data

MDPI

Article

Optimizing Parkinson's Disease Prediction: A Comparative Analysis of Data Aggregation Methods Using Multiple Voice Recordings via an Automated Artificial Intelligence Pipeline

Zhengxiao Yang [1,†], Hao Zhou [1,†], Sudesh Srivastav [2], Jeffrey G. Shaffer [2], Kuukua E. Abraham [3], Samuel M. Naandam [4] and Samuel Kakraba [2,5,*]

[1] Biostatistics and Data Science Graduate Program, Celia Scott Weatherhead School of Public Health and Tropical Medicine, Tulane University, 1440 Canal St., New Orleans, LA 70112, USA; zyang17@tulane.edu (Z.Y.); hzhou13@tulane.edu (H.Z.)
[2] Department of Biostatistics and Data Science, Celia Scott Weatherhead School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA 70112, USA; ssrivas@tulane.edu (S.S.); jshaffer@tulane.edu (J.G.S.)
[3] Department of Mathematics and Statistics, Minnesota State University, Mankato, MN 60001, USA; kuukua.abraham@mnsu.edu
[4] Department of Mathematics, University of Cape Coast, Cape Coast 00233, Ghana; snaandam@ucc.edu.gh
[5] Tulane Center for Aging, School of Medicine, Tulane University, 1440 Canal St., New Orleans, LA 70112, USA
* Correspondence: skakraba@tulane.edu; Tel.: +1-504-988-2475
† These authors contributed equally to this work.

Abstract: Patient-level grouped data are prevalent in public health and medical fields, and multiple instance learning (MIL) offers a framework to address the challenges associated with this type of data structure. This study compares four data aggregation methods designed to tackle the grouped structure in classification tasks: post-mean, post-max, post-min, and pre-mean aggregation. We developed a customized AI pipeline that incorporates twelve machine learning algorithms along with the four aggregation methods to detect Parkinson's disease (PD) using multiple voice recordings from individuals available in the UCI Machine Learning Repository, which includes 756 voice recordings from 188 PD patients and 64 healthy individuals. Seven performance metrics—accuracy, precision, sensitivity, specificity, F1 score, AUC, and MCC—were utilized for model evaluation. Various techniques, such as Bag Over-Sampling (BOS), cross-validation, and grid search, were implemented to enhance classification performance. Among the four aggregation methods, post-mean aggregation combined with XGBoost achieved the highest accuracy (0.880), F1 score (0.922), and MCC (0.672). Furthermore, we identified potential trends in selecting aggregation methods that are suitable for imbalanced data, particularly based on their differences in sensitivity and specificity. These findings provide meaningful implications for the further exploration of grouped imbalanced data.

Keywords: Parkinson's disease (PD); machine learning (ML); artificial intelligence (AI); multiple instance learning (MIL); data aggregation; classification; supervised learning; comparative study

Academic Editor: Florentino Fdez-Riverola

Received: 2 December 2024
Revised: 29 December 2024
Accepted: 30 December 2024
Published: 2 January 2025

Citation: Yang, Z.; Zhou, H.; Srivastav, S.; Shaffer, J.G.; Abraham, K.E.; Naandam, S.M.; Kakraba, S. Optimizing Parkinson's Disease Prediction: A Comparative Analysis of Data Aggregation Methods Using Multiple Voice Recordings via an Automated Artificial Intelligence Pipeline. *Data* **2025**, *10*, 4. https://doi.org/10.3390/data10010004

1. Introduction

Healthcare has experienced significant strides in personalized medicine, predictive analytics, medical imaging diagnostics, and optimizing clinical workflows, leading to improved patient outcomes and more efficient healthcare delivery due to the application of artificial intelligence (AI) [1–5]. In the dynamic field of AI, comprehensive research and

accumulated expertise have consistently demonstrated a fundamental principle: there is no universal, one-size-fits-all approach that can optimally solve every computational challenge or problem across diverse scenarios [6–8]. When dealing with diverse structures of data, it is essential to customize methodologies based on the data's intrinsic characteristics to provide reliable outcomes. This work concentrates on a particular form of imbalanced grouped data, where each group encompasses multiple individual samples. This data structure is common in fields such as public health and medicine, where individual patients would have multiple records and diagnoses need to be made based on these records. This work is both relevant and essential, especially considering the swift progress in AI applications in these fields [4]. Additionally, handling unbalanced data remains to be a common difficulty that requires customized solutions to ensure solid and significant analytical results [9].

Multiple instance learning (MIL) is a specialized machine learning approach designed to handle data grouped into "bags" containing multiple instances [6]. It is particularly useful in scenarios where labels are assigned to bags rather than individual instances, such as in medical research where patients (bags) have various test results (instances) [6,7]. MIL's key feature is its ability to work with partially labeled data, making it suitable for situations where traditional supervised learning methods fall short. The primary goal of MIL is to develop techniques that can effectively analyze and make predictions based on these collections of unlabeled instances within labeled bags, enabling more effective analysis in scenarios where instance-level labels are unavailable or impractical to obtain [10,11].

In addressing multiple instance learning challenges, researchers typically employ two main approaches: the bag-based approach and the instance-based approach. The bag-based method focuses on consolidating data at the bag level [12–15]. Essentially, it simplifies the problem by treating each bag as a single, cohesive unit rather than a collection of separate instances. This approach involves representing each bag (a collection of instances) as a single vector. By doing so, it becomes possible to apply conventional supervised learning models to these bag-level representations. One effective technique within this approach is the use of neural network embedding to extract features for each bag [10]. While the bag-based approach offers simplicity and ease of implementation, it comes with a significant drawback: the potential loss of valuable information from individual instances within each bag. This limitation is particularly concerning in medical research and practice. It risks overlooking subtle but potentially vital patterns or indicators that exist at the instance level [10], which could be essential for accurate diagnosis, personalized treatment, or in-depth medical research [10].

The instance-based approach focuses on labeling individual instances within each bag, allowing for the application of supervised learning models at the instance level. Once these models predict labels for each instance, the collective results are used to determine the overall label for the bag. One innovative technique within this approach involves clustering instances into several groups, effectively creating distinct classes. This clustering strategy provides a systematic way to assign labels to all instances, enabling a more granular analysis of the data. By preserving the detailed information of each instance, this approach can potentially capture nuanced patterns that might be lost in bag-level aggregation, though it may also introduce additional complexity in terms of computation and result interpretation [16]. The instance-based approach preserves the unique details of each instance but labeling instances requires a careful design to capture their nature [10]. Nevertheless, both approaches have distinct advantages and disadvantages, and the selection primarily depends on the characteristics of the data and the specific objectives of the project.

To evaluate the effectiveness of bag-based and instance-based approaches in handling multiple instances, we developed two distinct data aggregation strategies tailored to a specific dataset. Our comprehensive AI-driven analysis employs a custom-designed

pipeline optimized for supervised learning, with a focus on classification tasks. This sophisticated workflow comprises three core components: data preprocessing, model fitting, and evaluation, with each component adaptable to either data aggregation strategy. We rigorously tested twelve AI algorithms, assessing their performance across seven diverse evaluation metrics. To ensure the robustness and reliability of our findings, we incorporated several advanced techniques, including data augmentation to expand our dataset, cross-validation for thorough model testing, and grid search for optimal hyperparameter tuning. This methodical approach allows for a thorough comparison of the two strategies, providing valuable insights into their relative strengths and weaknesses in multiple instance learning scenarios.

2. Parkinson Disease and Dataset Description

Parkinson's disease (PD) is a chronic, progressive neurological disorder that primarily affects motor function [17]. It is marked by the progressive deterioration of dopamine-producing neurons in the substantia nigra, a region near the brain's base. The loss of neurons hinders the brain's capacity to regulate bodily movements, leading to a trifecta of hallmark motor symptoms: tremors, muscle rigidity, and bradykinesia [18–22]. In addition to movement abnormalities, PD frequently entails cognitive deterioration and several non-motor symptoms, including depression, sleep disturbances, and anosmia. The disorder profoundly affects patients' quality of life and presents an escalating global health concern, with an estimated 500,000 to 1 million individuals impacted in the United States alone and a consistently increasing global occurrence [23–29]. Numerous research studies have consistently linked the persistent accumulation of cytotoxic intracellular and extracellular protein aggregates to various neurodegenerative disorders (NDs), including PD [30–33]. This established connection has catalyzed extensive exploration into potential pharmacological interventions, with a significant focus on developing and identifying novel non-steroidal anti-inflammatory drugs (NSAIDs) that show promise in inhibiting protein aggregation associated with NDs. These innovative compounds include Aspirin, various quinoline analogs, TDZD analogs such as PNR886 and PNR962, and combretastatin-A4 analog PNR502, among others [30–33]. These compounds are being designed to target not only PD, but also a broader spectrum of NDs. The pursuit of such multi-faceted therapeutic approaches reflects the growing understanding of shared pathological mechanisms across different NDs and the potential for more comprehensive treatment strategies that could address multiple aspects of these complex disorders simultaneously. However, despite these advancements, no intervention has been able to cure PD to date, reinforcing the critical need for continued research in this field. Timely diagnosis and comprehensive care are vital for preserving patients' independence and quality of life as PD advances [34–40].

The dataset used in this research was obtained from the UCI Machine Learning Repository, specifically the Parkinson's Disease Classification dataset (https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+classification, accessed on 1 September 2024) [41]. The objective of this study is to develop AI models capable of predicting Parkinson's disease (PD) using voice recordings as input. We utilized a comprehensive dataset consisting of 756 voice samples collected from 252 individuals, encompassing 188 PD patients (107 males and 81 females, aged 33–87) and 64 healthy controls (23 males and 41 females, aged 41–82). Each participant provided three sustained phonations of the vowel /a/, recorded at a sampling rate of 44.1 kHz. A unique identifier was assigned to link samples from the same individual. The dataset includes binary labels (1 for PD, 0 for healthy) and an extensive set of 754 features for each voice recording. These features, derived from advanced speech signal processing algorithms, comprise Time Frequency Features, Mel Frequency Cepstral Coefficients (MFCCs), Wavelet Transform-based Features,

Vocal Fold Features, and TWQT (Tunable Q-factor Wavelet Transform) features. Despite the acoustic limitations of the data, the vocal features captured by sustained vowel sounds were sufficient to distinguish patients from healthy individuals, enabling the development of sophisticated AI models for disease prediction [42]. Extensive research has revealed that vocal biomarkers, derived from patients' speech patterns and acoustic characteristics, offer a promising avenue for the identification and diagnosis of Parkinson's disease (PD). This method stands out for its reliability and non-invasive nature. Even in the early stages of the disease, when speech abnormalities are often subtle, advanced vocal analysis techniques can detect these changes. This capability not only suggests a powerful tool for early detection but also presents an opportunity for the ongoing monitoring of disease progression. By leveraging sophisticated analysis of speech, healthcare professionals may be able to identify PD earlier and track its development more effectively, potentially leading to improved patient outcomes through timely intervention and personalized treatment strategies [43,44].

PD frequently manifests in speech abnormalities, including reduced volume (hypophonia), slowed speech rate, imprecise articulation, and voice tremor [45–47]. These vocal changes stem from the progressive degeneration of neural pathways governing motor control in the brain. Leveraging advanced artificial intelligence algorithms and speech processing technologies to analyze these vocal biomarkers offers a promising avenue for early PD detection and diagnosis [48]. This non-invasive method may facilitate earlier diagnosis and intervention strategies, thereby enabling more timely and targeted treatment with the potential to halt the progression of Parkinson's disease and enhance the quality of life for individuals with the condition [49]. The application of AI in vocal analysis has shown remarkable potential, with some studies demonstrating the ability of artificial neural networks (ANNs) to detect PD based on voice samples of vowels, with test accuracy rates as high as 86.47% [50]. By utilizing these technologies, clinicians could potentially intervene at the pre-clinical stage of PD, a critical period when neuroprotective therapies might be most effective in preserving neurological function.

In this research, we leveraged the meticulously preprocessed features from the original dataset to construct a bespoke multi-algorithm AI pipeline specifically designed for PD classification. Our approach builds upon this solid foundation, emphasizing the implementation and refinement of various AI algorithms. The primary objective of our study is to enhance the accuracy of PD detection while eliminating the need for additional signal processing steps. By focusing on algorithm application and optimization within our custom pipeline, we aim to streamline the classification process and potentially improve its effectiveness, contributing to the advancement of non-invasive diagnostic tools for PD.

3. Methods (Workflow Description)

This study was implemented using Python 3.10.12, with scikit-learn 1.6.0 [51] serving as the primary tool for data preprocessing and AI tasks.

3.1. Data Preprocessing

The data preprocessing stage involved two key steps: first, splitting the data into training and test sets, and second, conducting data augmentation to address the issue of class imbalance.

3.1.1. Train–Test Split

To assess the generalizability of our AI pipeline, we implemented a strategic 70:30 split of the dataset into training and test sets [52]. This division resulted in a training set of 528 voice recordings from 176 subjects and a test set of 228 recordings from 76 subjects.

Crucially, we employed stratified sampling based on the diagnosis status to maintain a consistent distribution of PD cases (approximately 75%) and healthy controls across both sets. This stratification is essential for ensuring reliable performance evaluation, especially given the imbalanced nature of our dataset. By preventing the overrepresentation of the majority class in either set, this approach enhances the robustness of our model assessment and helps guarantee that our AI pipeline's performance is accurately evaluated across diverse data subsets [53].

3.1.2. Bag Over-Sampling (BOS)

The dataset exhibits a mild class imbalance, with Parkinson's disease (PD) cases outnumbering healthy controls in a ratio of approximately 3:1. Although this level of imbalance may not significantly impact the classification performance of our models, it remains an important consideration, particularly when contemplating real-world applications. For instance, in population-based screening scenarios, the class distribution is likely to be reversed, with healthy individuals substantially outnumbering those with PD. Addressing this imbalance is crucial for ensuring our models can generalize effectively to diverse real-world situations, maintain fairness in predictions, and potentially optimize performance across various population distributions [54].

Data augmentation was implemented on the training set to enable the AI models to place more emphasis on the minority class during the training phase. Specifically, we use Bag Over-Sampling (BOS) [55], which is a generalization of the Synthetic Minority Over-Sampling Technique (SMOTE) [56] in MIL. Since this is an interpolation method in Euclidean space, all features of the voice recordings are first standardized to mitigate the influence of different feature scales.

In the BOS algorithm, a new bag is synthesized based on two existing bags. Since there is no information showing that the three recordings from each subject have any inherent order or weight, we assumed that the recordings are equivalent and should be treated with equal importance for each subject [41]. Considering two existing bags, (i.e., subject) B_i and B_j, an instance (i.e., a voice recording) is randomly sampled from each of them, say x_p^i and x_q^j. Then, an instance in the new bag B_{ij} can be generated as follows:

$$x_{pq}^{ij} = x_p^i + \delta_{pq} \times \left(x_q^j - x_p^i \right), \tag{1}$$

where $\delta_{pq} \in [0,\ 1]$ is a random number. This process assigns equal sampling probabilities to the three instances in a bag, ensuring equivalent treatment. This step is repeated three times and then a new bag B_{ij} containing three instances is synthesized.

To select two bags for synthesis, the K-nearest neighbor method is used. For a randomly selected bag, the K-nearest neighbor ($K = 2$ in our case) is identified, and new bags are then synthesized based on this bag and each of its neighbors, respectively. Since this method requires the distance between bags, we defined it as the distance between their centroids, which are calculated as the mean vector of the three instances in a bag. The over-sampling process is repeated for the minority class until the ratio of the two classes reaches 1:1, thereby completing the data augmentation process.

3.2. Handling Multiple Instances

In this PD dataset, each subject provided three voice recordings, resulting in multiple instances that need to be addressed through MIL [57]. Therefore, we designed two customized data aggregation strategies, one following the instance-based approach and another following the bag-based approach, to handle these multiple instances.

3.2.1. Post-Aggregation Strategy

Our exploratory data analysis (EDA) revealed a high degree of similarity among the three voice recordings from each subject. When clustered, these recordings typically grouped together, suggesting they would likely receive the same label. This observation led us to adopt a straightforward strategy: assigning the subject's overall label to all three of their voice recordings. This strategy eliminates the need for a separate clustering process and provides practical meaning to the labels. Consequently, we labeled all recordings from Parkinson's disease (PD) patients as diseased (1) and those from healthy individuals as non-diseased (0). This allows AI algorithms to directly predict whether a voice recording indicates the presence of disease, simplifying the classification task while maintaining clinical relevance.

Once predictions for the voice recordings are generated, they must be aggregated to produce the final predictions for each subject, a process known as post-aggregation. We proposed three aggregation methods: mean, min, and max. These methods calculate the mean, minimum, and maximum of the predicted probabilities of being diseased from the three voice recordings to represent the overall likelihood of each subject having Parkinson's disease (PD). This aggregation process ensures that we capture a comprehensive assessment of the subject's condition based on their voice recordings. These methods can be represented as follows:

$$P_i^{mean} = \frac{p_{i,1} + p_{i,2} + p_{i,3}}{3}, \tag{2}$$

$$P_i^{min} = \min\{p_{i,1},\ p_{i,2},\ p_{i,3}\}, \tag{3}$$

$$P_i^{max} = \max\{p_{i,1},\ p_{i,2},\ p_{i,3}\}, \tag{4}$$

where $p_{i,j}$ is the predicted probability of being diseased for the jth voice recording of the ith subject, and P_i^{mean}, P_i^{min}, and P_i^{max} are the three kinds of overall probabilities of having PD for the ith subject derived from the three post-aggregation methods. During this process, we avoid assigning any different weights to the three voice recordings from each subject to ensure the equivalent treatments. The overall probability is then compared to a threshold of 0.5 to determine whether the subject is classified as having PD or not.

The selection of specific aggregation methods—mean, min, and max—was based on varying priorities. The post-min aggregation method requires all three voice recordings to be classified as diseased in order to predict that the subject has Parkinson's disease (PD). This method is ideal for situations where it is crucial to avoid misdiagnosing healthy individuals as diseased. Conversely, the post-max aggregation method predicts a subject as having PD if any one of their three voice recordings is classified as diseased, making it suitable for scenarios where it is vital not to overlook individuals with the disease. The post-mean aggregation method, on the other hand, does not favor either class and operates similarly to a voting system, providing a balanced assessment. This strategy allows us to leverage all available data points, enabling the model to learn from the nuances of each recording. However, it is important to note that assuming each voice recording has a definitive label indicating whether it is diseased may not be entirely accurate, which could limit the model's effectiveness and necessitates validation in practical applications.

3.2.2. Pre-Aggregation Strategy

In the pre-aggregation strategy, we employed the bag-based idea by aggregating the features of the three voice recordings from each subject into a single vector to represent the subject. Since the recordings are assumed to be equally important, we simply aggregate the features of the three voice recordings by taking their mean value. In other words, the centroid of the three instances is used to represent the entire bag. This produces a

composite feature vector for each patient, subsequently utilized to train AI models on the subject level. This strategy facilitates the direct modeling of the subjects, but at the cost of omitting certain details from the individual recordings. By simply mean-aggregating the three recordings into one, we may lose some information on variations across recordings that could be useful for diagnosis.

3.3. Artificial Intelligence (AI) Algorithms

In the modeling stage, we applied a range of AI classification algorithms to predict Parkinson's disease (PD). The inclusion of multiple machine learning algorithms in the pipeline serves to assess whether the aggregation methods yield consistent results across different algorithms, thereby evaluating their generalizability. This approach allows us to determine the robustness of our findings and ensures that the predictive performance is not overly dependent on a single algorithm. The algorithms employed include Logistic Regression [58], Decision Tree [59], Random Forest [60], Gradient Boosting [61], XGBoost [62], LightGBM [63], K-nearest neighbor (KNN) [64], Support Vector Machine (SVM) [65], Naive Bayes [66], AdaBoost [67], and Multi-layer Perceptron (MLP) [68].

In addition, we included a stacking classifier, which consists of two base estimators (KNN and SVM) and a final estimator (Logistic Regression). The stacking model leverages the predictive power of base models, with the final model learning to optimally combine the predictions of these base estimators [69,70].

3.3.1. Hyperparameter Tuning

We employed cross-validation and grid search for hyperparameter tuning to minimize the risk of overfitting while enhancing the generalizability and stability of our AI models [71,72]. For each AI algorithm, we delineate a spectrum of hyperparameter values and use 5-fold cross-validation to identify the best hyperparameter combinations, where models are evaluated based on accuracy.

In particular, we implemented group 5-fold cross-validation when applying the post-aggregation strategy, ensuring that all voice recordings from the same subject are kept within the same fold to prevent data leakage [73]. The final model is then trained on the entire training set using the optimal hyperparameters identified in this process.

3.3.2. Model Evaluation

Once the final models have been developed through hyperparameter tuning, we evaluate the performance of each AI model using a range of evaluation metrics. This comprehensive assessment is designed to capture the various aspects of classification performance, providing a well-rounded understanding of how effectively each model predicts outcomes. These metrics include accuracy, precision, sensitivity, F1 score, Area Under the Receiver Operating Characteristic Curve (AUC), and Matthews Correlation Coefficient (MCC) [74,75]. By standard conventions, accuracy measures the overall correctness of the model. Mathematically,

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \tag{5}$$

where TP refers to true positives, TN denotes true negatives, FP represents false positives, and FN refers to false negatives.

Precision measures the proportion of positive identifications that are accurate:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{6}$$

Sensitivity and specificity measure the proportions of actual positives and negatives that are accurately identified, respectively:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{8}$$

The F1 score is the harmonic mean of precision and sensitivity, providing a balance between the two metrics:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \tag{9}$$

The Area Under the Receiver Operating Characteristic Curve (AUC) quantifies a model's capacity to differentiate between classes, with a higher AUC reflecting a superior performance. Finally, the Matthews Correlation Coefficient (MCC) provides a balanced measure, even in the presence of class imbalance:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{10}$$

Each AI model was trained on the training set and evaluated on the test set by these performance metrics. Subsequently, comparisons between models were made to identify the best-performing AI model for PD classification.

3.4. Workflow

To clearly illustrate our customized multi-algorithm AI pipeline for predictive modeling, we have included a detailed flowchart that outlines the methods employed in this study. Supplementary Material for a fully reproducible workflow using the provided Python scripts, including all necessary packages and code snippets, is available in our GitHub repository: https://github.com/Durixas/Parkinson-s-Disease-Prediction-Code-and-Data-Repository (accessed on 2 December 2024). The entire workflow is summarized in Figure 1.

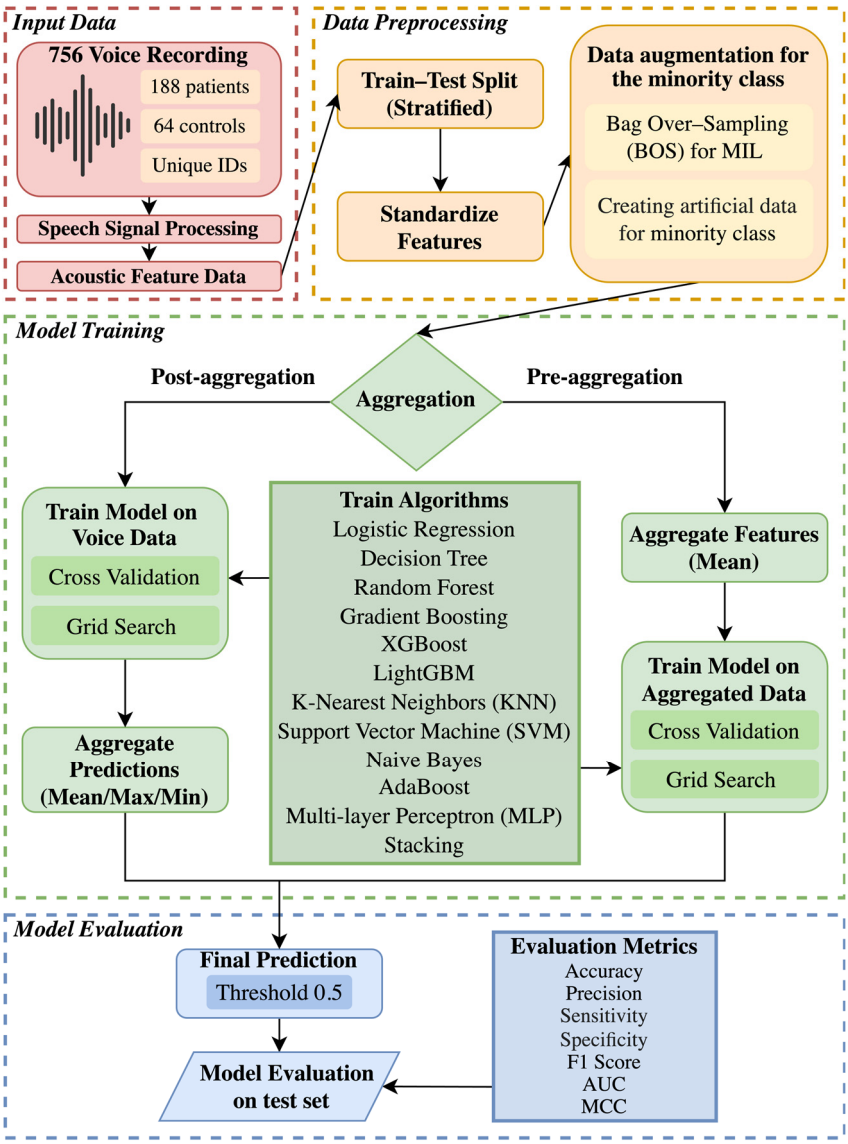

Figure 1. Workflow of the multi-algorithm AI pipeline. The pipeline starts with input data and progresses through preprocessing steps, including stratified train–test splitting, feature standardization, and data augmentation. It then diverges into two primary strategies: post-aggregation, where machine learning models are trained at the voice level with subsequent aggregation of predictions, and pre-aggregation, where features are aggregated at the subject level prior to model training. For both aggregation strategies, twelve (12) machine learning algorithms are utilized. The process concludes with an evaluation phase that assesses model performance on the test set using seven different metrics.

4. Results

We present the evaluation results of the models trained using the four aggregation methods—post-aggregation (mean/min/max) and pre aggregation via our customized multi-algorithm AI pipeline. The models were evaluated on the test set using the seven-performance metrics: accuracy, precision, sensitivity, specificity, F1 score, AUC, and MCC. The data were randomly split into training and test sets 10 times, and the pipeline was run accordingly. The average metrics of different AI algorithms for each aggregation method are summarized in Tables 1–4.

Table 1. Model performance on the test set with post-mean aggregation ranked by accuracy.

Model Name	Acc	Prec	Se	Sp	F1	AUC	MCC
XGBoost	0.880	0.904	0.942	0.695	0.922	0.907	0.672
LightGBM	0.871	0.890	0.946	0.647	0.917	0.909	0.642
MLP	0.866	0.887	0.942	0.637	0.913	0.866	0.625
GBDT	0.859	0.878	0.946	0.600	0.910	0.899	0.606
SVM	0.855	0.861	0.965	0.526	0.909	0.854	0.586
Stacking	0.854	0.857	0.968	0.511	0.909	0.863	0.579
AdaBoost	0.843	0.910	0.881	0.732	0.894	0.898	0.601
Random Forest	0.836	0.868	0.925	0.568	0.894	0.878	0.545
Logistic Regression	0.824	0.888	0.877	0.663	0.881	0.866	0.542
Decision Tree	0.772	0.855	0.840	0.568	0.845	0.752	0.415
KNN	0.749	0.892	0.758	0.721	0.817	0.792	0.436
Naive Bayes	0.733	0.895	0.732	0.737	0.803	0.811	0.420

Table 2. Model performance on the test set with post-min aggregation ranked by accuracy.

Model Name	Acc	Prec	Se	Sp	F1	AUC	MCC
GBDT	0.843	0.905	0.886	0.716	0.894	0.892	0.599
LightGBM	0.842	0.928	0.858	0.795	0.891	0.904	0.616
XGBoost	0.833	0.924	0.849	0.784	0.883	0.900	0.598
Stacking	0.824	0.884	0.882	0.647	0.882	0.858	0.530
Random Forest	0.822	0.888	0.877	0.658	0.880	0.862	0.537
SVM	0.822	0.888	0.875	0.663	0.881	0.845	0.534
MLP	0.803	0.907	0.823	0.742	0.862	0.860	0.528
AdaBoost	0.783	0.926	0.774	0.811	0.842	0.878	0.527
Logistic Regression	0.733	0.921	0.705	0.816	0.796	0.861	0.462
Naive Bayes	0.680	0.911	0.637	0.811	0.746	0.803	0.393
Decision Tree	0.676	0.884	0.656	0.737	0.748	0.702	0.350
KNN	0.667	0.907	0.621	0.805	0.735	0.750	0.373

Table 3. Model performance on the test set with post-max aggregation ranked by accuracy.

Model Name	Acc	Prec	Se	Sp	F1	AUC	MCC
LightGBM	0.858	0.853	0.981	0.489	0.912	0.907	0.591
MLP	0.853	0.847	0.981	0.468	0.909	0.854	0.575
XGBoost	0.851	0.852	0.972	0.489	0.908	0.908	0.570
AdaBoost	0.849	0.864	0.949	0.547	0.904	0.893	0.570
Logistic Regression	0.841	0.850	0.958	0.489	0.900	0.846	0.540
GBDT	0.841	0.837	0.979	0.426	0.902	0.909	0.535
Random Forest	0.830	0.833	0.968	0.416	0.895	0.876	0.503
SVM	0.829	0.820	0.989	0.347	0.897	0.848	0.495
Stacking	0.820	0.810	0.993	0.300	0.892	0.840	0.463
Decision Tree	0.797	0.814	0.947	0.347	0.875	0.655	0.393
Naive Bayes	0.795	0.874	0.853	0.621	0.861	0.773	0.467
KNN	0.789	0.839	0.891	0.484	0.863	0.741	0.414

Figure 2 presents a visual comparison of the seven metrics for the best-performing models—those with the highest accuracy—across the four aggregation methods. The error bars denote the 95% confidence intervals for the means of each metric. The upper bounds of some intervals are truncated at 1, given that 1 is the maximum possible value for these metrics.

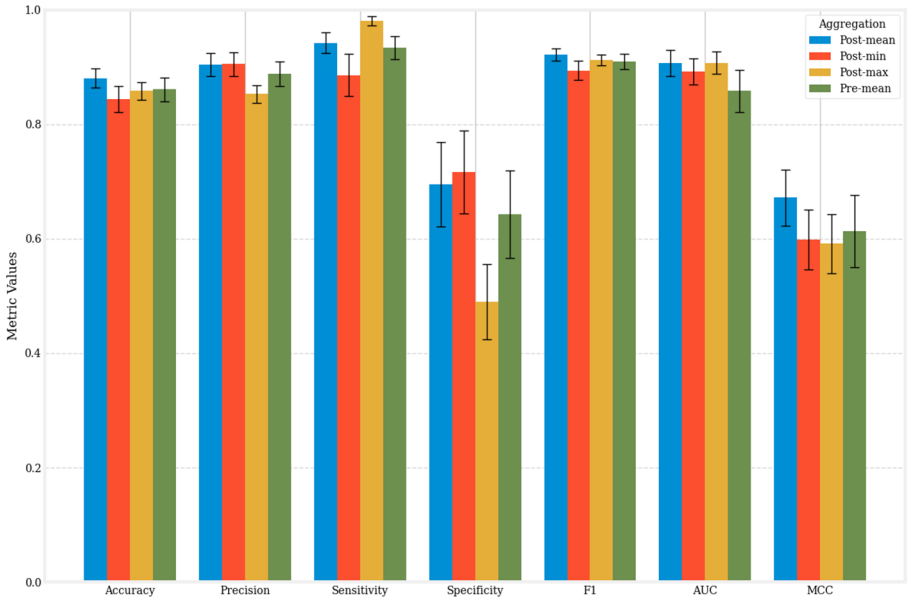

Figure 2. Comparison of best models on test set across different aggregation methods. For each aggregation method, the model with the highest accuracy was identified as the best performer. Overall, the post-mean aggregation method achieved a superior classification performance based on accuracy, F1 score, AUC, and MCC. In contrast, the post-min aggregation exhibited higher precision and specificity but lower sensitivity, while the post-max aggregation showed the opposite trend. The error bars on the graphs represent the 95% confidence intervals of means using the Standard Error of the Mean (SEM).

Figure 3 offers a visual comparison of the averages of seven metrics for the twelve machine learning algorithms across the four aggregation methods.

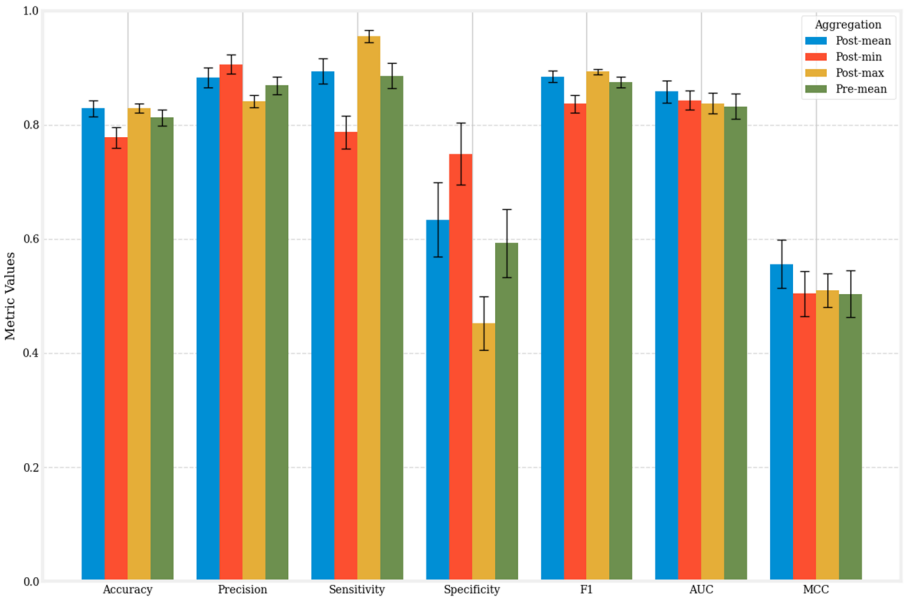

Figure 3. Comparison of average performance on the test set across different aggregation methods. The mean values of seven metrics were calculated for each aggregation method across twelve AI algorithms to represent average performance. In line with Figure 2, the post-mean aggregation method achieved the highest overall classification performance in terms of accuracy, AUC, and MCC. Furthermore, the post-min aggregation demonstrated higher precision and specificity but lower sensitivity, while the post-max aggregation displayed the opposite trend. The error bars on the graphs represent the 95% confidence intervals of means using the Standard Error of the Mean (SEM).

To further investigate the effect of aggregation methods on classification performance, we performed a comprehensive comparison of each AI algorithm across the four aggregation methods within our customized multi-algorithm AI pipeline. Figure 4 displays the performance metrics for each algorithm across these four aggregation strategies.

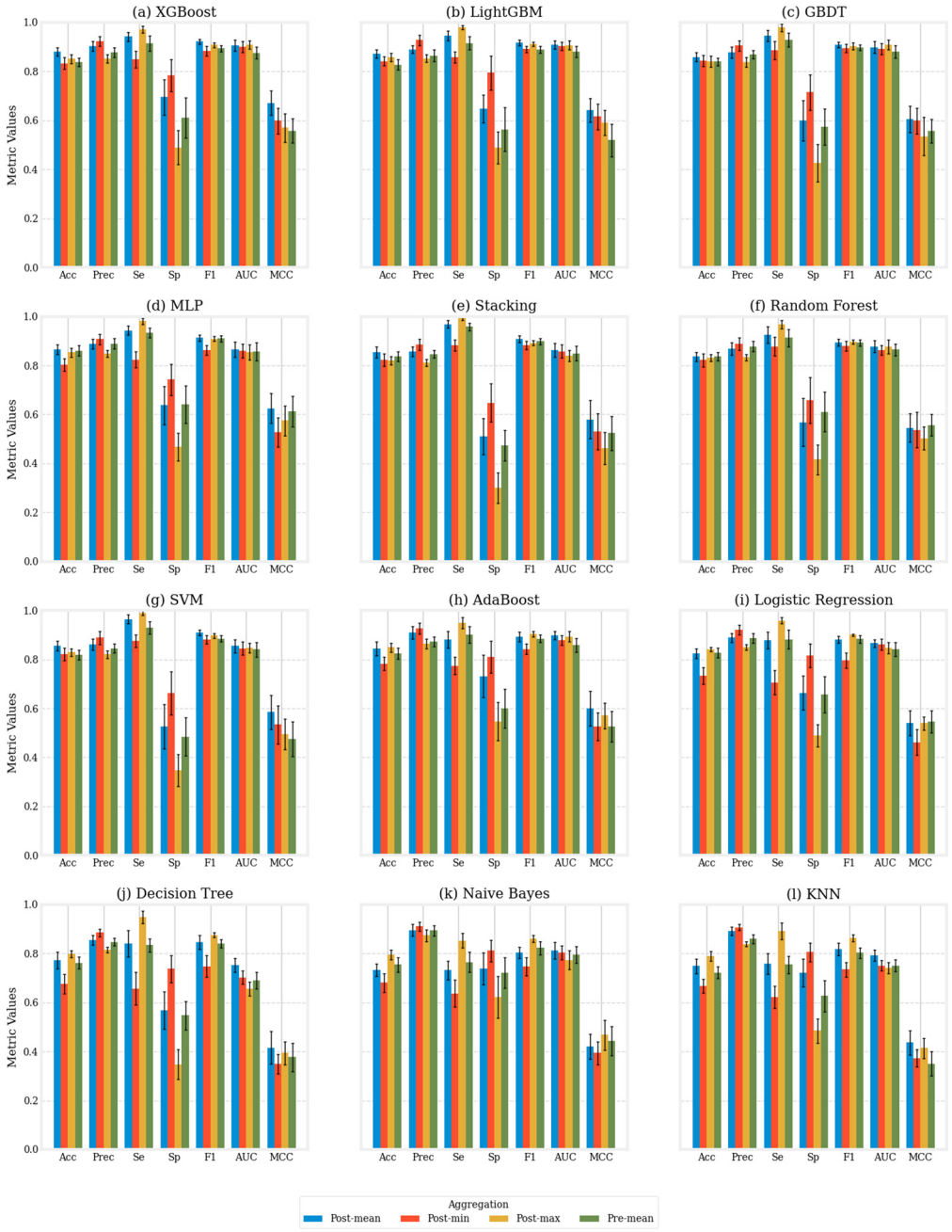

Figure 4. Comparison of AI algorithms' performance on the test set across different aggregation methods. The twelve AI algorithms were ranked according to their average accuracy across the four aggregation methods. The results show that more complex algorithms, such as MLP and boosting models, generally outperformed simpler algorithms like Naive Bayes, Decision Tree, and KNN. Among the higher-ranked algorithms, post-mean aggregation consistently proved to be the most effective method, while post-max aggregation was particularly noteworthy for the lower-performing algorithms. The error bars on the graphs represent the 95% confidence intervals of means using the Standard Error of the Mean (SEM).

Table 4. Model performance on the test set with pre-mean aggregation ranked by test accuracy.

Model Name	Acc	Prec	Se	Sp	F1	AUC	MCC
MLP	0.861	0.888	0.933	0.642	0.909	0.858	0.613
GBDT	0.841	0.869	0.930	0.574	0.897	0.880	0.556
XGBoost	0.838	0.878	0.914	0.611	0.894	0.876	0.558
Random Forest	0.837	0.878	0.912	0.611	0.893	0.864	0.557
Stacking	0.837	0.846	0.958	0.474	0.898	0.849	0.524
LightGBM	0.828	0.865	0.916	0.563	0.888	0.880	0.519
Logistic Regression	0.826	0.888	0.882	0.658	0.883	0.841	0.546
AdaBoost	0.825	0.873	0.900	0.600	0.885	0.858	0.526
SVM	0.818	0.845	0.930	0.484	0.885	0.840	0.476
Decision Tree	0.762	0.847	0.833	0.547	0.840	0.691	0.376
Naive Bayes	0.754	0.893	0.765	0.721	0.822	0.795	0.444
KNN	0.722	0.860	0.754	0.626	0.802	0.750	0.350

5. Discussion

5.1. Findings

Previous research that employed similar AI algorithms reported accuracies between 0.66 and 0.90 for PD diagnosis using different datasets [9,76–88]. Notably, in the post-mean aggregation method, the XGBoost algorithm achieved peak performance with an accuracy of 0.880, precision of 0.904, sensitivity of 0.942, specificity of 0.695, F1 score of 0.922, AUC of 0.907, and MCC of 0.672.

Our results demonstrate that the post-mean aggregation method is the overall best aggregation method, as evidenced by the mean accuracy, F1 score, AUC, and MCC presented in Figure 3. The other three methods display varying performances across these metrics without demonstrating significant superiority. Figure 2 provides an additional comparison of the top-performing machine learning algorithms across different aggregation methods.

Figure 4 ranks the machine learning models based on their mean accuracy across the four aggregation methods, revealing that higher-performing models tend to be more complex, such as MLP and boosting algorithms, with post-mean aggregation achieving the best results. Conversely, for the lower-performing models, post-max aggregation tends to demonstrate a superior performance.

The evaluation metrics indicate that post-min aggregation generally results in higher specificity, but lower sensitivity, compared to post-mean aggregation. For instance, the top-performing GBDT algorithm achieved an accuracy of 0.843, with a sensitivity of 0.886 and a specificity of 0.716. This distinction is evident when the metrics are analyzed in detail. To explain this phenomenon, it is important to examine how the aggregation method affects the balance between true positives (*TP*), true negatives (*TN*), false negatives (*FN*), and false positives (*FP*). The post-min aggregation method employs a conservative threshold by focusing on the lowest prediction scores. This method emphasizes the accurate classification of negative samples (i.e., *TN*) while minimizing false positives (*FP*), resulting in higher specificity. However, this method also has the drawback of increasing false negatives (*FN*) and decreasing true positives (*TP*), which ultimately leads to reduced sensitivity.

In contrast, post-max aggregation prioritizes the maximum prediction scores, leading to the opposite effect: it emphasizes the identification of positive cases, thereby enhancing sensitivity (0.981 for LightGBM). However, this method also decreases specificity (0.489 for LightGBM) because it is more prone to misclassifying negative samples as positives, resulting in an increase in false positives (*FP*). This may explain why post-max aggregation is more effective for lower-performing models. By applying straightforward formula-based reasoning, we can infer that in an imbalanced dataset with more positive cases than

negative ones, sensitivity has a greater influence on accuracy than specificity. For models like MLP or boosting algorithms, the sensitivity achieved through post-mean aggregation is already close to 1, and post-max aggregation does not result in a significant increase in sensitivity. However, for models that underperformed, the enhancement in sensitivity provided by post-max aggregation becomes crucial for improving overall accuracy.

These differences underscore the importance of selecting between post-min and post-max aggregation based on specific requirements. In applications where it is essential to prevent misclassifying negatives as positives—such as situations where false positives could lead to serious consequences (e.g., unnecessary medical interventions) —post-min aggregation would be the preferable choice. Conversely, when the goal is to maximize the identification of positive cases, such as in screening for rare but critical conditions, post-max aggregation may be the more suitable option [89,90].

Another important consideration is the inherent class imbalance in the dataset, with positive samples outnumbering negative ones at a ratio of 3:1. This type of imbalance leads machine learning models to prioritize positive samples, resulting in higher sensitivity but lower specificity [89]. Although the data augmentation performed by BOS mitigates the imbalance to some extent, this trend persists. In this context, even post-min aggregation shows high sensitivity and low specificity, though it maintains the highest specificity among the four aggregation methods. Post-max aggregation exacerbates this trend, increasing mean sensitivity by only 0.168 while decreasing mean specificity by 0.296 compared to post-min aggregation. This illustrates the potential limitations of post-max aggregation, particularly in scenarios where specificity is critical [91].

The pre-aggregation strategy, in contrast to post-aggregation, exhibits no advantages in any metric. One possible explanation is that simply taking the mean of features before model training results in substantial information loss. Despite our study utilizing a relatively large volume of data for PD voice analysis [51,92,93], any loss of information is potentially detrimental [51,92,93].

Several rival methods for Parkinson's disease (PD) prediction using voice analysis can be compared to the proposed AI pipeline. Clustering-based approaches, such as the hierarchical clustering applied by Tsanas and Arora, have shown promise in identifying PD subtypes based on voice characteristics [94]. Deep learning methods, particularly Convolutional Neural Networks (CNNs), have been explored with success, as demonstrated by Shen et al.'s hybrid model combining CNN, RNN, Multiple Kernel Learning, and Multi-layer Perceptron, achieving 91.11% accuracy in PD diagnosis [95,96]. Support Vector Machines (SVMs) have also proven effective, with Little et al. establishing an early benchmark of 91.4% accuracy in classifying PD patients' voice recordings. Combined machine learning approaches, like Shen et al.'s hybrid method, have shown robust performance in distinguishing between healthy controls and PD patients [96]. These rival methods offer various strengths in feature selection, dimensionality reduction, and classification, with the choice of method depending on factors such as dataset size, feature complexity, and the desired interpretability of results.

5.2. Future Directions

The current design of the pre-aggregation method in our research can be improved, necessitating further exploration. To preserve intraindividual variability, more advanced techniques are required to extract maximum information from each voice recording during the pre-aggregation phase. One potential strategy is to concatenate various descriptive statistics, such as mean and variance, from the features of multiple voice recordings to create subject-level features. It may capture more information beyond the mean and has the potential to enhance classification performance. Additionally, neural networks could

be effective if a specialized encoder is developed to extract features from multiple voice recordings [97–103].

While our pipeline shows promising results, certain limitations might be addressed in future studies to enhance its generalizability. The data imbalance identified in this study stems from the case–control design, which maintains a 3:1 ratio of cases to controls. However, in practical applications, such class imbalances may not always be present. For instance, in population-based research, the number of individuals with a disease can be significantly lower than that of healthy individuals. This potential reversal in class distribution requires modifications to the pipeline for effective adaptation. Possible adjustments include implementing additional resampling techniques or recalibrating classification thresholds to strike a balance between sensitivity and specificity.

We acknowledge that our approach to voice analysis for Parkinson's disease detection uses fewer parameters than typical phonoscopy methods. While our feature selection aimed for simplicity and effectiveness based on previous research, we recognize its potential limitations in capturing the full spectrum of PD-related voice changes. To enhance our method, we propose that future studies incorporate a wider range of voice characteristics, conduct comparative analyses with more comprehensive approaches, and collaborate with experts in phonoscopy and speech pathology. These steps aim to improve the model's robustness and diagnostic accuracy while balancing simplicity and effectiveness. Furthermore, developing aggregation strategies that maintain robustness across varying class distributions by automatically adjusting for class imbalance could improve the pipeline's applicability in real-world scenarios.

6. Conclusions

This study developed a multi-algorithm AI pipeline for non-invasive Parkinson's disease screening using voice recordings, emphasizing the importance of data aggregation strategies in multiple instance learning with imbalanced classes. The post-mean aggregation method performed best, achieving high accuracy and MCC scores. While promising for early PD detection, we stress the need to consider potential confounding factors like nasopharyngeal and lung diseases. It is appropriate that a comprehensive approach combining thorough medical evaluation with advanced analytical techniques as presented here is considered for reliable PD diagnosis using voice analysis. The findings from this work have broader implications for applying aggregation methods in medical diagnostics for improved healthcare outcomes.

Supplementary Materials: A fully reproducible workflow using the provided Python scripts, including all necessary packages and code snippets, is available in our GitHub repository: https://github.com/Durixas/Parkinson-s-Disease-Prediction-Code-and-Data-Repository (accessed on 2 December 2024). These scripts are designed to facilitate reproducibility and allow users to replicate the machine learning classification tasks outlined in our study.

Author Contributions: S.K. planed and developed the conceptual framework for this study. Under S.K.'s direct supervision and with contributions from S.S. and J.G.S., Z.Y. and H.Z. designed and implemented the artificial intelligence pipeline for predictive modeling used in this study. The manuscript was written by Z.Y., H.Z., S.S., J.G.S., S.M.N., K.E.A. and S.K. S.K. was the corresponding author for the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Parkinson's Disease classification dataset used in this study is accessible at the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+classification, accessed on 1 September 2024) [42]. All data have been anonymized to

protect privacy. For additional assistance in reproducing results or customizing analysis workflows, users are encouraged to contact the corresponding author at the provided email address.

Acknowledgments: The authors are thankful to the office of the president, provost and the Dean of the Celia Scott Weatherhead School of Public Health and Tropical Medicine for the wonderful support.

Conflicts of Interest: The authors declare that there are no conflicts of interest or competing interests that could be perceived to bias this work. This includes, but is not limited to, financial relationships, personal relationships, professional affiliations, or intellectual property interests that might have influenced the research, its interpretation, or its presentation.

References

1. Olawade, D.B.; David-Olawade, A.C.; Wada, O.Z.; Asaolu, A.J.; Adereni, T.; Ling, J. Artificial intelligence in healthcare delivery: Prospects and pitfalls. *J. Med. Surg. Public Health* **2024**, *3*, 100108. [CrossRef]
2. Alowais, S.A.; Alghamdi, S.S.; Alsuhebany, N.; Alqahtani, T.; Alshaya, A.I.; Almohareb, S.N.; Aldairem, A.; Alrashed, M.; Bin Saleh, K.; Badreldin, H.A.; et al. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Med. Educ.* **2023**, *23*, 689. [CrossRef] [PubMed]
3. Maleki Varnosfaderani, S.; Forouzanfar, M. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. *Bioengineering* **2024**, *11*, 337. [CrossRef]
4. Javaid, M.; Haleem, A.; Pratap Singh, R.; Suman, R.; Rab, S. Significance of machine learning in healthcare: Features, pillars and applications. *Int. J. Intell. Netw.* **2022**, *3*, 58–73. [CrossRef]
5. Siddique, S.; Chow, J.C.L. Machine Learning in Healthcare Communication. *Encyclopedia* **2021**, *1*, 220–239. [CrossRef]
6. Sterkenburg, T.F.; Grünwald, P.D. The no-free-lunch theorems of supervised learning. *Synthese* **2021**, *199*, 9979–10015. [CrossRef]
7. Xu, Y.; Liu, X.; Cao, X.; Huang, C.; Liu, E.; Qian, S.; Liu, X.; Wu, Y.; Dong, F.; Qiu, C.-W.; et al. Artificial intelligence: A powerful paradigm for scientific research. *Innovation* **2021**, *2*, 100179. [CrossRef]
8. Datta, S.D.; Islam, M.; Rahman Sobuz, M.H.; Ahmed, S.; Kar, M. Artificial intelligence and machine learning applications in the project lifecycle of the construction industry: A comprehensive review. *Heliyon* **2024**, *10*, e26888. [CrossRef]
9. Salmi, M.; Atif, D.; Oliva, D.; Abraham, A.; Ventura, S. Handling imbalanced medical datasets: Review of a decade of research. *Artif. Intell. Rev.* **2024**, *57*, 273. [CrossRef]
10. Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; Gagnon, G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.* **2018**, *77*, 329–353. [CrossRef]
11. Foulds, J.; Frank, E. A review of multi-instance learning assumptions. *Knowl. Eng. Rev.* **2010**, *25*, 1–25. [CrossRef]
12. Ilse, M.; Tomczak, J.M.; Welling, M. Chapter 22—Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*; Zhou, S.K., Rueckert, D., Fichtinger, G., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 521–546. [CrossRef]
13. Asif, A.; Minhas, F.u.A.A. An embarrassingly simple approach to neural multiple instance classification. *Pattern Recognit. Lett.* **2019**, *128*, 474–479. [CrossRef]
14. Luan, T.; Gu, S.; Tang, X.; Zhuge, W.; Hou, C. Multi-Instance Learning with One Side Label Noise. *ACM Trans. Knowl. Discov. Data* **2024**, *18*, 122. [CrossRef]
15. Møllersen, K.; Hardeberg, J.Y.; Godtliebsen, F. A Probabilistic Bag-to-Class Approach to Multiple-Instance Learning. *Data* **2020**, *5*, 56. [CrossRef]
16. Herold, F.; Törpel, A.; Hamacher, D.; Budde, H.; Zou, L.; Strobach, T.; Müller, N.G.; Gronwald, T. Causes and Consequences of Interindividual Response Variability: A Call to Apply a More Rigorous Research Design in Acute Exercise-Cognition Studies. *Front. Physiol.* **2021**, *12*, 682891. [CrossRef]
17. Zhou, Z.-H.; Zhang, M.-L. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowl. Inf. Syst.* **2007**, *11*, 155–170. [CrossRef]
18. Aarsland, D.; Batzu, L.; Halliday, G.M.; Geurtsen, G.J.; Ballard, C.; Ray Chaudhuri, K.; Weintraub, D. Parkinson disease-associated cognitive impairment. *Nat. Rev. Dis. Primers* **2021**, *7*, 47. [CrossRef]
19. Ramesh, S.; Arachchige, A. Depletion of dopamine in Parkinson's disease and relevant therapeutic options: A review of the literature. *AIMS Neurosci.* **2023**, *10*, 200–231. [CrossRef]
20. Radad, K.; Moldzio, R.; Krewenka, C.; Kranner, B.; Rausch, W.-D. Pathophysiology of non-motor signs in Parkinson's disease: Some recent updating with brief presentation. *Explor. Neuroprotect. Ther.* **2023**, *3*, 24–46. [CrossRef]
21. Chaudhuri, K.R.; Schapira, A.H.V. Non-motor symptoms of Parkinson's disease: Dopaminergic pathophysiology and treatment. *Lancet Neurol.* **2009**, *8*, 464–474. [CrossRef]
22. Rana, A.Q.; Ahmed, U.S.; Chaudry, Z.M.; Vasan, S. Parkinson's disease: A review of non-motor symptoms. *Expert Rev. Neurother.* **2015**, *15*, 549–562. [CrossRef] [PubMed]

23. Park, A.; Stacy, M. Non-motor symptoms in Parkinson's disease. *J. Neurol.* **2009**, *256*, 293–298. [CrossRef] [PubMed]
24. Thach, A.; Jones, E.; Pappert, E.; Pike, J.; Wright, J.; Gillespie, A. Real-world assessment of the impact of "OFF" episodes on health-related quality of life among patients with Parkinson's disease in the United States. *BMC Neurol.* **2021**, *21*, 46. [CrossRef] [PubMed]
25. Zhu, J.; Cui, Y.; Zhang, J.; Yan, R.; Su, D.; Zhao, D.; Wang, A.; Feng, T. Temporal trends in the prevalence of Parkinson's disease from 1980 to 2023: A systematic review and meta-analysis. *Lancet Healthy Longev.* **2024**, *5*, e464–e479. [CrossRef] [PubMed]
26. Dorsey, E.R.; Bloem, B.R. The Parkinson Pandemic—A Call to Action. *JAMA Neurol.* **2018**, *75*, 9–10. [CrossRef]
27. Schiess, N.; Cataldi, R.; Okun, M.S.; Fothergill-Misbah, N.; Dorsey, E.R.; Bloem, B.R.; Barretto, M.; Bhidayasiri, R.; Brown, R.; Chishimba, L.; et al. Six Action Steps to Address Global Disparities in Parkinson Disease: A World Health Organization Priority. *JAMA Neurol.* **2022**, *79*, 929–936. [CrossRef]
28. Savica, R.; Grossardt, B.R.; Bower, J.H.; Ahlskog, J.E.; Rocca, W.A. Time Trends in the Incidence of Parkinson Disease. *JAMA Neurol.* **2016**, *73*, 981–989. [CrossRef]
29. Kowal, S.L.; Dall, T.M.; Chakrabarti, R.; Storm, M.V.; Jain, A. The current and projected economic burden of Parkinson's disease in the United States. *Mov. Disord.* **2013**, *28*, 311–318. [CrossRef]
30. Rong, S.; Xu, G.; Liu, B.; Sun, Y.; Snetselaar, L.G.; Wallace, R.B.; Li, B.; Liao, J.; Bao, W. Trends in Mortality from Parkinson Disease in the United States, 1999–2019. *Neurology* **2021**, *97*, e1986–e1993. [CrossRef]
31. Ayyadevara, S.; Balasubramaniam, M.; Kakraba, S.; Alla, R.; Mehta, J.L.; Shmookler Reis, R.J. Aspirin-Mediated Acetylation Protects Against Multiple Neurodegenerative Pathologies by Impeding Protein Aggregation. *Antioxid. Redox Signal.* **2017**, *27*, 1383–1396. [CrossRef]
32. Bowroju, S.K.; Mainali, N.; Ayyadevara, S.; Penthala, N.R.; Krishnamachari, S.; Kakraba, S.; Shmookler Reis, R.J.; Crooks, P.A. Design and Synthesis of Novel Hybrid 8-Hydroxy Quinoline-Indole Derivatives as Inhibitors of Aβ Self-Aggregation and Metal Chelation-Induced Aβ Aggregation. *Molecules* **2020**, *25*, 3610. [CrossRef] [PubMed]
33. Kakraba, S.; Ayyadevara, S.; Mainali, N.; Balasubramaniam, M.; Bowroju, S.; Penthala, N.R.; Atluri, R.; Barger, S.W.; Griffin, S.T.; Crooks, P.A.; et al. Thiadiazolidinone (TDZD) Analogs Inhibit Aggregation-Mediated Pathology in Diverse Neurodegeneration Models, and Extend Life- and Healthspan. *Pharmaceuticals* **2023**, *16*, 1498. [CrossRef] [PubMed]
34. Kakraba, S.; Ayyadevara, S.; Penthala, N.R.; Balasubramaniam, M.; Ganne, A.; Liu, L.; Alla, R.; Bommagani, S.B.; Barger, S.W.; Griffin, W.S.T.; et al. A Novel Microtubule-Binding Drug Attenuates and Reverses Protein Aggregation in Animal Models of Alzheimer's Disease. *Front. Mol. Neurosci.* **2019**, *12*, 310. [CrossRef]
35. Tod, A.M.; Kennedy, F.; Stocks, A.-J.; McDonnell, A.; Ramaswamy, B.; Wood, B.; Whitfield, M. Good-quality social care for people with Parkinson's disease: A qualitative study. *BMJ Open* **2016**, *6*, e006813. [CrossRef]
36. Paulsen, J.S.; Nance, M.; Kim, J.-I.; Carlozzi, N.E.; Panegyres, P.K.; Erwin, C.; Goh, A.; McCusker, E.; Williams, J.K. A review of quality of life after predictive testing for and earlier identification of neurodegenerative diseases. *Prog. Neurobiol.* **2013**, *110*, 2–28. [CrossRef]
37. Alanazi, M.D.S.; Ateeq, H.A.A.A.; Aldhefri, A.B.; Alanazi, A.H.; Almutairi, M.A.A.; Almogamas, A.M.A.; Al-Qahtani, A.Y.; Khabrani, H.A. Parkinson's Disease: Neurotransmitter Imbalance, Motor Dysfunction, and Nursing Interventions for Quality of Life. *J. Int. Crisis Risk Commun. Res.* **2024**, *7*, 269–282.
38. Bužgová, R.; Kozáková, R.; Bar, M. The effect of neuropalliative care on quality of life and satisfaction with quality of care in patients with progressive neurological disease and their family caregivers: An interventional control study. *BMC Palliat. Care* **2020**, *19*, 143. [CrossRef]
39. Rees, R.N.; Acharya, A.P.; Schrag, A.; Noyce, A.J. An early diagnosis is not the same as a timely diagnosis of Parkinson's disease. *F1000Research* **2018**, *7*. [CrossRef]
40. Dodel, R.C.; Berger, K.; Oertel, W.H. Health-Related Quality of Life and Healthcare Utilisation in Patients with Parkinson's Disease. *Pharmacoeconomics* **2001**, *19*, 1013–1038. [CrossRef]
41. Goldman, J.G.; Volpe, D.; Ellis, T.D.; Hirsch, M.A.; Johnson, J.; Wood, J.; Aragon, A.; Biundo, R.; Di Rocco, A.; Kasman, G.S.; et al. Delivering Multidisciplinary Rehabilitation Care in Parkinson's Disease: An International Consensus Statement. *J. Park. Dis.* **2024**, *14*, 135–166. [CrossRef]
42. Sakar, C.; Serbes, G.; Gunduz, A.; Nizam, H.; Sakar, B. Parkinson's Disease Classification [Dataset]. 2018. Available online: https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+classification (accessed on 1 September 2024).
43. Wang, M.; Zhao, X.; Li, F.; Wu, L.; Li, Y.; Tang, R.; Yao, J.; Lin, S.; Zheng, Y.; Ling, Y.; et al. Using sustained vowels to identify patients with mild Parkinson's disease in a Chinese dataset. *Front. Aging Neurosci.* **2024**, *16*, 1377442. [CrossRef] [PubMed]
44. Aich, S.; Kim, H.C.; Younga, K.; Hui, K.L.; Al-Absi, A.A.; Sain, M. A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson's Disease. In Proceedings of the 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang, Republic of Korea, 17–20 February 2019; pp. 1116–1121.

45. Ho, A.K.; Iansek, R.; Marigliani, C.; Bradshaw, J.L.; Gates, S. Speech impairment in a large sample of patients with Parkinson's disease. *Behav. Neurol.* **1998**, *11*, 131–137. [CrossRef] [PubMed]
46. Vandana, V.P.; Darshini, J.K.; Vikram, V.H.; Nitish, K.; Kumar, P.P.; Ravi, Y. Speech Characteristics of Patients with Parkinson's Disease-Does Dopaminergic Medications Have a Role? *J. Neurosci. Rural Pract.* **2021**, *12*, 673–679. [CrossRef] [PubMed]
47. Skodda, S. Aspects of speech rate and regularity in Parkinson's disease. *J. Neurol. Sci.* **2011**, *310*, 231–236. [CrossRef]
48. Tabari, F.; Berger, J.I.; Flouty, O.; Copeland, B.; Greenlee, J.D.; Johari, K. Speech, voice, and language outcomes following deep brain stimulation: A systematic review. *PLoS ONE* **2024**, *19*, e0302739. [CrossRef]
49. Krasko, M.N.; Hoffmeister, J.D.; Schaen-Heacock, N.E.; Welsch, J.M.; Kelm-Nelson, C.A.; Ciucci, M.R. Rat Models of Vocal Deficits in Parkinson's Disease. *Brain Sci.* **2021**, *11*, 925. [CrossRef]
50. Iyer, A.; Kemp, A.; Rahmatallah, Y.; Pillai, L.; Glover, A.; Prior, F.; Larson-Prior, L.; Virmani, T. A machine learning method to process voice samples for identification of Parkinson's disease. *Sci. Rep.* **2023**, *13*, 20615. [CrossRef]
51. Berus, L.; Klancnik, S.; Brezocnik, M.; Ficko, M. Classifying Parkinson's Disease Based on Acoustic Measures Using Artificial Neural Networks. *Sensors* **2019**, *19*, 16. [CrossRef]
52. Pedregosa, F. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825.
53. Tan, J.; Yang, J.; Wu, S.; Chen, G.; Zhao, J. A critical look at the current train/test split in machine learning. *arXiv* **2021**, arXiv:2106.04525.
54. Szeghalmy, S.; Fazekas, A. A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. *Sensors* **2023**, *23*, 2333. [CrossRef] [PubMed]
55. Zheng, W.; Jin, M. The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study. *SN Comput. Sci.* **2020**, *1*, 71. [CrossRef]
56. Cao, P.; Ren, F.; Wan, C.; Yang, J.; Zaiane, O. Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis. *Comput. Med. Imaging Graph.* **2018**, *69*, 112–124. [CrossRef] [PubMed]
57. Chawla, N.; Bowyer, K.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *arXiv* **2002**, arXiv:1106.1813. [CrossRef]
58. Liu, L. Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning. In Proceedings of the 2018 International Conference on Robots & Intelligent System (ICRIS), Changsha, China, 26–27 May 2018; pp. 157–160.
59. Charbuty, B.; Abdulazeez, A. Classification Based on Decision Tree Algorithm for Machine Learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 20–28. [CrossRef]
60. Liu, Y.; Wang, Y.; Zhang, J. New Machine Learning Algorithm: Random Forest. In *Information Computing and Applications*; Springer: Berlin/Heidelberg, Germany, 2012.
61. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2013**, *7*, 21. [CrossRef]
62. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
63. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Neural Information Processing Systems*; Curran Associates Inc.: San Francisco, CA, USA, 2017.
64. Jodas, D.S.; Passos, L.A.; Adeel, A.; Papa, J.P. PL-kNN: A Python-based implementation of a parameterless k-Nearest Neighbors classifier. *Softw. Impacts* **2023**, *15*, 100459. [CrossRef]
65. Abdullah, D.M.; Abdulazeez, A.M. Machine Learning Applications based on SVM Classification A Review. *Qubahan Acad. J.* **2021**, *1*, 81–90. [CrossRef]
66. Lowd, D.; Domingos, P.M. Naive Bayes models for probability estimation. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005.
67. Schapire, R.E. The Boosting Approach to Machine Learning: An Overview. In *Nonlinear Estimation and Classification*; Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B., Eds.; Springer: New York, NY, USA, 2003; pp. 149–171.
68. Orrù, P.F.; Zoccheddu, A.; Sassu, L.; Mattia, C.; Cozza, R.; Arena, S. Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal Pump in the Oil and Gas Industry. *Sustainability* **2020**, *12*, 4776. [CrossRef]
69. Mienye, I.D.; Sun, Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access* **2022**, *10*, 99129–99149. [CrossRef]
70. Jakhar, A.K.; Gupta, A.; Singh, M. SELF: A stacked-based ensemble learning framework for breast cancer classification. *Evol. Intell.* **2024**, *17*, 1341–1356. [CrossRef]
71. Shekar, B.H.; Dagnew, G. Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. In Proceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, 25–28 February 2019.
72. Browne, M.W. Cross-Validation Methods. *J. Math. Psychol.* **2000**, *44*, 108–132. [CrossRef] [PubMed]

73. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
74. Wardhani, N.W.S.; Rochayani, M.Y.; Iriany, A.; Sulistyono, A.D.; Lestantyo, P. Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data. In Proceedings of the 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Tangerang, Indonesia, 23–24 October 2019.
75. Halimu, C.; Kasem, A.; Newaz, S.M. Empirical Comparison of Area under ROC curve (AUC) and Mathew Correlation Coefficient (MCC) for Evaluating Machine Learning Algorithms on Imbalanced Datasets for Binary Classification. In Proceedings of the 3rd International Conference on Machine Learning and Soft Computing, Da Lat, Vietnam, 25–28 January 2019.
76. Pah, N.D.; Motin, M.A.; Kempster, P.; Kumar, D.K. Detecting Effect of Levodopa in Parkinson's Disease Patients Using Sustained Phonemes. *IEEE J. Transl. Eng. Health Med.* **2021**, *9*, 4900409. [CrossRef]
77. Ngo, Q.C.; Motin, M.A.; Pah, N.D.; Drotár, P.; Kempster, P.; Kumar, D. Computerized analysis of speech and voice for Parkinson's disease: A systematic review. *Comput. Methods Programs Biomed.* **2022**, *226*, 107133. [CrossRef]
78. Ali, L.; Zhu, C.; Zhang, Z.; Liu, Y. Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network. *IEEE J. Transl. Eng. Health Med.* **2019**, *7*, 2000410. [CrossRef]
79. Arora, S.; Tsanas, A. Assessing Parkinson's Disease at Scale Using Telephone-Recorded Speech: Insights from the Parkinson's Voice Initiative. *Diagnostics* **2021**, *11*, 1892. [CrossRef]
80. Azadi, H.; Akbarzadeh-T, M.-R.; Shoeibi, A.; Kobravi, H.R. Evaluating the Effect of Parkinson's Disease on Jitter and Shimmer Speech Features. *Adv. Biomed. Res.* **2021**, *10*, 54. [CrossRef]
81. Viswanathan, R.; Arjunan, S.P.; Bingham, A.; Jelfs, B.; Kempster, P.; Raghav, S.; Kumar, D.K. Complexity Measures of Voice Recordings as a Discriminative Tool for Parkinson's Disease. *Biosensors* **2020**, *10*, 1. [CrossRef]
82. Gunduz, H. Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets. *IEEE Access* **2019**, *7*, 115540–115551. [CrossRef]
83. Polat, K.; Nour, M. Parkinson disease classification using one against all based data sampling with the acoustic features from the speech signals. *Med. Hypotheses* **2020**, *140*, 109678. [CrossRef] [PubMed]
84. Sakar, C.O.; Serbes, G.; Gunduz, A.; Tunc, H.C.; Nizam, H.; Sakar, B.E.; Tutuncu, M.; Aydin, T.; Isenkul, M.E.; Apaydin, H. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Appl. Soft Comput.* **2019**, *74*, 255–263. [CrossRef]
85. Sakar, B.E.; Isenkul, M.E.; Sakar, C.O.; Sertbas, A.; Gurgen, F.; Delil, S.; Apaydin, H.; Kursun, O. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE J. Biomed. Health Inform.* **2013**, *17*, 828–834. [CrossRef] [PubMed]
86. Naranjo, L.; Pérez, C.J.; Martín, J.; Campos-Roca, Y. A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications. *Comput. Methods Programs Biomed.* **2017**, *142*, 147–156. [CrossRef] [PubMed]
87. Moro-Velázquez, L.; Gómez-García, J.A.; Godino-Llorente, J.I.; Villalba, J.; Orozco-Arroyave, J.R.; Dehak, N. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease. *Appl. Soft Comput.* **2018**, *62*, 649–666. [CrossRef]
88. García, A.M.; Carrillo, F.; Orozco-Arroyave, J.R.; Trujillo, N.; Vargas Bonilla, J.F.; Fittipaldi, S.; Adolfi, F.; Nöth, E.; Sigman, M.; Fernández Slezak, D.; et al. How language flows when movements don't: An automated analysis of spontaneous discourse in Parkinson's disease. *Brain Lang.* **2016**, *162*, 19–28. [CrossRef]
89. Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [CrossRef]
90. Sasse, E.A. Objective evaluation of data in screening for disease. *Clin. Chim. Acta* **2002**, *315*, 17–30. [CrossRef]
91. Trevizan, B.; Chamby-Diaz, J.; Bazzan, A.L.C.; Recamonde-Mendoza, M. A comparative evaluation of aggregation methods for machine learning over vertically partitioned data. *Expert Syst. Appl.* **2020**, *152*, 113406. [CrossRef]
92. Naranjo, L.; Pérez, C.J.; Campos-Roca, Y.; Martín, J. Addressing voice recording replications for Parkinson's disease detection. *Expert Syst. Appl.* **2016**, *46*, 286–292. [CrossRef]
93. Holmes, R.J.; Oates, J.M.; Phyland, D.J.; Hughes, A.J. Voice characteristics in the progression of Parkinson's disease. *Int. J. Lang. Commun. Disord.* **2000**, *35*, 407–418. [CrossRef] [PubMed]
94. Tsanas, A.; Arora, S. Large-scale Clustering of People Diagnosed with Parkinson's Disease using Acoustic Analysis of Sustained Vowels: Findings in the Parkinson's Voice Initiative Study. In Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing, Valletta, Malta, 24–26 February 2020.
95. Lee, S.; Hussein, R.; Ward, R.; Jane Wang, Z.; McKeown, M.J. A convolutional-recurrent neural network approach to resting-state EEG classification in Parkinson's disease. *J. Neurosci. Methods* **2021**, *361*, 109282. [CrossRef] [PubMed]

96. Shen, M.; Mortezaagha, P.; Rahgozar, A. Explainable Artificial Intelligence to Diagnose Early Parkinson's Disease via Voice Analysis. *medRxiv* **2024**. [CrossRef]
97. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
98. Hou, J.C.; Wang, S.S.; Lai, Y.H.; Tsao, Y.; Chang, H.W.; Wang, H.M. Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2018**, *2*, 117–128. [CrossRef]
99. Ye, F.; Yang, J. A Deep Neural Network Model for Speaker Identification. *Appl. Sci.* **2021**, *11*, 3603. [CrossRef]
100. Chen, L.; Chen, J. Deep Neural Network for Automatic Classification of Pathological Voice Signals. *J. Voice* **2022**, *36*, e215–e288. [CrossRef]
101. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [CrossRef]
102. Chaiani, M.; Selouani, S.A.; Boudraa, M.; Sidi Yakoub, M. Voice disorder classification using speech enhancement and deep learning models. *Biocybern. Biomed. Eng.* **2022**, *42*, 463–480. [CrossRef]
103. Yadav, S.P.; Zaidi, S.; Mishra, A.; Yadav, V. Survey on Machine Learning in Speech Emotion Recognition and Vision Systems Using a Recurrent Neural Network (RNN). *Arch. Comput. Methods Eng.* **2022**, *29*, 1753–1770. [CrossRef]