Leader-Enhanced Optimization for Robust Neural Ensemble Learning in Medical Audio Diagnosis

Ziang Zhao
Department of Computer Science
Royal Holloway, University of London
Surrey, the United Kingdom
ziang.zhao@rhul.ac.uk

Arjun Panesar
DDM Health Ltd.
University of Warwick
Coventry, the United Kingdom
arjun@ddm.health

Li Zhang*
Department of Computer Science
Royal Holloway, University of London
Surrey, the United Kingdom
li.zhang@rhul.ac.uk

Yonghong Yu
College of Tongda
Nanjing University of Posts and Telecommunications
Nanjing, China
yuyh@njupt.edu.cn

***Abstract*—Medical audio analysis offers a non-invasive and scalable solution for disease screening, but its predictions are often affected by substantial uncertainty arising from inter-subject variability, recording conditions, and limited data availability. While ensemble models are commonly used to improve performance, their effectiveness critically depends on how individual models are combined, particularly under noisy evaluation settings. This paper presents a leader-enhanced evolutionary search framework for robust ensemble weight optimization in medical audio analysis. We formulate ensemble construction as a constrained swarm-based optimization problem with noisy objectives and propose four new Particle Swarm Optimization (PSO) variants for base classifier weight identification. where Genetic Algorithm (GA), Differential Evolution (DE), Simulated Annealing (SA), and Covariance Matrix Adaptation Evolution Strategy (CMA-ES) are used for swarm leader enhancement. Five audio deep networks and transformer models are adopted as base classifiers, including ResNet, RegNet, MobileViT, Whisper, and wav2vec2, with either spectrograms or waveforms as inputs to increase base model diversity. Experiments conducted on diabetes and Parkinson's disease voice/speech datasets demonstrate that the proposed framework consistently outperforms individual base models and conventional ensemble baselines. Ablation studies further reveal that leader-enhancement strategies using GA, DE, SA and CMA-ES effectively balance exploration and stability in adaptive ensemble weight search. These empirical results indicate that evolutionary ensemble optimization provides a practical and reliable approach for robust medical audio classification.**

***Index Terms*—Medical Audio classification, Swarm intelligent algorithms, Ensemble Optimization.**

I. Introduction

Medical audio analysis has emerged as a promising non-invasive tool for disease screening and monitoring, including respiratory disorder [1], diabetes [2], and Parkinson's disease detection [3]. Despite encouraging progress, predictions derived from medical audio data are inherently uncertain due to multiple factors, such as environmental noise, patient-specific artifacts, and recording device inconsistencies [4]. Most existing approaches rely on a single predictive model, such as a convolutional neural network (CNN), a recurrent architecture, or a classical machine learning classifier [5]. However, this paradigm suffers from structural limitations in clinical audio settings. A single model inevitably encodes inductive biases tied to its architecture and feature representation. When the test sample distribution deviates from the training instance distribution, these biases can lead to prediction errors.

To mitigate such limitations, ensemble learning has gained prominence as a strategy to combine multiple base models, leveraging their complementary strengths for robust prediction [6]. Ensemble modeling provides a principled mechanism for mitigating medical risks. Instead of using ensemble learning as a simple performance-boosting technique, it can be more appropriately treated as a form of risk control. By aggregating multiple base predictors, ensembles are able to reduce variance and improve robustness against distributional shifts. This perspective is important in clinical applications, where overconfident but incorrect predictions can have serious consequences. In this context, the objective of ensemble design is not only to maximize average accuracy, but also to achieve stable and reliable predictions under uncertainty.

Despite their advantages, the effectiveness of ensembles critically depends on how individual models are combined. Fixed or heuristic weighting schemes fail to acknowledge dataset-specific noise characteristics and inter-model dependencies. Moreover, performance evaluation in medical audio tasks is noisy, as metrics can fluctuate across subjects or sampling strategies. As a result, we formulate ensemble construction as a constrained swarm-based optimization problem with noisy objectives (i.e., stochastic validation fluctuations due to limited samples and speaker variability), where the goal is to identify robust combinations of models and weights that perform well across different conditions.

In this work, we propose a leader-enhanced evolutionary

This work is supported by the Innovate UK Knowledge Transfer Partnerships (No. 10143791). Corresponding author: Li Zhang.

search framework for robust model ensembling in clinical audio analysis. Four Particle Swarm Optimization (PSO) variant methods are proposed for identifying optimal base classifier weights, where Genetic Algorithm (GA), Differential Evolution (DE), Simulated Annealing (SA), and Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [7], are used for swarm leader enhancement. The new PSO methods equipped with these leader-enhancement mechanisms are able to handle noisy biased evaluation results and to work within the constraints of the weight space. Experiments on diabetes and Parkinson's disease audio datasets demonstrate that the proposed approach consistently improves both predictive performance and stability, compared to single models and traditional average weighting strategy.

The main contributions of this paper are summarized as follows:

- We formulate clinical audio ensemble construction as a constrained swarm-based optimization problem with noisy objectives, motivated by the uncertainty inherent in medical audio data.
- Four hybrid evolutionary search mechanisms are proposed, where GA, DE, SA, and CMA-ES, are used for swarm leader adjustment and search efficiency improvement of the original PSO method.
- Extensive experiments on two medical audio tasks demonstrate that leader-enhanced PSO variants consistently improve accuracy and weight stability.

Due to institutional ethics policies and project non-disclosure agreements, the diabetes dataset and implementation code are available upon request, subject to industrial partner approval.

II. Related Work

A. Medical Audio-Based Disease Identification

Growing research has focused on disease detection using speech, coughing, and respiratory sounds due to its non-invasive nature and low deployment cost [8, 9]. For lung disease identification, deep learning models based on convolutional and recurrent architectures have been widely applied to respiratory and coughing sound analysis, achieving promising results on public datasets such as Coswara [10, 11].

Medical studies link diabetic neuropathy to vocal cord impairment, motivating audio-based diabetes detection [12,13]. Audio-based diabetes detection has therefore gained attention as a non-invasive alternative to biochemical tests. Several studies report statistically significant differences in prosodic and spectral speech features between diabetic and healthy subjects [14,15]. Deep learning and hybrid embedding models have been applied to large-scale smartphone-recorded speech datasets, achieving moderate predictive performance while revealing substantial variability across demographic groups [2,16]. These findings reinforce the need for robust modeling strategies that explicitly address uncertainty.

In Parkinson's disease analysis, voice impairment has long been recognized as an early symptom. Recent studies employ deep neural networks and hybrid ensemble frameworks to classify Parkinson's disease and to predict clinical severity scores from sustained phonation and text-reading speech recordings [17, 18]. These works report high accuracy under controlled evaluation settings, but also highlight challenges related to speaker variability, limited sample sizes, and cross-dataset generalization.

B. Deep Model Ensemble Methods

Ensemble learning is a well-established paradigm for improving generalization in machine learning. Classical approaches such as bagging, boosting, and stacking reduce variance and bias by aggregating multiple learners [6]. In deep learning, ensembles of independently trained networks, snapshot ensembles, and heterogeneous architectures have been shown to outperform single models across a wide range of tasks, including medical image and audio analysis.

In medical applications, ensembles are particularly attractive due to their improved robustness and implicit uncertainty reduction. Recent studies have applied weighted averaging and stacking-based ensembles to medical audio tasks, reporting consistent performance gains [19]. However, most existing approaches rely on fixed or manually tuned weights, which limit their adaptability to different applications.

C. Evolutionary Algorithms for Ensemble Weight Optimization

Evolutionary algorithms have been widely used for model selection, hyperparameter tuning, and ensemble weight optimization in machine learning [20]. PSO, GA, and DE have all been applied to ensemble construction, demonstrating advantages over gradient-based or heuristic methods [21, 22]. These approaches are suitable when objective functions are non-differentiable, noisy, or expensive to evaluate.

Recent work has proposed hybrid evolutionary strategies that combine multiple optimizers to improve search efficiency and solution quality. Such approaches have been successfully applied to disease classification tasks, including diabetes prediction using UCI datasets with pre-defined feature vectors [23], where PSO combined with Grey Wolf Optimization (GWO) helps navigate complex and noisy objective spaces for hyperparameter search.

D. Limitations of Existing Approaches

Existing ensemble optimization methods in clinical audio analysis exhibit two main limitations. First, they largely ignore predictive uncertainty and instability, instead purely focusing on mean performance metrics. This is risky in clinical scenarios, where reliable decision-making is as important as accuracy. Second, current evolutionary ensemble methods rarely exploit structured search guidance. Leader selection is typically static or implicit, which can lead to premature convergence or inefficient exploration of the noisy optimization landscapes.

These limitations motivate the proposed leader-enhanced evolutionary search strategy, which addresses noisy objective evaluations and introduces adaptive leader mechanisms to improve robustness and convergence behavior in clinical audio ensemble optimization.

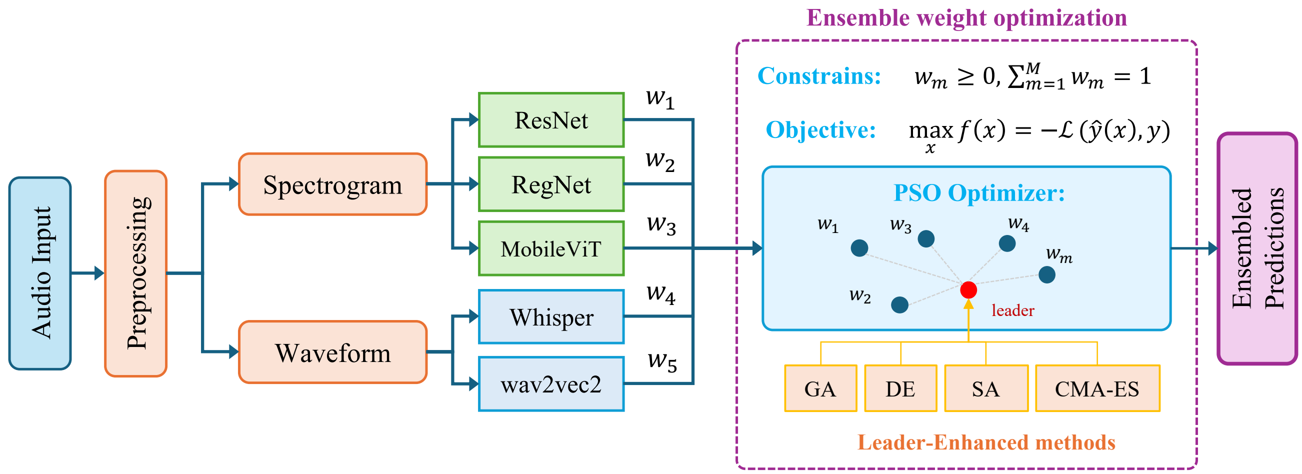

Fig. 1. The architecture of the proposed model, where GA, DE, SA, and CMA-ES-enhanced PSO variants are used for ensemble weight optimization.

III. Proposed Method

Firstly, several CNNs and audio transformers are selected as the base classifiers, i.e., ResNet, RegNet, MobileViT, Whisper, and wav2vec2, owing to their complementary local and global feature learning behaviors. Adaptive ensemble construction is subsequently performed, where the four PSO variants with evolutionary leader enhancement strategies are used to generate base classifier weighting factors to fine-tune their effects for determining ensemble prediction outcomes. These hybrid PSO variants with GA, SA, DE, and CMA-ES based leader enhancement mitigate local optima and enhance leader exploitation. We introduce each key step in detail below.

A. Audio Preprocessing

In this research, public and privately collected voice/speech datasets are employed for diabetes and Parkinson's disease diagnosis. Audio preprocessing is first performed for noise removal, audio length clipping/padding, and spectrogram transformation. Detailed processes are provided below.

1) Spectrogram Transformation: For each audio signal, a Short-Time Fourier Transform (STFT) was computed to obtain a magnitude spectrogram. A fixed hop length of 512 samples was used to control the temporal resolution. The magnitude spectrogram was subsequently transformed to a logarithmic decibel (dB) scale, which compresses the dynamic range. Each spectrogram was rendered at a fixed resolution and saved as an image, which served as the input to the image-based models, such as ResNet, RegNet and MobileViT.

2) Waveform Processing: All raw audio waveforms were also directly used as model inputs for waveform-based models. Given an input audio signal, all recordings were first converted to mono and resampled to a fixed sampling rate of 16 kHz, to align with pre-trained models and ensure cross-dataset consistency. As the length of the audio is different, the waveform was zero-padded to a fixed duration of 30 seconds for batch-loading. The processed audio waveforms are used as the input to audio encoders such as wav2vec2 and Whisper.

B. Base Model Construction

According to the above two types of audio input representations (spectrograms and waveforms), two categories of base models were employed: image-based models operating on spectrograms and waveform-based models operating directly on raw audio signals.

For image-based inputs, convolutional and hybrid convolution–transformer architectures were adopted, including ResNet [24], RegNet [25], and MobileViT [26]. These models take spectrograms as the input and are capable of capturing both local time-frequency patterns and longer-range contextual dependencies. In this study, ResNet-18, RegNet-Y80, and MobileViT-Small were selected as representative image-based base classifiers. All models were adapted to the binary disease classification tasks by re-training the final classification layer, while keeping the rest of the backbone architectures unchanged.

For waveform-based inputs, large-scale pre-trained speech models, Whisper [27] and wav2vec2 [28], were used. These models process normalized raw waveforms through convolutional frontends followed by transformer encoders, producing high-level speech representations directly from the input signal. On top of each pre-trained encoder, a simple two-layer fully connected network was appended as the classification head.

C. Proposed PSO-based Ensemble Weight Optimization

After generating the base classifier pool, we discuss ensemble model construction with the proposed PSO-optimized ensemble weights. Standard PSO may stagnate around sub-optimal leaders under noisy validation, which motivates our leader-enhancement mechanisms. Four leader-enhanced variants are developed: PSO-GA, PSO-SA, PSO-DE, and PSO-CMA. These methods are used for optimal weight generation to fine-tune the effects of each classifier for ensemble prediction. The illustration of the proposed method is shown in Figure 1.

Each optimization iteration follows a unified procedure: (1) swarm evaluation and position update via standard PSO, (2) local refinement of the global best particle using a designated evolutionary operator (e.g., GA, DE, SA, or CMA), (3) softmax projection of the optimized unconstrained solution onto the probability simplex to obtain ensemble weights.

D. Weighted Ensemble Construction

We introduce the proposed weighted ensemble prediction strategy in detail below. Assume M trained medical audio classification models. For an input sample, each model predicts a class probability vector $\mathbf{o}_m \in \mathbb{R}^C$, and the final ensemble prediction $\hat{\mathbf{y}}$ is computed as a weighted combination:

$$\hat{\mathbf{y}} = \sum_{m=1}^{M} w_m \mathbf{o}_m \tag{1}$$

where $\mathbf{w} = (w_1, \ldots, w_M)$ are generated by the optimization process:

$$w_m \geq 0, \quad \sum_{m=1}^{M} w_m = 1 \tag{2}$$

Precisely several proposed PSO variants are used to devise these weighting factors w_m to fine-tune the effects of each base classifier for ensemble result formulation. The objective function for ensemble weight optimization is defined as follows.

$$\max_{\mathbf{x}} \; f(\mathbf{x}) = -\mathcal{L}\big(\hat{\mathbf{y}}(\mathbf{x}), \mathbf{y}\big) \tag{3}$$

where $\mathcal{L}$ denotes the validation loss and $\mathbf{y}$ is the ground-truth label. This objective function is used to measure ensemble model performance with the proposed PSO-optimized weights. Such a process helps to determine the effectiveness of each set of base classifier weight settings recommended by each individual particle. We subsequently introduce the proposed PSO variant for base classifier weight optimization.

After the optimization process converges, the softmax operation is applied to the optimized weights w_m to obtain the final ensemble weights w'_m:

$$w'_m = \frac{\exp(x_m)}{\sum_{j=1}^{M} \exp(x_j)} \tag{4}$$

E. Original PSO

We employ PSO as the foundational algorithm, which maintains a swarm of N particles. Each particle i has a position $\mathbf{x}_i$ and a velocity $\mathbf{v}_i$. At iteration t, particles are updated according to the following operations:

$$\begin{cases} \mathbf{v}_i^{t+1} = \omega \mathbf{v}_i^t + c_1 \mathbf{r}_1 \odot (\mathbf{p}_i - \mathbf{x}_i^t) + c_2 \mathbf{r}_2 \odot (\mathbf{g} - \mathbf{x}_i^t) \\ \mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{v}_i^{t+1} \end{cases} \tag{5}$$

where $\mathbf{p}_i$ is the personal best of particle i, $\mathbf{g}$ is the global best particle, ω is the inertia weight, c_1 and c_2 are acceleration coefficients, and $\mathbf{r}_1$ and $\mathbf{r}_2$ are random vectors sampled from $[0,1]^M$.

F. The Proposed PSO Variants with Leader-Enhanced Evolutionary Refinement

The standard PSO search process rapidly converges towards the global best, which may lead to premature convergence in a rugged fitness landscape. To address this issue, we introduce a *leader enhancement* stage that explicitly refines the global best solution using auxiliary evolutionary operators. Four leader-enhanced strategies are considered.

1) PSO with Genetic Algorithm Enhancement (PSO-GA): The proposed PSO-GA method employs genetic operators to evolve the leader through elite selection and adaptive mutation. At each iteration, we select the top-K particles based on the fitness scores of the personal best solutions as the elite population $\mathcal{E}$.

For each enhancement trial, offspring solutions are generated through the following crossover operation:

$$\mathbf{c} = \alpha \mathbf{e}_i + (1-\alpha)\mathbf{e}_j, \quad \alpha \sim \mathcal{U}(0,1) \tag{6}$$

where $\mathbf{e}_i, \mathbf{e}_j \in \mathcal{E}$ are randomly selected parents, α is the crossover rate. A Gaussian mutation is then applied:

$$\mathbf{c} \leftarrow \mathbf{c} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_t^2 \mathbf{I}) \tag{7}$$

where the mutation strength $\sigma_t = \sigma_0(1 - t/T)$ decays with initial strength $\sigma_0 = 0.1$ over T iterations. If the generated offspring solutions are fitter than the current leader $\mathbf{g}$, they are used to replace the swarm leader.

2) PSO with Simulated Annealing Enhancement (PSO-SA): The proposed PSO-SA algorithm incorporates SA for swarm leader enhancement. Specifically, it adopts stochastic hill-climbing through the Metropolis criterion of the SA to escape local optima. At each iteration, multiple perturbations are generated:

$$\mathbf{c} = \mathbf{g} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{8}$$

where σ is set equal to the current PSO velocity scale. The candidate is accepted according to the Metropolis criterion:

$$P_{accept} = \begin{cases} 1, & \Delta f > 0 \\ \exp(\Delta f / \tau_t), & \text{otherwise} \end{cases} \tag{9}$$

where $\Delta f = f(\mathbf{c}) - f(\mathbf{g})$ is the fitness improvement and the temperature τ_t follows a geometric cooling schedule:

$$\tau_t = \max(\alpha_{cool} \cdot \tau_{t-1}, \tau_{min}) \tag{10}$$

τ_0 and τ_{min} are initial and minimum temperature, α_{cool} is cooling rate. If accepted, the global best $\mathbf{g}$ is replaced with the vector $\mathbf{c}$.

3) PSO with Differential Evolution Enhancement (PSO-DE): The proposed PSO-DE algorithm adopts the mutation operations of DE to refine the global best position. At each iteration, the leader is refined by a differential mutation operation using two randomly selected swarm particles $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}$:

$$\mathbf{m} = \mathbf{g} + F(\mathbf{x}_{i_1} - \mathbf{x}_{i_2}) \tag{11}$$

$$\mathbf{c}_j = \begin{cases} \mathbf{m}_j, & \text{if } r_j < \mathrm{CR} \\ \mathbf{g}_j, & \text{otherwise} \end{cases} \tag{12}$$

where $\mathbf{m}$ is the mutant vector, i_1, i_2 are randomly selected particle indices, F is the mutation scaling factor, $r_j \sim \mathcal{U}(0,1)$, and CR is the crossover rate. The vector $\mathbf{c}$ replaces the global best $\mathbf{g}$ if it has a better fitness score.

4) PSO with Covariance Matrix Adaptation Enhancement (PSO-CMA): The proposed PSO-CMA algorithm integrates the Covariance Matrix Adaptation Evolution Strategy to provide adaptive, second-order refinement of the global best. At each iteration, the current global best $\mathbf{g}$ is used as the initial mean of a local CMA-ES optimizer. A multivariate Gaussian search distribution is initialized:

$$\mathcal{N}(\mathbf{m}_0 = \mathbf{g}, \sigma_0^2 \mathbf{C}_0) \tag{13}$$

where σ_0 denotes the initial step size and $\mathbf{C}_0$ is the identity covariance matrix. CMA-ES is then executed for a fixed number of internal iterations to refine the leader locally.

At each CMA-ES step, a population of candidate solutions $\{\mathbf{c}_k\}$ is sampled and evaluated using the same fitness function $f(\cdot)$. Based on the ranked fitness values, CMA-ES updates its internal parameters (mean, step-size, and covariance matrix), thereby adapting the search direction and scale:

$$\mathbf{m}_{t+1} = \sum_{k=1}^{\mu} w_k \mathbf{c}_{k:\lambda} \tag{14}$$

where $\mathbf{c}_{k:\lambda}$ denotes the k-th best solution among λ samples and w_k are recombination weights. After termination, the best solution $\mathbf{g}_{\text{CMA}}$ found during the CMA-ES procedure is evaluated. If it is fitter than the current PSO leader, the global best $\mathbf{g}$ is updated with $\mathbf{g}_{\text{CMA}}$.

In summary, the aforementioned PSO variants, PSO-GA, PSO-SA, PSO-DE, and PSO-CMA, are used to identify optimal weighting factors to fine-tune the effects of the CNN and audio transformer base models for adaptive ensemble construction. We subsequently evaluate base and ensemble model efficiency using speech/voice diabetes and Parkinson's disease datasets.

IV. Experimental Studies

A. Dataset Details

Diabetes dataset. We collected diabetes speech and voice datasets contributed by 609 speakers in our studies. In each recording session, participants produced three types of audio clips: (i) slow and fast counting from 1 to 20, and (ii) reading the sentence *"Hello, how are you? What is my glucose right now?"*. Each clip has a length of 10–20 seconds. Participants self-reported whether they had been diagnosed with diabetes, which was used as the ground-truth labels. In total, 1646 audio recordings were collected across the three audio types, comprising 1178 positive and 468 negative samples.

Parkinson's dataset. We employed Mobile Device Voice Recordings at King's College London (MDVR-KCL) dataset [29], which consists of 37 smartphone-recorded voice recordings from 21 healthy controls (HC) and 16 Parkinson's disease (PD) participants in a clinical environment. The participants were asked to make a call to complete two tasks: (i) Text Reading: reading out the phonetically balanced text *"The North Wind and the Sun"* and an additional technical text *"Tech. Engin. Computer applications in geography snippet"* depending on participant condition. (ii) Spontaneous Dialogue: engaging in spontaneous conversation with the examiner on neutral topics. All participants were clinically annotated using established PD severity scales. The recording length ranges from about 1.2 to 3.6 minutes. All audio recordings had a 44.1 kHz sampling rate and a 16-bit resolution. Each recording was segmented into non-overlapping 20-second clips to expand the dataset, resulting in 278 text reading recordings and 1730 spontaneous dialogue audio clips.

For both datasets, we conducted a speaker-independent split, where a speaker in the training set will not be included in the test set. An 80-20 split is used for each dataset for model training and testing. Stratified sampling was applied to the dataset to preserve the original class distribution in both subsets. We summarized the dataset distribution in Table I.

TABLE I
The distribution of the datasets.

Aspect	**Diabetes (Ours)**			**Parkinson's (MDVR-KCL)**	
Audio type	Slow Counting	Fast Counting	Speech	Read Text	Spontaneous Dialogue
Train	479	473	361	214	1388
Test	122	120	91	64	342
Total	601	593	452	278	1730

B. Parameter Settings for Base Models

Base classifiers (ResNet, RegNet, MobileViT, Whisper, wav2vec2) use spectrogram or waveform inputs to maximize diversity. Specifically, these base models were trained independently using identical data splits and training protocols. Pre-trained weights released by Hugging Face were used for model initialization. No joint training or co-adaptation between models was performed.

During training, the batch size was set to 16 for all models. Optimization was performed using the Adam optimizer with an initial learning rate of 0.002 and a multi-step learning rate scheduler. Each model was trained for up to 5,000 iterations, with early stopping applied if no improvement in validation performance was observed for 20 consecutive epochs.

C. Parameter Settings for Ensemble Weight Optimization

The proposed PSO-GA, PSO-DE, PSO-SA, and PSO-CMA are used for base classifier weight optimization for ensemble construction. All optimization algorithms were evaluated after softmax projection onto the probability simplex. Specifically, all the search methods were executed with a swarm size of $N = 20$ with 50 iterations. Each search method is performed 10 runs. Moreover, all particle positions were initialized as $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with zero initial velocities. All the search parameters were fixed with an inertia weight, $\omega = 0.5$, and acceleration

TABLE II
RESULTS OF FIVE BASE NETWORKS AND OPTIMIZED WEIGHTED ENSEMBLES.

Dataset		Diabetes (Ours)									Parkinson's (MDVR-KCL)					
		Slow Counting			Fast Counting			Speech			Read Text			Spontaneous Dialogue		
Metrics		Acc	Recall	AUC	Acc	Recall	AUC	Acc	Recall	AUC	Acc	Recall	AUC	Acc	Recall	AUC
Base Model	ResNet	0.680	0.849	0.647	0.667	0.881	0.545	0.714	0.926	0.613	0.859	0.895	0.737	0.927	0.900	0.974
	RegNet	0.689	0.906	0.628	0.691	0.980	0.530	0.703	0.911	0.555	0.765	0.421	0.697	0.818	0.600	0.851
	MobileViT	0.623	0.872	0.523	0.642	0.869	0.527	0.710	0.912	0.664	0.875	0.631	0.803	0.927	0.900	0.982
	Whisper	0.697	0.976	0.656	0.692	0.984	0.580	0.700	0.809	0.609	0.828	0.895	0.894	0.763	0.750	0.778
	wav2vec2	0.663	0.930	0.525	0.667	0.916	0.564	0.725	0.985	0.609	0.844	0.631	0.836	0.836	0.750	0.872
Ensemble	PSO	0.704	0.976	0.653	0.700	0.980	0.576	0.736	0.941	0.617	0.859	0.684	**0.908**	0.945	0.900	0.980
	PSO-GA	0.704	0.980	**0.692**	0.708	**0.985**	0.569	0.747	0.926	0.616	0.875	0.736	0.871	0.963	0.950	**0.981**
	PSO-SA	0.729	0.977	0.673	**0.717**	0.976	**0.626**	**0.780**	0.897	**0.661**	**0.891**	0.737	0.883	**0.981**	**0.950**	0.980
	PSO-DE	**0.738**	0.980	0.665	0.700	0.980	0.561	0.758	**0.985**	0.629	0.843	**0.947**	0.870	0.945	0.900	0.980
	PSO-CMA	0.721	**0.988**	0.668	0.708	0.980	0.564	0.769	0.970	0.592	0.890	0.789	0.907	0.963	0.900	0.978

coefficients, $c_1 = c_2 = 1.2$, across all experiments. For leader-enhanced variants, additional evolutionary refinement steps were applied to the global best particle at each iteration.

In PSO-GA, $K_{\text{GA}} = 5$ elite particles were selected based on personal best fitness, and 10 offspring solutions were generated per iteration using crossover and Gaussian mutation operations with an initial standard deviation of 0.1. In PSO-SA, simulated annealing was applied with an $\tau_0 = 1.0$, $\tau_{min} = 10^{-3}$, and $\alpha = 0.95$, $K_{\text{SA}} = 5$ annealing steps per iteration. In PSO-DE, differential evolution was performed with $F = 0.5$, $\text{CR} = 0.7$, and $K_{\text{DE}} = 10$ refinement trials per iteration. In PSO-CMA, the leader refinement was conducted using CMA-ES initialized at the current global best, with step size $\sigma_0 = 0.5$ and $K_{\text{CMA}} = 10$ internal CMA-ES iterations per PSO iteration. The CMA-ES internal hyperparameters followed the default configuration of the existing implementation [7]. Fitness evaluation was based on negative binary cross-entropy loss computed from the ensemble outputs.

Model performance was evaluated using accuracy, recall, and area under the ROC curve (AUC). Accuracy reflects overall classification performance, recall emphasizes sensitivity to the positive (disease) class, and AUC provides a threshold-independent measure of discriminative ability. These metrics were used for both individual and optimized ensemble model performance monitoring.

V. Evaluation Results

We present the evaluation results for five base networks and ensemble models with the optimized ensemble weights identified by the four proposed PSO variants across two medical audio analysis tasks. Table II shows the detailed mean performance over 10 runs.

A. Main performance

Across all datasets and tasks, ensemble methods consistently outperform or match the strongest individual base model, particularly in terms of accuracy and AUC. This proves that weight optimization via evolutionary search improves robustness over single-model predictions, even when base models already exhibit high recall.

1) Diabetes dataset (Ours): For the slow counting task, individual models achieve AUC scores between 0.523 and 0.656. All ensembles with optimized weights devised by all the search methods improve upon this range, showing a clear gain over the best base model (Whisper, 0.656). As an example, PSO-GA and PSO-SA attain the best AUC scores of 0.692 and 0.673, respectively, indicating robust discriminative capabilities of the resulting ensemble models. The proposed genetic operators and SA operations account for improved optimal weight search behaviors to best exploit complementary local-global feature learning of base networks.

For the fast counting task, ensemble methods also improve model performance. PSO-SA achieves the best overall performance with the highest accuracy and AUC scores of 0.717 and 0.626, respectively. This demonstrates that leader-enhanced strategies can better exploit complementary information under faster speech conditions, taking advantage of variability and diversity of individual models.

For the speech task, ensemble advantages become more pronounced. PSO-SA achieves the highest accuracy (0.780) and AUC (0.661), outperforming all individual models by a clear margin. PSO-DE obtains the highest recall (0.985). These observations indicate that different leader strategies are able to boost distinctive search behaviors with optimized ensemble weights exploiting different ensemble operating characteristics.

2) Parkinson's dataset (MDVR-KCL): On the text reading task, base models show large performance variation, especially in recall. Ensemble approaches with optimized weights substantially stabilize performance. PSO-SA achieves the highest accuracy (0.891), while standard PSO and PSO-CMA obtain the highest AUC scores of 0.908 and 0.907, respectively, demonstrating improved robustness in decision confidence.

On the spontaneous dialogue task, ensemble methods consistently outperform individual models across all metrics. PSO-SA achieves the best accuracy (0.981) and recall (0.950), while PSO-GA achieves the highest AUC (0.981). These results indicate that evolutionary weight optimization is effective for unconstrained, highly variable speech scenarios.

3) Comparison among leader-enhanced strategies: No single leader-enhanced variant dominates all metrics and tasks. PSO-SA shows the most consistent gains in accuracy across datasets, PSO-GA often outputs the best AUC, and PSO-DE tends to focus on recall improvement. This complementary behavior supports the motivation for leader-enhanced evolutionary search: different optimization strategy shows different robustness criteria under different medical audio scenarios.

Overall, the results demonstrate that leader-enhanced PSO-based ensemble optimization strategies improve model performance, especially in accuracy and AUC. The optimized weighting factors are able to identify a reasonable trade-off between local and global context learning in CNN- and transformer-based models for robust medical audio analysis.

B. Ablation Study

This section investigates the effect of different optimization strategies on ensemble weight learning. Experiments are conducted on the MDVR-KCL Spontaneous Dialogue subset.

1) Ablation on Base Optimization Methods: Table III compares different standalone optimization algorithms for ensemble weight search over 10 runs. We also include a uniform averaging ensemble strategy as the baseline, which serves as a lower bound, with all base classifiers contributing equally with 0.2 as weighting factors without any weight optimization applied. In Table III, from left to right, the optimal weights yielded by each search method are for ResNet, RegNet, MobileViT, Whisper, and wav2vec2, respectively.

Among all base optimizers, PSO achieves the highest accuracy (0.945) and a strong AUC (0.980), clearly outperforming GA, DE, and CMA. This indicates that PSO is better suited for continuous, low-dimensional weight optimization problems. SA attains the highest AUC (0.981) with slightly lower accuracy, suggesting improved ranking quality at the expense of decision-level optimality. In contrast, GA and DE exhibit close accuracy and AUC, and converge to weight distributions dominated by a single-based model (RegNet). CMA shows the highest weight variance and collapses the ensemble into an almost single-model solution (MobileViT), indicating that aggressive adaptation without sufficient diversity control can harm ensemble robustness. Overall, these results justify the selection of PSO as the base optimizer in our method.

2) Ablation on Leader-Enhanced Optimization: Table IV evaluates the impact of incorporating leader-enhanced strategies into PSO over 10 runs. All enhanced variants consistently outperform or match the standard PSO baseline, proving that leader enhancement provides effective improvement rather than incidental gains. PSO-GA and PSO-CMA improve accuracy (0.963). PSO-SA achieves the highest accuracy (0.981) with the lowest standard deviation (0.067), indicating that SA promotes stable refinement around high-quality leaders.

In short, the proposed optimized weighted ensemble fusions take advantage of sufficient emphasis of fine-grained local and global pattern representations yielded by distinctive base classifiers and result in robust voice diagnosis of diverse challenging diabetic and Parkinson's disease conditions.

VI. Conclusion and Future Work

This paper presents a leader-enhanced evolutionary search framework for robust ensemble construction in medical audio analysis. By formulating ensemble weight learning as a constrained swarm-based optimization problem with noisy objectives, the proposed method addresses the uncertainty inherent in clinical audio data. Four PSO variant models, i.e., PSO-SA, PSO-DE, PSO-SA, and PSO-CMA, are proposed to fine-tune effects of the five base CNN and transformer audio classifiers (ResNet, RegNet, MobileViT, Whisper, and wav2vec2) to formulate ensemble prediction. The empirical results indicate that the optimized ensemble weights identified by all the proposed algorithms are able to fine-tune effects of complementary local and global feature learning characteristics of the base classifiers. The resulting optimized fusion strategies are able to strike better classification boundaries for fuzzy ambiguous audio disease diagnosis, outperforming individual base and conventional average ensemble baselines.

Ablation studies further prove the importance of both the base optimizer and leader-enhancement mechanisms in achieving robust ensemble behavior. Overall, our experiments demonstrate that evolutionary ensemble optimization is a practical and effective solution for medical audio prediction.

Future work will explore adaptive ensemble weighting at finer granularity, uncertainty-aware optimization objectives aligned with clinical risk, validation on multi-center datasets, and integration with self-supervised or diffusion audio models to enhance robustness and generalizability.

References

[1] L. Zhang, C. P. Lim, Y. Yu, and M. Jiang, "Sound classification using evolving ensemble models and Particle Swarm Optimization," *Appl. Soft Comput.*, vol. 116, Feb. 2022.

[2] J. M. Kaufman, A. Thommandram, and Y. Fossat, "Acoustic Analysis and Prediction of Type 2 Diabetes Mellitus Using Smartphone-Recorded Voice Segments," *Mayo Clinic Proceedings: Digital Health*, vol. 1, pp. 534–544, Dec. 2023.

[3] A. Tsanas, M. Little, P. McSharry, and L. Ramig, "Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 884–893, Apr. 2010.

[4] M. Khedimi, T. Zhang, C. Dehmani, X. Zhao, and Y. Geng, "A Unified Deep Learning Ensemble Framework for Voice-Based Parkinson's Disease Detection and Motor Severity Prediction," *Bioengineering*, vol. 12, p. 699, June 2025.

[5] M. G. Di Cesare, D. Perpetuini, D. Cardone, and A. Merla, "Machine Learning-Assisted Speech Analysis for Early Detection of Parkinson's Disease: A Study on Speaker Diarization and Classification Techniques," *Sensors*, vol. 24, p. 1499, Jan. 2024.

[6] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, Oct. 2022. arXiv:2104.02395 [cs].

[7] N. Hansen and A. Ostermeier, "Completely Derandomized Self-Adaptation in Evolution Strategies," *Evolutionary Computation*, vol. 9, pp. 159–195, June 2001.

TABLE III
ABLATION ON THE BASE OPTIMIZATION METHODS USING MDVR-KCL (AVERAGE BASELINE REFERS TO THE TRADITIONAL MEAN AVERAGE ENSEMBLE)

	Acc	Recall	AUC	ensemble weights	ensemble weights std.
Average (baseline)	0.836	0.850	0.931	[0.200, 0.200, 0.200, 0.200, 0.200]	0
PSO	**0.945**	**0.900**	0.980	[0.031, 0.020, 0.082, 0.365, 0.503]	0.197
GA	0.872	0.900	0.940	[0.186, 0.631, 0.138, 0.015, 0.031]	0.224
SA	0.927	0.900	**0.981**	[0.206, 0.442, 0.075, 0.001, 0.276]	**0.154**
DE	0.865	0.900	0.941	[0.187, 0.688, 0.029, 0.012, 0.084]	0.252
CMA	0.854	0.900	0.944	[0.004, 0.001, 0.992, 0.002, 0.001]	0.396

TABLE IV
ABLATION ON THE ENHANCED OPTIMIZATION METHODS USING MDVR-KCL

	Acc	Recall	AUC	ensemble weights	ensemble weights std.
PSO	0.945	0.900	0.980	[0.031, 0.02, 0.082, 0.365, 0.503]	0.197
PSO-GA	0.963	0.950	**0.981**	[0.249, 0.414, 0.048, 0.030, 0.258]	0.144
PSO-SA	**0.981**	**0.950**	0.980	[0.219, 0.282, 0.237, 0.082, 0.180]	**0.067**
PSO-DE	0.945	0.900	0.980	[0.229, 0.445, 0.067, 0.069, 0.190]	0.138
PSO-CMA	0.963	0.900	0.978	[0.189, 0.462, 0.038, 0.190, 0.121]	0.142

[8] C. Wall, L. Zhang, Y. Yu, A. Kumar, and R. Gao, "A Deep Ensemble Neural Network with Attention Mechanisms for Lung Abnormality Classification Using Audio Inputs," *Sensors*, vol. 22, p. 5566, July 2022.

[9] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, and T. Niesler, "Automatic cough classification for tuberculosis screening in a real-world environment," *Physiological Measurement*, vol. 42, p. 105014, Oct. 2021.

[10] S. R. Chetupalli, P. Krishnan, N. Sharma, A. Muguli, R. Kumar, V. Nanda, L. M. Pinto, P. K. Ghosh, and S. Ganapathy, "Multi-Modal Point-of-Care Diagnostics for COVID-19 Based on Acoustics and Symptoms," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 199–210, 2023.

[11] J. Shen, X. Zhang, Y. Lu, P. Ye, P. Zhang, and Y. Yan, "Novel audio characteristic-dependent feature extraction and data augmentation methods for cough-based respiratory disease classification," *Computers in Biology and Medicine*, vol. 179, p. 108843, Sept. 2024.

[12] R. Ravi and D. R. Gunjawate, "Effect of diabetes mellitus on voice: a systematic review," *Practical Diabetes*, vol. 36, pp. 177–180, Sept. 2019.

[13] A. Gangani, L. Zhang, and A. Panesar, "Deep Fusion Networks with Hybrid Attention Mechanisms for Voice Diabetes Detection," in *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 5491–5498, IEEE, Oct. 2025.

[14] J. Guo, W. Peng, S. Hu, D. Lu, and S. Chen, "A Novel Machine Learning-Driven Voice and Clinical Biomarkers Framework for Robust Prediction of Type 2 Diabetes Mellitus," *Journal of Voice*, p. S089219972500400X, Oct. 2025.

[15] D. Chitkara and R. Sharma, "Voice based detection of type 2 diabetes mellitus," in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 83–87, Feb. 2016.

[16] J. Kaufman, J. Jeon, J. Oreskovic, and Y. Fossat, "Linear effects of glucose levels on voice fundamental frequency in type 2 diabetes and individuals with normoglycemia," *Scientific Reports*, vol. 14, p. 19012, Aug. 2024.

[17] F. Abedinzadeh Torghabeh, S. A. Hosseini, and E. Ahmadi Moghadam, "Enhancing Parkinson's disease severity assessment through voice-based wavelet scattering, optimized model selection, and weighted majority voting," *Medicine in Novel Technology and Devices*, vol. 20, p. 100266, Dec. 2023.

[18] M. K. Reddy and P. Alku, "Exemplar-Based Sparse Representations for Detection of Parkinson's Disease From Speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1386–1396, 2023.

[19] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," *Healthcare*, vol. 11, p. 1808, June 2023.

[20] K. Kondrotas, L. Zhang, C. P. Lim, H. Asadi, and Y. Yu, "Audio-Visual Emotion Classification Using Reinforcement Learning-Enhanced Particle Swarm Optimisation," *IEEE Transactions on Affective Computing*, vol. 16, pp. 3434–3451, Oct. 2025.

[21] L. Zhang, D. Zhao, C. P. Lim, H. Asadi, H. Huang, Y. Yu, and R. Gao, "Video Deepfake classification using particle swarm optimization-based evolving ensemble models," *Knowledge-Based Systems*, vol. 289, p. 111461, Apr. 2024.

[22] S. Slade, L. Zhang, H. Huang, H. Asadi, C. P. Lim, Y. Yu, D. Zhao, H. Lin, and R. Gao, "Neural Inference Search for Multiloss Segmentation Models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 15113–15127, Nov. 2024.

[23] H. Ulutas, R. B. Günay, and M. E. Sahin, "Detecting diabetes in an ensemble model using a unique PSO-GWO hybrid approach to hyperparameter optimization," *Neural Computing and Applications*, vol. 36, pp. 18313–18341, Oct. 2024.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

[25] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing Network Design Spaces," Mar. 2020. arXiv:2003.13678 [cs].

[26] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," Mar. 2022. arXiv:2110.02178 [cs].

[27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022. arXiv:2212.04356 [eess].

[28] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 2020. arXiv:2006.11477 [cs].

[29] H. Jaeger, Dhaval Trivedi, and M. Stadtschnitzer, "Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls," May 2019.