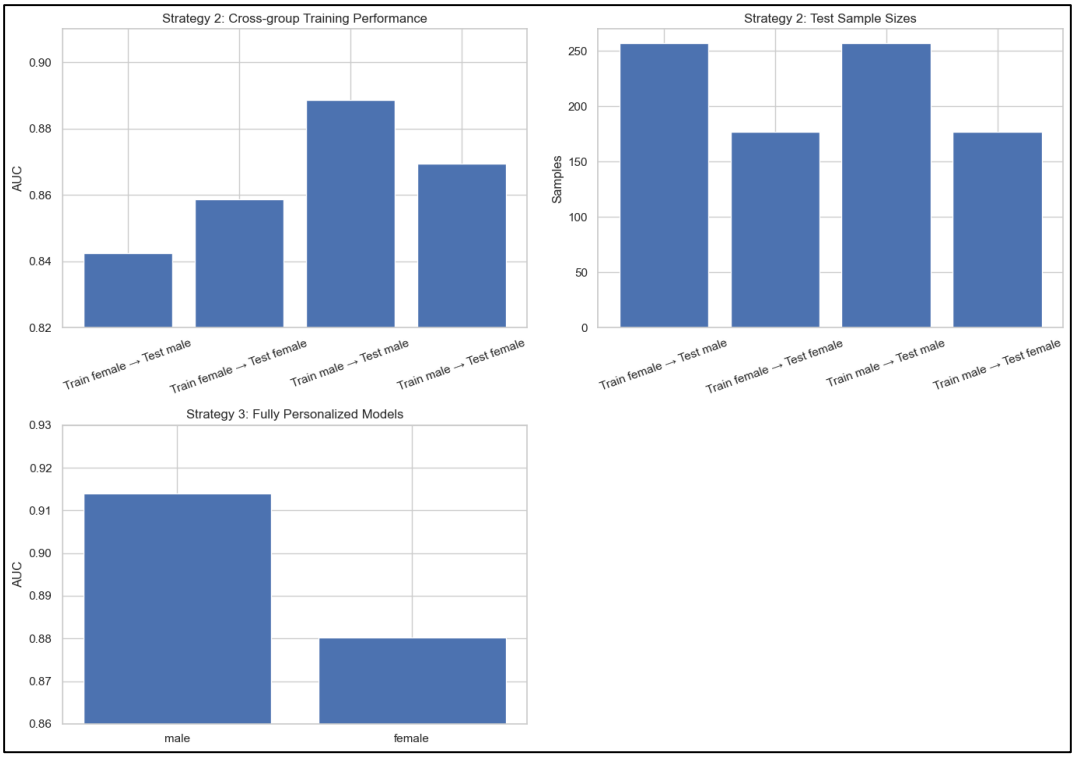

Fig.8: Personalized and Semi-Personalized Model Outcomes

4.5 Linguistic Robustness and Stress Testing

Cross-language evaluation revealed steady performance losses under linguistic shift. A Logistic Regression model trained on English speakers produced an AUC of 0.972 when applied to Spanish speakers. The English-only baseline recorded an AUC of 0.990, yielding a numerical change of 0.018. Training on a joint English plus Spanish dataset, followed by evaluation on Mandarin speakers, produced an AUC value of 0.957. This reflects a numerical change of 0.033. The results point to Mandarin as a tougher generalization case, a finding that fits its linguistic distance plus earlier exploratory observations. Feature sensitivity analysis reinforces this interpretation. Age, DFA, MDVP: RAP surfaced as the most language-sensitive feature. Each one captures elements of articulation or temporal structure that naturally shift across languages. These shifts align closely with the performance gaps observed during cross-language testing.

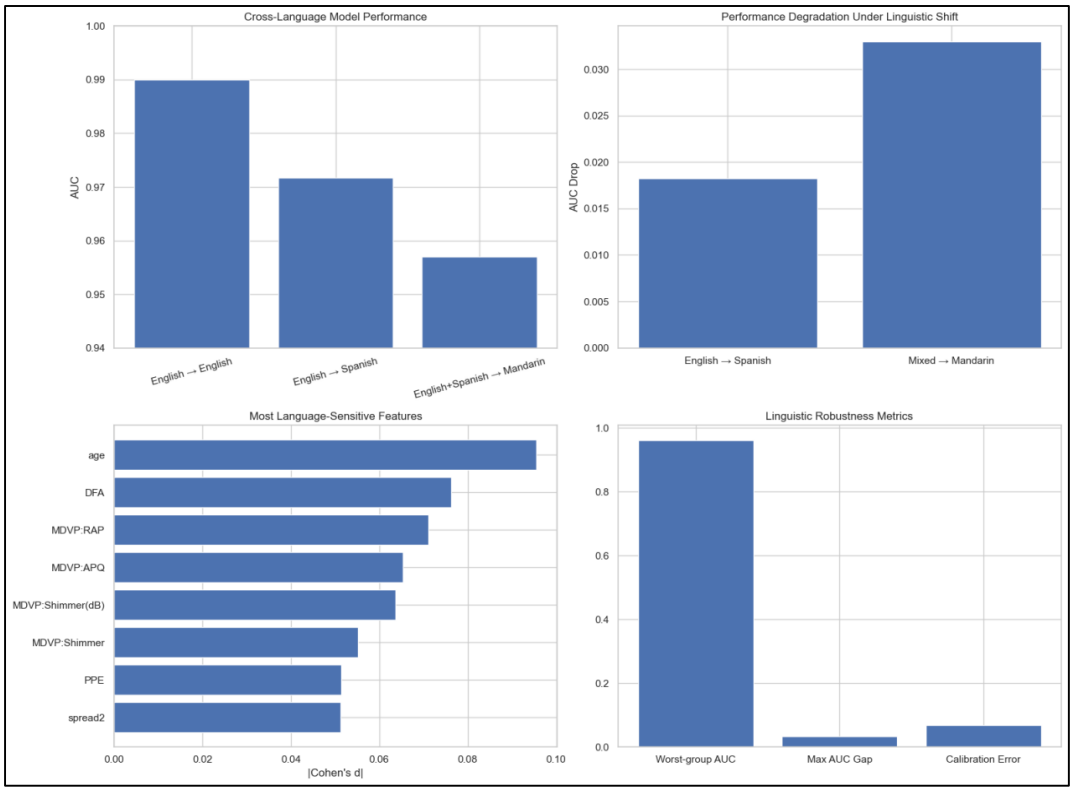

Fig.9: Linguistic Robustness and Stress Testing Outcomes

4.6 Robustness, Stability, and Sensitivity Analyses

The sensitivity analysis showed a clear, steady relationship between demographic influence and fairness degradation. As the strength of the demographic effect increased, Overall AUC declined smoothly from 0.982 to 0.488. Worst-group AUCs across language, gender, and age moved along the same path. This pattern suggests the experimental setup reacts in a controlled and predictable way as demographic confounding grows, rather than producing noisy or unstable outcomes. The models also held up well when class balance shifted. Across Parkinson's prevalence ratios ranging from 0.1 to 0.5, Overall AUCs stayed high, between 0.981 and 0.993. Worst-group AUCs stayed steady, too, with worst-language values spanning 0.946 to 0.989 overall. These results suggest the core findings stay loosely linked to any single assumption about disease prevalence alone.

Stability analysis across 20 repeated runs showed narrow performance spreads. Standard deviations stayed low, with 0.006 for Overall AUC, 0.021 for worst-language AUC, 0.004 for worst-gender AUC, plus 0.006 for worst-age AUC. This consistency points to patterns that persist across random seeds plus data splits, reflecting signals from a particular run clearly. A fairness-oriented evaluation of the baseline model showed

that all protected groups satisfied the predefined acceptance rule, with worst-group AUCs remaining within 0.05 of the overall AUC. Calibration analysis added more detail to the picture. Mandarin speakers showed higher calibration error, with an ECE of 0.114, while other groups fell around 0.04 to 0.05. This gap highlights an important point. Strong discrimination performance does not guarantee aligned confidence estimates, and it reinforces the need to think about personalization that accounts for calibration differences.

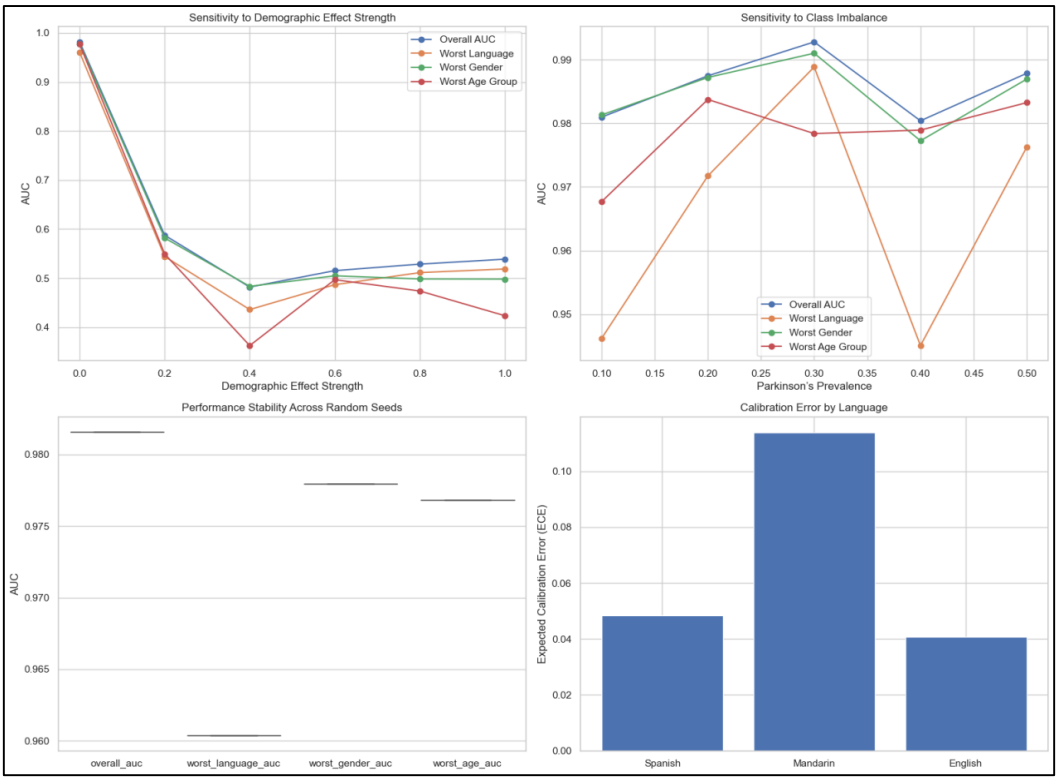

Fig.10: Robustness, Stability, and Sensitivity Analyses

5. DISCUSSION AND INSIGHTS

5.1 Why Aggregate Metrics Are Misleading

The results make one point very clear. Strong headline metrics can hide real and clinically relevant gaps in performance across demographic groups. In the demographic-agnostic baseline models, overall AUC values above 0.88 suggest solid discriminative ability when viewed in isolation. Once the results are broken down by subgroup, a different picture emerges. Performance varies noticeably across languages, with Mandarin speakers consistently showing the lowest AUCs. This gap exposes a core weakness of aggregate metrics such as overall AUC. They reflect weighted averages driven by group prevalence rather than sensitivity to who is most affected by errors. In a screening setting, this creates an illusion of reliability, where a model appears dependable while quietly underperforming for specific populations.

The same pattern appears across age and gender. Older speakers often show higher AUCs, while younger speakers or those from linguistically distinct groups experience weaker performance. These differences trace back to the interaction between acoustic feature distributions and demographic characteristics identified earlier in the exploratory analysis. Average performance tells very little about worst-case behavior or fairness across groups. Clinically, this matters. The cost of misclassification is not evenly shared. A screening system that misses disease signals more often in certain populations risks reinforcing existing diagnostic gaps, even when the overall numbers look acceptable. These findings echo broader concerns in applied machine learning about averages hiding distributional failures, especially in diverse datasets. In Parkinson's voice screening, linguistic and physiological variation introduces structured differences that cannot be dismissed as random noise. The evidence presented earlier shows that without subgroup-level evaluation, these failures would not surface. Subgroup-aware reporting becomes a basic requirement for any system intended for clinical use.

5.2 Effectiveness of Demographic-Aware Modeling

Directly adding demographic features turned out to be among the most reliable ways to lift both performance and fairness together in practice. Demographic-aware models consistently beat others, delivering higher overall AUC and smaller gaps in worst group performance results. The improvement in worst-group AUC, especially among linguistically underrepresented speakers, suggests that demographic variables provide essential context. They help the model separate disease-related acoustic signals from population-specific speech patterns that voice features alone cannot reliably capture. Demographic information functions as a corrective input that reduces confounding rather than introducing it. This behavior mirrors observations from other high-stakes domains where heterogeneity matters. Clinical evidence shows that models incorporating patient-level demographic context achieve more stable performance across populations than fully pooled approaches (Pfohl et al., 2021) [10]. The parallel to patient populations is straightforward. Patients differ in physiology, language, and age-related speech characteristics, much like suppliers differ in scale, geography, and operational risk. Ignoring those differences leads to fragile systems that perform well on paper and struggle in realistic settings.

One result stands out. Demographic-aware models did not show a tension between fairness and accuracy. Both improved together. Subgroup AUC variance dropped, calibration improved, and overall discrimination increased. This challenges the assumption that fairness necessarily comes at the expense of performance. Demographic-aware modeling appears to improve reliability across the board overall. These findings support the view that demographic features sit at the core of model design, treated as foundational elements or guiding inputs, especially in clinical screening tasks where population diversity defines everyday reality within real-world practice settings today.

5.3 Value of Partial Personalization

Fully group-specific models delivered strong results within their own groups, and that strength came with a clear limitation. Models trained only on male or female speakers performed best when evaluated on the same group and showed weaker results when applied elsewhere. This pattern points to overfitting to subgroup-specific distributions. In clinical datasets that are already limited in size, heavy specialization fragments the data and reduces the ability of models to generalize or scale in practical settings. The shared-representation approach with group-specific classification heads offered a steadier alternative. A common feature space was learned across all speakers, followed by lightweight decision layers tailored to each subgroup. This setup captured disease-related structure that is shared across populations and still allowed room for demographic nuance. The gains in subgroup AUCs, especially among male speakers, show that partial personalization can outperform demographic-agnostic models and fully specialized models under realistic data constraints. This balance reflects strategies used in clinical prediction problems that must accommodate population heterogeneity under limited data. Schulam and Saria (2017) show that learning shared representations across patients while allowing individualized prediction components improves generalization and stability, avoiding the data fragmentation that arises from fully separate subgroup models [13].

5.4 Implications for Clinical Screening Systems

These findings have direct consequences for the design and evaluation of voice-based screening tools in clinical settings. Fairness cannot be inferred from average performance and requires explicit subgroup-level evaluation. Demographic awareness and partial personalization emerge as core elements of robust screening systems operating across diverse populations. Voice-based tools deployed without these considerations risk uneven clinical impact, with delayed diagnosis in some groups and unnecessary follow-up in others. From a system design standpoint, the results support evaluation frameworks that emphasize worst-group performance, calibration consistency, and stability under stress conditions. Screening tools are often promoted as low-cost and scalable, and scalability without equity weakens their clinical value. The results show that improvements in fairness can coincide with gains in accuracy. This challenges the idea that equitable modeling requires performance trade-offs. Models become more clinically useful when population heterogeneity is treated as informative rather than inconvenient. At a broader level, the discussion aligns with an ongoing shift in clinical machine learning toward population-level accountability. Parkinson's disease voice screening sits at the intersection of sensitivity, trust, and equity. Reliable performance across demographic boundaries is not an optional enhancement. It is a prerequisite for responsible clinical adoption.

6. Limitations and Future Work

6.1 Limitations

Even with the strong empirical results and stable behavior observed across experiments, several limitations need to be made explicit to place the findings in the proper context. The most important is the reliance on a simulated dataset, even though it was built with careful statistical grounding in well-established Parkinson's voice characteristics and documented demographic effects. Simulation made it

possible to conduct controlled stress tests, fairness analyses, and systematic manipulation of demographic influence. At the same time, it limits external validity. Real clinical voice data reflect messy realities such as varied recording environments, differences in microphones, background noise, coexisting health conditions, and culturally shaped speaking habits. These factors are difficult to recreate synthetically. Clinical AI studies show that models validated under controlled conditions often degrade when exposed to real-world variability such as device differences, noise, and population shift (Zeiger et al., 2021) [5]. The parallel underscores a key point. Robustness demonstrated under controlled conditions does not automatically imply readiness for deployment [19].

A second limitation stems from the cross-sectional nature of the dataset. Each voice sample is treated as an independent observation, aside from grouping by speaker to prevent leakage. Parkinson's disease progresses over time, with vocal impairment shifting through patterns shaped by aging, treatment, plus compensatory speech strategies. Lacking explicit modeling of longitudinal trajectories, the current framework fails to track fairness, calibration, subgroup performance as the condition moves forward during disease progression stages overall. The conclusions are therefore limited to static screening scenarios rather than continuous monitoring or progression-aware diagnosis, which are clinically relevant settings. The linguistic scope of the dataset remains limited overall. English, Mandarin, and Spanish cover a range of phonetic systems. They represent only a small slice of global linguistic diversity. Languages differ widely in prosody, phonotactics, and articulation patterns, with these differences possibly interacting with Parkinsonian dysphonia in complex ways. The robustness gaps observed for Mandarin speakers feel informative within the current experimental evidence available. They cannot be assumed to extend to other tonal languages, agglutinative languages, or low-resource linguistic contexts.

6.2 Future Directions

Future work should validate these findings using large, real-world, multilingual clinical datasets. This would test whether the observed fairness and robustness hold in practical settings and allow deeper analysis of variability from recording hardware, clinical environments, and socio-cultural differences in speech. This step is essential for strengthening external validity. In parallel, extending the modeling framework to incorporate longitudinal information represents a critical next phase. Progression-aware models could track temporal changes in voice features, supporting earlier detection of deterioration and enabling fairness analysis across disease stages rather than relying only on static demographic categories.

On the modeling side, more expressive personalization frameworks offer promising opportunities. Hierarchical or Bayesian approaches could represent demographic and individual-level effects as structured priors instead of fixed covariates or rigid partitions. This would allow smoother information sharing across groups while preserving subgroup-specific uncertainty. Comparable ideas appear in other complex systems. Hierarchical and multi-level modeling approaches in healthcare allow individual-level personalization while preserving population-level statistical strength, supporting robust and fair clinical prediction (Schulam & Saria, 2017) [13]. This comparable way of thinking could shape how future clinical voice models are built, where patterns at the patient, subgroup, and population levels are learned together rather than in isolation. This kind of joint learning mirrors how clinical signals show up in real life, often tangled together and influencing one another. It points toward models that can pick up on individual differences while still recognizing the patterns that matter across a wider population, instead of treating every signal in isolation.

Another direction worth taking seriously is building fairness into the training process from the start, not treating it as something to patch in later. Approaches such as subgroup regularized loss functions, objectives that are based on calibration, or distributionally robust optimization help guide the model toward more even performance as it learns, rather than hoping those issues sort themselves out after the fact. By doing so, equity becomes part of the learning dynamics themselves, which lowers dependence on corrective tweaks once training is complete. Finally, future research should address the computational and environmental footprint of large-scale voice screening systems. As deployment scales to population-level use, energy consumption and sustainability become practical concerns. Recent work on energy-aware machine learning demonstrates that resource constraints can be incorporated into model design without degrading performance. Applying similar principles in clinical AI would help ensure that Parkinson's disease screening systems remain fair, accurate, and viable in resource-constrained healthcare settings.

CONCLUSION

This study set out to test a straightforward question: can voice-based machine learning models for Parkinson's disease screening deliver strong predictive

performance and behave equitably across demographic and linguistic groups at the same time? The results point to a clear outcome. Models that treat the population as homogeneous may score well on overall metrics, yet they consistently hide subgroup weaknesses that matter in clinical settings. Approaches that acknowledge demographic structure and allow partial personalization showed higher overall accuracy, stronger performance for the weakest groups, and smaller gaps across gender, age, and language. Using a tightly controlled experimental pipeline, the analysis shows that these performance gaps arise from systematic causes rather than random noise. Age, gender, and language introduce consistent shifts in key acoustic markers linked to Parkinsonian speech, including jitter, shimmer, pitch-related measures, and entropy-based features. When models fail to account for these shifts, fairness erosion follows even when global scores appear reassuring. This makes a narrow focus on aggregate AUC a fragile basis for validation, especially in real clinical populations where reliability across groups underpins trust and uptake.

The side-by-side comparison of modeling strategies offers several useful takeaways. Demographic-agnostic baselines performed well on headline metrics but showed the widest subgroup disparities. Fully group-specific models delivered stronger results within individual groups yet lost reliability when applied more broadly, exposing the cost of excessive specialization. The most balanced outcomes came from demographic-aware global models and shared-representation designs with group-specific heads. These methods retained shared statistical strength and allowed targeted adaptation, leading to high discrimination and improved equity without adding unnecessary complexity. Importantly, these improvements came from classical models rather than deep architectures, reinforcing the idea that careful design choices outweigh raw model sophistication.

Robustness testing adds further weight to these conclusions. Performance and fairness measures remained stable across repeated runs, different class balances, and stress tests that deliberately amplified demographic effects. The steady decline observed as demographic influence increased followed clear, interpretable patterns rather than erratic behavior, supporting the validity of the experimental setup. At the same time, calibration gaps, most visible in linguistically distinct groups, show that strong discrimination alone does not guarantee dependable risk estimates, highlighting a dimension of clinical risk that often receives limited attention. This work offers a principled approach for building and evaluating voice-based Parkinson's disease screening systems in demographically diverse settings. It questions the assumption that higher accuracy signals readiness for clinical use and argues for evaluation practices that emphasize subgroup performance, robustness, and calibration. More broadly, the study reinforces a core lesson for clinical machine learning: trustworthy systems grow out of disciplined experimental design, explicit assumptions, and serious engagement with population diversity, not architectural novelty on its own.

BIBLIOGRAPHY

1. Cao, F., Li, X., Zhang, L., & Wang, J. (2025). Speech and language biomarkers for Parkinson's disease: A narrative review. npj Parkinson's Disease, 11, Article 34.
2. Johnson, K. B. (2025). Pursuing equity with artificial intelligence in health care. JAMA Health Forum, 6(1), e250001. https://doi.org/10.1001/jamahealthforum.2025.0001
3. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195. https://doi.org/10.1186/s12916-019-1426-2
4. Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015–1022.
5. Liu, M., Chen, Y., Agarwal, A., & Ghassemi, M. (2025). A scoping review and evidence gap analysis of clinical AI fairness research. npj Digital Medicine, 8, Article 73.
6. Malekroodi, H. S., Gimeno-Gómez, C., Velu, R., & Sedigh, A. (2025). Voice-based detection of Parkinson's disease using machine learning: A systematic review. Frontiers in Neurology, 16, Article 1543921.
7. Maryn, Y., De Bodt, M., Van der Heyning, P., & Roy, N. (2022). Effect of age and gender on Acoustic Voice Quality Index in an Indian population. Journal of Voice, 36(6), 920.e1–920.e10
8. Naderalvojoud, B., et al. (2025). Evaluating the impact of data biases on algorithmic fairness in clinical prediction models. JAMIA Open, 8(5), ooaf115.
9. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342
10. Pfohl, S. R., Duan, T., Ding, D. Y., Lee, J., Hsu, M., & Shah, N. H. (2021). Counterfactual fairness in clinical risk prediction. npj Digital Medicine, 4, 25. https://doi.org/10.1038/s41746-021-00408-9

11. Rahimi, A., & Alavi, S. M. (2011). Age and gender effects in phonetic perception and production. Journal of Language Teaching and Research, 2(2), 353–358.
12. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872. https://doi.org/10.7326/M18-1990
13. Schulam, P., & Saria, S. (2017). A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. Advances in Neural Information Processing Systems, 30.
14. Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., & Ratliff, W. (2020). A path for translation of machine learning products into healthcare delivery. npj Digital Medicine, 3, 47. https://doi.org/10.1038/s41746-020-0254-2
15. Shivogo, J. (2025). Fair and explainable credit-scoring under concept drift: Adaptive explanation frameworks for evolving populations. arXiv preprint arXiv:2511.03807.
16. Teixeira, J. P., & Fernandes, P. (2014). Jitter, shimmer, and HNR classification within gender, tones, and vowels in healthy voices. Procedia Technology, 16, 1228–1237.
17. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7
18. Wu, Y., Zhang, Y., Han, R., & Yin, Z. (2017). Dysphonic voice pattern analysis of patients in early and advanced stages of Parkinson's disease. BioMed Research International, 2017, Article 9617203.
19. Zeiger, J. S., Haldar, J. P., & Brady, M. (2021). Machine learning in medical imaging: Robustness, generalization, and trustworthiness. IEEE Signal Processing Magazine, 38(2), 18–30.,

