Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Leakage-Safe Benchmarking of Tempered Fractional Optimization and Swarm-Driven Feature Construction for Heart Disease Prediction on Structured Clinical Data

Intissar DABBACHI
University of Sfax
Sabeur MASMOUDI
University of Sfax
Imed Bouzida
imed.bouzida@gmail.com

Abu Dhabi University
Omar NAIFAR
University of Sfax

Research Article

Keywords: heart disease prediction, tempered fractional optimization, particle swarm optimization, tabular machine learning, leakage-safe evaluation, ensemble learning, calibration, medical artificial intelligence

Posted Date: May 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9447779/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Leakage-Safe Benchmarking of Tempered Fractional Optimization and Swarm-Driven Feature Construction for Heart Disease Prediction on Structured Clinical Data

Intissar DABBACHI[1], Sabeur MASMOUDI[2], Imed Bouzida [*3,4,5], and Omar NAIFAR[1,3]

[1]Control and Energy Management Laboratory, National Engineering School of Sfax, University of Sfax, Sfax, 3038, Tunisia.

[2]Laboratory of Advanced Technologies for Medicine and Signals (ATMS), National Engineering School of Sfax, University of Sfax, Sfax, 3038, Tunisia.

[3]Higher Institute of Applied Sciences and Technology of Kairouan, University of Kairouan, Kairouan, Tunisia.

[4]Abu Dhabi University, Abu Dhabi, United Arab Emirates

[5]Laboratory of Probability and Statistics, Sfax University, Sfax, Tunisia

Abstract

Prediction of heart disease from structured clinical data requires both predictive performance and methodological rigor. This paper presents a leakage-safe benchmark of tempered fractional optimization and swarm-driven feature construction for heart disease prediction. The proposed framework combines a tempered fractional gradient-based logistic learner with particle swarm optimization (PSO)-driven nonlinear feature construction and is evaluated against tuned strong baselines, including elastic-net logistic regression, Extra Trees, histogram-based gradient boosting, XGBoost, LightGBM, and a stacking ensemble. The experimental protocol nests preprocessing, hyperparameter tuning, and PSO construction strictly within training data to avoid leakage. Across 15 repeated stratified hold-out splits, the proposed `TFGD_PSO` consistently improves over plain `TFGD`. However, the stacking ensemble achieves the strongest overall performance (mean ROC-AUC 0.9318 ± 0.0146, mean F1-score 0.8917 ± 0.0210, mean MCC 0.7564 ± 0.0463), while Extra Trees reaches the highest mean PR-AUC. `TFGD_PSO` remains competitive and computationally attractive. The study provides a rigorous benchmark showing that PSO-enhanced tempered fractional learning improves over plain tempered fractional optimization, while tuned tree-based ensembles remain the strongest predictors for this dataset.

Keywords: heart disease prediction; tempered fractional optimization; particle swarm optimization; tabular machine learning; leakage-safe evaluation; ensemble learning; calibration; medical artificial intelligence.

*Corresponding author(s). Email: `imed.bouzida@gmail.com`

1 Introduction

Cardiovascular diseases (CVDs) are the number one cause of mortality in the world with a mortality rate of 19.8 million deaths in 2022, this is estimated to be 32 percent of all the deaths in the world [1]. The Global Burden of Disease 2021 report attests that cardiovascular disorders are one of the most impactful causes of death and life loss in all regions [2]. Such statistics highlight the pressing requirement of precise, scaleable and powerful procedures of early cardiovascular risk assessment, especially using variables of structured clinical data that are routinely acquired.

Machine learning (ML) has gained a lot of interest in enhancing cardiovascular risk prediction beyond the traditional statistical scores. According to recent systematic reviews and meta-analyses, it is possible to obtain high discrimination (e.g., area under the receiver operating characteristic curve, AUC) with ML models that are trained on structured electronic health records and clinical tabular data [3, 4]. Nevertheless, the literature indicates that methodological heterogeneity, inconsistent reporting, limited external validation and lack of assurance regarding applicability to different populations and clinical contexts have remained [3, 5, 6]. Therefore, even though the area of cardiovascular prediction using machine learning is promising, the rigorous evaluation pipelines continue to be lacking and until the superiority of algorithms or their clinical implications can be proven beyond doubts.

Another complexity is that heart disease prediction based on structured clinical data is essentially a *tabular learning* problem. Random forests, gradient boosting and stacking, which are tree-based ensemble methods that are well-tuned to data have been shown to be highly reliable baselines in this context, especially when the data is medium-sized and features a mixture of numerical and categorical variables [7, 8]. Massive tabular benchmarks have demonstrated that deep learning does not necessarily outperform classical ensemble-based approaches on such data, due to the tendency of the inductive biases of tree-based learners to match the irregular structure and mixed types of features and moderate sized samples that many biomedical tables exhibit [7]. This fact is also applicable to cardiovascular prediction, where ensemble algorithms like stacking have the ability to combine the complementary decision boundaries of various base learners and often result in state-of-the-art performance [8].

In addition to the predictive performance metric, a larger amount of methodological literature has revealed that the phenomenon of data leakage persists as a common cause of overoptimistic tendencies in the science of machine-learned procedures, both in the biology and health domains [9, 10]. The leakage makes the leakage of information in the test data (or in the future in time series) compromise the performance estimates to reflect realistic performance in the real world deployment. Simultaneously, clinical AI research is commonly characterized by a relative deficiency of rigorous real-world testing and too many tools promoted as potentially useful, in practice, fail to be sufficiently evaluated in a prospective commentary [11, 6]. When predicting structured cardiovascular models, it implies that the benchmarking studies should necessarily avoid leakage and test the models repeatedly when they are used in realistic train-test separation.

In this landscape, it is yet another (relatively unexplored) avenue of direction in structured clinical prediction: learning strategies based on optimisation. In fractional-order optimisation

new rules of update are introduced, which at least in theory can be used to enhance the stability and robustness of convergence by using nonlocal information based on past gradients [12, 13]. Recently, it has been suggested to use tempered fractional gradient descent (TFGD), which utilizes fractional gradient memory but tempering with exponential, in order to keep long-range information and limit the contribution of outdated gradients [14]. Simultaneously, particle swarm optimisation (PSO) models of swarm intelligence provide the researcher with a adaptable means of nonlinear feature building and interaction exploration, and have long been acknowledged as an effective means of search in optimisation and feature engineering [15]. The usefulness of the combination of tempered fractional learning and PSO based feature construction in cardiovascular prediction under a leakage safe experiment design is not determined, though.

In this work, we present a leakage-safe benchmark of tempered fractional optimisation and swarm-driven feature construction for heart disease prediction on structured clinical data. Our proposed framework combines a logistic regression model trained with the TFGD optimizer (plain `TFGD`) with a nested PSO-based nonlinear feature construction module (`TFGD_PSO`). All tuning and feature-construction steps are nested strictly within the training data of each repetition, preventing information leakage from the test set. We compare the proposed methods against tuned strong baselines, including elastic-net logistic regression, Extra Trees, histogram-based gradient boosting, XGBoost, LightGBM, and a stacking ensemble. The benchmark uses repeated stratified hold-out splits (15 repetitions), multiple discrimination and calibration metrics, and paired statistical tests with Holm–Bonferroni correction.

The main contributions of this paper are as follows:

1. We design a leakage-safe benchmarking pipeline for heart disease prediction that nests preprocessing-sensitive optimisation steps, PSO-based feature construction, and hyperparameter selection strictly within the training data.

2. We introduce and evaluate a PSO-enhanced tempered fractional learning framework (`TFGD_PSO`) for structured cardiovascular prediction, showing that it consistently improves over plain `TFGD`.

3. We contrast the suggested approach with tuned strong tabular baselines and show that in addition to the tree-based ensemble-based approaches, which are overall the strongest in this dataset, there exists a competitor that is computationally inexpensive and still convincing, the TFGD-PSO approach.

4. We give a pragmatic evaluation of predictive performance, calibration, runtime and feature behaviour, thus showing tempered fractional learning is an option in the machine learning clinical tabular toolbox and not the over-sold general winner.

The rest of the paper is structured in the following way. Section 2 presents literature on related studies on the prediction of heart disease, tabular machine learning, leakage in medical AI, and fractional/swarm-based optimisation. This Section 3 contains the dataset, preprocessing procedures, the tempered fractional optimizer, nested PSO feature-construction mechanism, base models, and the experimental design. Section 4 presents the benchmark results, statistical

deviance, calibration analysis, trade-offs in the run-time, and feature analysis. Section 5 is a discussion on practical and methodological implications and Section 6 is a conclusion to the paper.

2 Related Work

2.1 Machine learning for heart disease prediction

Heart disease prediction by machine learning has developed very fast with the last one decade inspired by the fact that it is necessary to more accurately, and earlier in time, detect cardiovascular risk based on structured clinical variables. Recent reviews indicate that the range of approaches covered by the literature is quite wide, i.e. it includes not only traditional classifiers, i.e. logistic regression, support vector machines, decision trees, and random forests, but also hybrids and deep learning models [4, 16]. Besides atomic proposals of algorithms, there has been growing attention in the field to feature engineering, feature selection and optimization-informed improvements as a means of enhancing predictive performance and robustness [4, 17].

Simultaneously, the evidence base is still heterogeneous. Extensive survey of heart disease prediction research finds that various papers typically tend to use varied datasets, inconsistent preprocessing algorithms, unequal class distribution and various evaluation criteria hence cannot be directly compared [16, 5]. Furthermore, the systematic reviews are founded on electronic health records that indicate that despite the promising discrimination in machine learning models, calibration, methodological transparency, and mobility between populations are inconsistent [3, 4]. These observations encourage the necessity to have performance-oriented benchmarks in a more rigorous and reproducible methodology.

2.2 Strong baselines for structured tabular clinical data

Routinely measured heart disease prediction is technically a tabular learning task. In the case of such data, powerful tree-based ensemble algorithms still remain one of the strongest baselines. Big comparative research studies on tabular benchmarks have found that decision-tree assembly can achieve higher or equivalent results with more intricate neural architecture designs, especially on medium-sized datasets with combined gauge control and non-quantitative variables, moderate sample size, and non-uniform decision limits [7]. The results are extremely applicable in medical tabular prediction, where inductive biases of tree-based models can fit organically with non-smooth decision rule heterogeneous structured variables.

In this regard, ensemble learning is particularly relevant because various tabular learners tend to pick complementary structures. The classical stacked generalization model suggested by Wolpert continues to be among the most dominating methods of integrating numerous predictors via a meta-learner [8]. Ensemble and meta-ensemble techniques are also appealing in the medical domain where improved stability in predictions can be achieved even without large-scale representation learning [16, 18]. As a result, the comparison of any newly proposed optimization-based or feature-construction heart disease prediction method with tuned strong baselines, including Extra Trees, gradient boosting, XGBoost, LightGBM, and stacking, should be evaluated as opposed to weak single-model comparisons only.

2.3 Leakage, validation, and implementation challenges in medical AI

One of the most important issues with the machine-learning-based science is that excellent performance reported does not always reflect to the solid real-world validity. An expanding methodological literature has demonstrated that such phenomenon as data leakage is a common cause of overoptimistic findings. The failures of leakage reproducibility had been established by Kapoor and Narayan in several scientific areas and a taxonomy of leakage mechanisms was provided which could influence model development, model tuning and the model assessment [9]. More recently, Bernett et al. gave specific recommendations on how to detect and prevent leakage in more complex biological machine learning systems and that leakage can be subtle and may be challenging to notice when datasets have some latent dependencies or correlated pre-processing operations [10]. In the medical AI case, poor prospective assessment and dearth of evidence of implementation adds to the problem. The latest scoping review of the randomized controlled clinical trials on the use of AI indicated that most of the studies have been limited to single institutions, often use surrogate outcomes, and do not comment on the overall feasibility of generalizability [11]. On the same note, external validation is not prevalent although it is crucial to the behavior of predictive models when they are not used in their development sample [6]. Regulatory authorization and lack of conceivable clinical validation of AI tools have also been considered a problem, where it has been reported that recent findings are that not all systems that have been authorized have undergone appropriate testing on actual patient case data [19]. Lastly, implementation-focused reviews suggest that although predictive tools may be providing good results on the retrospective, their adoption in healthcare organizations is usually hampered by such factors as the integration of workflow, monitoring, fairness, infrastructure, and regulatory ambiguity [20, 18]. Such concern renders leakage-safe benchmarking and strict repetitive assessment of particular importance in cardiovascular prediction research.

2.4 Fractional optimization and swarm intelligence in learning systems

The aspect of fractional-order optimization has become an appealing option to the use of the conventional first-order approaches through the introduction of the memory-driven dynamics of the optimization process. Fractional methods are also capable of using weighted information of previous updates rather than using solely the current gradient which may enhance stability and convergence in challenging optimization landscapes [12, 13]. Modern optimization Survey Surveys Fractional calculus has been applied in modern optimization: particularly in gradient-based learning, adaptive filtering, swarm intelligence, and hybrid optimization systems, but also open questions concerning parameter choice, computation cost, and convergence behavior are highlighted [13]. In this family, more recently, it has been suggested that fractional memory be combined with exponential tempering by use of a method called as tempered fractional gradient descent (TFGD), which maintains the useful history, but minimizes the effects of old gradients [14]. This is especially attractive in machine learning applications where stable gradient aggregation can be used on the predictors which are correlated and the gradients are noisy. Parallel to it, particle swarm optimization (PSO) is amongst the most popular metaheuristics of nonlinear search, optimization, and construction of features that are based on conceptual simplicity and flexibility [15]. Swarm-based procedures have common applications in the biomedical field either

to select features or interactively search or retrieve them however, when feature construction is not appropriately embedded in the training data, then methodologically, such swarm-based procedures can be problematic. Considered altogether, the available literature indicates that four research threads are very relevant to structured heart disease prediction, namely, strong tabular baselines, leakage-safe evaluation, fractional optimization, and swarm-based feature construction. These threads are however normally learnt on a case basis. The majority of cardiovascular ML articles compare traditional models with deep ones, whereas methodology articles regarding the fractional optimization and swarm intelligence are also developed concurrently [4, 13, 14]. The gap in this analysis is the motivation behind the current research, which compares a tempered fractional learner and an improved variant of the one using PSO in the framework of a leak-safe experimental setup to tuned strong tabular baselines.

3 Materials and Methods

3.1 Study design overview

The study was designed as a leakage-safe benchmark for heart disease prediction from structured clinical data. The complete workflow is summarised in Figure 1. Briefly, the pipeline consists of four main stages: (i) data loading, cleaning, and feature engineering; (ii) repeated stratified train–test splitting; (iii) leakage-safe model development inside the training data, including nested hyperparameter tuning and nested PSO-based feature construction for the proposed method; and (iv) final evaluation on an untouched test set using multiple clinically relevant discrimination, threshold-based classification, calibration, and correlation-based performance metrics. The benchmark compares the proposed tempered fractional method and its PSO-enhanced variant against tuned strong tabular baselines and a stacking ensemble.

Algorithm 1 summarises the full leakage-safe repeated benchmarking procedure used in this study. For each repetition, the dataset is first divided into an outer training split and an untouched outer test split using stratified sampling. All preprocessing, hyperparameter tuning, and PSO-based feature construction are then performed strictly within the outer training data, thereby preventing information leakage from the test set. The procedure includes tuning and fitting the proposed `TFGD` and `TFGD_PSO` models, tuning the classical tabular baselines through inner cross-validation, constructing the stacking ensemble from the strongest tuned base models, and evaluating all trained models on the corresponding outer test split.The measures that were taken during repetitions are then combined in order to come up with summary statistics and comparative statistics.

3.2 Dataset and prediction target

The experiments were carried out on the publicly accessible Heart Failure Prediction Dataset, which is downloaded on the Kaggle and stored locally as heart.csv to be used as a benchmark. The dataset is available at `https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction`. The dataset contains a binary target variable, `HeartDisease`, and the following eleven original predictors:

Algorithm 1 Leakage-safe repeated benchmark for heart disease prediction

Require: Dataset $\mathcal{D} = (X, y)$, number of repetitions R, test fraction ρ
Ensure: Repetition-wise metrics, summary statistics, statistical comparisons

for $r = 1$ to R **do**
    Split $\mathcal{D}$ into outer training set $\mathcal{D}_{\text{train}}^{(r)}$ and outer test set $\mathcal{D}_{\text{test}}^{(r)}$ using stratified sampling
    Fit all preprocessing steps using only $\mathcal{D}_{\text{train}}^{(r)}$
    Tune plain `TFGD` on an inner split of $\mathcal{D}_{\text{train}}^{(r)}$
    Train final `TFGD` on full $\mathcal{D}_{\text{train}}^{(r)}$
    Tune and train `TFGD_PSO` using nested PSO feature construction inside $\mathcal{D}_{\text{train}}^{(r)}$
    for each classical baseline model **do**
        Perform randomized hyperparameter search with inner stratified cross-validation on $\mathcal{D}_{\text{train}}^{(r)}$
        Refit the tuned model on full $\mathcal{D}_{\text{train}}^{(r)}$
    end for
    Select the top three classical models by inner cross-validated ROC-AUC
    Train a stacking ensemble using the selected base models and a logistic regression meta-learner
    for each trained model **do**
        Evaluate on $\mathcal{D}_{\text{test}}^{(r)}$
        Record Accuracy, Precision, Recall, Specificity, Balanced Accuracy, F1-score, MCC, ROC-AUC, PR-AUC, and Brier score
    end for
end for
Compute mean and standard deviation of each metric across repetitions
Identify the best model by mean ROC-AUC
Perform paired t-tests, Wilcoxon signed-rank tests, and Holm–Bonferroni correction
Generate plots and post hoc analyses

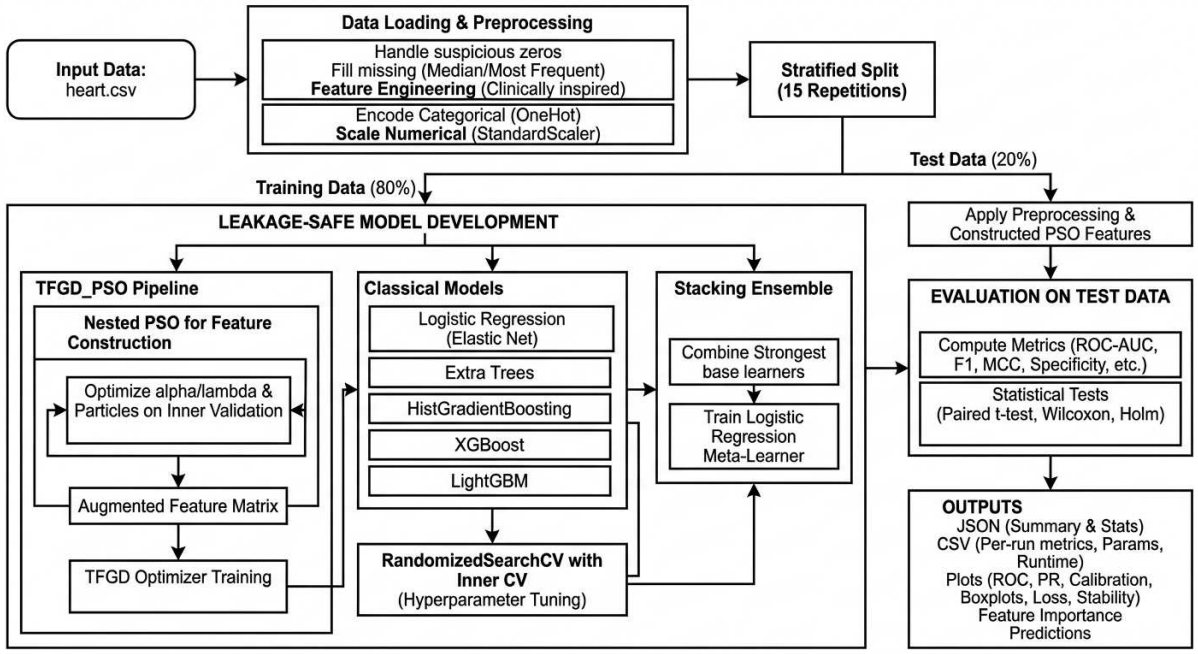


Figure 1: Overview of the leakage-safe benchmarking pipeline. All preprocessing-sensitive operations, including hyperparameter tuning and PSO-based feature construction, are restricted to the training data within each repetition, thereby preventing information leakage from the test set.

- `Age`,
- `Sex`,
- `ChestPainType`,
- `RestingBP`,
- `Cholesterol`,
- `FastingBS`,
- `RestingECG`,
- `MaxHR`,
- `ExerciseAngina`,
- `Oldpeak`,
- `ST_Slope`.

The prediction task is binary classification, where the goal is to estimate the presence or absence of heart disease.

3.3 Data cleaning and clinically inspired feature engineering

3.3.1 Handling suspicious zero values

Following common practice for this dataset, zero values in `Cholesterol` and `RestingBP` were treated as clinically suspicious rather than physiologically meaningful. Two binary indicator variables were therefore created:

- `Cholesterol_was_zero`,
- `RestingBP_was_zero`.

After creating these indicators, zero entries in `Cholesterol` and `RestingBP` were replaced with missing values and imputed using the median of the corresponding variable.

3.3.2 Clinically inspired engineered variables

To enrich the structured feature space while avoiding target leakage, a set of hand-crafted clinically motivated predictors was added:

- `Age_SexMale`,
- `MaxHR_Age_Ratio`,
- `Chol_Age_Ratio`,
- `BP_Age_Ratio`,
- `Oldpeak_ExerciseAngina`,
- `FastingBS_Chol`,
- `Oldpeak_MaxHR`,
- `ChestPain_ExerciseAngina`,
- `STSlope_ExerciseAngina`.

These variables were constructed directly from the original clinical inputs and are target-safe, because they do not use the output label in any way.

3.3.3 Preprocessing for classical models

For the classical tabular baselines, numerical variables were median-imputed and standardised, while categorical variables were imputed using the most frequent category and then one-hot encoded. This preprocessing was implemented inside a `ColumnTransformer` pipeline and therefore fitted only on training data within each repetition.

3.3.4 Preprocessing for TFGD-based models

For the tempered fractional models, categorical variables were converted to numerical codes using mappings learned from the current training split only. Unseen categories in the corresponding test split were mapped to a fallback code (-1). The resulting numeric matrix was standardised using `StandardScaler` fitted only on the training split.

3.4 Tempered fractional logistic regression

3.4.1 Base logistic model

Let $\mathbf{x}_i \in \mathbb{R}^d$ denote the input vector for the i-th patient and let $y_i \in \{0, 1\}$ denote the corresponding class label. The probabilistic output is modelled as

$$\hat{y}_i = \sigma(\mathbf{x}_i^\top \mathbf{w} + b), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathbf{w}$ is the weight vector, and b is the bias term.

The empirical binary cross-entropy loss is

$$\mathcal{L}(\mathbf{w}, b) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]. \tag{2}$$

3.4.2 Tempered fractional gradient mechanism

The proposed optimizer extends conventional gradient descent by using a finite-memory tempered fractional aggregation of historical gradients. Let $\alpha > 0$ denote the fractional order, $\lambda > 0$ the tempering parameter, and $M_{\max} = 30$ the maximum allowed memory length in our implementation. The unnormalised memory weights are defined as

$$w_j = \frac{\Gamma(\alpha + j)}{\Gamma(\alpha)\Gamma(j + 1)} \exp(-\lambda j), \qquad j = 0, 1, \ldots, m - 1, \tag{3}$$

and are normalised as

$$\tilde{w}_j = \frac{w_j}{\sum_{\ell=0}^{m-1} w_\ell}. \tag{4}$$

Let $\nabla_{\mathbf{w}} \mathcal{L}_{k-j}$ and $\nabla_b \mathcal{L}_{k-j}$ denote the mini-batch gradients of the loss with respect to the weight vector and bias term, respectively, computed j steps in the past. The effective aggregated gradients at iteration k are

$$\mathbf{G}_k^{(\mathbf{w})} = \sum_{j=0}^{r_k} \tilde{w}_j \nabla_{\mathbf{w}} \mathcal{L}_{k-j}, \tag{5}$$

and

$$G_k^{(b)} = \sum_{j=0}^{r_k} \tilde{w}_j \nabla_b \mathcal{L}_{k-j}, \tag{6}$$

where $r_k = \min(k, m - 1)$ is the available history length.

The model parameters are updated according to

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \mathbf{G}_k^{(\mathbf{w})}, \tag{7}$$

$$b_{k+1} = b_k - \eta G_k^{(b)}, \tag{8}$$

where η is the learning rate.

3.4.3 Finite-memory implementation

To control computational cost, a finite history buffer was used. In practice, the effective memory length m was capped by $M_{\max} = 30$ and adjusted as a function of λ through

$$m = \min\left(M_{\max}, \max\left(5, \left\lceil \frac{5}{\lambda} \right\rceil\right)\right), \tag{9}$$

so that smaller tempering values permit longer effective memory while preserving computational tractability.

3.5 Nested PSO-driven feature construction

3.5.1 Particle representation

For the proposed `TFGD_PSO` model, additional nonlinear interaction features were generated through particle swarm optimisation (PSO). Each particle encodes a set of candidate feature constructions. A single constructed feature is defined by a tuple

$$(i_1, i_2, o), \tag{10}$$

where i_1 and i_2 are indices of source variables in the standardised numerical feature matrix and o is an operator selected from a predefined set.

3.5.2 Operator set

The operator set consists of:

- addition $(f_1 + f_2)$,
- subtraction $(f_1 - f_2)$,
- multiplication $(f_1 \cdot f_2)$,
- protected division $\big(f_1/(f_2 + \varepsilon)\big)$,
- log-ratio $\big(\log(|f_1| + \varepsilon) - \log(|f_2| + \varepsilon)\big)$,
- product-over-sum $\left(\frac{f_1 f_2}{f_1 + f_2 + \varepsilon}\right)$,

where $\varepsilon = 10^{-8}$ is a small constant introduced for numerical stability.

Each generated feature is standardised before being appended to the original numerical matrix.

3.5.3 Nested leakage-safe PSO optimisation

PSO optimisation was performed strictly within the training data of each outer repetition. Specifically, the current training split was subdivided into an inner training subset and an inner validation subset. The validation AUC of a TFGD model trained on the particle-augmented inner training subset was used as the PSO fitness function. This design ensures that the outer test data never influence feature construction.

The PSO settings used in the benchmark were:

- number of particles: 8,
- number of generations: 6,
- number of constructed features per particle: 6.

3.6 Hyperparameter tuning for TFGD and TFGD_PSO

For plain `TFGD`, the hyperparameters (α, λ) were selected on an inner split of the current outer training data using validation ROC-AUC as the selection criterion. The search grids were

$$\alpha \in \{0.3, 0.5, 0.6, 0.7, 0.9\}, \qquad \lambda \in \{0.1, 0.3, 0.5, 0.7, 1.0\}. \tag{11}$$

For `TFGD_PSO`, the same grid search for (α, λ) was first carried out on an inner split of the current training matrix. The best configuration was then used to run nested PSO on that same training-only inner split, producing the particle used for final feature augmentation in the outer repetition.

Algorithm 2 details the nested training procedure for the proposed `TFGD_PSO` model. The outer training data are first divided into an inner training subset and an inner validation subset. A grid search over the fractional order α and tempering parameter λ is then used to identify the best `TFGD` configuration according to validation ROC-AUC. Using the selected hyperparameters, a nested PSO procedure searches for nonlinear feature constructions that improve validation performance without accessing the outer test data. The best particle returned by PSO is then used to augment the full outer training matrix, and the final `TFGD` model is trained on this enriched feature space. In this way, the proposed method combines memory-aware optimisation with leakage-safe nonlinear interaction discovery inside a unified training pipeline.

3.7 Classical baseline models

The benchmark included the following tuned strong baselines:

- elastic-net logistic regression (`LogRegEN`),
- Extra Trees,
- histogram-based gradient boosting (`HistGB`),
- XGBoost (if available in the execution environment),
- LightGBM (if available in the execution environment).

Each baseline was implemented as a pipeline that combined the preprocessing transformer with the classifier.

Algorithm 2 Training procedure for `TFGD_PSO`

Require: Outer training matrix X_{train}, labels y_{train}, search grids $\mathcal{A}$ for α and Λ for λ
Ensure: Trained `TFGD_PSO` model

Split $(X_{\text{train}}, y_{\text{train}})$ into inner training set $(X_{\text{in}}, y_{\text{in}})$ and inner validation set $(X_{\text{val}}, y_{\text{val}})$ using stratified sampling
Initialise best validation AUC to $-\infty$
for each $\alpha \in \mathcal{A}$ **do**
    for each $\lambda \in \Lambda$ **do**
        Train a `TFGD` model on $(X_{\text{in}}, y_{\text{in}})$
        Compute validation ROC-AUC on $(X_{\text{val}}, y_{\text{val}})$
        if validation ROC-AUC improves **then**
            Store current (α, λ) as best hyperparameters
        end if
    end for
end for
Run nested PSO on $(X_{\text{in}}, y_{\text{in}})$ with validation on $(X_{\text{val}}, y_{\text{val}})$ using the best (α, λ)
Obtain the best particle and construct augmented feature matrices:

$$X_{\text{train}}^{+} = \Phi_{\text{PSO}}(X_{\text{train}}), \qquad X_{\text{test}}^{+} = \Phi_{\text{PSO}}(X_{\text{test}})$$

Train the final `TFGD` model using the best (α, λ) on the full augmented outer training matrix X_{train}^{+}
Return the trained `TFGD_PSO` model

3.8 Hyperparameter tuning

Hyperparameters for the classical baselines were optimised using `RandomizedSearchCV` with:

- 4-fold stratified cross-validation,
- 12 random search iterations per model,
- ROC-AUC as the primary tuning metric.

The search spaces covered the main complexity and regularisation parameters for each learner. For example, for Extra Trees we searched over number of trees (250–900), maximum depth ($\{\texttt{None}, 4, 6, 8, 10, 15\}$), minimum samples per leaf (1–6), and feature subsampling strategies ($\{\texttt{sqrt}, \texttt{log2}, \texttt{None}, 0.5, 0.8\}$). The complete search ranges for all classical baselines are available from the corresponding author upon reasonable request.

3.9 Stacking ensemble

In addition to the single tuned baselines, a stacking ensemble was constructed inside each outer repetition. The three strongest classical base learners were selected according to their mean cross-validated ROC-AUC obtained during hyperparameter tuning. These three tuned base models were then refit on the full outer training split and combined using a logistic regression meta-learner. The final stacked predictor was evaluated on the corresponding outer test split.

3.10 Repeated hold-out evaluation protocol

The full benchmark used repeated stratified hold-out evaluation with:

- number of repetitions: 15,
- test size per repetition: 20%,
- seeds: $42, 43, \ldots, 56$.

For each repetition:

1. the dataset was split into an outer training set (80%) and an outer test set (20%) using stratification on the binary target;
2. all preprocessing, hyperparameter tuning, and PSO feature construction were performed using only the outer training data;
3. the final tuned models were trained on the full outer training split;
4. performance was evaluated once on the untouched outer test split.

This design ensured that no test-set information entered preprocessing, feature engineering, PSO search, or hyperparameter optimisation.

3.11 Evaluation metrics

Model performance was evaluated using:

- Accuracy,
- Precision,
- Recall (Sensitivity),
- Specificity,
- Balanced Accuracy,
- F1-score,
- Matthews Correlation Coefficient (MCC),
- ROC-AUC,
- PR-AUC,
- Brier score.

Specificity was computed as

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{12}$$

and balanced accuracy as

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}. \tag{13}$$

MCC was used because it provides a balanced summary even when class distributions or error types differ. Brier score was used to assess the quality of probabilistic predictions.

3.12 Statistical analysis

For each model and metric, the mean and standard deviation across the 15 repetitions were computed. The best model was selected by highest mean ROC-AUC, with mean F1-score used as a tie-breaker if necessary.

Pairwise comparisons between the best model and all competing models were performed using:

- paired t-tests,
- Wilcoxon signed-rank tests,
- Holm–Bonferroni correction for multiple testing.

In addition, effect sizes were quantified using Cohen's d based on the repetition-wise distributions of the compared metrics.

3.13 Outputs and post hoc analyses

The benchmark pipeline exported:

- repetition-level metrics,
- summary statistics,
- best hyperparameters per repetition,
- runtime measurements,
- last-run prediction files,
- ROC and precision–recall curves,
- calibration plots,
- loss curves for TFGD-based models,
- permutation importance for the strongest classical model,
- PSO feature-frequency plots,
- constructed features and coefficients for the last repetition.

These outputs were used for the final comparative, interpretability, calibration, and runtime analyses presented in the Results and Discussion sections.

4 Results

4.1 Overall predictive performance

Table 1 reports the mean and standard deviation of the main performance metrics across the 15 repeated stratified hold-out repetitions. Among all tested methods, the stacking ensemble achieved the strongest overall performance, with the highest mean ROC-AUC of 0.9318 ± 0.0146, the highest mean F1-score of 0.8917 ± 0.0210, the highest mean MCC of 0.7564 ± 0.0463, and the lowest mean Brier score of 0.0964 ± 0.0104, while Extra Trees achieved the highest mean PR-AUC of 0.9347 ± 0.0157.

The upper competitiveness was based on the most powerful single classical baselines targeting the most powerful applications, namely Extra Trees, HistGB, and XGBoost. Conversely, the proposed `TFGD_PSO` did get better than the plain tuned up to baseline non-ensembles, but it did not get better than the vigorous tuned up to ensembles.

Table 1: Main predictive results over 15 repeated stratified hold-out repetitions. Values are reported as mean ± standard deviation.

Model	AUC	F1-score	PR-AUC	MCC	Brier
TFGD	0.9205 ± 0.0187	0.8722 ± 0.0155	0.9268 ± 0.0202	0.7160 ± 0.0341	0.1089 ± 0.0120
TFGD_PSO	0.9224 ± 0.0183	0.8774 ± 0.0182	0.9285 ± 0.0202	0.7232 ± 0.0404	0.1055 ± 0.0128
LogRegEN	0.9278 ± 0.0146	0.8884 ± 0.0155	0.9324 ± 0.0142	0.7498 ± 0.0356	0.1026 ± 0.0094
ExtraTrees	0.9307 ± 0.0148	0.8910 ± 0.0148	$\mathbf{0.9347 \pm 0.0157}$	0.7510 ± 0.0356	0.1009 ± 0.0090
HistGB	0.9301 ± 0.0159	0.8855 ± 0.0202	0.9287 ± 0.0226	0.7376 ± 0.0475	0.0965 ± 0.0111
XGBoost	0.9301 ± 0.0149	0.8884 ± 0.0199	0.9256 ± 0.0247	0.7484 ± 0.0467	0.0976 ± 0.0093
LightGBM	0.9292 ± 0.0153	0.8843 ± 0.0193	0.9268 ± 0.0211	0.7394 ± 0.0440	0.1004 ± 0.0133
Stacking	$\mathbf{0.9318 \pm 0.0146}$	$\mathbf{0.8917 \pm 0.0210}$	0.9324 ± 0.0186	$\mathbf{0.7564 \pm 0.0463}$	$\mathbf{0.0964 \pm 0.0104}$

Figure 2 gives a concise visual explanation of AUC and F1-score between models, whereas Figure 3 gives the distributions in terms of repetition. These values prove that the most powerful models are concentrated in a very slim upper-performance region, Stacking and ExtraTrees being the highest in this range, and that TFGD_PSO is continually going upward compared to the raw TFGD.

4.2 Effect of PSO on tempered fractional learning

Compared to TFGD, one can find that the PSO-based construction of features enhanced the proposed tempered fractional learner in all the main metrics.ics. The PSO-enhanced version of plain TFGD had a mean gain around +0.0019 of AUC, +0.0052 of F1-score, and +0.0072 of MCC, as well as decreasing the mean Brier score by 0.1089 to 0.1055.. Such advances reveal that nonlinear interactions generated by PSO have a positive contribution to the tempered fractional optimisation pipeline.

4.3 Statistical comparison against the best model

Stacking was the best model in terms of ROC-AUC.. Table 2 gives the pairwise results between Stacking and all competing models.s. The most evident gap was found between Stacking and plain TFGD, in particular, F1-score.re. Conversely, the discrepancies between Stacking and the

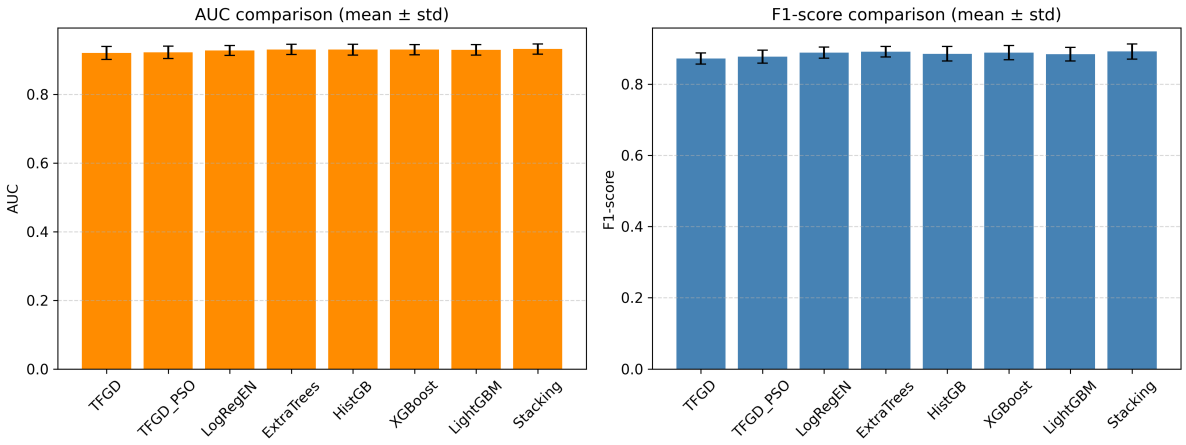


Figure 2: Comparison of mean ROC-AUC and mean F1-score across models, with error bars representing one standard deviation over the 15 repetitions. The stacking ensemble achieved the strongest overall average performance, while `TFGD_PSO` consistently improved over plain `TFGD`.

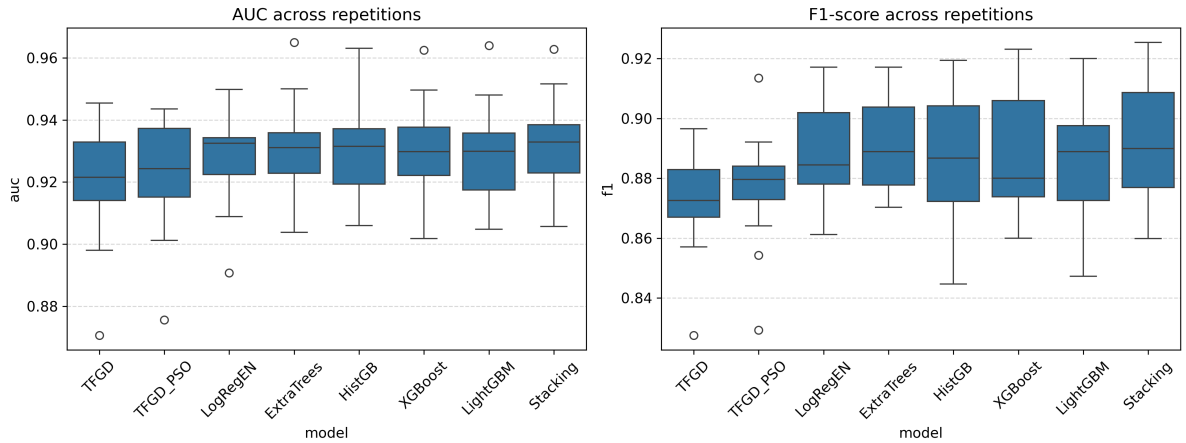


Figure 3: Distribution of ROC-AUC and F1-score across the 15 repeated stratified hold-out repetitions. The strongest tuned baselines and the stacking ensemble remain concentrated in a narrow upper band, whereas `TFGD` and `TFGD_PSO` show slightly lower centers with modest variability.

strongest classical baselines were minimal and most of them were not statistically significant following correction.In the case of TFGD-PSO, unadjusted tests preferred to use Stacking but the AUC and F1-score differences were not significant after holm-Bonferonni correction, which shows that TFGD-PSO reduces the gap between itself and the strongest ensemble more efficiently compared to simple TFGD.. For TFGD-PSO , the uncorrected tests favoured Stacking , but the differences in AUC and F1-score did not remain significant after Holm–Bonferroni adjustment, indicating that TFGD-PSO narrows the gap to the strongest ensemble more effectively than plain TFGD .

Table 2: Paired comparisons between the best model (`Stacking`) and competing models over the 15 repetitions. Reported are mean differences in favour of `Stacking`, paired t-test p-values, and Holm-adjusted p-values for ROC-AUC and F1-score.

Comparison	AUC			F1-score		
	Diff.	p	Holm	Diff.	p	Holm
Stacking vs TFGD	0.0114	0.00793	0.06343	0.0160	0.00090	0.00723
Stacking vs TFGD_PSO	0.0095	0.01229	0.07374	0.0109	0.01658	0.11604
Stacking vs LogRegEN	0.0047	0.00972	0.06807	0.0004	0.89653	1.00000
Stacking vs ExtraTrees	0.0029	0.09691	0.38765	0.0008	0.85644	1.00000
Stacking vs HistGB	0.0013	0.32584	0.65168	-0.0028	0.30527	1.00000
Stacking vs XGBoost	0.0013	0.18699	0.56099	-0.0003	0.92460	1.00000
Stacking vs LightGBM	0.0043	0.01449	0.07374	0.0030	0.27685	1.00000

For completeness, Figure 4 visualises the Holm-corrected p-values for AUC and F1-score, while Figure 5 shows repetition-wise paired differences for the two most relevant contrasts: `Stacking` vs `TFGD_PSO`, and `TFGD_PSO` vs `TFGD`. These plots visually reinforce that the most consistent beneficial shift occurs in the contrast `TFGD_PSO` − `TFGD`, whereas the gap between `Stacking` and `TFGD_PSO` is smaller and less decisive across repetitions.

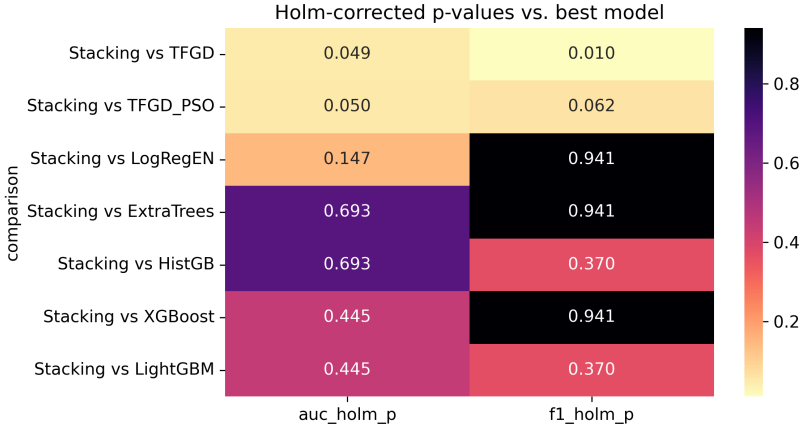


Figure 4: Holm-corrected p-values for pairwise comparisons against the best model (`Stacking`). The strongest evidence of difference is observed for plain `TFGD`, whereas the upper-performing classical baselines and `TFGD_PSO` are much closer to `Stacking`.

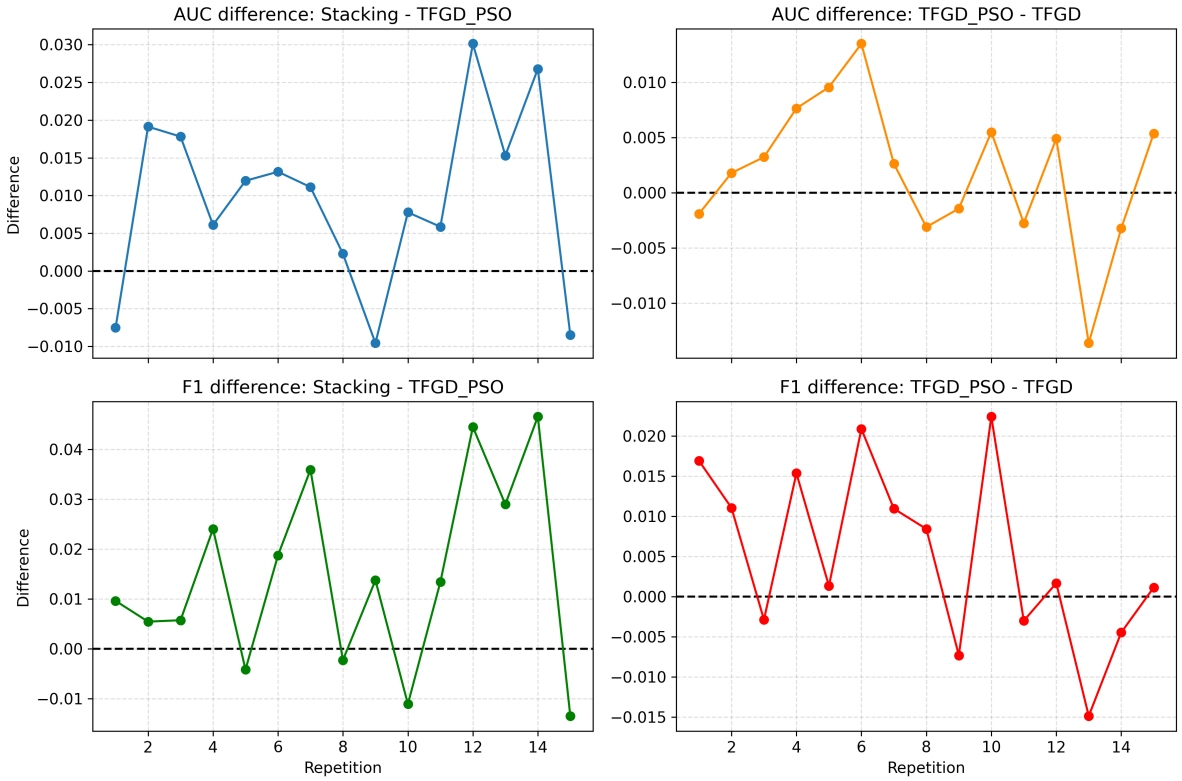


Figure 5: Repetition-wise paired differences in AUC and F1-score for the contrasts `Stacking` − `TFGD_PSO` and `TFGD_PSO` − `TFGD`. The latter contrast remains predominantly positive, confirming that PSO enhancement benefits the tempered fractional learner.

4.4 ROC and precision–recall behaviour

Figure 6 shows the ROC and precision–recall curves for the final repetition. The strongest models occupy a narrow upper region in both representations, with `Stacking` and the best classical baselines producing closely overlapping high-performing envelopes. The proposed `TFGD_PSO` remains consistently above plain `TFGD`, visually reinforcing the performance gains observed in the summary statistics.

4.5 Calibration and probabilistic quality

Calibration-related behaviour was assessed numerically through the Brier score and visually through the calibration plot. Figure 7 shows the calibration curves of the strongest calibrated models in the final repetition. The top ensembles were not very far off the diagonal in most of the probability range, meaning that they were not only discriminative, but also probabilistically valid. This is in alignment with the Brier-score, with "Stacking" having the lowest mean Brier score with the next best being HistGB and XGBoost.

4.6 Runtime and practical trade-off

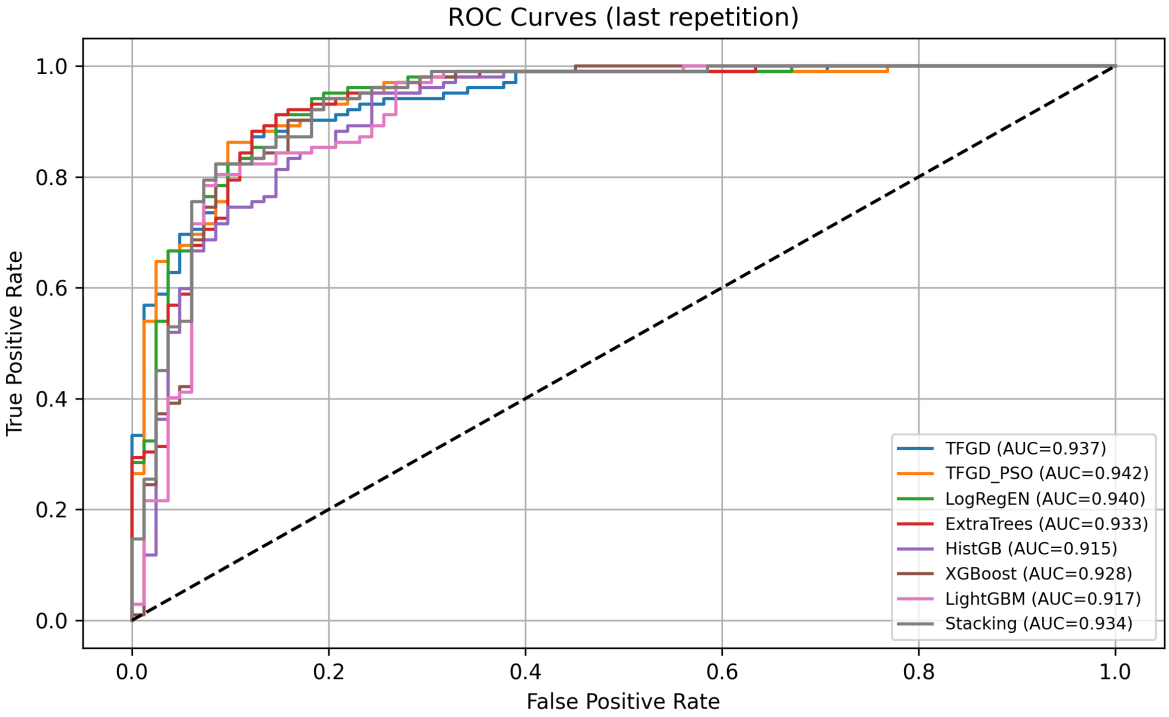
The analysis of the runtime is summarised in Table 3 and Figure 8. The simple version of the TFGD was by far the quickest version, but `TFGD_PSO` was still relatively lightweight when compared to the heaviest strong baseline models. Extra Trees evidently was the most costly base, whereas, `Stacking` was median and unpredictable. These findings suggest that the algorithm, namely, the LP algorithm, is at a sensible practical point in the performance cost space: it is not the most accurate raw predictor, however, it is much cheaper than the most expensive tuned baselines and does not deteriorate as much as plain `TFGD` does.

Table 3: Mean runtime per repetition for each model.

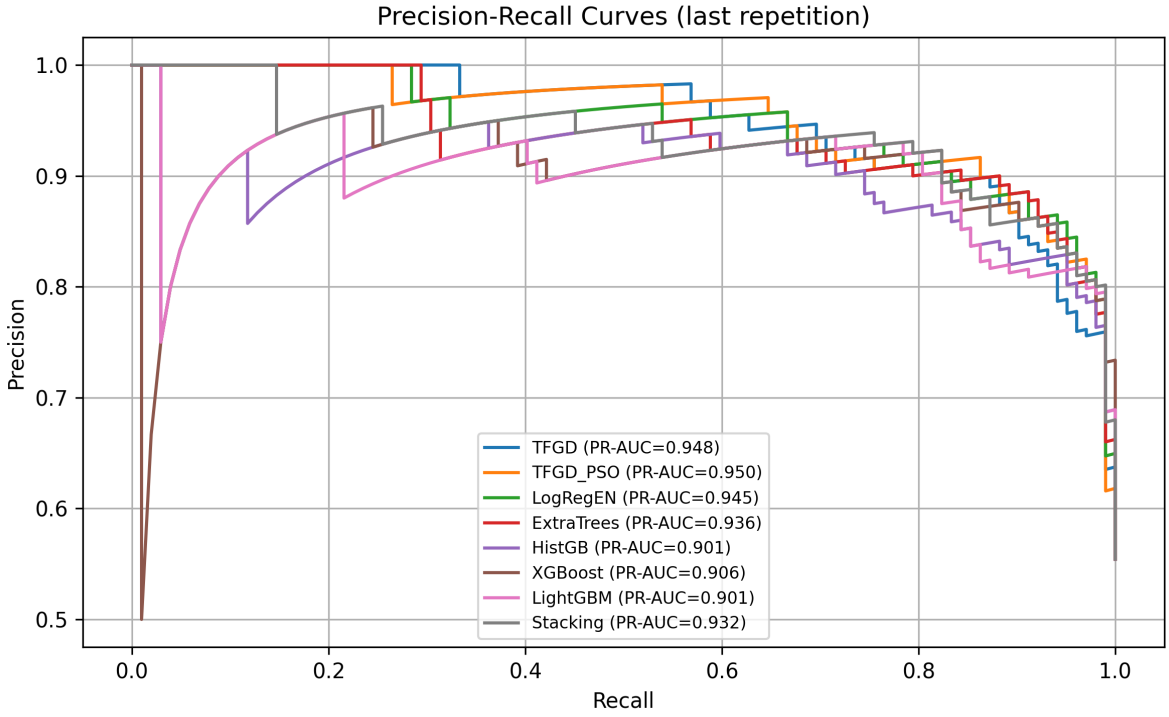
Model	Mean runtime (s)
TFGD	0.19
TFGD_PSO	6.56
LogRegEN	19.57
ExtraTrees	55.55
HistGB	8.86
XGBoost	8.76
LightGBM	11.77
Stacking	7.75

4.7 Class-specific error behaviour

Figure 9 displays the normalised confusion matrices of both the models used to compare normalised confusion matrices of `Stacking` and `TFGD_PSO` on the last repetition. The two models both demonstrated balanced performance at the class-level, although the overall best ensemble demonstrated a little more favourable error allocation, which is consistent with the fact that it



(a) ROC curves for the final repetition.



(b) Precision–recall curves for the final repetition.

Figure 6: Discrimination behaviour of the compared models on the final repetition. The strongest tuned baselines and `Stacking` occupy the top region of both ROC and precision–recall space, while `TFGD_PSO` remains consistently above plain `TFGD`.

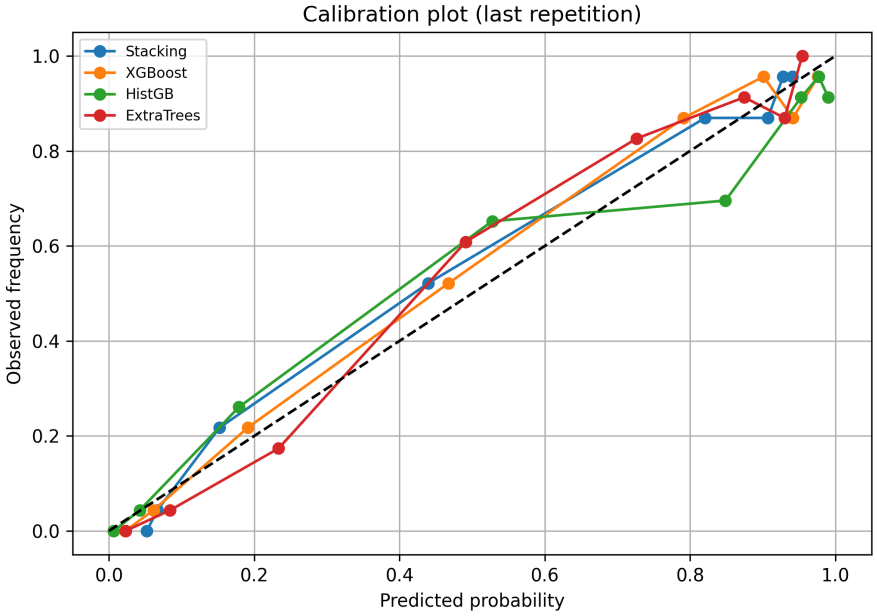


Figure 7: Calibration curves for the best-calibrated models in the final repetition. The strongest ensemble models remain reasonably close to the diagonal, indicating useful probability estimates in addition to strong ranking performance.

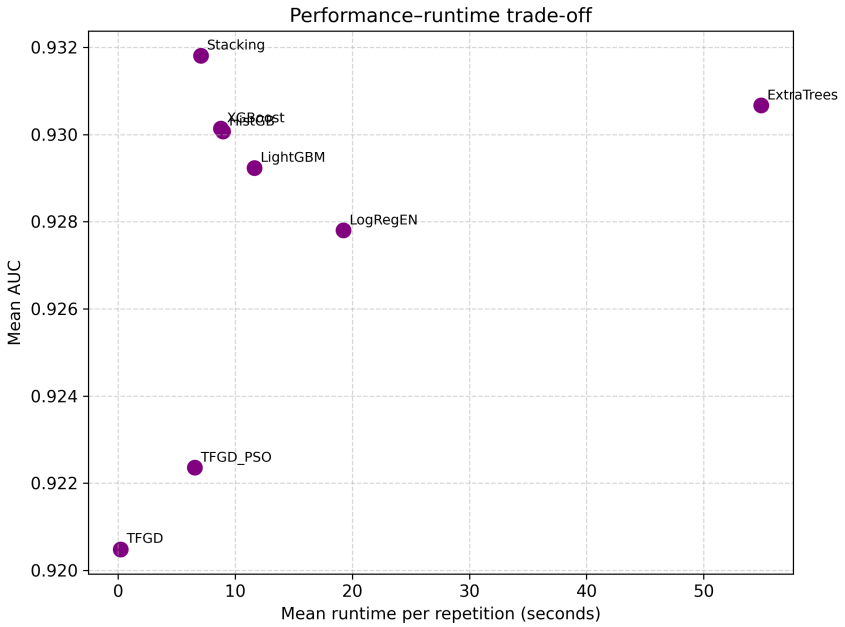


Figure 8: Performance–runtime trade-off. While the strongest AUC values are achieved by ensemble-based models, `TFGD_PSO` offers a more favourable computational profile than the heaviest tuned baselines, particularly Extra Trees.

demonstrated better overall specificity, balanced accuracy, and MCC. This figure gives a literal comparison between the ultimate model and the ultimate proposed method.

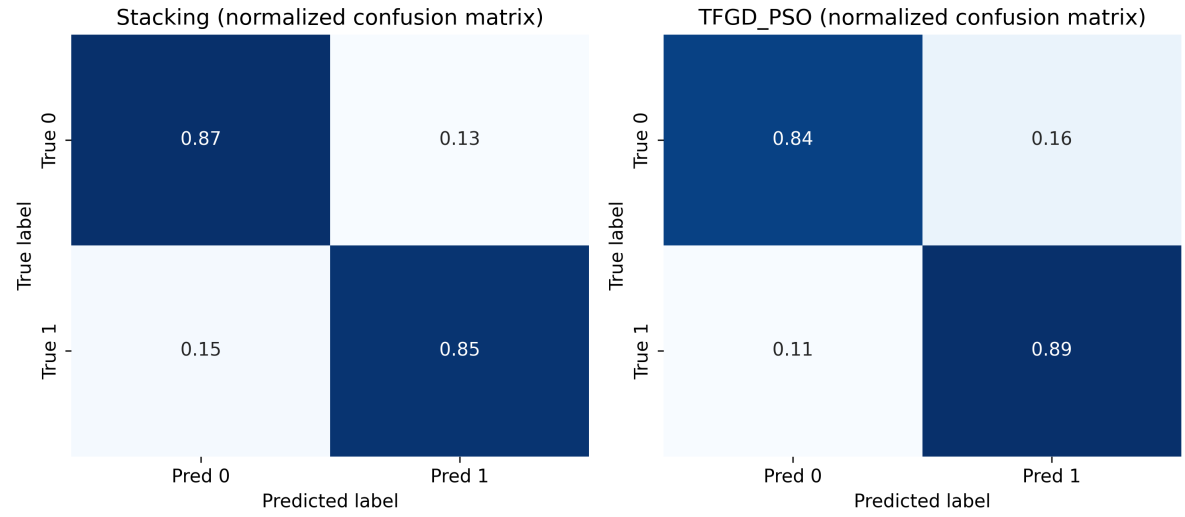


Figure 9: Normalised confusion matrices for the final repetition. The stacking ensemble provides the strongest overall class-wise classification profile, while `TFGD_PSO` remains competitive as the best proposed method.

4.8 Permutation importance of the strongest classical model

Permutation-based importance analysis of the strongest classical model highlighted a clinically coherent set of influential predictors. The largest importance was assigned to `Chol_Age_Ratio`, followed by `Cholesterol`, `FastingBS_Chol`, `BP_Age_Ratio`, `MaxHR_Age_Ratio`, `Oldpeak`, and chest-pain categories. This importance profile is consistent with clinically meaningful cardiovascular risk structure, emphasising age-normalised lipid burden, glucose-related interactions, haemodynamic burden, exercise markers, and chest-pain phenotypes. Figure 10 (shown as Figure 10 in the PDF) presents the permutation importance for the final repetition.

4.9 PSO-generated feature behaviour

The PSO-generated features in the final repetition included several plausible nonlinear interactions, such as `MaxHR * FastingBS`, `ChestPain_ExerciseAngina / Age`, and `Age_SexMale * Cholesterol_was_zero`. However, one constructed feature was degenerate by construction, namely `FastingBS_Chol - FastingBS_Chol`, which received coefficient 0. Across the 15 repetitions, PSO feature-frequency analysis showed only modest recurrence: the two most frequent interactions appeared twice each, whereas all others appeared only once. This suggests that PSO acts as a useful exploratory interaction generator, but not as a highly stable feature selector. Figure 11 (shown as Figure 11 in the PDF) illustrates the frequency of PSO-constructed features.

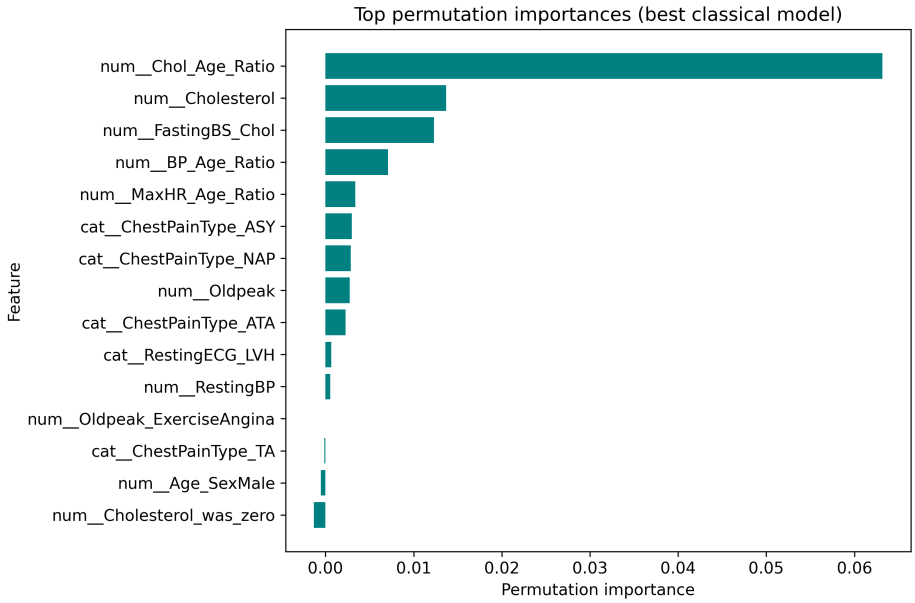


Figure 10: Permutation importance of the strongest classical model in the final repetition. Age-normalised cholesterol-related variables, glucose-related interaction terms, haemodynamic burden, and chest-pain categories dominate the importance ranking.

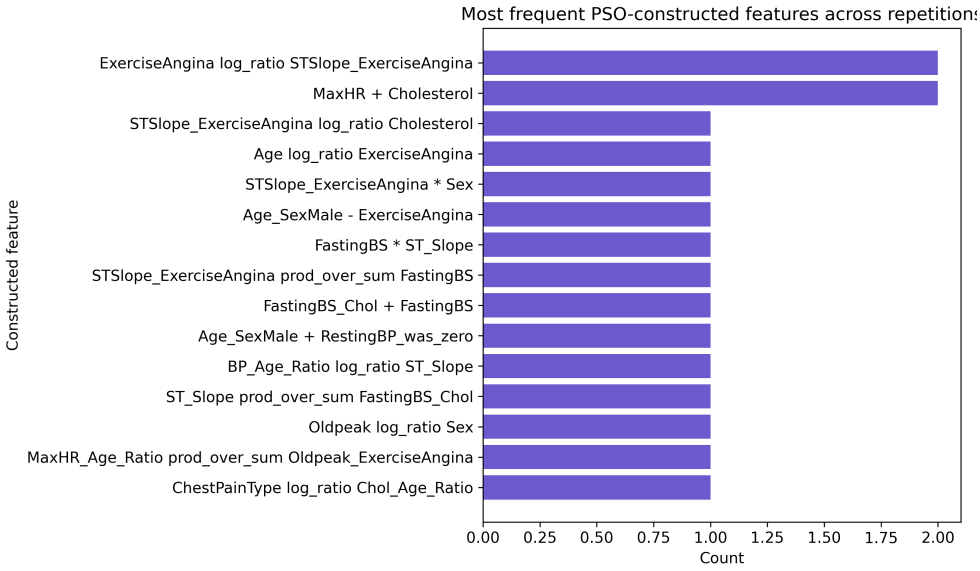
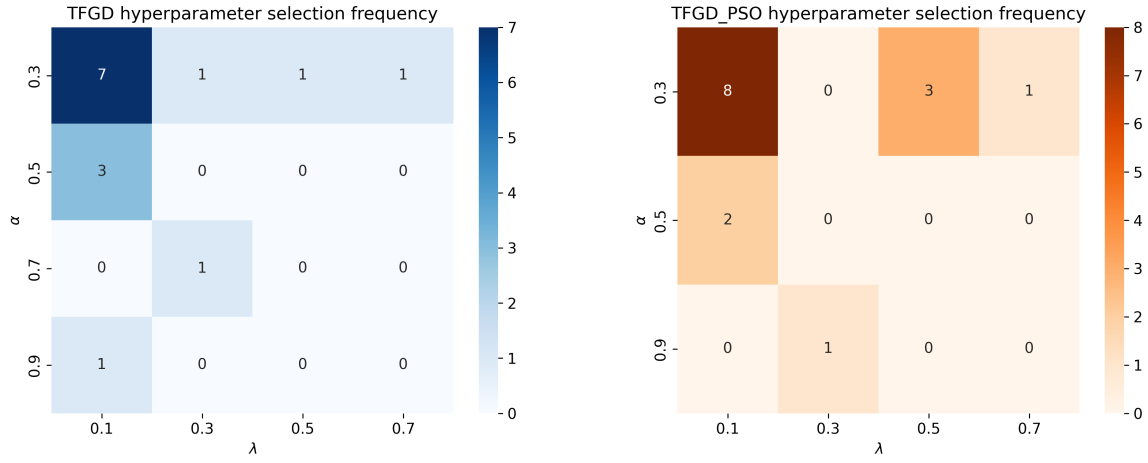


Figure 11: Frequency of PSO-constructed features across the 15 repetitions. Only two interactions recur more than once, indicating that the PSO mechanism behaves more like an exploratory interaction generator than a highly stable feature selector.

4.10 Hyperparameter selection patterns

Figures 12a and 12b summarise the hyperparameter selection frequencies for `TFGD` and `TFGD_PSO`, respectively. For plain `TFGD`, lower tempering values were frequently preferred, especially $\lambda = 0.1$, often together with smaller fractional orders such as $\alpha = 0.3$. For `TFGD_PSO`, $\alpha = 0.3$ and $\lambda = 0.1$ or 0.5 appeared repeatedly, although some repetitions selected alternative settings. This pattern suggests that the proposed tempered fractional learner benefits from mild fractional behaviour and relatively persistent memory on this structured clinical dataset.



(a) `TFGD`.

(b) `TFGD_PSO`.

Figure 12: Hyperparameter selection frequencies across the 15 repetitions. Lower tempering values were frequently selected, while `TFGD_PSO` additionally showed repeated preference for $\alpha = 0.3$ together with $\lambda = 0.1$ or 0.5.

4.11 Optimisation behaviour

Figure 13 shows the training loss curves of `TFGD` and `TFGD_PSO` for the final repetition. The PSO-enhanced variant converges along a smoother and generally lower trajectory, consistent with its improved classification and ranking metrics.

4.12 Summary of the main findings

The benchmark leads to four main findings. First, the strongest overall predictor on this dataset is the tuned stacking ensemble. Second, the PSO-enhanced tempered fractional learner consistently improves over plain tempered fractional learning across AUC, F1-score, PR-AUC, MCC, and Brier score. Third, the strongest classical and ensemble models rely on clinically meaningful predictors, particularly age-normalised cholesterol-related variables, glucose-related interactions, exercise markers, and chest-pain categories. Fourth, `TFGD_PSO` remains computationally attractive relative to heavier tuned baselines, even though it does not achieve the best raw predictive performance.

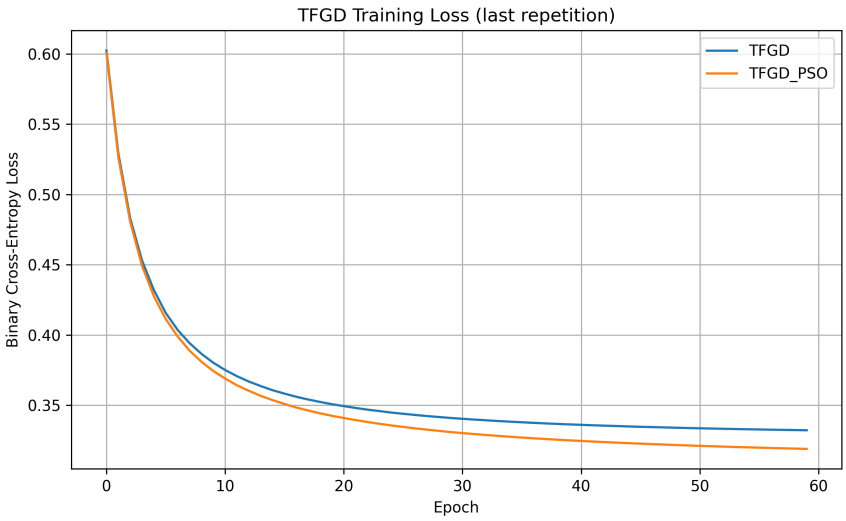


Figure 13: Training loss curves for `TFGD` and `TFGD_PSO` on the final repetition. The PSO-enhanced tempered fractional model follows a smoother and generally more favourable optimisation trajectory.

5 Discussion

The current study characterized the capability of tempered fractional learning and swarm-based feature construction to offer competitive performance on structured clinical data in predicting heart disease on a leakage-safe benchmarking protocol. The findings favor an amicable conclusion. On the one hand, the proposed `TFGD_PSO` consistently improved over plain `TFGD` across discrimination, threshold-based classification, calibration-related error, and MCC. On the other hand, the strongest overall predictor remained the tuned stacking ensemble, with strong tree-based baselines such as Extra Trees, HistGB, and XGBoost forming a tightly clustered upper-performance tier. Thus, the proposed method is competitive and methodologically meaningful, but it is not the strongest raw predictor on this dataset.

5.1 Why the PSO-enhanced tempered fractional model improves over plain TFGD

The first key finding is that swarm-driven nonlinear interaction construction improves the tempered fractional learner. Relative to plain `TFGD`, the PSO-enhanced variant achieved modest but consistent gains in AUC, F1-score, MCC, PR-AUC, and Brier score. This indicates that the PSO module enriches the feature space with interactions that are not captured by the plain linear tempered fractional classifier.

This improvement is plausible from both a modelling and an optimisation perspective. The base `TFGD` model operates as a logistic learner with a memory-aware gradient update rule, which can stabilise optimisation and improve convergence behaviour, but it still depends on the representational power of the available features.PSO enhances the expressive power of the ultimate input space by additionally incorporating nonlinear combinations among clinically relevant predictors with no fundamental more complicated classifier. The swarm component in this respect provides an optimisation dynamics complementing feature space enhancing tempered fractional

learning.

Nonetheless, the increase is not exceptional, but still moderate. This is probably an indication of the fact that the original structured feature space is already quite informative and the resulting model is linear in the augmented features. Therefore, PSO is beneficial but it is not effective enough to change the very nature of the learner in order to surpass the most powerful tuned ensemble baselines.

5.2 Why tree-based ensembles remain strongest on this dataset

One of the key findings of this benchmark is that tree-based ensembles and ensemble combinations are the best predictors in the studied structured clinical dataset. This observation is completely in line with the tabular-learning literature at large. Empirical benchmark studies have demonstrated decision-tree ensembles to be state-of-the-art on typical medium-sized tabular data sets and can frequently push more complex neural or optimisation-based methods out of the field, particularly where the data are mixed numeric and categorical, the sample sizes are small, and the decision boundaries are irregular [7]. This trend can be observed in the current work: Extra Trees, HistGB, XGBoost, and the stacking ensemble are all close to the top in terms of performance, and they are all better at mean AUC than the suggested tempered fractional models.

This is not surprising because of a number of reasons. To start with, tree based learners are inherently well adapted to heterogeneous structured variables, and are able to learn nonlinear thresholds and higher-order interactions without being explicitly constructed. Second, complementary inductive biases can be used by stacking ensemble combinations of several strong base learners [8]. Third, the present dataset is neither too large nor too small and is represented by classical clinical variables, which is exactly the regime where the tabular ensemble techniques are known to be especially competitive [7]. In turn, the fact that the proposed framework, which is denoted by the name of the algorithm, `TFGD_PSO`, is not as powerful as the most efficient methods of this category, should be viewed as an indication of the familiar weakness of an ensemble model in the specified environment of structured prediction.

5.3 Interpretation of the tempered fractional hyperparameter patterns

The repetition-wise hyperparameter selections provide useful insight into how the proposed optimiser behaves on this dataset. Both `TFGD` and `TFGD_PSO` frequently selected relatively small tempering values, especially $\lambda = 0.1$, while often favouring lower fractional orders such as $\alpha = 0.3$. This suggests that the model tends to benefit from mild fractional behaviour together with relatively persistent gradient memory.

One interpretation is that the structured cardiovascular dataset contains correlated predictors and moderately smooth decision boundaries, for which stable gradient averaging is beneficial. Lower tempering values reduce the rate at which historical gradients are discounted, thereby preserving a stronger memory effect, while smaller or moderate fractional orders may smooth the optimisation trajectory without making it excessively inert. This pattern is consistent with the intended role of tempered fractional optimisation, namely to balance nonlocal

gradient memory against the need to prevent outdated updates from dominating the optimisation process [14, 13]. Although the present study does not provide a full theoretical characterisation of these selected regimes, the empirical consistency of the low-λ preference suggests that persistent but controlled memory can be advantageous for structured medical classification tasks.

5.4 Interpretation of the PSO feature behaviour

The PSO analysis reveals both strengths and limitations of the feature-construction mechanism. On the positive side, several of the generated interactions are clinically plausible, combining age, exercise-related variables, sex, cholesterol-related indicators, fasting blood sugar, and chest-pain/exercise-angina descriptors. This supports the claim that swarm-based search can discover meaningful nonlinear combinations in the structured cardiovascular feature space.

At the same time, the recurrence frequency of PSO-generated features across repetitions is low, and at least one degenerate feature was produced in the final repetition. This indicates that the PSO mechanism behaves more like an exploratory interaction generator than a stable feature selector. Such behaviour is not necessarily problematic, but it does affect how the method should be presented. The feature-construction component should therefore be described as a mechanism for discovering potentially useful nonlinear interactions under repeated resampling, rather than as a robust biomarker discovery tool. This distinction is particularly important in medical prediction studies, where interpretability claims must be stated carefully.

5.5 Clinical relevance of the strongest predictors

One of the strengths of the benchmark is that the best classical model is based on clinically coherent predictors. Age-normalised cholesterol related variables, glucose related interaction terms, haemodynamic burden, exercise related markers as well as categories of chest-pain were found by permutation importance to be the most influential features. It is promising since it suggests that the most successful models are not taking advantage of spurious relationships, but are taking advantage of predictors that have some meaning in the larger cardiovascular risk environment.

The clinical perspective is also useful in the calibration results. Ensemble models that were the most robust were not very far off the desired diagonal in the calibration plot and had lowest Brier scores. This implies that they do not only give good ranking performance but also comparably good probability estimates. Practically, this is important since heart disease prediction is not frequently applied to just rank patients; it is frequently applied to assist threshold-based decision-making, to assist triage, or assist risk communication. In line with this, the excellent calibration-related behaviour of the strongest ensembles supports the argument of their applicability in real-world structured cardiovascular prediction problems.

5.6 Methodological contribution of the leakage-safe benchmark

A major contribution of this study lies in its evaluation design. The entire benchmark was constructed to prevent information leakage by nesting all preprocessing-sensitive operations,

including hyperparameter tuning and PSO-based feature construction, strictly within the current training data. This is particularly important given the growing evidence that data leakage is a major source of overoptimistic performance in machine-learning-based science and biological applications [9, 10]. In the context of medical AI, where many studies rely on retrospective single-split evaluations and weak external validation, a repeated leakage-safe protocol is especially valuable [11, 6].

The present benchmark therefore contributes not only a comparative result, but also a methodological example. By evaluating all models under the same repeated stratified hold-out framework, with nested model development and paired statistical testing, the study provides a fairer and more reproducible basis for comparing optimisation-driven methods with strong ensemble baselines. This is one of the strongest aspects of the work and should remain central to how the paper is positioned.

5.7 Practical implications

Practically, the suggested "TFGD-PSO" lies in a healthy midpoint. It fails to perform the best absolute predictive, but it is always better than plain "TFGD" and also computationally appealing compared to a number of stronger baselines, particularly Extra Trees and repeatedly tuned elastic-net logistic regression. It implies that the TFGD-PSO can be appealing in the environment where a balance between a moderate computational cost, the smooth behaviour of optimisation, and feature-space interpretability are considered as well as the predictive accuracy.

The tuned ensemble models are however the strongest raw performers. In problems where the main goal is to be highly discriminative and provide well-calibrated probability estimates, stacking and close variants of strong tree-based baselines are still the methods of choice on this data. In this way, the empirical summation of the benchmark does not consist in the fact that tempered fractional learning competes with the contemporary ensemble techniques, but in the fact that it provides an alternative with its own competitive features in terms of optimisation and feature-construction.

5.8 Strengths and limitations

There are a number of strengths in this study. It employs a leakage-safe repeated evaluation scheme, contrasts the suggested scheme against tuned strong baselines instead of weak comparators, presents a wide range of metrics such as calibration-related measures and additional predictive results with runtime, feature-importance, and PSO-stability analysis. These elements combined make the comparison of leaderboards on a single- split basis look more believable and detailed.

There are also limitations of the study. First, the benchmark was done using one structured dataset of heart-disease, and no external cohort was employed. Second, there is little recurrence of interactions generated by PSO between repetitions and thus limits claims of stability in feature-selection. Third, the proposed method is not stronger than the best plain versions, despite its performance improvements, which is a limiting factor in terms of using it in applications where purely performance benefits are sought. These are not invalid to the contribution, but they clarify its extent and must be applied in future efforts.

5.9 Summary

To conclude, the benchmark demonstrates that the suggested "TFGD-PSO" is a significant upgrade on simple tempered fractional learning yet tuned ensemble techniques are the most powerful on structured clinical information to predict heart diseases. The work thus makes a smaller contribution as an assertion of universal predictive superiority to and a larger contribution as a rigorous and straight-forward benchmark which shows the behaviour of tempered fractional optimisation and swarm-driven over feature constructions compared to strong tabular baselines in the context of leakage-safe evaluation. This makes the research a scientifically reliable and reproducible contribution to the medical machine learning and not an overadvertised entry on the list of performance topping.

6 Conclusion

This paper presented a leakage-safe benchmark of tempered fractional optimization and swarm-driven feature construction for heart disease prediction on structured clinical data. The proposed `TFGD_PSO` consistently improved over plain `TFGD` across discrimination, threshold-based classification, and calibration-related metrics, indicating that PSO-based nonlinear interaction construction is beneficial within the tempered fractional learning framework. At the same time, the strongest overall performance was achieved by the tuned stacking ensemble, while strong tree-based baselines such as Extra Trees, HistGB, and XGBoost formed a highly competitive upper-performance tier.

The findings hence favour a moderate meaning. To begin with, the suggested swarm-improved tempered fractional model is a worthwhile methodological enhancement of its unsophisticated counterpart. Second, tree-based ensemble models are the most powerful predictors in this structured cardiovascular data and this is in line with the rest of the literature on tabular learning. Third, it offers a more stable foundation on which to compare optimisation-driven models and strong ensemble baselines because the leakage-safe repeated evaluation protocol in itself is an important contribution. The proposed method is competitive, computationally appealing, and methodologically interesting in the conditions of a rigorous experimental design, instead of being presented as a universal winner, as presented in the study.

There is also real world experience that is gained via the benchmark as opposed to leading in predictive performance. The most robust classical model was based on clinically coherent predictors, such as age-normalised cholesterol-related variables, glucose-related terms of interaction, exercise markers, and categories of chest-pain. Besides that, the PSO feature analysis revealed that swarm-based feature construction acts more like an exploration interaction generator but not a strongly stable feature selection. These observations can be used to explain the possibilities and the constraints of the proposed framework.

Future research must build on the current benchmark by applying it to other cardiovascular datasets and external cohorts, study more robust constrained swarm or evolutionary feature-construction processes, and apply calibration-conscious or multi-objective variations of tempered fractional learning. These guidelines can also enhance the medical machine learning practicality of the fractional optimization techniques and maintain the methodological rigor needed to make

the evaluation clinically credible.

Declarations

Clinical trial registration: Not applicable. This study is not a clinical trial. It is a retrospective methodological analysis using a publicly available, de-identified dataset.

Ethics approval: Not applicable. This study used a publicly available, de-identified dataset, and no ethics approval was required.

Consent to Participate declaration: not applicable.

Consent to Publish declaration: not applicable.

Code and data availability

The data that support the findings of this study are openly available in Kaggle, reference number [21]. No ethics approval was required for this retrospective analysis of de-identified public data.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

- **Intissar Dabbachi**: Conceptualization, methodology, software, formal analysis, investigation, data curation, writing – original draft preparation.
- **Sabeur Masmoudi**: Methodology, validation, formal analysis, writing – review and editing.
- **Imed Bouzida**: Methodology, validation, resources, writing – review and editing.
- **Omar Naifar**: Conceptualization, methodology, supervision, validation, resources, writing – review and editing, project administration, corresponding author.

All authors have read and approved the final manuscript.

Funding

No funding was received for this work.

References

[1] World Health Organization, Cardiovascular diseases (cvds), wHO Fact Sheet, updated 31 July 2025 (Jul. 2025).

URL https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] GBD 2021 Causes of Death Collaborators, Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021, The Lancet 403 (10440) (2024) 2100–2132. doi:10.1016/S0140-6736(24)00367-2.

[3] T. Liu, A. Krentz, L. Lu, V. Curcin, Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis, European Heart Journal – Digital Health 6 (1) (2025) 7–22. doi:10.1093/ehjdh/ztae080.

[4] R. Kumar, S. Garg, R. Kaur, M. G. M. Johar, S. Singh, S. V. Menon, P. Kumar, A. M. Hadi, S. A. Hasson, J. Lozanović, A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions, Frontiers in Artificial Intelligence 8 (2025) 1583459. doi:10.3389/frai.2025.1583459.

[5] C. Zhou, P. Dai, A. Hou, Z. Zhang, L. Liu, A. Li, F. Wang, A comprehensive review of deep learning-based models for heart disease prediction, Artificial Intelligence Review 57 (2024) 263. doi:10.1007/s10462-024-10899-9.

[6] P. Rockenschaub, E. M. Akay, B. G. Carlisle, A. Hilbert, J. Wendland, F. Meyer-Eschenbach, A.-F. Näher, D. Frey, V. I. Madai, External validation of ai-based scoring systems in the icu: a systematic review and meta-analysis, BMC Medical Informatics and Decision Making 25 (2025) 5. doi:10.1186/s12911-024-02830-7.

[7] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, in: Advances in Neural Information Processing Systems, Vol. 35, 2022.
URL https://papers.nips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html

[8] D. H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259. doi:10.1016/S0893-6080(05)80023-1.

[9] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns 4 (9) (2023) 100804. doi:10.1016/j.patter.2023.100804.

[10] J. Bernett, D. B. Blumenthal, D. G. Grimm, F. Haselbeck, R. Joeres, O. V. Kalinina, M. List, Guiding questions to avoid data leakage in biological machine learning applications, Nature Methods 21 (2024) 1444–1453. doi:10.1038/s41592-024-02362-y.

[11] R. Han, J. N. Acosta, Z. Shakeri, J. P. A. Ioannidis, E. J. Topol, P. Rajpurkar, Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review, The Lancet Digital Health 6 (5) (2024) e367–e373. doi:10.1016/S2589-7500(24)00047-5.

[12] J. Xie, S. Li, Training neural networks by time-fractional gradient descent, Axioms 11 (10) (2022) 507. doi:10.3390/axioms11100507.

[13] E. Fernandez, V. Huilcapi, I. Birs, R. Cajo, The role of fractional calculus in modern optimization: a survey of algorithms, applications, and open challenges, Mathematics 13 (19) (2025) 3172. doi:10.3390/math13193172.

[14] O. Naifar, Tempered fractional gradient descent: Theory, algorithms, and robust learning applications, Neural Networks 193 (2025) 108005. doi:10.1016/j.neunet.2025.108005.

[15] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of the IEEE International Conference on Neural Networks, 1995, pp. 1942–1948. doi:10.1109/ICNN.1995.488968.

[16] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, K. M. Kaky, A review of machine learning's role in cardiovascular disease prediction: Recent advances and future challenges, Algorithms 17 (2) (2024) 78. doi:10.3390/a17020078.

[17] H. A. Al-Alshaikh, P. Prabu, R. C. Poonia, A. K. J. Saudagar, M. Yadav, H. S. Al-Sagri, A. A. AlSanad, Comprehensive evaluation and performance analysis of machine learning in heart disease prediction, Scientific Reports 14 (2024) 7819. doi:10.1038/s41598-024-58489-7.

[18] A. Rajagopal, S. Ayanian, A. J. Ryu, R. Qian, S. R. Legler, E. A. Peeler, M. Issa, T. J. Coons, K. Kawamoto, Machine learning operations in health care: A scoping review, Mayo Clinic Proceedings: Digital Health 2 (3) (2024) 421–437. doi:10.1016/j.mcpdig.2024.06.009.

[19] S. Chouffani El Fassi, A. Abdullah, Y. Fang, S. Natarajan, A. Bin Masroor, N. Kayali, S. Prakash, G. E. Henderson, Not all ai health tools with regulatory authorization are clinically validated, Nature Medicine 30 (2024) 2718–2720. doi:10.1038/s41591-024-03203-3.

[20] L. M. Preti, V. Ardito, A. Compagni, F. Petracca, G. Cappellaro, Implementation of machine learning applications in health care organizations: Systematic review of empirical studies, Journal of Medical Internet Research 26 (2024) e55897. doi:10.2196/55897.

[21] Fedesoriano, Heart failure prediction dataset, accessed: 2026-04-07 (2021).
URL https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction