RESEARCH PAPER

A Nested Cross-Validated Evaluation of A Hybrid Arithmetic Optimization Algorithm on Clinical, MRI and Speech Data for Xgboost-Based Alzheimer's Disease Prediction

Manisha M. Chaudhari[1*], Mukta Agrawal[1]

[1]Department of Computer Science, Indus University, Ahmedabad, Gujarat, India.

ABSTRACT

Metaheuristic hyperparameter tuning is popular in machine-learning studies of Alzheimer's disease (AD), yet new optimizers are rarely compared with established tuners under an evaluation that keeps tuning and testing apart. This study asks whether a hybrid arithmetic optimization algorithm (HybAOA), which switches between arithmetic-optimization moves, sine–cosine moves and center-based opposition, improves XGBoost for AD prediction. HybAOA was compared with default settings, grid search, random search, the tree-structured Parzen estimator (TPE) and five other metaheuristics, each given 110 evaluations of a six-dimensional search space, under nested cross-validation on clinical records of 2,149 patients, on 4,556 MRI slices from 200 subjects (split by subject) and on picture-description recordings of 79 AD and 79 healthy speakers (eGeMAPS acoustic features). On the clinical data all tuners reached a test AUC of 0.948–0.951 (HybAOA 0.950 ± 0.009, balanced accuracy 0.942), as did untuned XGBoost (0.950), whose log-loss, however, was higher (0.256 versus 0.200). On MRI, tuning raised the slice-level AUC from 0.726–0.749 (defaults) to 0.770–0.776 (HybAOA 0.770; subject-level 0.862). On speech, the search methods reached AUCs of 0.665–0.737 (HybAOA 0.702) and untuned XGBoost 0.757, but with the worst log-loss (0.726 versus 0.654). No difference between HybAOA and any other search method was significant (Wilcoxon–Holm and corrected resampled t-tests). On shifted benchmark functions HybAOA ranked sixth of eight optimizers and below random search, and a structural-bias probe showed that its arithmetic moves are attracted to the bounds of the search space, a bias that the opposition step reduces at a cost in convergence. In control experiments, pairing clinical records with MRI slices of the same diagnosis from unrelated subjects raised the AUC of uninformative clinical features from 0.497 to 0.907, and removing the interviewer's speech left a signature that a classifier detected (AUC 0.755). For XGBoost on these data, tuning mainly improves the probabilities and the choice of optimizer does not matter; credible multimodal and speech studies need every modality from the same participants and identically processed recordings.

Keywords: Alzheimer's disease, XGBoost, Hyperparameter optimization, Arithmetic optimization algorithm, Metaheuristics, Nested cross-validation, Speech analysis, Structural bias, Data leakage.

How to cite this article: Chaudhari MM, Agrawal M, A Nested Cross-Validated Evaluation of A Hybrid Arithmetic Optimization Algorithm on Clinical, MRI and Speech Data for Xgboost-Based Alzheimer's Disease Prediction. Int J Drug Deliv Technol. 2026;16(90s): 944-980; DOI: 10.25258/ijddt.16.90s.94

ORCID IDs: Manisha M. Chaudhari: [0000-0002-9673-736X], Mukta Agrawal: [0009-0008-1865-6127]

Source of support: Nil

Conflict of interest: None

Data Availability Statement

The clinical dataset [54] and the MRI image collection [55] analysed in this study are publicly available from the cited sources. The speech recordings belong to the ADReSSo 2021 challenge release of the Pitt corpus [33,38] and are available to DementiaBank members through TalkBank.

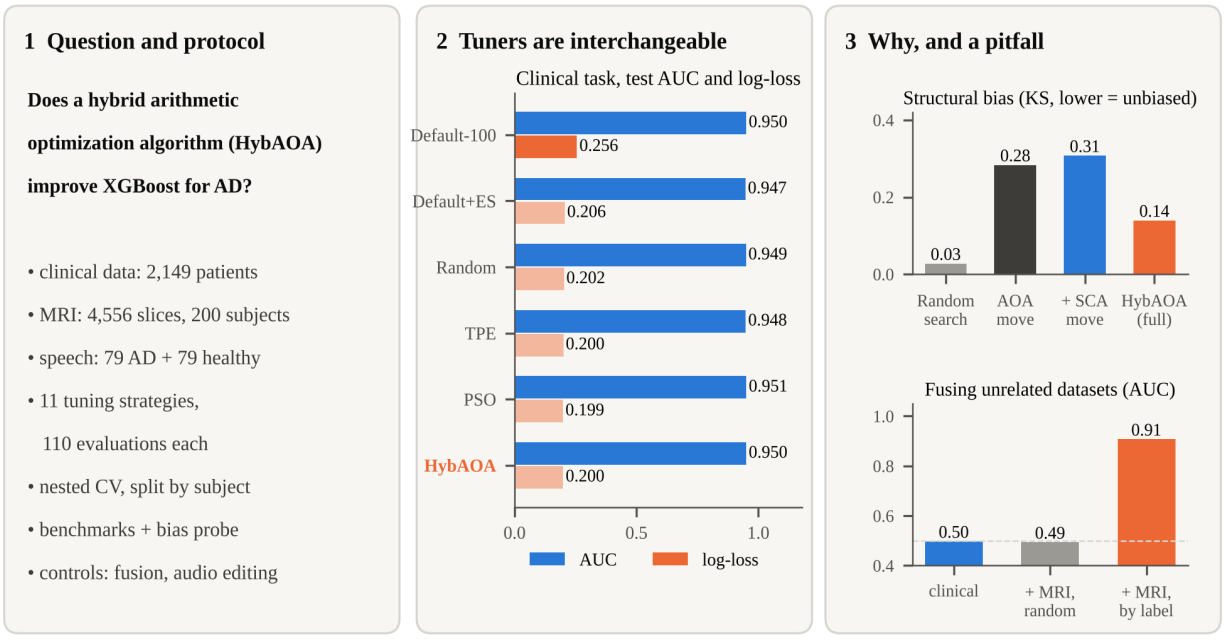
Highlights

- A hybrid arithmetic optimization algorithm (HybAOA) for XGBoost tuning was compared with nine tuning strategies under nested cross-validation.
- HybAOA-tuned XGBoost (AUC 0.950 clinical, 0.770 MRI, 0.702 speech) was statistically indistinguishable from grid, random, Bayesian and other metaheuristic tuners.

Corresponding Author: Manisha M. Chaudhari, Research Scholar, Department of Computer Science, Indus University, Ahmedabad, Gujarat, India. ORCID: [0000-0002-9673-736X] chaudharimanisha.21.rs@indusuni.ac.in

- Tuning mainly improved probability quality; on speech, untuned XGBoost discriminated as well as any tuned model.
- On shifted benchmark functions HybAOA ranked below random search, and its arithmetic moves showed a structural bias towards the search-space bounds.
- Label-based pairing of unrelated datasets inflated AUC to 0.907, and class-specific audio editing was itself detectable.

Graphical Abstract:



1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia, a syndrome that affected an estimated 57 million people worldwide in 2019 and is projected to affect about 153 million by 2050 [1]. In the United States alone, about 6.9 million people aged 65 years or older were living with Alzheimer's dementia in 2024 [2]. Diagnosis rests on clinical examination, cognitive screening instruments such as the Mini-Mental State Examination (MMSE) [3], functional assessment, neuroimaging and, increasingly, fluid biomarkers, combined according to consensus criteria [4]. Because each of these steps needs specialist time, machine-learning (ML) models that estimate the probability of AD from routinely collected data have attracted sustained interest, and several reviews document many such models built from structured clinical records, magnetic resonance imaging (MRI), speech and their combinations [5–7]. For tabular clinical data, gradient-boosted decision trees remain among the strongest and most practical model families [8–11]. XGBoost [9] in particular combines second-order boosting with shrinkage, row and column subsampling and explicit L2 regularization, and it is widely used in AD prediction studies [12]. Its accuracy depends on a handful of hyperparameters (learning rate, tree depth, subsampling ratios, minimum child weight and regularization strength) whose influence varies between datasets [13]. Hyperparameter optimization (HPO) is therefore a routine part of model development, and the methods available range from grid and random search [14] to sequential model-based (Bayesian) optimization such as the tree-structured Parzen estimator (TPE) [15,16], and to population-based metaheuristics [17].

Metaheuristics are especially popular in applied AD studies. Recent examples tune support vector machines with modified grey wolf optimizers [18], select features with parrot, grey wolf and differential-evolution optimizers [19], or combine oversampling with hyperparameter tuning of XGBoost [12]. New optimizers, or hybrids of existing ones, are typically introduced with a single headline accuracy and compared, if at all, with one or two other metaheuristics. The optimization community has

pointed out three reasons for caution. First, no optimizer can be best on all problems [20], so a claim of superiority needs evidence on the problem at hand. Second, many recent algorithms re-describe known search operators under new metaphors [21,22]. Third, population-based algorithms can carry a structural bias, a tendency to return particular regions of the search space regardless of the objective [23], which can masquerade as good performance when the optimum happens to lie in that region.

How the tuned model is evaluated matters at least as much as how it is tuned. When hyperparameters are selected on the same folds that are used to report performance, the estimate is optimistically biased [24,25]. Data leakage, in which information about the test samples reaches the training process, is a recurring cause of irreproducible ML results [26,27]; in AD imaging, splitting two-dimensional (2-D) MRI slices rather than subjects between training and test sets is a well-documented example [28,29]. A further pitfall arises in multimodal studies that assemble clinical, imaging and speech data from different public sources. When no participant appears in more than one source, records can only be combined by matching diagnostic labels [30], and the auxiliary modality then carries the label by construction rather than through biology.

This work started as a multimodal extension of the authors' earlier studies on clinical, image, voice and transcript data [31,32], in which clinical, MRI and speech features were to be concatenated and classified by XGBoost tuned with a hybrid arithmetic optimization algorithm (HybAOA). Auditing the three public datasets showed that they share no participants (none of the 2,149 clinical patient identifiers corresponds to an image or audio file), so early fusion was abandoned. The audit also showed that the first copy of the speech data could not be used, because its folders labelled AD and control contained byte-identical recordings, and that a second copy paired AD recordings from which the interviewer's speech had been removed with unedited control recordings. The speech analysis therefore used the original, identically processed interviews of 79 AD and 79 healthy speakers from the ADReSSo corpus [33]. The study was redesigned around a question that the available data can answer: does tuning XGBoost with HybAOA improve AD prediction from clinical, imaging and speech data, analysed separately, compared with default settings and with established tuners when every method is evaluated under the same nested, leakage-controlled protocol?

The main contributions are as follows.

- A complete, reproducible specification of HybAOA for XGBoost tuning, including its operator-switching rule, and a comparison with nine other tuning strategies (default settings, grid search, random search, TPE, particle swarm optimization, differential evolution, the grey wolf optimizer, the sine cosine algorithm and the original arithmetic optimization algorithm) at an equal budget of 110 objective evaluations.
- A nested cross-validation (CV) protocol in which tuning uses only the outer training part, MRI slices are split by recovered subject, and every data-dependent transform is fitted inside the fold that uses it, applied to four tasks: clinical data with all 32 predictors, clinical data without the MMSE score, MRI slices represented by ResNet-18 embeddings, and picture-description recordings represented by eGeMAPS acoustic features.
- An analysis of optimizer behavior on seven shifted benchmark functions, a structural-bias probe and an ablation of the three HybAOA operators, which explain the results obtained on the clinical, MRI and speech tasks.
- Two control experiments on how data assembly can create or distort accuracy: one quantifies the apparent accuracy created when unrelated clinical and imaging datasets are fused by matching diagnostic labels, and the other shows that removing the interviewer's speech leaves an acoustic signature that a classifier detects, so recordings of the two classes must be processed identically.
- Practical recommendations for tuning and evaluating XGBoost-based AD classifiers, together with the limitations of the public datasets that are widely used for this purpose.

Section 2 reviews related work. Section 3 describes the data, the optimizers, the evaluation protocol and the auxiliary experiments. Section 4 reports the results, Section 5 discusses them, and Section 6 concludes.

2. Related work

2.1 Machine learning for Alzheimer's disease prediction

ML models for AD have been built from structured clinical and demographic variables, structural and functional neuroimaging, genetics, speech and language, and combinations of these sources [5–7]. Structural MRI is the most studied modality; reviews of convolutional neural networks (CNNs) for AD classification report a wide range of accuracies and

show that a substantial fraction of published results are affected by data leakage, most often because slices or scans from the same subject appear on both sides of the train–test split [28]. When the split is made at the subject level, accuracies of 2-D slice classifiers fall markedly [29]. Clinical and cognitive variables, on the other hand, are low-dimensional and well suited to tree ensembles, and cognitive and functional test scores usually dominate their predictions. Because instruments such as the MMSE are part of the clinical work-up that produces the diagnosis [3,4], models that use them partly reproduce the labelling process; this is acceptable for decision support but has to be kept in mind when interpreting accuracy.

Multimodal models aim to exploit complementary information, for example from imaging, clinical records and genetics [34,35]. Such studies draw on cohorts in which every participant contributes every modality, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [36] or the Open Access Series of Imaging Studies (OASIS) [37]. A different line of work combines modalities from heterogeneous public datasets that were collected from different people; one recent study fused MRI, PET, clinical and speech-derived features from such sources with a label-consistent feature-level fusion strategy [30]. The authors' own earlier studies used image and text data [31] and voice and transcript data [32] for AD prediction.

Speech is attractive because it can be recorded cheaply and without discomfort. Most studies use the Cookie Theft picture descriptions of the Pitt corpus in DementiaBank [38], in which linguistic features of the transcripts identify AD well [39]. The ADReSS and ADReSSo challenges released subsets of this corpus with fixed training and test partitions, the first balanced for age and gender [33,40]. Acoustic features alone are weaker than linguistic ones: the acoustic baseline of ADReSS reached a test accuracy of 62.5 % [40]; on the ADReSSo training set, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [41] gave a leave-one-subject-out accuracy of 74.1 % and an audio-only ensemble reached 67.6 % on the test set [42], whereas the challenge baseline that combined acoustic and linguistic features reached 78.9 % [33]. Because recording conditions can differ between groups, acoustic classifiers are also exposed to shortcut learning, in which a model separates the classes by how the recordings were made rather than by how the speakers talk [43].

2.2 Hyperparameter optimization

HPO treats the validation loss of a learning algorithm as a black-box function of its hyperparameters. Grid search evaluates a fixed lattice and scales poorly with dimension; random search is a strong baseline because only a few hyperparameters usually matter [14]. Sequential model-based methods fit a surrogate of the loss and choose the next configuration by an acquisition rule; TPE models the densities of good and poor configurations separately and is the default sampler of the Optuna framework [15,16]. Best practice recommends nested resampling for honest performance estimates, early stopping for iterative learners, and comparisons against random search at equal budget [17].

Population-based metaheuristics are an alternative that needs no surrogate. Particle swarm optimization (PSO) moves particles towards their own and the swarm's best positions [44,45]; differential evolution (DE) perturbs candidates with scaled differences of other candidates [46]; the grey wolf optimizer (GWO) moves candidates towards the three best solutions [47]; the sine cosine algorithm (SCA) oscillates candidates around the best solution with sine and cosine factors [48]; and the arithmetic optimization algorithm (AOA) generates candidates from the best solution using multiplication and division for exploration and addition and subtraction for exploitation, controlled by two time-varying schedules [49]. Opposition-based learning evaluates the 'opposite' of a candidate to accelerate search [50,51], and center-based sampling exploits the fact that points near the center of the search space are, on average, closer to an unknown optimum [52]. Hybrid algorithms combine such operators; HybAOA, studied here, combines the AOA move, an SCA move and a center-based random opposition step.

Three critiques of this literature motivate the design of the present study. The no-free-lunch theorems imply that superiority must be shown on the problem of interest rather than assumed [20]. Several popular algorithms have been shown to be re-descriptions of earlier ideas [21,22]. Finally, structural bias, measured by running an algorithm on an objective that carries no information and testing whether its final positions are uniformly distributed, reveals preferences for the center or the bounds of the search space [23]. An algorithm with a boundary preference will look excellent on problems whose optimum lies on a bound and poor elsewhere.

2.3 Metaheuristic-tuned models in Alzheimer's disease studies

Table 1 summarizes representative studies that tuned AD classifiers with metaheuristics or fused heterogeneous sources, together with the authors'

earlier work. Two patterns stand out. The reported accuracies are high (often above 90 %), and the validation scheme is frequently not reported in the abstract or is a single train–test split, so the contribution of the optimizer cannot be separated from that of the split. Moreover, optimizers are rarely compared with random search or a Bayesian tuner at equal budget, which is the comparison that would show whether the added complexity pays off [17].

2.4 Research gap and questions

No study, to our knowledge, has evaluated an AOA-family hybrid for tuning an AD classifier against random search, a Bayesian tuner and standard metaheuristics at equal budget under nested CV, nor examined whether such an optimizer has a structural bias that could explain its behavior. The present study addresses four research questions:

1. RQ1: Does HybAOA tuning improve the test performance of XGBoost relative to default settings on clinical, MRI and speech data?
2. RQ2: Is HybAOA better than established tuners (grid, random, TPE, PSO, DE, GWO, SCA, AOA) at the same evaluation budget?
3. RQ3: How do HybAOA's three operators contribute, and does the algorithm have a structural bias?
4. RQ4: How do two data-assembly errors, label-consistent fusion of unrelated datasets and class-specific editing of speech recordings, affect the measured accuracy?

3. Materials and methods

3.1 Study design

Figure 1 summarizes the design. Three public data sources were audited. Because they share no participants, the clinical, MRI and speech data were analysed as separate prediction tasks, and the speech task used the copy of the recordings in which both classes had been processed identically. For each task, thirteen tuning strategies (ten main strategies and three HybAOA ablations) were compared under nested CV, with identical folds, budgets and random seeds for every strategy. Three auxiliary experiments, on benchmark functions, on a negative control for data fusion and on a control for speech preprocessing, explain and qualify the main results. Reporting follows the TRIPOD+AI recommendations where applicable [53].

3.2 Datasets

3.2.1 Clinical dataset (D1)

The clinical dataset [54] contains one record for each of 2,149 patients with a binary diagnosis (AD: 760, 35.4 %; no AD: 1,389, 64.6 %) and 32 predictors: demographic variables (age, gender, ethnicity, education level), lifestyle variables (body-mass index, smoking, alcohol consumption, physical activity, diet quality, sleep quality), medical history (family history of AD, cardiovascular disease, diabetes, depression, head injury, hypertension), clinical measurements (systolic and diastolic blood pressure, total, LDL and HDL cholesterol, triglycerides), cognitive and functional assessments (MMSE, functional assessment, memory complaints, behavioral problems, activities of daily living [ADL]) and symptoms (confusion, disorientation, personality changes, difficulty completing tasks, forgetfulness). A patient identifier and a constant administrative field were removed. There are no missing values and no duplicate records.

Figure 2 shows the class balance and the discriminative power of single predictors. Only five predictors, all cognitive or functional assessments, reach a single-variable area under the ROC curve (AUC) above 0.58; functional assessment (0.72) and ADL (0.70) are the strongest, followed by MMSE (0.64) and memory complaints (0.63). The remaining 27 predictors are close to chance level. The dataset also shows signatures of simulation: 13 of its 15 continuous variables are compatible with uniform distributions, diastolic pressure equals or exceeds systolic pressure in 186 records, and LDL plus HDL cholesterol exceeds total cholesterol in 578 records. The data were analysed as provided, and these properties are taken into account when the results are interpreted.

Two clinical tasks were defined. D1-all uses all 32 predictors and is the primary clinical task, comparable with published work on the same dataset. D1-noMMSE removes the MMSE score, which is part of the diagnostic process [3,4], to check that the conclusions do not depend on this label-related predictor.

3.2.2 MRI dataset (D2)

The MRI collection [55] is a widely redistributed set of axial 2-D brain MRI slices (176 × 208 pixels, grayscale) in four folders: non-demented, very mild, mild and moderate dementia. As received it held 7,840 files but only 4,556 distinct images: 3,277 image hashes occurred in two class folders (2,560 shared between the non-demented and very-mild folders, 717 between the mild and moderate folders). Each image was kept once, in the folder indicated by its original filename series. The collection has no subject identifiers, yet consecutive slices of one brain

are highly correlated. Pseudo-subjects were therefore recovered by linking slices whose downsampled images had a Pearson correlation of at least 0.97 and taking connected components; for the part of the collection whose filenames encode the subject, this procedure reproduced the 200 known subjects exactly, with no group mixing subjects or classes. The final set contains 4,556 slices from 200 pseudo-subjects (non-demented: 3,200 slices, 100 subjects; very mild: 448, 70; mild: 896, 28; moderate: 12, 2). Because the moderate class has only two subjects, the MRI task (D2-MRI) was defined as binary: non-demented (3,200 slices, 100 subjects) against demented (1,356 slices, 100 subjects). **Figure 3** shows example slices and the class composition.

3.2.3 Speech dataset (D3)

The speech recordings come from the ADReSSo 2021 challenge [33], which re-released Cookie Theft picture descriptions from the Pitt corpus of DementiaBank [38]: each participant describes a line drawing to an interviewer, and the recording contains both voices. Three copies were received, and auditing them decided which could be used. The first held 90 WAV files in two folders labelled AD and control (45 each); every file in the AD folder was byte-identical to one in the control folder, and all 45 belonged to speakers of the healthy group, so the labels carried no information. The second paired 79 AD recordings from which the interviewer's turns had been removed (participant-only versions) with 79 unedited healthy recordings. Because one class had been edited and the other had not, a classifier could separate the classes partly by the editing itself (Section 3.13). The third copy supplied the unedited original interviews of the same 79 AD speakers. The final dataset therefore consists of the original full interviews of 79 AD and 79 healthy speakers, one recording per speaker and processed identically. They correspond to all 79 healthy and 79 of the 87 AD recordings of the ADReSSo diagnosis training set; the challenge test set was not available.

The originals were WAV files sampled at 44.1 kHz (AD: 63 mono 16-bit, 7 stereo 16-bit and 9 stereo 32-bit floating-point files; healthy: 64 mono 16-bit, 14 stereo 16-bit and 1 mono floating-point file). All were converted with the same FFmpeg command to 16 kHz, one channel and 16-bit FLAC. After conversion the two groups had similar levels (mean RMS level −25.4 dBFS for AD and −24.9 dBFS for healthy speakers) but different durations (median 73.4 s and 67.8 s; mean 84.9 s and 68.8 s). The speech task (D3-speech) is binary: AD (79) against healthy (79). **Figure 1** shows where the speech data enter the design, and **Table 2** summarizes the three sources.

3.3 Why the datasets were not fused

Early fusion concatenates, for each participant, the feature vectors of all modalities. It requires that the rows of the clinical table, the MRI slices and the recordings refer to the same people. Matching the 2,149 clinical patient identifiers against the names of all image and audio files found no match, and the MRI collection and speech corpus come from different sources than the clinical table. A pipeline that looks up each patient's image and recording by identifier therefore receives no data for these modalities and, if it substitutes zero vectors for missing files, silently reduces to a clinical-only model. The only alternative way to align unrelated records is to pair them by diagnosis, which injects the label into the auxiliary features. Section 3.12 describes an experiment that measures the size of this effect.

3.4 Feature representations

For D1 the 32 (or 31) predictors were used as provided. Tree ensembles are invariant to monotone transformations of individual predictors, so no scaling or encoding beyond the existing integer codes was applied, and no imputation was needed. For D2 each slice was converted to three identical channels, resized to 224 × 224 pixels, normalized with the ImageNet channel statistics and passed through a ResNet-18 network [56] pretrained on ImageNet-1k [57] (weights from the 'ResNet strikes back' A1 training recipe [58]). The 512-dimensional output of the global average-pooling layer was used as the slice embedding. The network was not fine-tuned, so the embeddings do not depend on any label. Principal component analysis (PCA) reduced the embeddings to 32 components, which retain about 80 % of the variance; the PCA was fitted on the training part of every fold in which it was used, never on validation or test slices. For D3, the 88 functionals of eGeMAPS (version 02) [41] were computed for each whole recording with openSMILE [59]. They summarize frequency-related parameters (pitch, jitter, formant frequencies and bandwidths), energy and amplitude parameters (loudness, shimmer, harmonics-to-noise ratio), spectral balance (spectral slopes, alpha ratio, Hammarberg index, relative formant energies, MFCC 1–4, spectral flux) and temporal parameters (rate of loudness peaks, number and length of voiced and unvoiced segments), each as a mean, a coefficient of variation or percentiles over the recording. The feature definitions are fixed and use no labels, so they were computed once for all recordings; no transcripts or linguistic features were used.

3.5 XGBoost and the search space

All models were binary XGBoost classifiers (logistic objective, histogram tree method with 128 bins, one thread per job). Six hyperparameters were tuned (**Table 3**): the learning rate η, the maximum tree depth, the row subsampling ratio, the column subsampling ratio per tree, the minimum child weight and the L2 regularization weight λ. Each optimizer searched the unit hypercube [0,1]6, and a candidate x was decoded into hyperparameters as follows:

(1)

$$\eta = 0.02 \cdot 15^{x_1}, \quad d = \mathrm{round}(2 + 8x_2), \quad s_r = 0.5 + 0.5x_3,$$

$$s_c = 0.3 + 0.7x_4, \quad w_{\min} = 20^{x_5}, \quad \lambda = 10^{4x_6 - 3}$$

so that the learning rate, minimum child weight and λ are searched on logarithmic scales. The number of boosting rounds was not a search dimension; it was set by early stopping. For a candidate x, the objective was the mean validation log-loss over K = 3 inner folds,

(2)

$$f(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_k(\mathbf{x}), \qquad \mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i)\ln(1 - p_i)]$$

where each inner model was trained for at most 500 rounds with early stopping after 25 rounds without improvement of the validation log-loss, and pi is the predicted probability of AD. The rounded mean of the three early-stopping rounds was stored with the objective value and used when the selected configuration was refitted. Log-loss was chosen as the objective because it is a proper scoring rule that rewards both discrimination and calibration and, unlike accuracy, does not depend on a decision threshold.

3.6 The hybrid arithmetic optimization algorithm

HybAOA maintains a population of N candidate vectors in [0,1]D and the best vector found so far, b. Each iteration t = 0, …, T − 1 updates three control quantities, the math optimizer accelerated (MOA) function, the math optimizer probability (MOP) and a small step size ε:

(3)

$$\mathrm{MOA}(t) = 0.2 + 0.7\,\frac{t}{T}, \qquad \mathrm{MOP}(t) = 1 - \left(\frac{t}{T}\right)^{1/\alpha}, \qquad \varepsilon(t) = \frac{1}{T + t + 1}$$

with α = 5. MOA grows from 0.2 to 0.9 and therefore shifts the arithmetic move from exploration to exploitation, while MOP decays from 1 towards 0 and controls the size of the exploratory steps. Each candidate i has a stagnation counter τi, which is reset to zero when the candidate improves and incremented otherwise. The operator applied to candidate i depends on τi and on the limit L = max(1, ⌊T/2N⌋):

(4)

$$\mathrm{operator}_i = \begin{cases} \text{AOA move} & (\tau_i < L), \\ \text{SCA move} & (L \le \tau_i < 2L), \\ \text{opposition} & (\tau_i \ge 2L) \end{cases}$$

The AOA move generates every coordinate j of the new candidate from the corresponding coordinate bj of the best vector. With r1 ~ U(0,1), r2 ~ U(0,2π), r3 ~ U(0,2) and μ = 0.5,

(5)

$$x_j' = b_j \, \mu r_3 / (\mathrm{MOP} + \varepsilon) \quad \text{if } r_1 > \mathrm{MOA},\ r_2 > 0.5 \ \text{(division)}$$

$$x_j' = b_j \, \mathrm{MOP} \, \mu r_3 \quad \text{if } r_1 > \mathrm{MOA},\ r_2 \le 0.5 \ \text{(multiplication)}$$

$$x_j' = b_j + \varepsilon \, \mu r_3 \quad \text{if } r_1 \le \mathrm{MOA},\ r_2 > 0.5 \ \text{(addition)}$$

$$x_j' = b_j \, \mathrm{MOP} - \varepsilon \, \mu r_3 \quad \text{otherwise (subtraction)}$$

Because r2 is drawn on [0, 2π] but compared with 0.5, the division and addition branches are selected with probability 1 − 0.5/(2π) ≈ 0.92; this detail of the original implementation was retained so that the algorithm is reproduced exactly. Note that the multiplicative branches scale bj towards zero and the division branch can push it beyond 1, after which the coordinate is clipped to the bound; this property is examined in Section 4.2. The SCA move perturbs the current candidate xi around b using a random amplitude r1 ~ U(0,1) drawn once per move and, for every coordinate, r2 ~ U(0,2π), r3 ~ U(0,2) and r4 ~ U(0,1):

(6)

$$x_j' = x_{ij} + 2 r_1 \, \phi(r_2) \, | \, r_3 b_j - x_{ij} |, \qquad \phi = \sin \ \text{if } r_4 < 0.5, \ \text{otherwise } \cos$$

The center-based random opposition step reflects a stagnating candidate through the center c of the search box with a random stretch ρ, which restarts it in the opposite region [50,52]:

(7)

$$\mathbf{x}' = \mathbf{c} + \rho \, (\mathbf{c} - \mathbf{x}_i), \qquad \mathbf{c} = (\mathbf{l} + \mathbf{u})/2, \qquad \rho \sim U(0.8,\ 1.2)$$

Every new candidate is clipped to the bounds l = 0 and u = 1 and evaluated, and it replaces xi only if it is better (greedy selection). The global best b is refreshed once per iteration, after all candidates have been processed. **Figure 4** shows the flow of the algorithm, and the steps are:

5. Draw N candidates uniformly in [0,1]D, evaluate them, set b to the best and all counters τi to zero.
6. For t = 0, …, T − 1, compute MOA, MOP and ε from Eq. (3).
7. For each candidate i, choose the operator from τi with Eq. (4) and generate x' with Eq. (5), (6) or (7).
8. Clip x' to [0,1]D, evaluate it, and keep it if it improves on xi (τi ← 0), otherwise τi ← τi + 1.
9. After all candidates, replace b by the best member of the population if it is better.
10. Return b after T iterations; the budget is N + NT evaluations.

With N = 10 and T = 10 the budget is 110 evaluations and L = 1, so a candidate uses the AOA move after an improvement, the SCA move after one failure and the opposition step after two or more consecutive failures. To quantify the role of each operator, three ablated variants were also run: HybAOA without the SCA move (AOA move, then opposition), HybAOA without opposition (AOA move, then SCA move) and the AOA move alone.

3.7 Comparator tuning strategies

Table 4 lists the strategies compared with HybAOA. Two baselines involve no search: XGBoost defaults with 100 rounds, which is what many applied studies use, and the defaults with the number of rounds chosen by early stopping in the inner folds. Grid search evaluated the 108 configurations of **Table 3** in random order. Random search drew configurations uniformly in the unit hypercube. TPE was run with Optuna's multivariate sampler and 10 random start-up trials [15,16]. The population-based methods used N = 10 candidates and the parameter values recommended by their authors: PSO with constriction coefficients [45], DE/rand/1/bin [46], GWO [47], SCA with a linearly decreasing amplitude [48] and AOA following its reference implementation [49]. All population-based methods searched the same normalized space with clipping at the bounds, and every strategy except the two baselines and grid search (108 points) used exactly 110 evaluations.

3.8 Nested cross-validation protocol

Figure 5 illustrates the protocol. The outer loop provides the performance estimate. For D1-all it was stratified 5-fold CV repeated twice with different shuffles (10 outer folds), and the same scheme was used for D3-speech, where every speaker contributes a single recording; for D1-noMMSE and D2-MRI it was a single 5-fold partition (5 outer folds), stratified for D1 and stratified and grouped by pseudo-subject for D2, so that no subject contributed slices to both the training and the test part. Within each outer training part, the inner loop (3-fold CV, grouped by subject for D2) defined the objective of Eq. (2). Every strategy received the same outer and inner folds and the same random seed, and the PCA for D2 was refitted inside every inner and outer training set. After tuning, the selected configuration was retrained on the whole outer training part with the stored number of rounds and scored once on the outer test fold. The test folds were never used for any choice.

3.9 Performance measures

Threshold-free measures were the AUC, the average precision (AP), the log-loss and the Brier score, BS = (1/n)Σ(pi − yi)2. At the default threshold of 0.5, accuracy, balanced accuracy (the mean of sensitivity and specificity), sensitivity (recall for AD), specificity, precision, the F1 score and the Matthews correlation coefficient (MCC) were computed. All measures were computed per outer test fold and summarized as mean ± standard deviation (SD) over folds. For D2, slice-level measures were complemented by subject-level measures obtained by averaging the predicted probabilities of all slices of a subject. Wall-clock tuning time per outer fold was recorded on identical single-thread jobs.

3.10 Statistical analysis

For each task, the strategies were compared on the outer-fold test log-loss and AUC with the Friedman test [60,61], treating outer folds as blocks, followed by the Nemenyi post-hoc procedure and critical-difference (CD) analysis [61,62]. HybAOA was compared with every other strategy by two-sided Wilcoxon signed-rank tests [63] with Holm correction [64]. Because CV training sets overlap, which makes naive paired t-tests anti-conservative, the pairwise contrasts were also tested with the corrected resampled t-test [65,66],

(8)

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{J} + \frac{n_2}{n_1}\right)\hat{\sigma}_d^2}}$$

where d̄ and σ̂d2 are the mean and variance of the J fold-wise differences (J = 10 for D1-all and D3-speech and 5 for the other tasks) and n2/n1 = 1/4 is the ratio of test to training size. A difference was called significant at α = 0.05 after correction. Given the small number of folds, the analysis emphasizes effect sizes and their spread rather than p-values alone [67].

3.11 Benchmark functions and structural bias

To characterize the optimizers independently of XGBoost, seven standard test functions were minimized: Sphere, Schwefel 2.22, Rosenbrock, Rastrigin, Ackley, Griewank and Levy, on their usual domains. Each function was shifted by a random offset drawn for every run, so that the optimum lies at a random interior point rather than at the center or a corner of the domain; the same shifted instance was given to every algorithm in a run. Two settings were used, each with 30 independent runs: D = 6 dimensions with 110 evaluations, which mirrors the HPO problem, and D = 10 with 1,010 evaluations (N = 10, 100 iterations). TPE was run only in the first setting because its cost grows with the number of trials. Algorithms were ranked within each function and run, and HybAOA was compared with every other algorithm by Wilcoxon signed-rank tests over the 30 paired runs of each function, recorded as wins, ties and losses at α = 0.05.

Structural bias was probed following [23]: each algorithm minimized a function that returns independent uniform random numbers, so that the objective carries no information about any location. With D = 6 and 1,010 evaluations, 100 runs per algorithm yielded 600 final coordinates, whose distribution was compared with U(0,1) by the Kolmogorov–Smirnov (KS) test [68]. An unbiased algorithm returns uniformly distributed positions; a large KS statistic indicates a preference for particular regions.

3.12 Negative-control fusion experiment

To quantify what label-consistent fusion of unrelated datasets does, every clinical record was paired with the ResNet-18 embedding (32 principal components) of one MRI slice in two ways: random pairing, in which the slice is drawn regardless of diagnosis, and label-consistent pairing, in which AD patients receive

a slice from the demented classes and non-AD patients a non-demented slice. Pairing was done either naively, drawing slices from the whole MRI pool for all patients, or subject-disjointly, giving training-fold and test-fold patients slices from disjoint sets of MRI subjects. Default XGBoost (100 rounds) was evaluated with stratified 5-fold CV repeated five times, with the pairing redrawn in every repeat. Two clinical feature sets were used: all 32 predictors, and the 27 predictors that remain after removing the five cognitive and functional assessments, which on their own are uninformative. Because the clinical and imaging data come from different people, any gain from pairing can only come from the pairing rule.

3.13 Speech preprocessing control

A second control asked whether removing the interviewer's speech leaves a trace that a classifier can detect, which matters because the second copy of the speech data had been edited for one class only. With the eGeMAPS features, default XGBoost (100 rounds) and stratified 5-fold CV repeated five times, four comparisons were made: (i) the full and participant-only versions of the same 79 AD recordings, with folds grouped by speaker so that the two versions of a recording were never split, a contrast that can be learned only from the editing; (ii) AD participant-only against healthy full recordings, the mismatched combination of the second copy; (iii) AD full against healthy full recordings, the matched design used for D3-speech; and (iv) AD participant-only against the 45 healthy participant-only recordings of the first copy, a matched but smaller design.

3.14 Implementation

All experiments were run in Python 3.13 with NumPy 2.5, SciPy 1.18, pandas 3.0, scikit-learn 1.9 [69], XGBoost 3.4, Optuna 5.0, PyTorch 2.14, torchvision 0.29 and the openSMILE Python package 2.6, with audio converted by FFmpeg, on two cores of an Intel Xeon processor at 2.1 GHz without a GPU. The optimizers were implemented from their published descriptions, HybAOA from the authors' original code, and each tuning job ran on one thread so that run times are comparable. Random seeds were fixed for folds, optimizers and models. SHAP values for the tuned models were computed with XGBoost's exact tree path-dependent algorithm [70,71].

4. Results

4.1 Optimizer behavior on benchmark functions

Figure 6 gives the mean rank of each optimizer on the seven shifted functions, and **Figure 7** shows median convergence curves for the larger setting. At the budget that mirrors the tuning problem (D = 6, 110 evaluations), TPE achieved the best mean rank (1.79), followed by PSO (2.31), GWO (2.66) and DE (3.85). Random search ranked fifth (5.48), HybAOA sixth (6.44), AOA seventh (6.47) and SCA last (6.99); the differences between the eight optimizers were highly significant (Friedman $p < 10^{-200}$). With D = 10 and 1,010 evaluations, PSO ranked first on every function (mean rank 1.17), followed by GWO (2.45) and DE (2.63), while random search (4.86), AOA (5.10), HybAOA (5.50) and SCA (6.30) formed a clearly weaker group. The convergence curves show that HybAOA and AOA improve quickly during the first one or two hundred evaluations and then stall, whereas PSO, GWO and DE keep improving by several orders of magnitude on the unimodal and moderately multimodal functions.

Table 5 lists pairwise wins, ties and losses of HybAOA over 30 paired runs per function. HybAOA lost to PSO, DE and GWO on all seven functions in both settings and to TPE on all seven functions in the first setting, and it lost to random search on all seven functions at the tuning-sized budget and on three (with four ties) at the larger budget. It was mostly tied with AOA and beat SCA on two and five functions, respectively. Against its own ablations, the full algorithm beat the variant without the SCA move on three and seven functions but lost to the variant without the opposition step on five and seven functions.

4.2 Structural bias

When the objective returned pure noise, random search produced uniformly distributed final positions (KS statistic 0.03, p = 0.77), as expected, whereas every population-based algorithm deviated significantly from uniformity (**Figure 8**, **Table 6**). The strongest deviations were found for SCA (0.49), GWO (0.37) and AOA (0.32). SCA and AOA returned 58 % and 47 % of all coordinates within 0.01 of a bound, and GWO concentrated 40 % of its coordinates in the lowest tenth of the range. The AOA move used by HybAOA was also strongly biased (0.28; 32 % of coordinates at a bound), and adding the SCA move did not change this (HybAOA without opposition: 0.31). The center-based opposition step, in contrast, reduced the deviation to 0.14 for the full algorithm and 0.12 without the SCA move, similar to PSO (0.11) and DE (0.11).

The ablation in **Figure 9** shows the corresponding trade-off on the benchmark functions. Among the four HybAOA variants, the one without the opposition step ranked best in both settings (mean rank 1.76 and 1.23 of 4), the AOA move alone came

second (2.37, 2.18), the full algorithm third (2.65, 2.80) and the variant without the SCA move last (3.22, 3.79). The SCA move therefore helps, while the opposition step, which removes most of the boundary bias, costs convergence because it repeatedly relocates stagnating candidates far from the incumbent.

4.3 Clinical task with all predictors (D1-all)

Table 7 and Figure 10a–b summarize the ten outer folds of D1-all. All eleven strategies reached almost the same discrimination: the mean test AUC ranged from 0.947 (Default+ES) to 0.951 (AOA), a spread of 0.004, whereas the SD of the AUC across outer folds was 0.008–0.013 for every strategy. Balanced accuracy at the 0.5 threshold lay between 0.940 and 0.943. HybAOA-tuned XGBoost reached an AUC of 0.950 ± 0.009, a balanced accuracy of 0.942 ± 0.009, a sensitivity of 0.914 and a specificity of 0.970. The one clear difference concerned the quality of the predicted probabilities. The 100-round default had a mean test log-loss of 0.256 ± 0.031, whereas all other strategies lay between 0.197 and 0.206, including the default configuration with early stopping alone (0.206), which needed a single objective evaluation and stopped after a median of 12 rounds. HybAOA's test log-loss was 0.200 ± 0.021. The Brier scores followed the same pattern but with smaller relative differences (0.048 for the 100-round default and 0.047 for HybAOA), because the Brier score penalizes confident errors less than log-loss. Tuning with HybAOA took 41 s per outer fold, compared with 36–76 s for the other search methods and less than 1 s for the defaults.

4.4 Clinical task without the MMSE score (D1-noMMSE)

Without the MMSE score (**Table 8**, Figure 10c–d), discrimination decreased for every strategy, and HybAOA's AUC fell from 0.950 to 0.901 and its balanced accuracy from 0.942 to 0.876. The six strategies again differed little in AUC (0.893–0.901) and balanced accuracy (0.835–0.876), and the 100-round default again had by far the worst log-loss (0.487, against 0.346–0.361 for the others). HybAOA had the highest mean balanced accuracy in this task (0.876 ± 0.010, sensitivity 0.897), ahead of AOA (0.870), TPE (0.867) and random search (0.857), but these differences were within the fold-to-fold variation and not significant after correction (Section 4.7). Removing the label-related predictor therefore lowers the absolute performance without changing the conclusion that the search methods perform alike.

4.5 MRI task (D2-MRI)

The MRI task was considerably harder (**Table 9**, Figure 10e–f), and here tuning mattered. With subject-grouped outer folds, the two untuned defaults had the lowest slice-level AUCs (0.726 with early stopping and 0.749 with 100 rounds), whereas the nine search methods reached 0.770–0.776, a spread much smaller than the fold-to-fold SD of about 0.032. Balanced accuracy at the 0.5 threshold ranged from 0.579 to 0.625. HybAOA reached a slice-level AUC of 0.770 ± 0.032 and a balanced accuracy of 0.618 ± 0.045 (sensitivity 0.332, specificity 0.903). Averaging the slice probabilities of each subject raised the AUC of all strategies, to 0.838–0.869 (HybAOA 0.862), with subject-level balanced accuracies of 0.564–0.679. As on the clinical data, the 100-round default had the worst slice-level log-loss (0.727), followed by the early-stopped default (0.552), against 0.509–0.514 for the search methods, which were statistically indistinguishable from one another (Section 4.7).

4.6 Speech task (D3-speech)

Table 10 and Figure 10g–h summarize the ten outer folds of D3-speech. Discrimination from the acoustic features was moderate and varied considerably between folds (SD of the AUC 0.040–0.107). The mean test AUC of the nine search methods ranged from 0.665 (SCA) to 0.737 (random search), and HybAOA reached 0.702 ± 0.102, with a balanced accuracy of 0.674 ± 0.082 (sensitivity 0.672, specificity 0.676). Unlike on the other tasks, the 100-round default had the highest mean AUC (0.757 ± 0.062; balanced accuracy 0.692), whereas the default with early stopping had the lowest (0.655): with inner validation folds of about 42 recordings, early stopping halted the default configuration ($\eta = 0.3$) after a median of 4 rounds, whereas the configurations selected by the search methods used medians of 19–46 rounds. The log-loss gave the reverse picture. The 100-round default had the worst mean log-loss and by far the largest spread (0.726 ± 0.177), while the other strategies lay between 0.642 (random search) and 0.686 (AOA), with HybAOA at 0.654 ± 0.087. Balanced accuracies lay between 0.622 and 0.698 and Brier scores between 0.220 and 0.239, so that on this task no strategy was clearly better than the others on every measure.

4.7 Statistical comparison

The Friedman test across outer folds found no difference in test AUC among the eleven strategies on D1-all ($\chi^2 = 11.7$, $p = 0.308$) but significant differences on D2-MRI ($\chi^2 = 26.0$, $p = 0.004$) and D3-speech ($\chi^2 = 26.1$, $p = 0.004$). On D2-MRI the two untuned defaults had the two worst mean AUC ranks; on D3-speech the 100-round default had the

best (2.5) and the early-stopped default, SCA and AOA the worst (7.6–7.7) (**Figure 11**). For test log-loss the differences were significant on D1-all (χ^2 = 36.8, $p < 0.001$) and D2-MRI (χ^2 = 26.6, $p = 0.003$), where the two defaults again ranked last, but not on D3-speech (χ^2 = 6.6, $p = 0.763$). The mean log-loss ranks of the nine search methods lay between 3.7 and 6.3 on D1-all, 3.2 and 6.2 on D2-MRI and 4.6 and 7.2 on D3-speech, well within one critical difference (4.77, 6.75 and 4.77 ranks). **Table 11** lists the pairwise contrasts between HybAOA and every other strategy. After Holm correction the Wilcoxon tests were significant only for the 100-round default on D1-all log-loss (difference −0.056); the corrected resampled t-test, which accounts for overlapping training sets, was significant only for the 100-round default on D1-all log-loss (difference −0.056), the 100-round default on D2-MRI log-loss (difference −0.214) and the default with early stopping on D2-MRI log-loss (difference −0.039); on D1-noMMSE it also separated HybAOA from the 100-round default (log-loss −0.142, $p < 0.001$; AUC +0.009, $p = 0.008$). On D3-speech no contrast was significant with either test; the largest were HybAOA's lower AUC relative to the 100-round default (−0.055, corrected t-test $p = 0.258$) and to random search (−0.036, $p = 0.577$). No contrast between HybAOA and another search method was significant on any metric or task, and the largest absolute difference in mean AUC between HybAOA and a search method was 0.037 (D3-speech, against SCA). With five outer folds the Wilcoxon test cannot reach $p < 0.05$ on D2-MRI or D1-noMMSE, which is why effect sizes and the corrected t-test are reported alongside it. Pooling the 30 outer folds of all four tasks for the six strategies run on every task gave mean log-loss ranks of HybAOA 2.57, AOA 2.77, Random search 3.10, TPE 3.10, Default+ES 4.10, Default-100 5.37 (Friedman $p < 0.001$) and mean AUC ranks of Random search 2.90, HybAOA 3.10, Default-100 3.33, AOA 3.33, TPE 3.57, Default+ES 4.77 ($p = 0.002$); the log-loss test was driven by the two defaults and the AUC test by the early-stopped default. In this pooled set HybAOA's test log-loss was on average 0.0015 higher than that of random search (adjusted $p = 0.860$), 0.0054 lower than that of TPE (adjusted $p = 0.421$) and 0.0097 lower than that of AOA (adjusted $p = 0.957$), and 0.1021 lower than that of the 100-round default (adjusted $p < 0.001$) and 0.0162 lower than that of the early-stopped default (adjusted $p = 0.006$); its AUC was 0.0249 higher than that of the early-stopped default (adjusted $p = 0.022$) and 0.0132 lower than that of the 100-round default (adjusted $p = 1.000$).

4.8 Convergence of the inner objective and computational cost

Figure 12 shows how quickly each tuner approached the best inner-CV log-loss reached in the same outer fold by any tuner. On D1-all, the number of the 10 outer folds in which a tuner came within 1 % of that value within its budget was 7 for TPE and PSO; 6 for GWO and SCA; 5 for AOA; 4 for grid search; 3 for DE and HybAOA; and 0 for random search. After 110 evaluations the median excess was smallest for SCA (0.59 %), PSO (0.62 %) and TPE (0.78 %), against 1.49 % for HybAOA and 3.41 % for random search. HybAOA thus optimized the inner objective less completely than the best metaheuristics and TPE, which is consistent with the slow late-stage convergence seen on the benchmark functions. These differences are small in absolute terms: the mean best inner log-loss of the nine search methods spanned only 0.2028–0.2088, and it was not significantly related to their test log-loss (Spearman $\rho = 0.45$, $p = 0.224$). The test log-loss of the refitted models was slightly lower than the inner estimate (differences −0.007 to −0.003), because the final models were trained on the whole outer training part rather than on two-thirds of it, so the inner estimates were mildly pessimistic rather than optimistic. On D2-MRI the picture was similar: after 110 evaluations the median excess over the fold-best inner log-loss was 0.58 % for HybAOA, between 0.05 % (TPE) and 0.89 % (grid search) for the search methods, and the best inner log-loss of the nine search methods spanned 0.5112–0.5144. Tuning time per outer fold (**Figure 13**) ranged from 36 s (grid search) to 76 s (SCA) on D1-all and from 37 s (PSO) to 59 s (grid search) on D2-MRI, with HybAOA at 41 s and 47 s; on the small D3-speech task every search method needed 8–12 s. Because every evaluation trains three XGBoost models, run time is governed by the learning rate and depth of the configurations a method visits rather than by the optimizer's own arithmetic, which is negligible. On D3-speech, TPE optimized the inner objective best (median excess 1.03 % after 110 evaluations, against 5.56 % for HybAOA and 4.53 % for random search), but here a better inner optimum did not mean a better model. The test log-loss exceeded the best inner log-loss by +0.061 to +0.111 for the nine search methods, so the inner estimates were optimistic rather than pessimistic, and the mean inner and test log-loss of the search methods were not positively related (Spearman $\rho = -0.35$, $p = 0.356$): TPE, with the lowest inner value (0.5681), had a test log-loss of 0.667, whereas random search, with a higher inner value (0.5812), had the lowest (0.642).

4.9 Selected hyperparameters

The tuners did not agree on a single configuration (**Table 12**, **Figure 14**), which is itself a sign of a flat objective near its optimum. Most selected deep trees (median depth 8–10) and column subsampling close to or at 1.0, with median learning rates between 0.020 and 0.065, while the L2 weight λ varied over three orders of magnitude without a visible effect on test performance. The share of selected coordinates lying exactly on a bound mirrored the structural biases of Section 4.2: it was 0 % for random search and 3 % for TPE, but 52 % for SCA, 42 % for AOA, 40 % for PSO and 27 % for HybAOA (grid search, whose lattice contains several bounds, reached 42 %). HybAOA chose the maximum depth of 10 in 7 of 10 folds and a column subsample of 1.0 in 7. On this problem the attraction to the upper bounds was harmless, because the unbiased tuners also favored deep trees and nearly complete column sampling; the bias therefore did not translate into worse test performance, but it would on a problem whose good configurations lie in the interior of the box. On D3-speech, too, random search and TPE rarely selected bound values (0 % and 2 % of coordinates), whereas SCA (40 %), AOA (28 %) and HybAOA (10 %) did so more often; AOA and HybAOA chose the largest learning rate of 0.3 in 5 and 4 of the 10 folds, random search, TPE and PSO in none.

4.10 Contribution of HybAOA's operators

Table 13 compares the four HybAOA variants on the five outer folds of the first repetition of D1-all. Their best inner log-loss ranged only from 0.2051 to 0.2079 (lowest for the AOA move alone), their test log-loss from 0.199 to 0.200 and their test AUC from 0.948 to 0.950. The operator-level differences that were clear on the benchmark functions therefore vanish on this tuning problem, consistent with its flat landscape: whichever operator drives the search, the selected configurations perform equally well on new patients. On D3-speech, where all ten outer folds were used (**Table 14**), the variants that minimized the inner objective best did worse on the test folds: without the opposition step the inner log-loss fell from 0.5837 to 0.5744, but the test log-loss rose from 0.654 to 0.702 and the AUC fell from 0.702 to 0.673; the AOA move alone behaved similarly (inner 0.5715, test log-loss 0.677), and removing the SCA move changed little (test log-loss 0.655, AUC 0.710). These differences are smaller than the fold-to-fold SDs, but their direction is consistent with the optimism of the inner estimates on this small dataset: the opposition step, which slows the optimization of the inner objective, also limited how closely the search fitted its noise.

4.11 Discrimination, calibration and errors

The ROC and precision–recall curves of the out-of-fold predictions overlap almost completely (**Figure 15**); for HybAOA the pooled AUC was 0.947 on D1-all. The reliability diagrams (**Figure 16**) and the error analysis locate the log-loss difference of the 100-round default. Its Brier score was almost the same as that of the tuned models (0.048 against 0.047 for HybAOA), but 56 of its 111 misclassified patients received a probability above 0.99 or below 0.01 for the wrong class, compared with 1 of 105 for the default with early stopping and 0 of 105 for HybAOA; the 2 % of patients with the largest losses accounted for 54 % of its total log-loss, against 38 % for HybAOA. At the 0.5 threshold, HybAOA-tuned XGBoost misclassified 65 of 760 AD patients and 40 of 1,389 non-AD patients on D1-all (first repetition; **Figure 17**). On D2-MRI the pooled AUC of HybAOA was 0.767. Here the 100-round default was again badly calibrated (Brier 0.214, 102 confident errors), whereas the tuned models gave smoother probabilities (Brier 0.173 for HybAOA). With 30 % demented slices, the 0.5 threshold favored the majority class: HybAOA detected 450 of 1,356 demented slices and 2,891 of 3,200 non-demented slices. At the subject level, 30 of 100 demented and 97 of 100 non-demented subjects were classified correctly, so that a threshold chosen on the inner folds, rather than the default 0.5, would be needed in any practical use. On D3-speech the pooled out-of-fold AUC of HybAOA was 0.643, against 0.708 for the 100-round default and 0.709 for random search. These values come from the first repetition only, whose partition was harder for HybAOA: its mean per-fold AUC was 0.637 in the first and 0.767 in the second repetition (100-round default 0.723 and 0.791; random search 0.740 and 0.734), which shows how much a single partition of 158 recordings can move the estimate. With so few recordings the reliability diagrams are noisy, and the Brier scores of the pooled predictions (0.234–0.250) were close to the 0.25 of a constant prediction of 0.5. At the 0.5 threshold HybAOA classified 49 of 79 AD and 50 of 79 healthy speakers correctly.

4.12 Model interpretation

SHAP values of the HybAOA-tuned models (**Figure 18**) confirm that the clinical predictions rest on the cognitive and functional assessments. The five largest mean absolute SHAP values belonged to FunctionalAssessment (1.33 log-odds), ADL (1.14), MemoryComplaints (0.81), MMSE (0.78) and BehavioralProblems (0.61), which together accounted for 84 % of the total; the next predictor, CholesterolLDL, reached only 0.09. Low functional assessment, ADL and MMSE scores and the presence of memory complaints or behavioral problems

increased the predicted log-odds of AD. The remaining 27 predictors, including age, blood pressure, cholesterol and lifestyle variables, contributed little, in line with their near-chance single-predictor AUCs (**Figure 2**). This concentration also explains why the tuning landscape is flat: once the trees can split on these five variables, the remaining hyperparameters mainly control how much noise from the other predictors is fitted.

On D3-speech the importance was spread over many features (**Figure 19**): the ten most important of the 88 eGeMAPS features accounted for 37 % of the total mean absolute SHAP value, whereas on D1-all five assessments accounted for 84 %. The largest contributions came from the variability of the second MFCC in voiced frames (0.30 log-odds), the 80th percentile of the pitch (0.22), the variability of the spectral flux in voiced frames (0.18), the number of voiced segments per second (0.15) and the median loudness (0.14). Higher values of the first three and lower values of the last two raised the predicted log-odds of AD (Spearman correlations between feature value and SHAP value +0.75, +0.70, +0.74, −0.75 and −0.51). Fewer voiced segments per second is compatible with slower or more hesitant speech, but every recording also contains the interviewer's voice, so these features describe the whole interview rather than the participant alone and should not be read as acoustic markers of AD.

4.13 Negative control: fusing datasets that share no participants

Table 15 and **Figure 20** report the negative control. Adding the MRI embedding of a randomly chosen slice to each clinical record changed nothing: AUC 0.951 with the 32 clinical predictors alone and 0.950 with the random MRI block, and 0.497 versus 0.493 for the 27 non-assessment predictors, which on their own are at chance level. Label-consistent pairing changed the picture. With all clinical predictors the AUC rose to 0.974 (naive pairing) and 0.968 (subject-disjoint pairing). With the uninformative clinical predictors it rose from chance to 0.907 (balanced accuracy 0.809) with naive pairing and to 0.750 with subject-disjoint pairing. The MRI block alone reached 0.920 and 0.754 under the two label-consistent schemes and 0.502 under random pairing. Because the clinical and imaging records come from different people, all of these gains are produced by the pairing rule: the image attached to a patient was chosen using the patient's diagnosis. The naive scheme adds slice-level leakage on top, since training and test patients can receive slices of the same MRI subject.

4.14 Speech preprocessing control

Table 16 and **Figure 21** report the preprocessing control. Default XGBoost separated the full and participant-only versions of the same 79 AD recordings with an AUC of 0.755 ± 0.050 (balanced accuracy 0.678), although the speakers and their diagnoses were identical. Removing the interviewer's turns therefore leaves an acoustic signature at least as easy to detect as the diagnosis, for which the matched comparison of full interviews gave 0.727 ± 0.081, consistent with the 0.757 of the 100-round default in the nested evaluation. In this sample the signature did not inflate the mismatched design: comparing edited AD recordings with unedited healthy recordings gave a lower AUC (0.672 ± 0.080) than the matched design. The matched comparison of participant-only recordings, which could use only the 45 healthy speakers whose edited recordings were available, gave 0.634 ± 0.111. A processing difference can add to or subtract from the diagnostic signal, depending on how the two are aligned in feature space, so a mismatched design yields an estimate whose bias cannot be predicted in either direction; only identically processed recordings, as used for D3-speech, give an interpretable estimate.

5. Discussion

5.1 Principal findings

This study set out to establish whether a hybrid arithmetic optimization algorithm improves XGBoost for AD prediction when it is evaluated as strictly as the alternatives. The answers to the four research questions are as follows.

- **RQ1 (HybAOA versus defaults).** On the clinical task, HybAOA did not improve discrimination over untuned XGBoost (AUC 0.950 against 0.950 for the 100-round default and 0.947 with early stopping), but it lowered the test log-loss of the 100-round default from 0.256 to 0.200, a gain that early stopping alone almost matched (0.206). On the MRI task, tuning mattered more: HybAOA raised the slice-level AUC from 0.749 (100-round default) and 0.726 (early stopping) to 0.770 and lowered the log-loss from 0.727 and 0.552 to 0.513. On the speech task, tuning did not improve discrimination: the 100-round default had the highest mean AUC (0.757, against 0.702 for HybAOA; difference not significant) but the worst and most variable log-loss (0.726 ± 0.177, against 0.654 ± 0.087), and early stopping alone under-fitted (AUC 0.655).
- **RQ2 (HybAOA versus other tuners).** HybAOA was neither better nor worse than the established tuners. Across the nine search

methods the mean test AUC spanned 0.948–0.951 on D1-all, 0.770–0.776 on D2-MRI and 0.665–0.737 on D3-speech, and no pairwise difference between HybAOA and another search method was significant after correction. Among the eleven strategies, HybAOA's mean ranks were 5.2 (log-loss) and 4.9 (AUC) on D1-all, 4.6 and 6.4 on D2-MRI, where its mean AUC (0.770) was the lowest of the nine search methods by a margin far below the fold-to-fold SD, and 5.6 and 5.8 on D3-speech. Pooled over all four tasks, its log-loss was 0.0015 higher than that of random search (adjusted $p = 0.860$) and similar to that of TPE and AOA.

- **RQ3 (operators and bias).** The arithmetic move at the core of HybAOA has a strong structural bias towards the bounds of the search space; the center-based opposition step removes most of that bias but slows convergence, and the SCA move helps. On shifted benchmark functions the full algorithm ranked below random search at the tuning-sized budget.
- **RQ4 (data-assembly errors).** Pairing clinical records with MRI slices that share the patient's diagnosis raised the AUC of an uninformative clinical feature set from 0.497 to 0.907 (to 0.750 when slice leakage was also removed), whereas pairing that ignored the diagnosis added nothing; label-consistent fusion therefore creates accuracy that does not exist. Removing the interviewer's speech left a signature that default XGBoost detected with an AUC of 0.755 among recordings of the same speakers. In this sample the mismatched design lowered rather than raised the AUC (0.727 → 0.672), which shows that class-specific processing makes an estimate uninterpretable rather than predictably optimistic.

5.2 Why HybAOA did not outperform simpler tuners

Four mechanisms explain the results, and a fifth applies to the small speech dataset. The first is the shape of the tuning problem. For XGBoost on these data, many configurations reach nearly the same validation loss: the best inner-CV log-loss found by the different tuners on D1-all spanned only 0.2028–0.2088, a range smaller than the fold-to-fold variation of the test log-loss. When the landscape is this flat near its optimum, any method that samples it reasonably finds a near-optimal configuration, and the remaining differences are dominated by noise in the inner CV estimate. This is consistent with earlier observations that only a few hyperparameters matter for boosted trees and that random search is hard to beat at moderate budgets [13,14,17].

The second mechanism is the geometry of the arithmetic operators. The division and multiplication branches of the AOA move scale each coordinate of the incumbent towards zero or beyond the upper bound, and the addition and subtraction branches take steps of fixed absolute size around it. These operations are not invariant to shifting or rescaling the search space, so their behavior depends on where the optimum lies relative to the origin and the bounds. In the normalized coordinates used for tuning, zero is the lower bound of every hyperparameter, and clipping turns overshoots into exact bound values. The structural-bias probe makes this visible: the AOA move alone placed 32 % of all coordinates at a bound when the objective carried no information. Such a bias is harmless, or even helpful, when the optimum lies on a bound, and harmful otherwise, and it can make an algorithm look excellent on benchmark functions whose optimum lies at a privileged point, which is why shifted functions were used here [23]. On these functions, where the optimum was moved to a random interior point, both AOA and HybAOA fell behind PSO, DE, GWO and TPE.

The third mechanism is the trade-off introduced by the opposition step. Reflecting a stagnating candidate through the center of the box reduced the structural bias from 0.31 to 0.14, but with $L = 1$ a candidate is relocated after only two unsuccessful moves, so much of the budget is spent far from the incumbent. The ablation shows the cost: removing the opposition step gave the best variant on the benchmark functions. The fourth mechanism is the budget. MOA and MOP are schedules over iterations, and with ten candidates and 110 evaluations the algorithm completes only ten iterations, too few for its exploration-to-exploitation transition to matter. Methods that learn from every evaluation, such as TPE, or that keep a velocity or difference memory, such as PSO and DE, use a small budget more efficiently.

The fifth mechanism is the noise of the inner estimate when data are scarce. On D3-speech each outer training part held about 126 recordings and each inner validation fold about 42, so the inner log-loss of a configuration was itself a noisy quantity, and searching it harder partly fitted that noise. The test log-loss exceeded the best inner log-loss by +0.061 to +0.111 across the search methods, the tuner with the lowest inner value did not have the lowest test log-loss, and the HybAOA variants that optimized the inner objective best were the worst on the test folds.

This is the selection bias described for model selection by cross-validation [24,25], and it means that on small datasets a more thorough optimizer can be a disadvantage. The 100-round default, which uses no inner estimate at all, had the highest mean AUC on this task but the least reliable probabilities.

5.3 What tuning does improve

The clearest benefit of tuning in this study concerned the probabilities rather than the ranking of patients. On D1-all the 100-round default had the same AUC as the tuned models but a test log-loss about 28 % higher, because 56 of its errors were made with a probability beyond 0.99 or below 0.01. On the MRI embeddings, where the signal is weaker, the defaults were also worse in AUC: with a learning rate of 0.3 the early-stopped default stopped after a median of 6 rounds and under-fitted, whereas the 100-round default over-fitted and produced 102 confident errors. On the speech features the 100-round default ranked speakers best on average but gave the least reliable probabilities (log-loss SD 0.177 across folds, against 0.087 for HybAOA), and with validation folds of about 42 recordings early stopping halted the default configuration after a median of 4 rounds. Boosted models are known to produce distorted probability estimates [72]; here, a high learning rate with a fixed number of rounds over-fitted the training data and produced over-confident predictions, which leaves the ranking of patients, and therefore the AUC, almost unchanged but penalizes the confident errors heavily in log-loss. For clinical decision support, calibrated probabilities matter as much as discrimination, because a predicted risk is used to decide on further testing [73]. The practical lesson is modest but useful: on the clinical data, early stopping on a validation split recovered almost all of the benefit of tuning, whereas on the weaker MRI signal a search over the learning rate and tree complexity was needed, and any of the nine search methods sufficed. On the small speech dataset, early stopping on validation folds of about 42 recordings was unreliable, and none of the strategies improved discrimination over the untuned default; there, the main benefit of tuning was a lower and more stable log-loss.

5.4 Implications for multimodal Alzheimer's disease studies

The negative control shows that a fused model can appear to benefit from an extra modality even when that modality comes from different people. With label-consistent pairing, the image block of every patient is drawn from the class that matches the patient's diagnosis, so the image features become a noisy copy of the label; the classifier learns to read the label back. The effect was largest when the clinical features were uninformative (AUC 0.497 → 0.907) and smaller, but still present, when they were strong (AUC 0.951 → 0.974). Because the gain is bounded by how well the auxiliary data separate the classes, it can easily be mistaken for a genuine multimodal benefit. A second, silent failure mode is identifier-based lookup across unrelated datasets: when no file matches, substituting zero vectors leaves a model that uses one modality while being described as multimodal. Both problems can be detected with simple checks: report how many participants have every modality, run a random-pairing control, and compare the fused model with the best single-modality model under the same folds.

Credible multimodal studies need cohorts in which every participant contributes every modality, such as ADNI [36] and OASIS [37], which have supported multimodal deep-learning models [34,35]. For speech, the ADReSS and ADReSSo challenge releases provide labelled recordings with fixed subject-level partitions [33,40]. The first two copies received for this study did not preserve these properties: one contained the same files in both class folders, and the other had been edited for one class only. The speech control shows why the second problem matters, because the editing step is itself detectable. Studies that assemble recordings from different processing pipelines should therefore compare durations, levels, sampling rates and channel layouts between classes and test whether the processing can be predicted from the features, in the same way as a random-pairing control tests a fused dataset.

5.5 Comparison with published results

On the clinical dataset, the tuned models reached a test accuracy of about 95 % (HybAOA 0.951), comparable to the 95.3 % reported for XGBoost with oversampling and tuning [12]. The SHAP and single-predictor analyses show that this level is reached almost entirely through five cognitive and functional assessment variables; without the MMSE score the AUC of HybAOA fell to 0.901, and without all five assessments the clinical predictors carried no signal (AUC 0.497 in the negative-control experiment). For the MRI collection, the subject-grouped slice-level AUC of about 0.77 (subject level 0.862) is far below the near-perfect accuracies that 2-D slice classifiers can obtain when slices of the same subjects appear in training and test sets, the leakage mechanism documented in [28,29]; results on this collection that are not split by subject should therefore not be compared with ours. Multimodal results obtained by label-consistent fusion of heterogeneous datasets,

such as the accuracy of 92.57 % and AUC of 0.973 reported in [30], are not comparable either: the negative control shows that the pairing rule alone can produce gains of this size, so a random-pairing control would be needed to attribute such results to complementary biological information. For speech, the mean nested test accuracies of 62.3–70.0 % from eGeMAPS features alone (HybAOA 67.4 %) are of the same order as the acoustic results reported for the challenge data, a test accuracy of 62.5 % for the acoustic baseline of ADReSS [40] and a leave-one-subject-out accuracy of 74.1 % for eGeMAPS on the ADReSSo training set [42], and below the 78.9 % of the ADReSSo baseline that also used linguistic features [33]. They are not directly comparable, because the challenge test set was not available here and the leave-one-subject-out estimate involved no nested tuning. Finally, the present nested estimates replace the single-split results of the authors' earlier studies [31,32] as the reference for this line of work.

5.6 Practical recommendations

- Always tune XGBoost with early stopping on an inner validation split, and search at least the learning rate and tree complexity with random search or TPE (50–100 evaluations) before considering a new metaheuristic.
- Compare any new optimizer with random search and a Bayesian tuner at equal budget, under nested CV, and report per-fold results and tuning time.
- Report a proper scoring rule (log-loss or Brier score) and a calibration plot, not only accuracy and AUC.
- Characterize new metaheuristics on shifted benchmark functions and with a structural-bias probe before applying them.
- Split imaging data by subject; when identifiers are missing, recover them before splitting.
- Fuse modalities only for participants who have all of them, and include a random-pairing control.
- Process the recordings of all classes with one identical pipeline, compare their durations, levels and formats by class, and check that the processing itself cannot be predicted from the features.

5.7 Limitations

The study has several limitations. The clinical dataset is public and convenient but shows signatures of simulation, and its strongest predictors are cognitive and functional assessments that are part of the diagnostic process; the absolute accuracy should therefore not be read as a clinical performance estimate. The MRI task uses 2-D slices from 200 recovered pseudo-subjects of a re-packaged collection without documented provenance, and the images were represented by frozen ImageNet features rather than a fine-tuned or 3-D network; a stronger image model could change the absolute results, although there is no reason to expect it to change the ranking of tuners. The speech task used 158 recordings, so its estimates are imprecise (fold-to-fold SD of the AUC up to 0.107). The recordings contain the interviewer's voice as well as the participant's, and the acoustic features were computed over whole recordings without speaker separation or transcripts, so part of the signal may come from the interviewer or from the course of the interview; the lower AUC of the participant-only comparison (0.634) is compatible with this, although that comparison had only 45 healthy speakers. The challenge test set was not available, and only acoustic features were used, whereas linguistic features from transcripts are usually more informative [33,39]. The number of outer folds (10 for D1-all and D3-speech, 5 for the other tasks) limits statistical power, so small true differences between tuners cannot be excluded; the conclusions rest on the consistency of the effect sizes rather than on non-significant p-values alone. Only one budget (110 evaluations), one population size and one six-dimensional search space were studied, and HybAOA was evaluated as implemented in the authors' code; other hybrid AOA designs may behave differently. Only XGBoost was tuned. Finally, no external validation cohort was available.

5.8 Future work

Three extensions follow directly. First, the protocol should be repeated on cohorts with paired clinical, imaging and speech data, where the value of fusion, and of tuning a fused model, can be measured honestly; for speech, this should include participant-only audio obtained identically for both classes, transcript-based features and the official challenge partition. Second, a translation- and scale-invariant reformulation of the arithmetic operators, for example applying them to differences between candidates rather than to absolute coordinates, could remove the boundary bias without the cost of opposition restarts; any such variant should be tested on shifted benchmarks and the structural-bias probe before being applied. Third, multi-fidelity tuners that allocate more rounds or data to promising configurations could reduce tuning cost further on larger cohorts.

6. Conclusion

A hybrid arithmetic optimization algorithm for tuning XGBoost was evaluated against nine alternatives under nested cross-validation on clinical, MRI and speech data for Alzheimer's disease prediction. HybAOA produced competitive models (clinical AUC 0.950, balanced accuracy 0.942; MRI subject-grouped slice AUC 0.770, subject-level AUC 0.862; speech AUC 0.702, balanced accuracy 0.674), but so did grid search, random search, TPE, PSO, DE, GWO, SCA and AOA, and no difference between these tuners was statistically or practically meaningful. Tuning lowered the log-loss, clearly on the clinical and MRI tasks, and improved discrimination on the MRI signal; early stopping provided most of the clinical-data benefit at a fraction of the cost, and on the small speech dataset no tuner discriminated better than untuned XGBoost, while the inner cross-validation estimates became optimistic. Benchmark functions and a structural-bias probe showed that the arithmetic moves of AOA-type algorithms are attracted to the bounds of the search space, which HybAOA's opposition step mitigates at the expense of convergence. Two control experiments showed how data assembly can distort accuracy: fusing unrelated clinical and imaging datasets by matching diagnoses inflated AUC from chance level to 0.907, and removing the interviewer's speech from one set of recordings left a signature that a classifier detected with an AUC of 0.755. For XGBoost-based AD prediction, the choice of optimizer matters far less than early stopping, subject-level splitting, honest nested evaluation, identical processing of all recordings and the use of cohorts in which every modality comes from the same participants.

Credit authorship contribution statement

Manisha M. Chaudhari: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. Mukta Agrawal: Supervision, Methodology, Validation, Writing – review & editing.

Ethics approval

This study is a secondary analysis of publicly available, de-identified data; no new data were collected from human participants and no ethics approval was required.

List of Figures

Figure 1. Study workflow. The audit showed that the three data sources share no participants, so the clinical, MRI and speech data were analysed as separate tasks under nested cross-validation (CV); the unusable first copies of the speech data were replaced by the original, identically processed interviews. Benchmark functions, a negative-control fusion experiment and a speech-preprocessing control complement the main comparison.

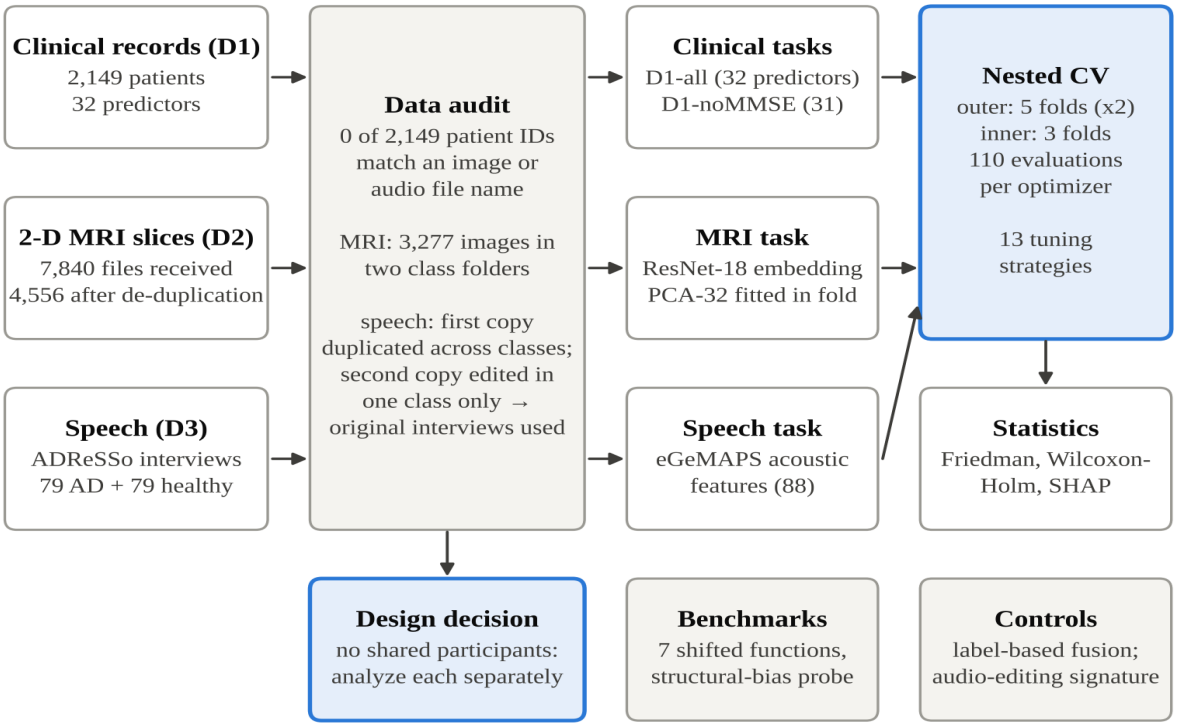


Figure 2. Clinical dataset (D1). (a) Diagnosis counts. (b, c) Distributions of the two strongest predictors by diagnosis. (d) AUC of each predictor used alone (values below 0.5 mirrored), for the 14 strongest of 32 predictors; orange bars mark the cognitive and functional assessments with AUC above 0.6.

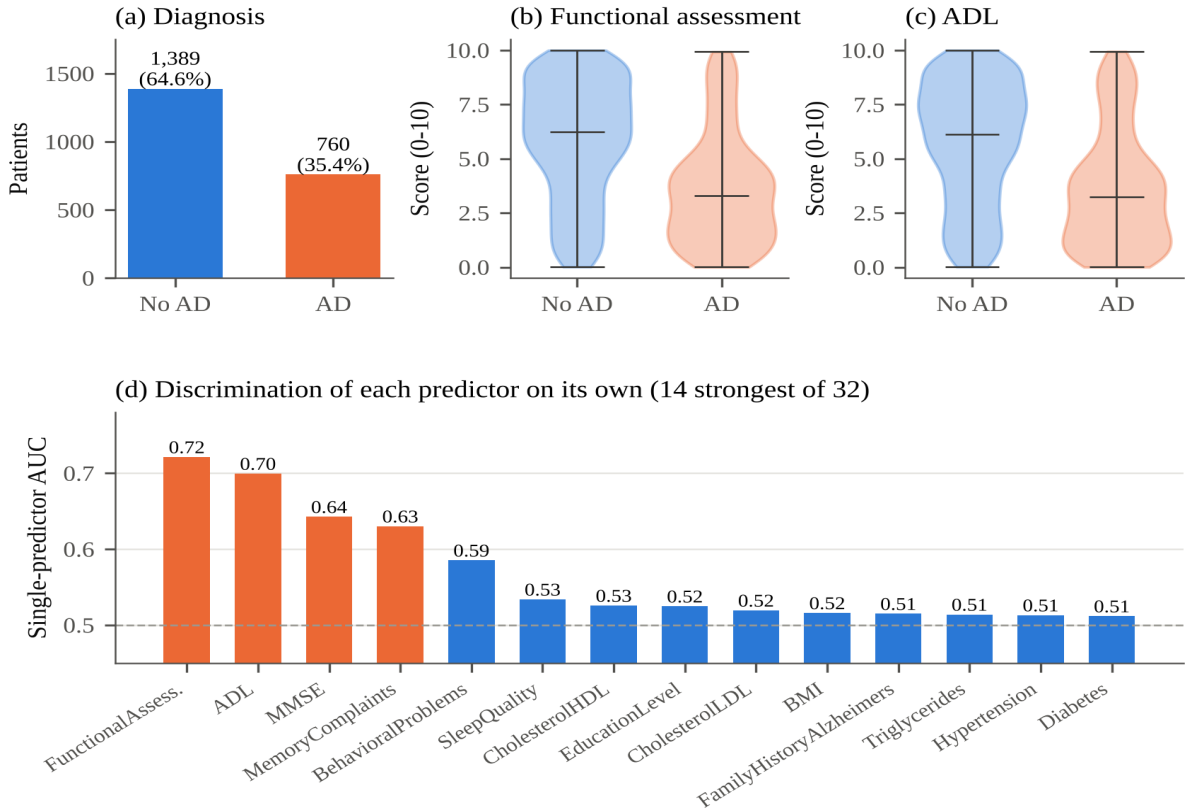


Figure 3. MRI dataset (D2). (a–d) Example axial slices from each class. (e) Slices per class after removing files duplicated across class folders. (f) Pseudo-subjects per class recovered by slice-similarity linkage.

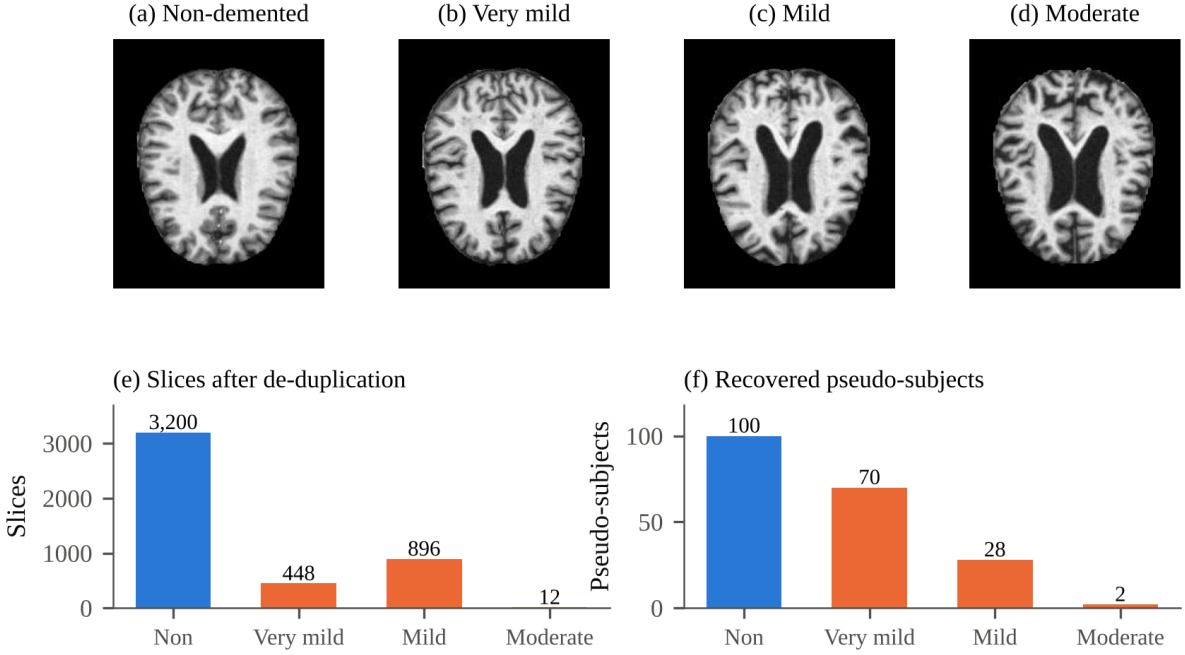


Figure 4. Flowchart of the hybrid arithmetic optimization algorithm (HybAOA). A per-candidate stagnation counter selects the arithmetic-optimization (AOA) move, the sine–cosine (SCA) move or the center-based random opposition step; selection is greedy and the global best is refreshed once per iteration.

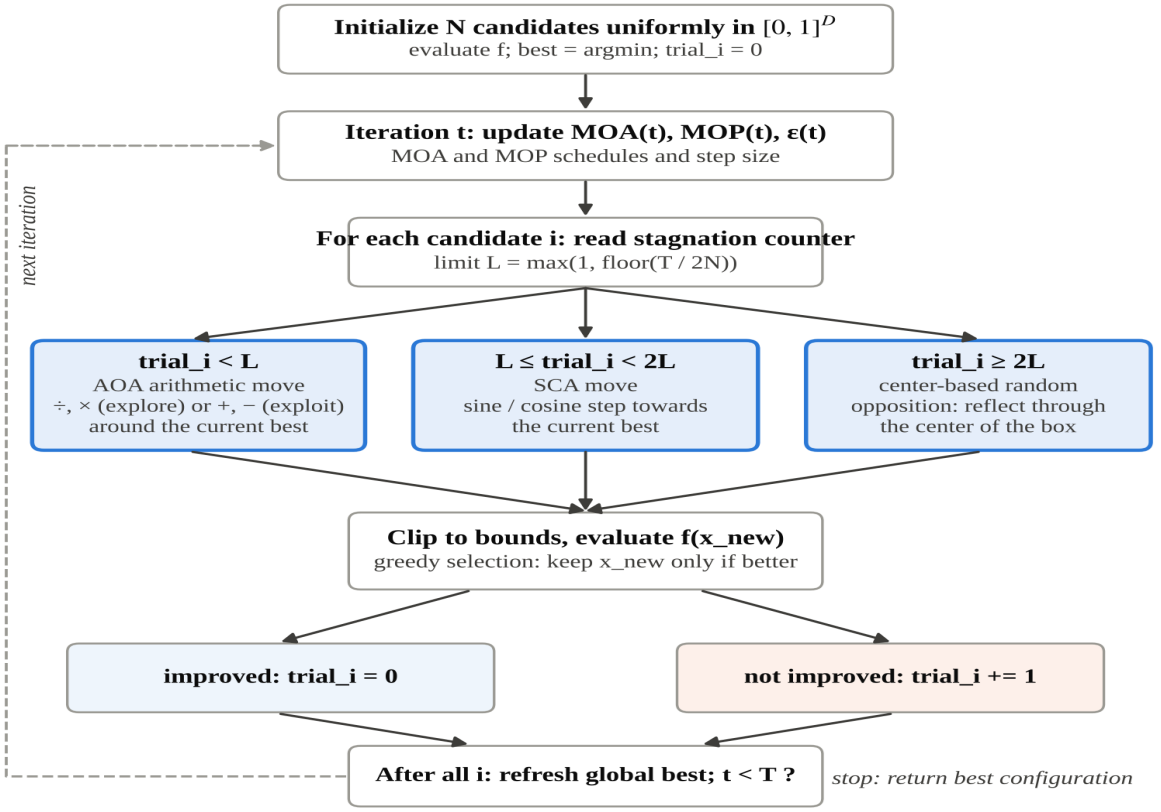


Figure 5. Nested cross-validation. Each outer test fold (orange) is predicted once by a model whose hyperparameters were chosen using only the outer training part, through 3-fold inner CV with early stopping; for MRI all splits are grouped by subject and PCA is refitted inside every training set.

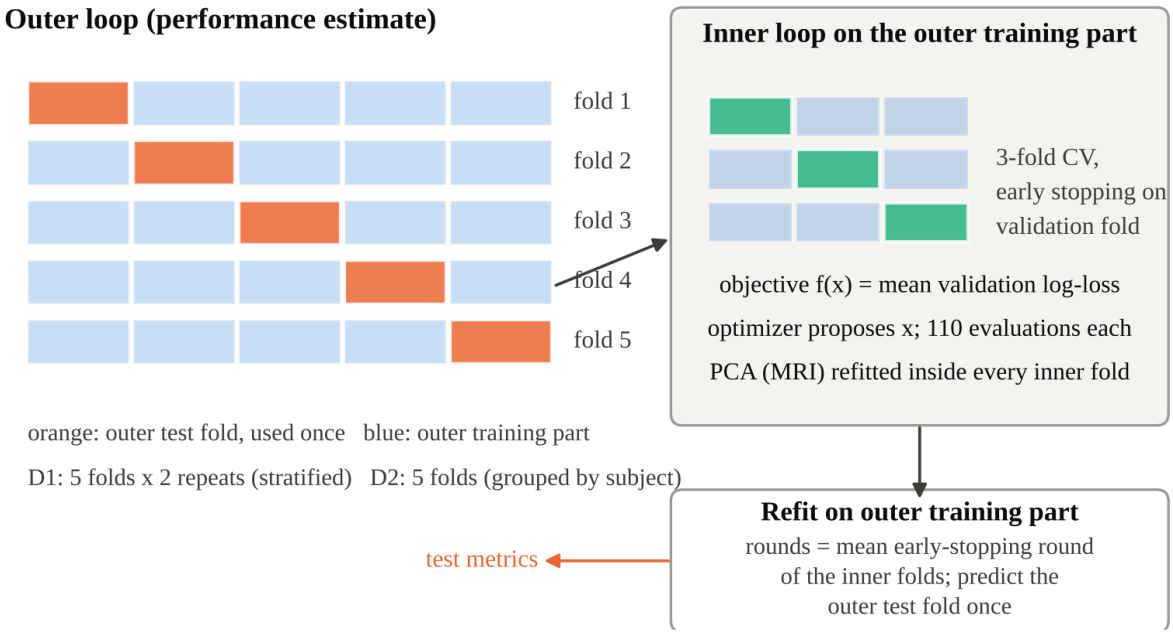


Figure 6. Mean rank (1 = best) of eight optimizers on seven randomly shifted benchmark functions, over 30 runs, for the tuning-sized setting (left) and a larger setting (right; TPE not run). HybAOA is highlighted.

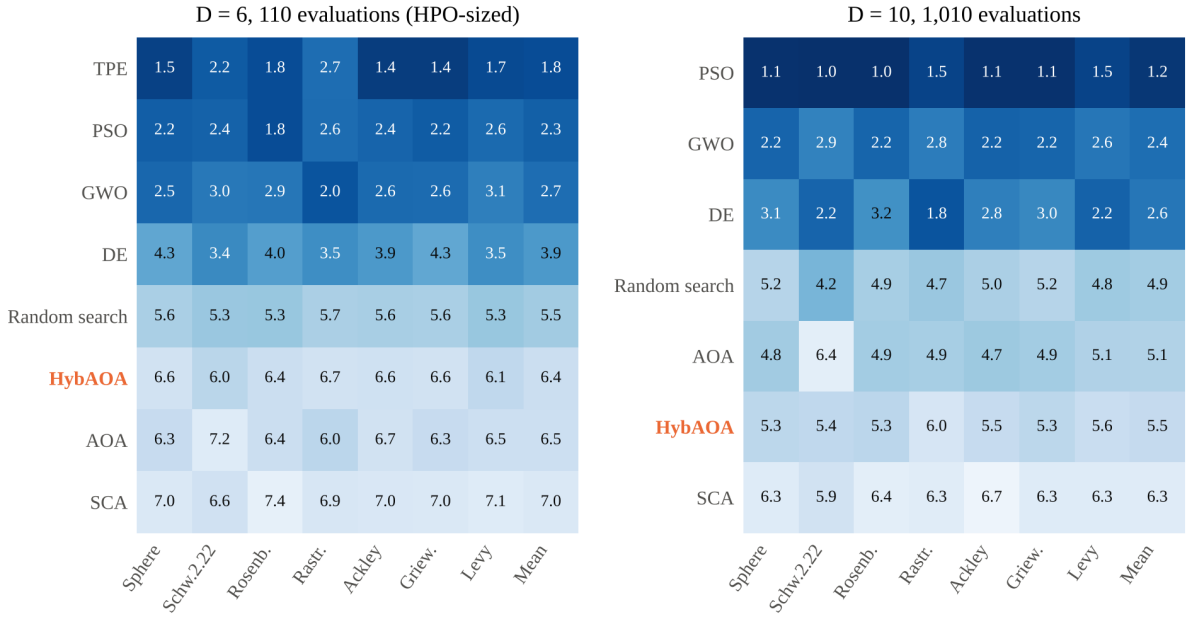


Figure 7. Median best-so-far value over 30 runs on the shifted benchmark functions (D = 10, 1,010 evaluations; logarithmic scale). HybAOA (orange) improves quickly at first and then stalls, like AOA, whereas PSO, GWO and DE continue to converge.

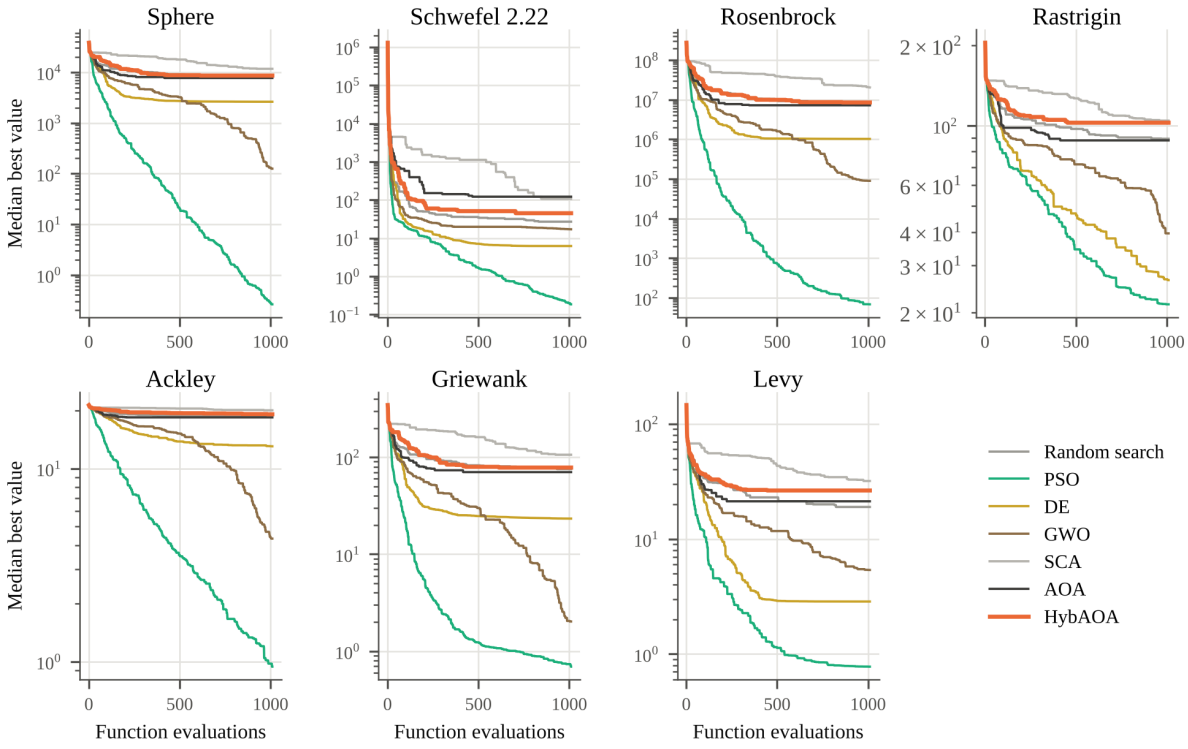


Figure 8. Structural-bias probe: final best positions (all six coordinates of 100 runs) when the objective returns uniform noise. An unbiased optimizer gives a uniform spread, as random search does; the Kolmogorov–Smirnov (KS) statistic against U(0,1) is printed above each algorithm. Arithmetic (AOA) and sine–cosine moves accumulate at the bounds; HybAOA's opposition step reduces this bias.

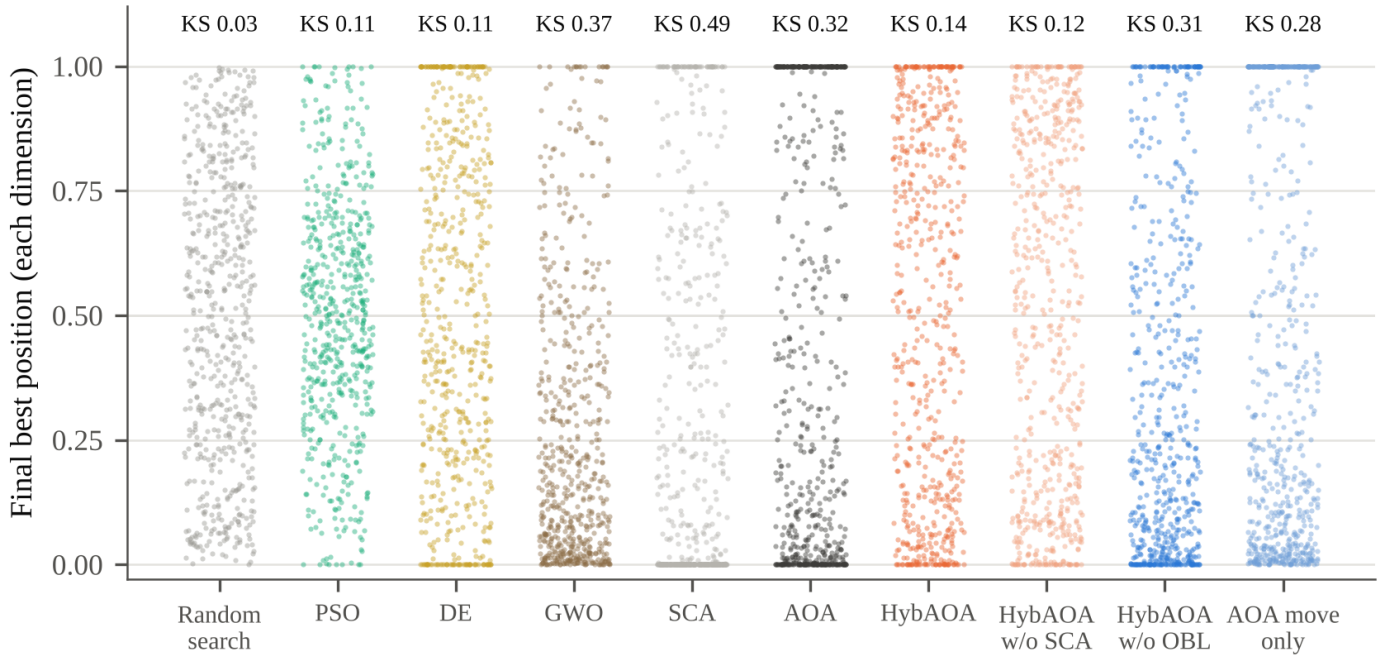


Figure 9. Ablation of HybAOA on the benchmark functions: mean rank among the four variants (1 = best) for both settings.

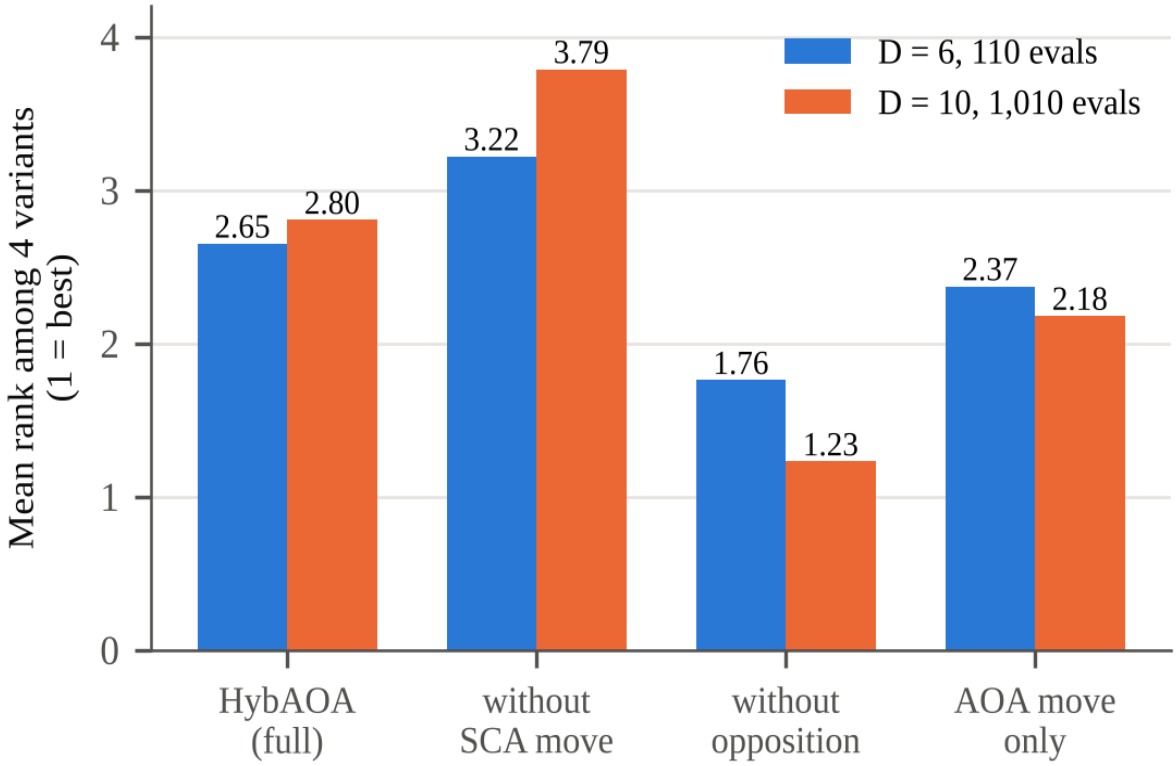


Figure 10. Outer-fold test AUC (left) and log-loss (right) of XGBoost under each tuning strategy for the four tasks. Small dots are individual outer folds; large dots and bars are means with 95 % confidence intervals. Strategies are sorted by mean performance (best at the top) and HybAOA is highlighted.

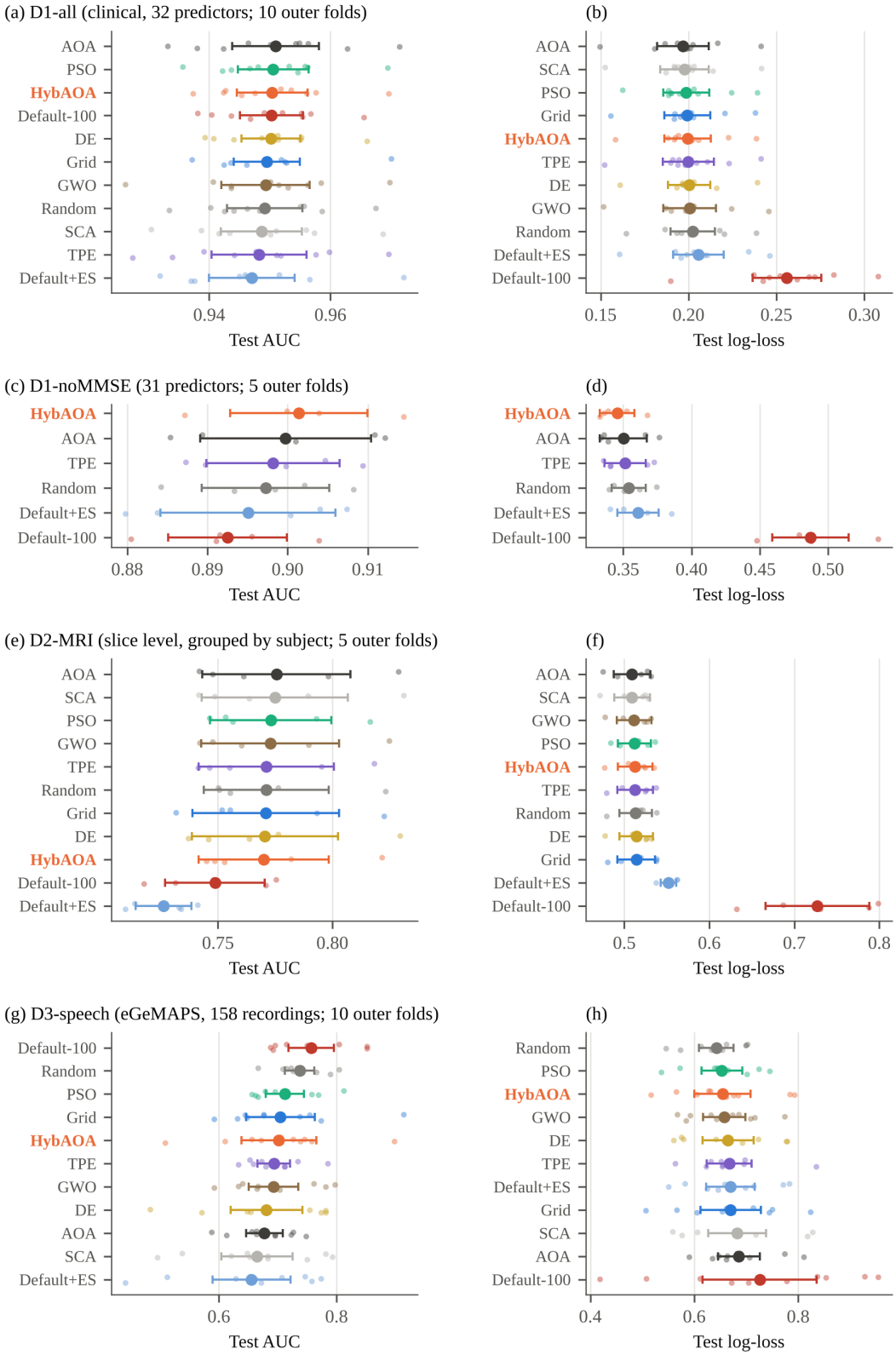


Figure 11. Mean ranks of the eleven tuning strategies across outer folds (1 = best) with the Nemenyi critical difference (CD) at α = 0.05, for test log-loss and AUC on D1-all (10 folds), D2-MRI (5 folds) and D3-speech (10 folds). Strategies whose mean ranks differ by less than the CD are not significantly different.

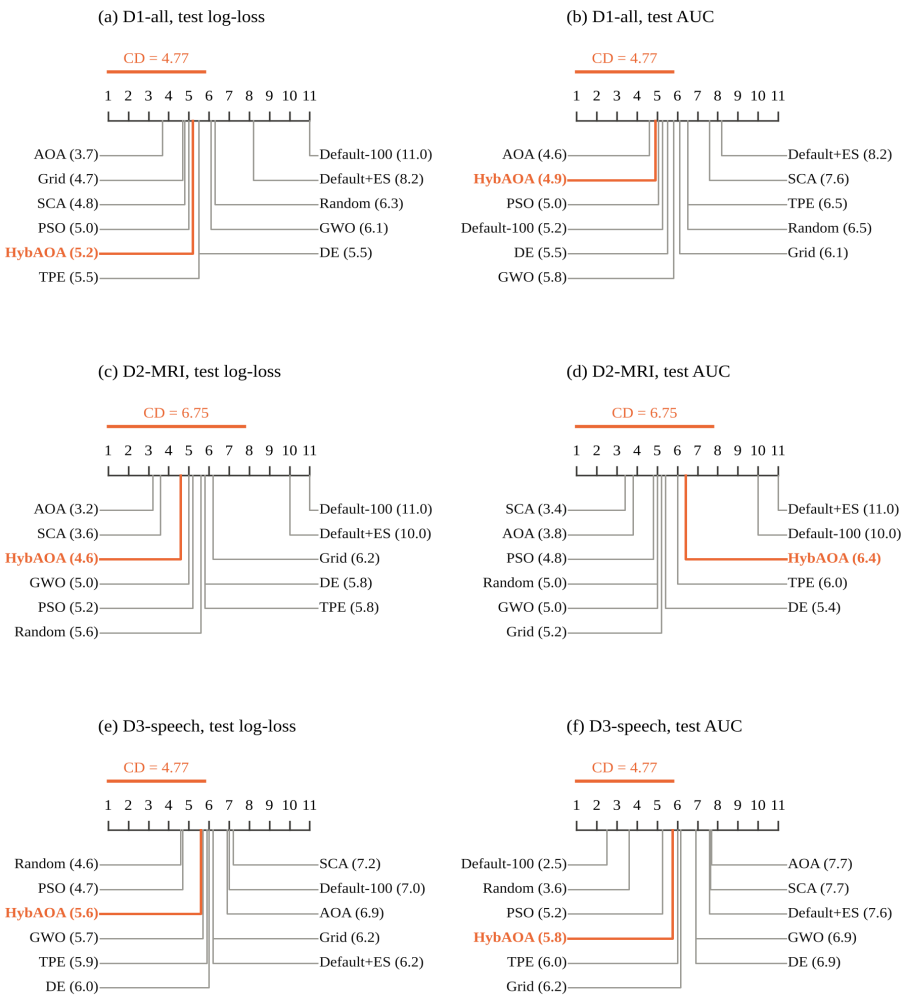


Figure 12. Median excess of the best inner-CV log-loss found so far over the best value found by any strategy in the same outer fold, as a function of the number of objective evaluations (logarithmic scale).

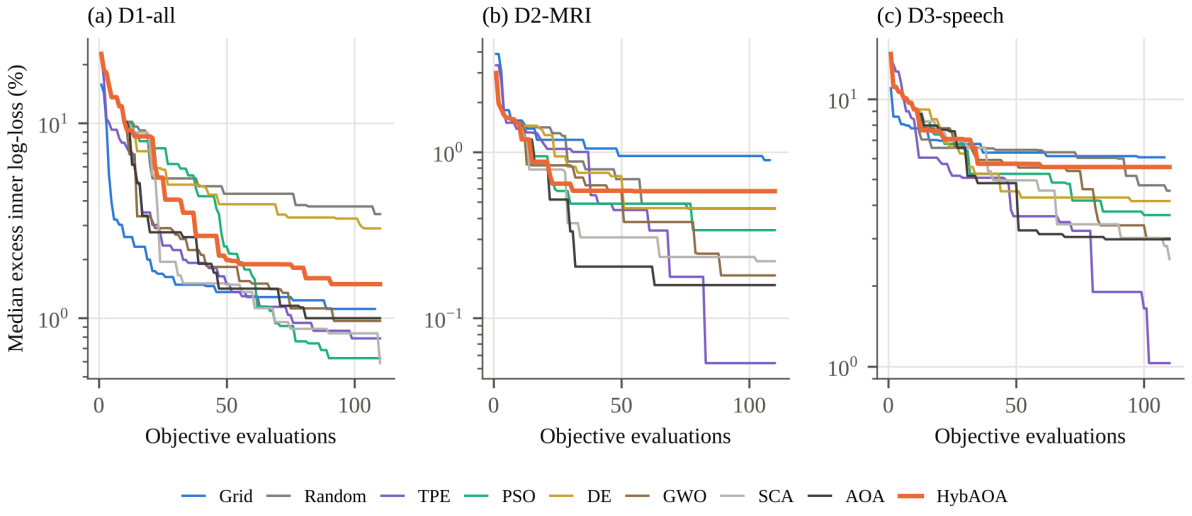


Figure 13. Mean tuning time per outer fold against mean test log-loss. On each task the nine search methods reach similar test log-loss; the 100-round default is fast but has the worst log-loss on every task, and the default with early stopping is clearly worse only on D2-MRI.

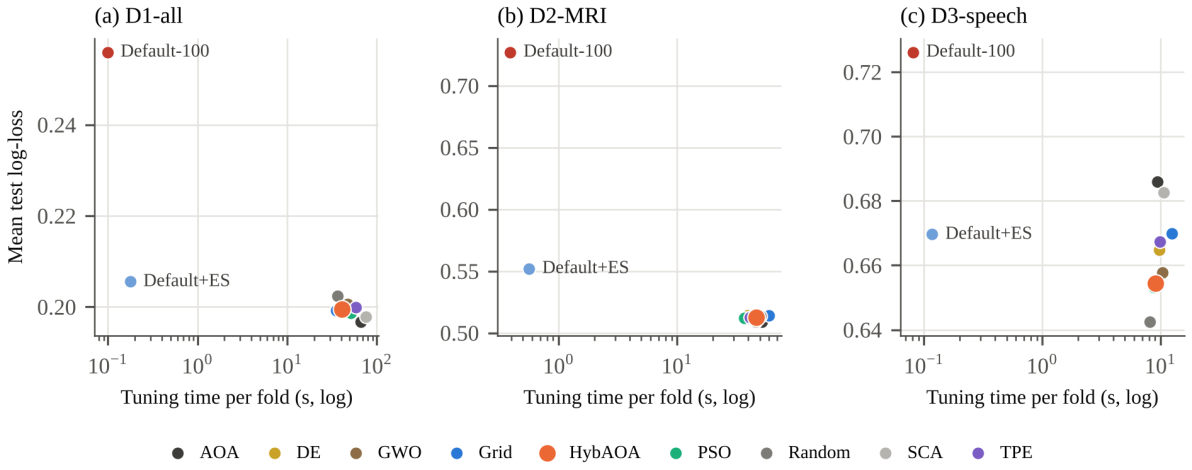


Figure 14. Hyperparameters selected on D1-all in each of the 10 outer folds by every tuning strategy. Dashed lines mark the bounds of the search space; η, minimum child weight and λ are shown on logarithmic scales.

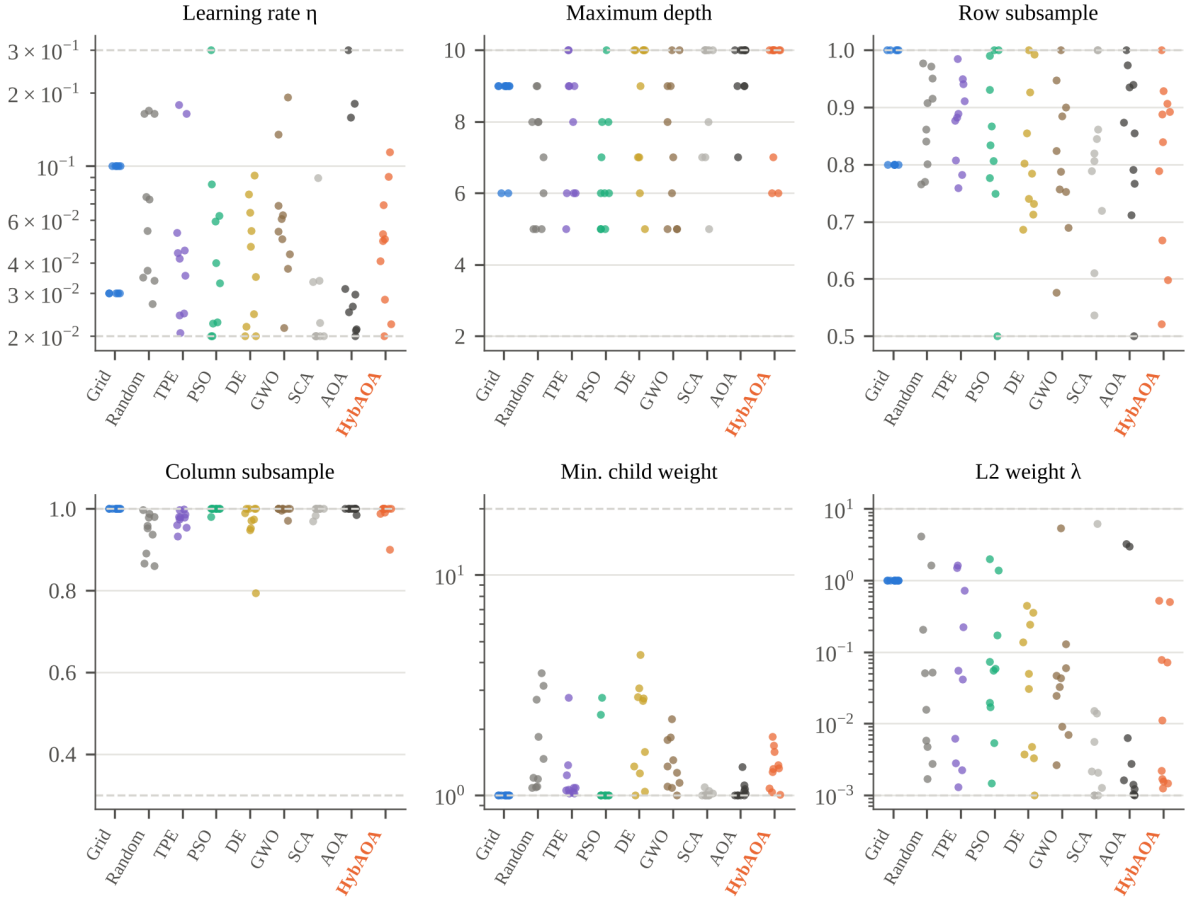


Figure 15. Receiver operating characteristic (top) and precision–recall curves (bottom) of out-of-fold predictions (first repetition) for HybAOA, the two untuned baselines, TPE and random search on D1-all (left), D2-MRI (middle) and D3-speech (right). Legends give the AUC and average precision (AP) of the pooled out-of-fold predictions.

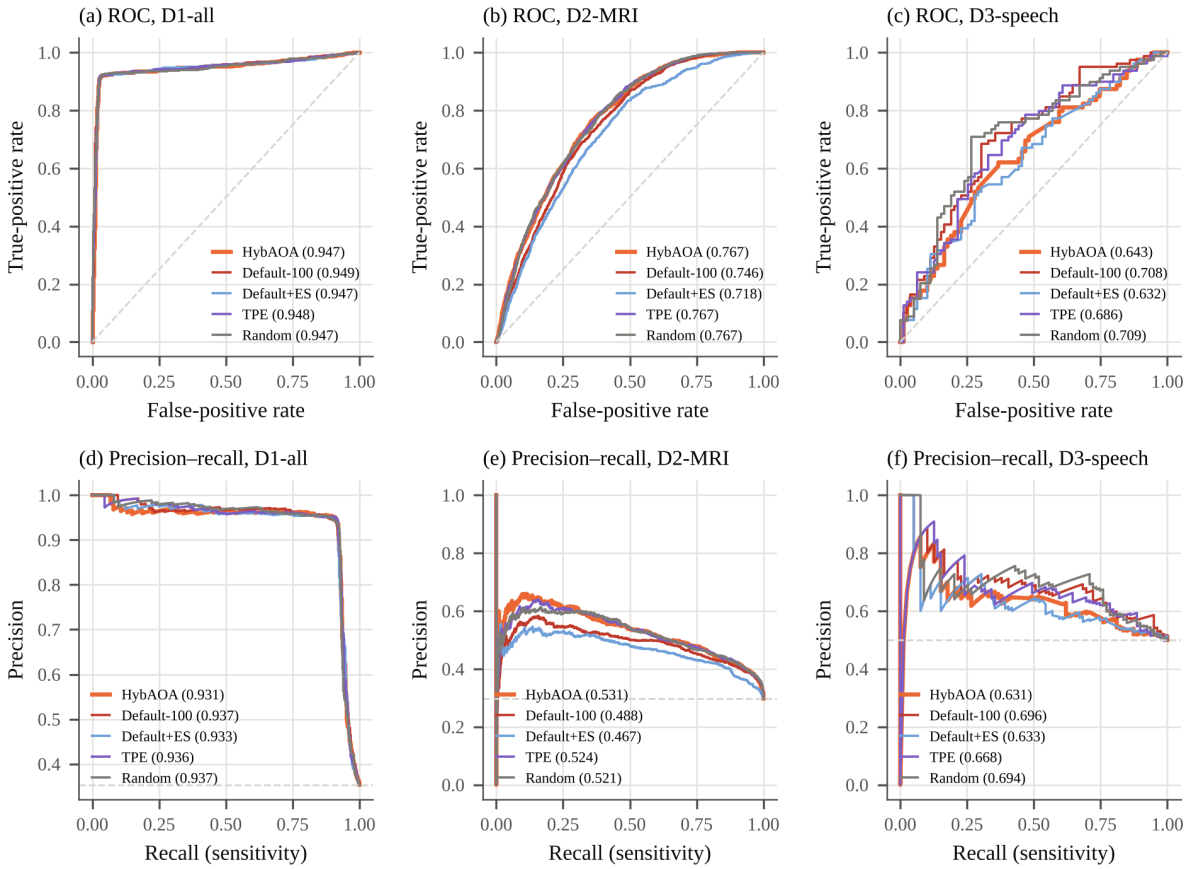


Figure 16. Reliability diagrams (10 quantile bins) of out-of-fold predictions (first repetition); legends give the Brier score. On D1-all all strategies are well calibrated on average, and the higher log-loss of the 100-round default comes from a small number of confident errors; on D2-MRI the 100-round default is clearly over-confident, whereas HybAOA and TPE give well-calibrated probabilities; on D3-speech, with 158 recordings, all curves are noisy and the 100-round default spreads its probabilities most widely.

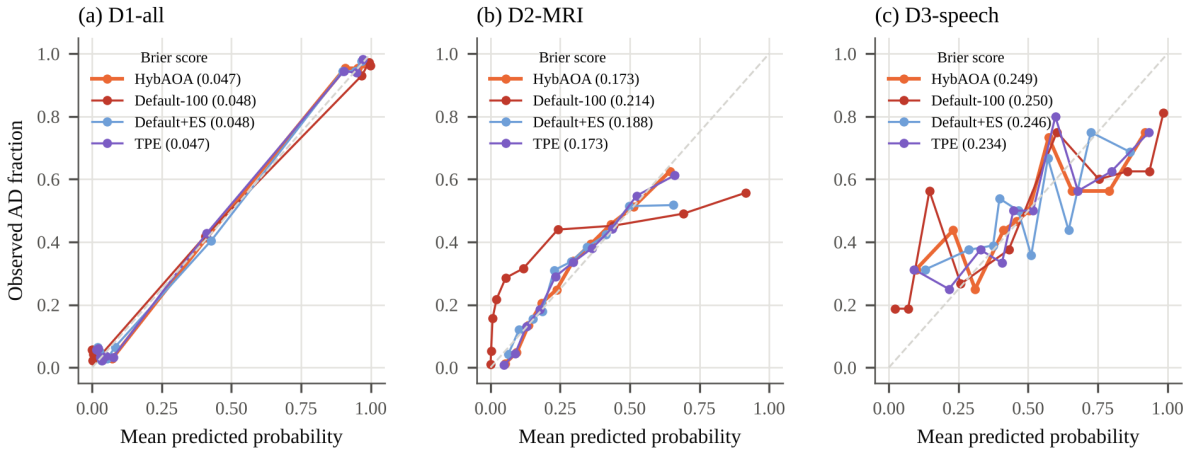


Figure 17. Confusion matrices of HybAOA-tuned XGBoost (out-of-fold predictions, first repetition, threshold 0.5) for the clinical tasks, for MRI slices, for MRI subjects (mean slice probability per subject) and for the speech recordings.

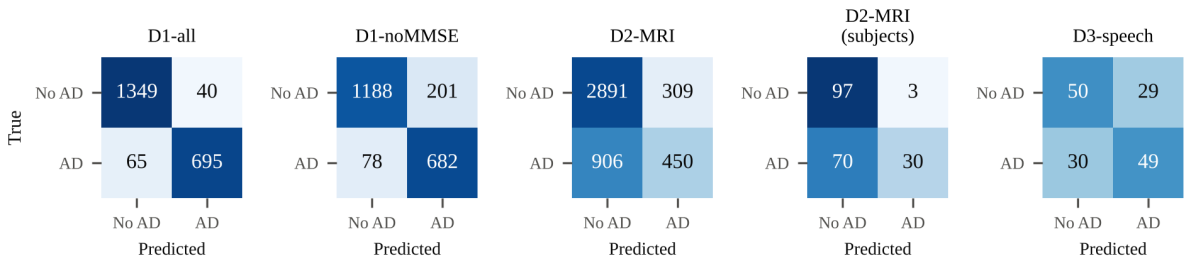


Figure 18. SHAP explanation of HybAOA-tuned XGBoost on D1-all (out-of-fold, first repetition). (a) Mean absolute SHAP value of the 15 most important predictors; orange bars are cognitive and functional assessments. (b) Distribution of SHAP values of the eight most important predictors, colored by the percentile of the predictor value.

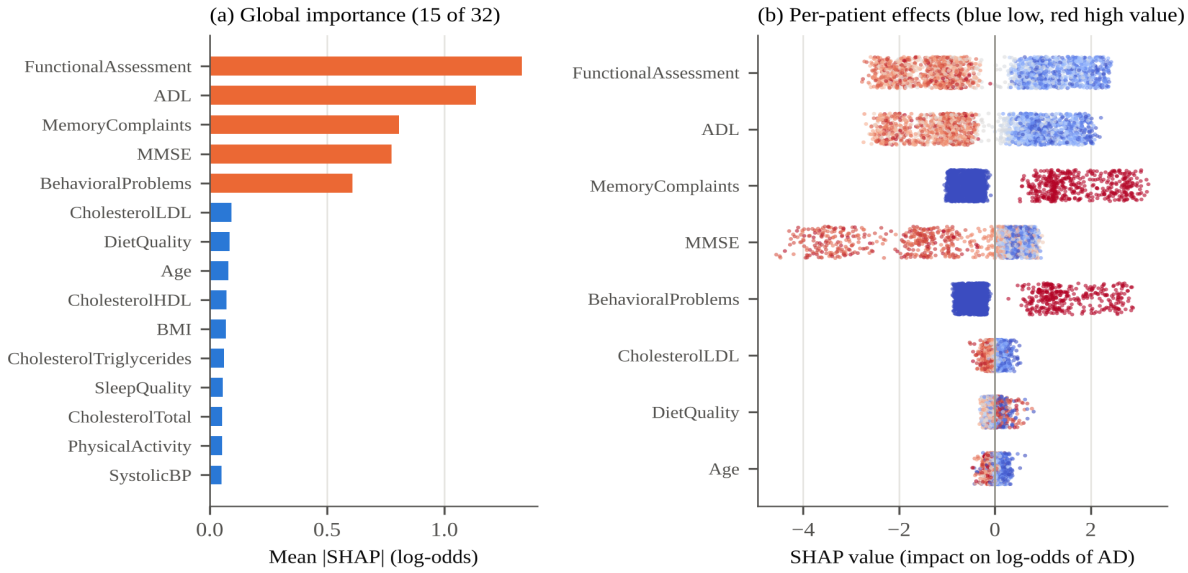


Figure 19. SHAP explanation of HybAOA-tuned XGBoost on D3-speech (out-of-fold, first repetition). (a) Mean absolute SHAP value of the 15 most important of the 88 eGeMAPS features (CoV, coefficient of variation; p80, 80th percentile; voiced, computed over voiced frames only). (b) Distribution of SHAP values of the eight most important features, colored by the percentile of the feature value. Every recording contains the interviewer's voice as well as the participant's.

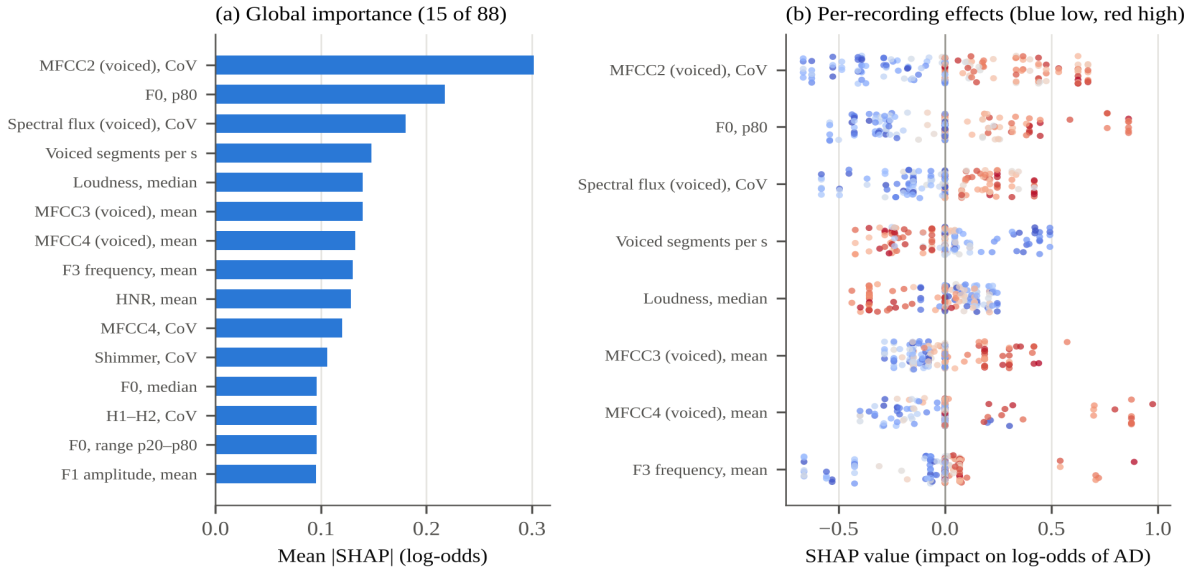


Figure 20. Negative control. Adding MRI embeddings of unrelated subjects to clinical records adds nothing when the pairing ignores the diagnosis, but raises AUC when slices are chosen to match the patient's diagnosis, most dramatically for clinical features that carry no signal on their own (b). Error bars are SDs over 25 folds.

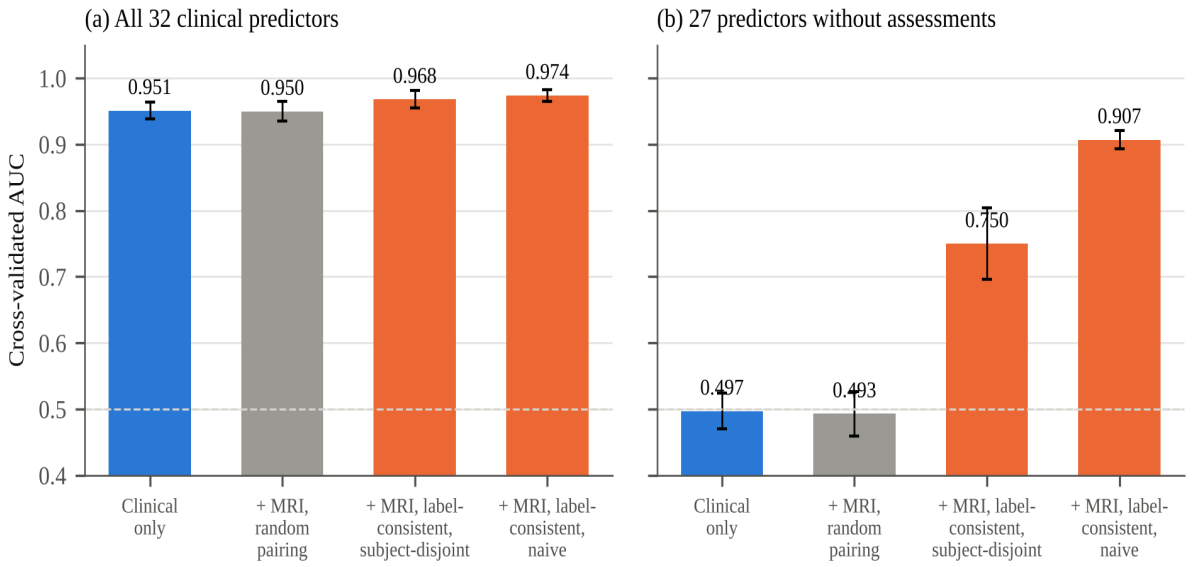
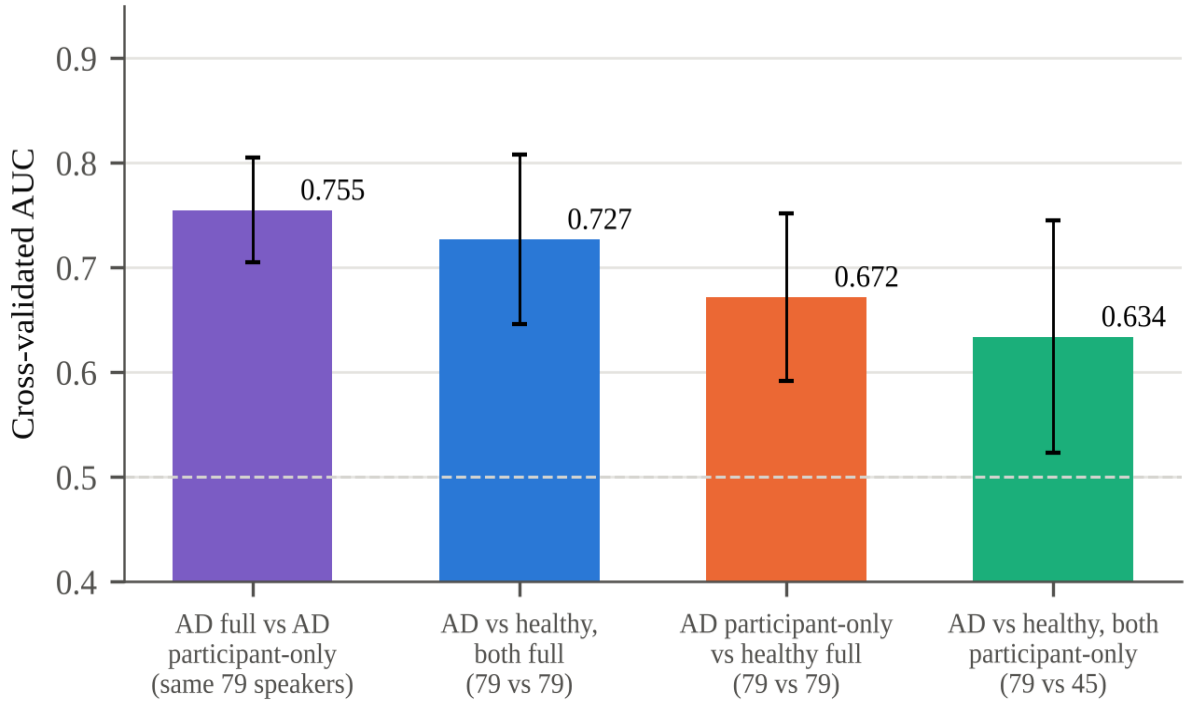


Figure 21. Speech preprocessing control. A classifier separates the full and participant-only versions of the same AD recordings (left), so the editing step leaves a detectable signature. In this sample, comparing edited AD recordings with unedited healthy recordings gave a lower AUC than the matched comparison of full interviews. Error bars are SDs over 25 folds; the dashed line marks chance level.



List of Tables

Table 1. Representative Alzheimer's disease studies that used metaheuristic tuning or heterogeneous data fusion, with the evaluation details reported.

Study	Data	Method and optimizer	Validation	Reported result
Mohammed et al. 2024 [18]	Gene expression (GEO), 569 samples	SVM tuned by crossover-enhanced GWO	Not stated; 10 repeated runs	Best accuracy 97.41 %, mean 94.03 %
Al-Tawil et al. 2025 [19]	AD dataset (source not stated in the abstract)	Parrot, GWO and DE feature selection; tree classifiers	Not stated in the abstract	Peak accuracy 0.94
Yahya et al. 2025 [12]	AD dataset (source not stated in the abstract)	XGBoost with oversampling and tuning	Not stated in the abstract	Accuracy 95.3 %
Lakshmi et al. 2026 [30]	MRI, PET, clinical and speech from heterogeneous datasets	Cross-attention CNN–BiLSTM, label-consistent feature-level fusion	Not stated in the abstract	Accuracy 92.57 %, AUC 0.973
Chaudhari and Agrawal 2024a [31]	Image and text datasets	3-D CNN and ANN	80:20 split	Training accuracy 93.57 %
Chaudhari and Agrawal 2024b [32]	Voice and transcripts	Neural classifiers	80:20 split	Test accuracy 60.61 % (audio), 72.73 % (text)
This study	Clinical (2,149 patients); MRI (200 subjects); speech (158 speakers)	XGBoost tuned by HybAOA and nine alternatives	Nested CV, subject-level for MRI	See Section 4

Table 2. Data sources, audit findings and the prediction tasks derived from them.

Source	Content as received	Audit finding	Use in this study
Clinical (D1) [54]	2,149 patients, 32 predictors, binary diagnosis (35.4 % AD)	No missing values; signatures of simulated data; MMSE is diagnosis-related	D1-all (32 predictors) and D1-noMMSE (31 predictors)
MRI slices (D2) [55]	7,840 JPEG slices in four class folders, no subject IDs	3,277 images duplicated across class folders; 4,556 unique slices from 200 recovered subjects	D2-MRI: non-demented vs demented, split by subject
Speech (D3) [33]	Copy 1: 90 WAV files (45 AD, 45 control); copy 2: 79 AD participant-only and 79 healthy full interviews; copy 3: 79 AD full interviews	Copy 1: every AD file identical to a control file; copy 2: the two classes edited differently; copy 3 matches the healthy recordings	D3-speech: 79 AD vs 79 healthy full interviews, identically converted; copy 2 used only in a control experiment
Cross-source	Patient IDs 4751–6899; image and audio file names	0 of 2,149 IDs match any image or audio file	No fusion; negative control only

Table 3. XGBoost hyperparameters, search ranges, XGBoost defaults and the values used by grid search.

Hyperparameter	Range (scale)	Encoding	Default	Grid values
Learning rate η	0.02–0.3 (log)	x_1	0.3	0.03, 0.1, 0.3
Maximum depth d	2–10 (integer)	x_2	6	3, 6, 9
Row subsample s_r	0.5–1.0 (linear)	x_3	1.0	0.6, 0.8, 1.0
Column subsample s_c	0.3–1.0 (linear)	x_4	1.0	0.5, 1.0
Minimum child weight w_{min}	1–20 (log)	x_5	1	1, 5
L2 weight λ	10^{-3}–10 (log)	x_6	1	1
Boosting rounds	≤ 500, early stopping (25)	–	100	early stopping

Table 4. Tuning strategies compared, with their settings (N = population size; budget = objective evaluations per tuning run).

Strategy	Type	Settings
Default (100 rounds)	No tuning	η 0.3, depth 6, $s_r = s_c = 1$, w_{min} 1, λ 1, 100 rounds
Default + early stopping	No tuning	Defaults; rounds from inner-fold early stopping (1 evaluation)
Grid search	Exhaustive	108 configurations (Table 3), random order
Random search [14]	Random	110 uniform draws in $[0,1]^6$
TPE [15,16]	Bayesian (model-based)	Optuna TPESampler, multivariate, 10 start-up trials, 110 trials
PSO [44,45]	Swarm	N 10, w 0.7298, $c_1 = c_2 = 1.49618$, velocity clamp 0.2
DE [46]	Evolutionary	N 10, DE/rand/1/bin, F 0.5, CR 0.9
GWO [47]	Swarm	N 10, a decreasing linearly from 2 to 0
SCA [48]	Math-inspired	N 10, amplitude decreasing linearly from 2 to 0
AOA [49]	Math-inspired	N 10, MOA 0.2→1, α 5, μ 0.499, reference implementation
HybAOA	Hybrid	N 10, T 10, α 5, μ 0.5, MOA 0.2→0.9, L = 1 (Section 3.6)
HybAOA ablations	Hybrid	Without SCA move; without opposition; AOA move only

Table 5. Pairwise comparison of HybAOA with each other optimizer on seven shifted benchmark functions: number of functions on which HybAOA wins / ties / loses (Wilcoxon signed-rank test over 30 paired runs, α = 0.05).

Competitor	D = 6, 110 evaluations	D = 10, 1,010 evaluations
Random search	0 / 0 / 7	0 / 4 / 3
TPE	0 / 0 / 7	–
PSO	0 / 0 / 7	0 / 0 / 7
DE	0 / 0 / 7	0 / 0 / 7
GWO	0 / 0 / 7	0 / 0 / 7
SCA	2 / 5 / 0	5 / 2 / 0
AOA	1 / 6 / 0	1 / 5 / 1
HybAOA w/o SCA	3 / 4 / 0	7 / 0 / 0
HybAOA w/o OBL	0 / 2 / 5	0 / 0 / 7
AOA move only	0 / 6 / 1	0 / 2 / 5

Table 6. Structural-bias probe (D = 6, 1,010 evaluations, 100 runs): Kolmogorov–Smirnov test of the final coordinates against U(0,1) and the share of coordinates at or near the bounds.

Algorithm	KS statistic	p-value	Within 0.01 of a bound (%)	In [0, 0.1) (%)	In (0.9, 1] (%)
Random search	0.027	0.77	1.8	12.0	10.0
PSO	0.114	$< 10^{-6}$	3.3	5.5	5.7
DE	0.110	$< 10^{-6}$	20.7	16.0	15.8
GWO	0.371	$< 10^{-6}$	13.5	40.5	4.2
SCA	0.489	$< 10^{-6}$	58.3	57.0	11.3
AOA	0.316	$< 10^{-6}$	47.2	41.2	24.5
HybAOA	0.139	$< 10^{-6}$	19.3	22.5	19.5
HybAOA w/o SCA	0.121	$< 10^{-6}$	14.5	21.3	18.7
HybAOA w/o OBL	0.307	$< 10^{-6}$	26.0	37.0	12.7
AOA move only	0.283	$< 10^{-6}$	32.3	35.0	24.2

Table 7. D1-all: test performance of XGBoost under each tuning strategy (mean ± SD over 10 outer folds of 5-fold CV repeated twice; threshold 0.5 for balanced accuracy, sensitivity, specificity and MCC) and mean tuning time per outer fold.

Strategy	AUC	Log-loss	Brier	Balanced accuracy	Sensitivity	Specificity	MCC	Time (s)
Default-100	0.950 ± 0.008	0.256 ± 0.031	0.048 ± 0.006	0.940 ± 0.009	0.911 ± 0.018	0.970 ± 0.008	0.888 ± 0.016	<1
Default+ES	0.947 ± 0.011	0.206 ± 0.023	0.049 ± 0.006	0.941 ± 0.009	0.912 ± 0.019	0.970 ± 0.007	0.889 ± 0.015	<1
Grid	0.949 ± 0.009	0.199 ± 0.021	0.047 ± 0.006	0.943 ± 0.007	0.913 ± 0.015	0.972 ± 0.008	0.893 ± 0.013	36
Random	0.949 ± 0.010	0.202 ± 0.020	0.048 ± 0.006	0.942 ± 0.009	0.913 ± 0.019	0.971 ± 0.008	0.891 ± 0.016	36
TPE	0.948 ± 0.013	0.200 ± 0.023	0.047 ± 0.006	0.943 ± 0.009	0.914 ± 0.019	0.972 ± 0.007	0.894 ± 0.015	58
PSO	0.951 ± 0.009	0.199 ± 0.021	0.047 ± 0.006	0.942 ± 0.007	0.914 ± 0.014	0.969 ± 0.007	0.890 ± 0.013	51
DE	0.950 ± 0.008	0.200 ± 0.019	0.048 ± 0.006	0.940 ± 0.009	0.909 ± 0.017	0.971 ± 0.007	0.888 ± 0.015	46
GWO	0.949 ± 0.012	0.201 ± 0.024	0.048 ± 0.007	0.942 ± 0.009	0.914 ± 0.018	0.970 ± 0.008	0.891 ± 0.016	48
SCA	0.949 ± 0.011	0.198 ± 0.022	0.047 ± 0.006	0.943 ± 0.009	0.914 ± 0.019	0.971 ± 0.007	0.892 ± 0.015	76
AOA	0.951 ± 0.012	0.197 ± 0.024	0.047 ± 0.006	0.942 ± 0.010	0.912 ± 0.020	0.971 ± 0.007	0.891 ± 0.016	66
HybAOA	0.950 ± 0.009	0.200 ± 0.021	0.047 ± 0.006	0.942 ± 0.009	0.914 ± 0.017	0.970 ± 0.008	0.892 ± 0.015	41

Table 8. D1-noMMSE: test performance of XGBoost without the MMSE predictor (mean ± SD over 5 outer folds).

Strategy	AUC	Log-loss	Brier	Balanced accuracy	Sensitivity	Specificity	MCC	Time (s)
Default-100	0.893 ± 0.008	0.487 ± 0.032	0.124 ± 0.009	0.835 ± 0.011	0.799 ± 0.034	0.871 ± 0.024	0.666 ± 0.020	<1
Default+ES	0.895 ± 0.012	0.361 ± 0.017	0.110 ± 0.007	0.855 ± 0.014	0.846 ± 0.034	0.863 ± 0.030	0.697 ± 0.028	<1
Random	0.897 ± 0.009	0.354 ± 0.014	0.107 ± 0.006	0.857 ± 0.012	0.854 ± 0.031	0.860 ± 0.026	0.699 ± 0.025	28
TPE	0.898 ± 0.009	0.351 ± 0.017	0.106 ± 0.008	0.867 ± 0.016	0.875 ± 0.015	0.858 ± 0.023	0.716 ± 0.034	48
AOA	0.900 ± 0.012	0.350 ± 0.020	0.106 ± 0.008	0.870 ± 0.017	0.882 ± 0.021	0.857 ± 0.022	0.720 ± 0.036	40
HybAOA	0.901 ± 0.010	0.346 ± 0.015	0.104 ± 0.006	0.876 ± 0.010	0.897 ± 0.011	0.855 ± 0.025	0.732 ± 0.023	28

Table 9. D2-MRI: test performance of XGBoost on ResNet-18 slice embeddings under each tuning strategy (mean ± SD over 5 subject-grouped outer folds). Slice-level measures use every slice; subject-level measures average the slice probabilities of each subject.

Strategy	Slice AUC	Slice log-loss	Slice balanced accuracy	Slice sensitivity	Slice specificity	Subject AUC	Subject balanced accuracy	Time (s)
Default-100	0.749 ± 0.025	0.727 ± 0.070	0.618 ± 0.029	0.389 ± 0.066	0.847 ± 0.026	0.838 ± 0.057	0.679 ± 0.065	<1
Default+ES	0.726 ± 0.014	0.552 ± 0.010	0.579 ± 0.023	0.255 ± 0.078	0.903 ± 0.037	0.847 ± 0.051	0.564 ± 0.050	<1
Grid	0.771 ± 0.036	0.514 ± 0.026	0.610 ± 0.043	0.316 ± 0.071	0.903 ± 0.027	0.864 ± 0.059	0.634 ± 0.094	59
Random	0.771 ± 0.031	0.513 ± 0.022	0.621 ± 0.033	0.343 ± 0.052	0.899 ± 0.025	0.862 ± 0.048	0.639 ± 0.061	51
TPE	0.771 ± 0.034	0.513 ± 0.024	0.623 ± 0.032	0.349 ± 0.064	0.897 ± 0.020	0.869 ± 0.050	0.642 ± 0.082	41
PSO	0.773 ± 0.030	0.512 ± 0.022	0.611 ± 0.038	0.319 ± 0.067	0.904 ± 0.014	0.867 ± 0.049	0.643 ± 0.101	37
DE	0.771 ± 0.036	0.514 ± 0.022	0.613 ± 0.041	0.323 ± 0.066	0.904 ± 0.020	0.863 ± 0.053	0.650 ± 0.046	39
GWO	0.773 ± 0.034	0.512 ± 0.023	0.611 ± 0.038	0.312 ± 0.067	0.909 ± 0.023	0.860 ± 0.053	0.639 ± 0.067	41
SCA	0.775 ± 0.036	0.509 ± 0.024	0.625 ± 0.043	0.349 ± 0.076	0.902 ± 0.017	0.860 ± 0.048	0.659 ± 0.087	46
AOA	0.776 ± 0.037	0.509 ± 0.025	0.624 ± 0.041	0.341 ± 0.071	0.906 ± 0.014	0.866 ± 0.061	0.649 ± 0.077	52
HybAOA	0.770 ± 0.032	0.513 ± 0.023	0.618 ± 0.045	0.332 ± 0.082	0.903 ± 0.017	0.862 ± 0.048	0.633 ± 0.082	47

Table 10. D3-speech: test performance of XGBoost on eGeMAPS acoustic features under each tuning strategy (mean ± SD over 10 outer folds of 5-fold CV repeated twice; 79 AD and 79 healthy speakers; threshold 0.5 for balanced accuracy, sensitivity, specificity and MCC) and mean tuning time per outer fold.

Strategy	AUC	Log-loss	Brier	Balanced accuracy	Sensitivity	Specificity	MCC	Time (s)
Default-100	0.757 ± 0.062	0.726 ± 0.177	0.227 ± 0.045	0.692 ± 0.062	0.697 ± 0.130	0.688 ± 0.151	0.396 ± 0.124	<1
Default+ES	0.655 ± 0.107	0.670 ± 0.075	0.233 ± 0.031	0.622 ± 0.108	0.612 ± 0.198	0.632 ± 0.089	0.246 ± 0.221	<1
Grid	0.705 ± 0.094	0.670 ± 0.094	0.229 ± 0.037	0.645 ± 0.089	0.640 ± 0.140	0.650 ± 0.165	0.300 ± 0.181	12
Random	0.737 ± 0.040	0.642 ± 0.054	0.220 ± 0.018	0.698 ± 0.047	0.671 ± 0.144	0.725 ± 0.152	0.411 ± 0.092	8
TPE	0.693 ± 0.045	0.667 ± 0.070	0.230 ± 0.020	0.645 ± 0.042	0.633 ± 0.127	0.657 ± 0.090	0.296 ± 0.083	10
PSO	0.712 ± 0.052	0.653 ± 0.064	0.226 ± 0.023	0.654 ± 0.048	0.639 ± 0.112	0.670 ± 0.094	0.312 ± 0.094	9
DE	0.681 ± 0.098	0.665 ± 0.079	0.232 ± 0.031	0.637 ± 0.076	0.631 ± 0.138	0.643 ± 0.139	0.282 ± 0.156	10
GWO	0.693 ± 0.068	0.658 ± 0.066	0.229 ± 0.025	0.637 ± 0.069	0.626 ± 0.136	0.649 ± 0.169	0.283 ± 0.140	10
SCA	0.665 ± 0.097	0.683 ± 0.090	0.239 ± 0.034	0.631 ± 0.069	0.587 ± 0.122	0.675 ± 0.144	0.268 ± 0.139	11
AOA	0.677 ± 0.051	0.686 ± 0.065	0.238 ± 0.020	0.626 ± 0.053	0.634 ± 0.125	0.618 ± 0.149	0.260 ± 0.109	9
HybAOA	0.702 ± 0.102	0.654 ± 0.087	0.226 ± 0.034	0.674 ± 0.082	0.672 ± 0.191	0.676 ± 0.153	0.361 ± 0.161	9

Table 11. HybAOA compared with every other strategy: mean difference in test log-loss and AUC (HybAOA minus comparator; negative log-loss and positive AUC differences favor HybAOA), followed in parentheses by the Holm-adjusted Wilcoxon p-value and the corrected resampled t-test p-value.

Comparator	D1-all log-loss	D1-all AUC	D2-MRI log-loss	D2-MRI AUC	D3-speech log-loss	D3-speech AUC
Default-100	−0.056 (0.020; <0.001)	0.000 (1.000; 0.971)	−0.214 (0.625; 0.004)	+0.021 (0.625; 0.134)	−0.072 (1.000; 0.396)	−0.055 (0.605; 0.258)
Default+ES	−0.006 (0.123; 0.105)	+0.003 (0.273; 0.188)	−0.039 (0.625; 0.040)	+0.044 (0.625; 0.066)	−0.015 (1.000; 0.795)	+0.046 (1.000; 0.480)
Grid	0.000 (1.000; 0.913)	+0.001 (1.000; 0.626)	−0.001 (1.000; 0.815)	−0.001 (1.000; 0.909)	−0.015 (1.000; 0.602)	−0.003 (1.000; 0.920)
Random	−0.003 (1.000; 0.544)	+0.001 (1.000; 0.361)	−0.001 (1.000; 0.812)	−0.001 (1.000; 0.716)	+0.012 (1.000; 0.733)	−0.036 (1.000; 0.577)
TPE	0.000 (1.000; 0.897)	+0.002 (1.000; 0.477)	0.000 (1.000; 0.979)	−0.001 (1.000; 0.813)	−0.013 (1.000; 0.740)	+0.008 (1.000; 0.860)
PSO	+0.001 (1.000; 0.693)	0.000 (1.000; 0.879)	+0.001 (1.000; 0.880)	−0.003 (1.000; 0.539)	+0.001 (1.000; 0.978)	−0.010 (1.000; 0.857)
DE	−0.001 (1.000; 0.732)	0.000 (1.000; 0.927)	−0.001 (1.000; 0.543)	−0.001 (1.000; 0.928)	−0.010 (1.000; 0.758)	+0.021 (1.000; 0.583)
GWO	−0.001 (1.000; 0.713)	+0.001 (1.000; 0.696)	+0.001 (1.000; 0.651)	−0.003 (1.000; 0.501)	−0.003 (1.000; 0.938)	+0.009 (1.000; 0.856)
SCA	+0.002 (1.000; 0.481)	+0.002 (0.334; 0.290)	+0.004 (1.000; 0.205)	−0.005 (1.000; 0.308)	−0.028 (0.645; 0.265)	+0.037 (0.756; 0.303)
AOA	+0.003 (0.844; 0.343)	−0.001 (1.000; 0.785)	+0.004 (1.000; 0.374)	−0.006 (1.000; 0.381)	−0.031 (1.000; 0.490)	+0.024 (1.000; 0.613)

Table 12. D1-all: hyperparameters selected by each tuning strategy across the 10 outer folds, as median [minimum–maximum], and the share of selected coordinates lying exactly on a bound of the search space.

Strategy	η	Depth	Row subsample	Column subsample	Min. child weight	λ	At bound (%)
Grid	0.065 [0.030–0.100]	9 [6–9]	0.90 [0.80–1.00]	1.00 [1.00–1.00]	1.0 [1.0–1.0]	1 [1–1]	42
Random	0.064 [0.027–0.169]	8 [5–9]	0.88 [0.77–0.98]	0.96 [0.86–1.00]	1.3 [1.1–3.6]	0.0329 [0.0017–4.14]	0
TPE	0.043 [0.021–0.178]	8 [5–10]	0.89 [0.76–0.98]	0.98 [0.93–1.00]	1.1 [1.0–2.8]	0.0487 [0.0013–1.61]	3
PSO	0.037 [0.020–0.300]	6 [5–10]	0.85 [0.50–1.00]	1.00 [0.98–1.00]	1.0 [1.0–2.8]	0.0569 [0.00146–2.01]	40
DE	0.041 [0.020–0.092]	10 [5–10]	0.79 [0.69–1.00]	0.98 [0.79–1.00]	2.1 [1.0–4.3]	0.0402 [0.001–0.446]	23
GWO	0.057 [0.022–0.192]	8 [5–10]	0.81 [0.58–1.00]	1.00 [0.97–1.00]	1.3 [1.0–2.2]	0.0378 [0.00265–5.4]	18
SCA	0.020 [0.020–0.090]	10 [5–10]	0.81 [0.54–1.00]	1.00 [0.97–1.00]	1.0 [1.0–1.1]	0.00211 [0.001–6.14]	52
AOA	0.028 [0.020–0.300]	10 [7–10]	0.86 [0.50–1.00]	1.00 [0.98–1.00]	1.0 [1.0–1.3]	0.00152 [0.001–3.26]	42
HybAOA	0.050 [0.020–0.114]	10 [6–10]	0.86 [0.52–1.00]	1.00 [0.90–1.00]	1.3 [1.0–1.8]	0.00668 [0.00124–0.517]	27

Table 13. D1-all ablation of HybAOA (first repetition, 5 outer folds): best inner-CV log-loss reached, and test log-loss, AUC and balanced accuracy of the refitted model (mean ± SD).

Variant	Inner log-loss	Test log-loss	Test AUC	Test balanced accuracy	Time (s)
HybAOA	0.206 ± 0.005	0.199 ± 0.006	0.948 ± 0.006	0.943 ± 0.003	44
HybAOA w/o SCA	0.208 ± 0.006	0.200 ± 0.006	0.950 ± 0.005	0.942 ± 0.006	37
HybAOA w/o OBL	0.206 ± 0.006	0.200 ± 0.010	0.948 ± 0.008	0.942 ± 0.004	57
AOA move only	0.205 ± 0.004	0.200 ± 0.008	0.949 ± 0.005	0.943 ± 0.003	65

Table 14. D3-speech ablation of HybAOA (all 10 outer folds): best inner-CV log-loss reached, and test log-loss, AUC and balanced accuracy of the refitted model (mean ± SD).

Variant	Inner log-loss	Test log-loss	Test AUC	Test balanced accuracy	Time (s)
HybAOA	0.584 ± 0.042	0.654 ± 0.087	0.702 ± 0.102	0.674 ± 0.082	9
HybAOA w/o SCA	0.582 ± 0.041	0.655 ± 0.081	0.710 ± 0.093	0.676 ± 0.077	8
HybAOA w/o OBL	0.574 ± 0.041	0.702 ± 0.125	0.673 ± 0.115	0.631 ± 0.096	10
AOA move only	0.571 ± 0.047	0.677 ± 0.079	0.685 ± 0.102	0.644 ± 0.095	9

Table 15. Negative-control fusion experiment: default XGBoost on clinical predictors combined with MRI embeddings of unrelated subjects (stratified 5-fold CV repeated five times; mean ± SD).

Clinical features	Added block	Pairing	MRI slice pool	AUC	Balanced accuracy	Accuracy
32 clinical	–	–	–	0.951 ± 0.013	0.940 ± 0.013	0.948
32 clinical	MRI (32 PCs)	random	naive	0.950 ± 0.015	0.941 ± 0.012	0.949
32 clinical	MRI (32 PCs)	label-consistent	naive	0.974 ± 0.009	0.944 ± 0.013	0.952
32 clinical	MRI (32 PCs)	label-consistent	subject-disjoint	0.968 ± 0.013	0.939 ± 0.013	0.948
27 clinical (no assessments)	–	–	–	0.497 ± 0.027	0.502 ± 0.012	0.588
27 clinical (no assessments)	MRI (32 PCs)	random	naive	0.493 ± 0.034	0.494 ± 0.019	0.583
27 clinical (no assessments)	MRI (32 PCs)	label-consistent	naive	0.907 ± 0.014	0.809 ± 0.022	0.835
27 clinical (no assessments)	MRI (32 PCs)	label-consistent	subject-disjoint	0.750 ± 0.054	0.635 ± 0.046	0.689
none	MRI (32 PCs)	random	naive	0.502 ± 0.025	0.500 ± 0.014	0.576
none	MRI (32 PCs)	label-consistent	naive	0.920 ± 0.012	0.817 ± 0.023	0.842
none	MRI (32 PCs)	label-consistent	subject-disjoint	0.754 ± 0.051	0.645 ± 0.044	0.697

Table 16. Speech preprocessing control: default XGBoost on eGeMAPS features (stratified 5-fold CV repeated five times, grouped by speaker in the first comparison; mean ± SD over 25 folds). Participant-only recordings are versions from which the interviewer's turns were removed.

Comparison	What separates the classes	AUC	Balanced accuracy
AD full interviews vs participant-only versions of the same 79 speakers	processing (not diagnosis)	0.755 ± 0.050	0.678 ± 0.060
AD participant-only (79) vs healthy full interviews (79)	diagnosis, mixed processing	0.672 ± 0.080	0.623 ± 0.084
AD full (79) vs healthy full interviews (79)	diagnosis, matched	0.727 ± 0.081	0.671 ± 0.069
AD participant-only (79) vs healthy participant-only (45)	diagnosis, matched	0.634 ± 0.111	0.595 ± 0.120

REFERENCES

1. Nichols, E., Steinmetz, J. D., Vollset, S. E., Fukutaki, K., Chalek, J., Abd-Allah, F., ... & Liu, X. (2022). Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. The Lancet Public Health, 7(2), e105-e125.

2. Holston, E. C. (2024). A Commentary: Invisibility of Older African-American Adults in Electrophysiological Research on Alzheimer's Disease. Ethnicity & Disease, 34(4), 181-184.

3. Sterr, P. A., Elbert, P. T., Berthold, D. I., Kölbel, D. S., Rockstroh, P. B., & Taub, P. E. (2002). Longer versus shorter daily constraint-induced movement therapy of chronic hemiparesis: an exploratory study. Archives of physical medicine and rehabilitation, 83(10), 1374-1377.

4. McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., ... & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & dementia, 7(3), 263-269.

5. Jo, T., Nho, K., & Saykin, A. J. (2019). Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. Frontiers in aging neuroscience, 11, 220.

6. Tanveer, M., Richhariya, B., Khan, R. U., Rashid, A. H., Khanna, P., Prasad, M., & Lin, C. T. (2020). Machine learning techniques for the diagnosis of Alzheimer's disease: A review. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(1s), 1-35.

7. Elazab, A., Wang, C., Abdelaziz, M., Zhang, J., Gu, J., Gorriz, J. M., ... & Chang, C. (2024). Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: Achievements and future directions. Expert Systems with Applications, 255, 124780.

8. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

9. Fouladvand, S., Noshad, M., Goldstein, M. K., Periyakoil, V. J., & Chen, J. H. (2023). Mild cognitive impairment: data-driven prediction, risk factors, and workup. AMIA Summits on Translational Science Proceedings, 2023, 167.

10. Fouladvand, S., Noshad, M., Goldstein, M. K., Periyakoil, V. J., & Chen, J. H. (2023). Mild cognitive impairment: data-driven prediction, risk factors, and workup. AMIA Summits on Translational Science Proceedings, 2023, 167.

11. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information fusion, 81, 84-90.

12. Yahya, F. N., Anshori, M., & Khudori, A. N. (2025). Evaluasi Performa XGBoost dengan Oversampling dan Hyperparameter Tuning untuk Prediksi Alzheimer. Techno. com, 24(1), 1.

13. Yahya, F. N., Anshori, M., & Khudori, A. N. (2025). Evaluasi Performa XGBoost dengan Oversampling dan Hyperparameter Tuning untuk Prediksi Alzheimer. Techno. com, 24(1), 1.

14. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

15. Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. Advances in neural information processing systems, 24.

16. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

17. Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations,

algorithms, best practices, and open challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13(2), e1484.

18. Mohammed, H. A. A., Jizany, A. A. K., Mahmood, I. M., & Kadhim, Q. K. (2024). Predicting Alzheimer's Disease Using a Modified Grey Wolf Optimizer and Support Vector Machine. Ingenierie des Systemes d'Information, 29(2), 669.

19. Al-Tawil A, Al-Muhtaseb W, Almazaydeh L, Fathi H (2025) Enhancing Alzheimer's disease diagnosis through metaheuristic feature selection and advanced classification techniques. Int J Electr Comput Eng 15:3382–3395, https://doi.org/10.11591/ijece.v15i3.pp3382-3395

20. Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67-82.

21. Sörensen, K. (2015). Metaheuristics—the metaphor exposed. International Transactions in Operational Research, 22(1), 3-18.

22. [22 Camacho Villalón, C. L., Stützle, T., & Dorigo, M. (2020, October). Grey wolf, firefly and bat algorithms: Three widespread algorithms that do not contain any novelty. In International conference on swarm intelligence (pp. 121-133). Cham: Springer International Publishing.

23. Kononova, A. V., Corne, D. W., De Wilde, P., Shneer, V., & Caraffini, F. (2015). Structural bias in population-based algorithms. Information Sciences, 298, 468-490.

24. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC bioinformatics, 7(1), 91.

25. Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, 11, 2079-2107.

26. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(4), 1-21.

27. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9).

28. Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., ... & Alzheimer's Disease Neuroimaging Initiative. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Medical image analysis, 63, 101694.

29. Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., ... & Alzheimer's Disease Neuroimaging Initiative. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Medical image analysis, 63, 101694.

30. Lakshmi, S. G., Kumar, C., Kumar, P., Kunwar, F. B., & Singh, S. K. (2026). Multimodal Alzheimer's Disease Classification From Heterogeneous Datasets Using Multihead Cross-Attention CNN–BiLSTM With BAM. Applied Computational Intelligence and Soft Computing, 2026(1), 8280381.

31. Chaudhari, M. M., & Agrawal, M. (2026). MACHINE LEARNING FOR ALZHEIMER'S DISEASE PREDICTION: A CRITICAL SURVEY AND LEAKAGE-AWARE COMPARATIVE STUDY ON TABULAR CLINICAL AND MRI DATASETS. ACTA INNOVATIONS, 1074-1097.

32. Chaudhari, M. M., & Agrawal, M. (2026). MACHINE LEARNING FOR ALZHEIMER'S DISEASE PREDICTION: A CRITICAL SURVEY AND LEAKAGE-AWARE COMPARATIVE STUDY ON TABULAR CLINICAL AND MRI DATASETS. ACTA INNOVATIONS, 1074-1097.

33. Luz, S., Haider, F., De la Fuente, S., Fromm, D., & MacWhinney, B. (2021). Detecting cognitive decline using speech only: The ADReSS O Challenge. MedRxiv, 2021-03.

34. Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). Multimodal deep learning models for early detection of Alzheimer's disease stage. Scientific reports, 11(1), 3254.

35. Golovanevsky, M., Eickhoff, C., & Singh, R. (2022). Multimodal attention-based deep learning for Alzheimer's disease diagnosis. Journal of the American Medical Informatics Association, 29(12), 2014-2022.

36. Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., ... & Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 27(4), 685-691.

37. Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle

aged, nondemented, and demented older adults. Journal of cognitive neuroscience, 19(9), 1498-1507.

38. Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of neurology, 51(6), 585-594.

39. Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's disease, 49(2), 407-422.

40. Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & MacWhinney, B. (2021). Alzheimer's dementia recognition through spontaneous speech. Frontiers in computer science, 3, 780169.

41. Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & MacWhinney, B. (2021). Alzheimer's dementia recognition through spontaneous speech. Frontiers in computer science, 3, 780169.

42. Chen, J., Ye, J., Tang, F., & Zhou, J. (2021, August). Automatic detection of Alzheimer's disease using spontaneous speech only. In Interspeech (Vol. 2021, p. 3830).

43. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.

44. Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks (Vol. 4, pp. 1942-1948). ieee.

45. Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks (Vol. 4, pp. 1942-1948). ieee.

46. Storn, R., & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11(4), 341-359.

47. Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in engineering software, 69, 46-61.

48. Mirjalili, S. (2016). SCA: a sine cosine algorithm for solving optimization problems. Knowledge-based systems, 96, 120-133.

49. Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., & Gandomi, A. H. (2021). The arithmetic optimization algorithm. Computer methods in applied mechanics and engineering, 376, 113609.

50. Tizhoosh, H. R. (2005, November). Opposition-based learning: a new scheme for machine intelligence. In International conference on computational intelligence for modelling, control and automation and international conference on intelligent agents, web technologies and internet commerce (CIMCA-IAWTIC'06) (Vol. 1, pp. 695-701). IEEE.

51. Rahnamayan, S., Tizhoosh, H. R., & Salama, M. M. (2008). Opposition-based differential evolution. IEEE Transactions on Evolutionary computation, 12(1), 64-79.

52. Rahnamayan, S., Tizhoosh, H. R., & Salama, M. M. (2008). Opposition-based differential evolution. IEEE Transactions on Evolutionary computation, 12(1), 64-79.

53. Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., ... & Logullo, P. (2024). TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. bmj, 385.

54. Dataset, P. P. (2024). Apple Leaf data https://www. kaggle. com/datasets/piantic/plantpathology-apple-dataset. Last accessed, 03-01.

55. Islam AFMS (2025) Alzheimer disease classification data (redistribution of the Kaggle "Alzheimer MRI 4 classes" images). IEEE DataPort, https://doi.org/10.21227/vnhm-3x95

56. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

57. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

58. Wightman, R., Touvron, H., & Jégou, H. (2021). Resnet strikes back: An improved training procedure in timm. arXiv preprint arXiv:2110.00476.

59. Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462).

60. Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the american statistical association, 32(200), 675-701.

61. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, 7(Jan), 1-30.

62. Nemenyi, P. B. (1963). Distribution-free multiple comparisons. Princeton University.

63. Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics bulletin, 1(6), 80-83.

64. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 65-70.

65. Nadeau, C., & Bengio, Y. (1999). Inference for the generalization error. Advances in neural information processing systems, 12.

66. Bouckaert, R. R., & Frank, E. (2004, May). Evaluating the replicability of significance tests for comparing learning algorithms. In Pacific-Asia conference on knowledge discovery and data mining (pp. 3-12). Berlin, Heidelberg: Springer Berlin Heidelberg.

67. Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage, 180, 68-77.

68. Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association, 46(253), 68-78.

69. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

70. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

71. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67.

72. Niculescu-Mizil, A., & Caruana, R. (2005, August). Predicting good probabilities with supervised learning. In Proceedings of the 22nd international conference on Machine learning (pp. 625-632).

73. Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., & Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. (2019). Calibration: the Achilles heel of predictive analytics. BMC medicine, 17(1), 230.