

A Leakage-Aware Machine Learning Framework for Multiclass Comorbidity Classification in Fibromyalgia: Integrating Genetic Feature Selection, Nested Cross-Validation, and Explainable AI

Asmaa A. Attwa

`asmaa.abdelhamed@fahst.bsu.edu.eg`

Beni-Suef University

Ahmed Yehia Ismaeel

Beni-Suef University

Ahmed A. Elngar

Beni-Suef University

Article

Keywords: Fibromyalgia, Disc Herniation, Inflammatory Bowel Syndrome, Multiclass Classification, Machine Learning, Genetic Algorithm, Bayesian Optimization

Posted Date: August 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10433360/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Leakage-Aware Machine Learning Framework for Multiclass Comorbidity Classification in Fibromyalgia: Integrating Genetic Feature Selection, Nested Cross-Validation, and Explainable AI

Asmaa A. Attwa¹, Ahmed Yehia Ismaeel² and Ahmed A. Elngar^{3,*}

¹ Faculty of Applied Health Sciences Technology, Beni-Suef University, Beni-Suef City, 62511, Egypt.

E-mail: asmaa.abdelhamed@fahst.bsu.edu.eg

² Faculty of Faculty of Medicine, Beni-Suef University, Beni-Suef City, 62511, Egypt.

E-mail: ahmed_yehia@med.bsu.edu.eg

³ Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt.

E-mail: elngar_7@yahoo.co.uk

* Corresponding author: elngar_7@yahoo.co.uk

Abstract

Fibromyalgia frequently co-occurs with disc herniation and inflammatory bowel syndrome, and the overlapping symptom profiles of these conditions complicate differential diagnosis. This paper proposes a machine learning framework for three-class comorbidity classification that distinguishes No Comorbidity, Disc Herniation, and Inflammatory Bowel Syndrome from multi-instrument clinical assessment data. Missing values are imputed with a k -nearest-neighbor pass, after which Mutual Information feature selection retains $K = 50$ features and a Genetic Algorithm refinement step converges to a 24-feature subset; both selection stages are computed once on the full analytical cohort ahead of the outer cross-validation split, a design choice this paper discloses and discusses explicitly rather than folding into an unqualified leakage-free claim. Multivariate Imputation by Chained Equations, z -score standardization, and Borderline Synthetic Minority Oversampling, with a standard Synthetic Minority Oversampling fallback, are the operations genuinely refit inside every outer training fold, so that no held-out patient contributes to the imputation statistics, scaling parameters, or synthetic minority samples used to train that fold's classifiers. Sixteen supervised classifiers spanning six algorithmic families were benchmarked under five-fold nested cross-validation and validated with the Friedman omnibus test ($\chi^2 = 42.25$, $p < 0.001$) and Bonferroni-corrected Wilcoxon signed-rank tests. The Weighted Soft Voting ensemble reached the strongest discriminative performance (accuracy = 0.864, F1 = 0.826, mean nested-CV macro-AUC = 0.866), while Gradient Boosting produced the most favorable probability calibration (expected calibration error = 0.187 ± 0.016 , Brier score = 0.103). Bayesian hyperparameter optimization through Optuna was applied to five base classifiers. SHAP TreeExplainer analysis identified Symptom Severity Scale item 13 and the Revised Fibromyalgia Impact Questionnaire physical-function subscale as the dominant predictive signals, with class-specific secondary signatures broadly consistent with known pathophysiology. LIME confirmed that these attributions hold at the level of individual predictions.

Keywords: Fibromyalgia; Disc Herniation; Inflammatory Bowel Syndrome; Multiclass Classification; Machine Learning; Genetic Algorithm; Bayesian Optimization.

1 Introduction

Chronic musculoskeletal and gastrointestinal conditions place a growing burden on health systems worldwide, contributing to disability, reduced quality of life, and rising healthcare costs [1]. Fibromyalgia is one of the more common chronic pain syndromes, affecting an estimated 2 to 4 percent of the general population and presenting with widespread musculoskeletal pain, fatigue, cognitive impairment, and unrefreshed sleep [2]. A substantial share of patients also present with intervertebral disc herniation, inflammatory bowel syndrome, psychiatric conditions, or metabolic disorders [3]. When several chronic conditions occur together, a pattern usually called multimorbidity, symptom burden increases, treatment planning becomes more complicated, and outcomes tend to be worse than for any single diagnosis on its own.

Intervertebral disc herniation is the leading cause of degenerative lumbar spine disease and a common indication for spinal surgery in high-income countries [4, 5]. It results from extrusion of the nucleus pulposus through fissures in the annulus, which compresses nerve roots and triggers local inflammation. Inflammatory bowel syndrome refers to chronic relapsing gastrointestinal disorders marked by visceral hypersensitivity, dysregulation of the gut-brain axis, and ongoing low-grade systemic inflammation [6, 7]. Both conditions share features with fibromyalgia, including widespread somatic sensitization and autonomic dysregulation, which makes differential diagnosis difficult in practice [8].

Routine clinical data in fibromyalgia cohorts add a further layer of difficulty. Multi-instrument psychometric batteries generate hundreds of variables per patient, while cohort sizes stay small because standardized clinical assessment is labor intensive. High dimensionality, small sample size, severe class imbalance, and overlapping symptom profiles together produce a classification problem that conventional statistical methods struggle to handle.

Machine learning offers a way to detect nonlinear patterns in heterogeneous clinical data that would be hard to capture with classical statistics [9, 10]. Even so, much of the published machine learning work on chronic disease classification has structural limitations that this study aims to address. Most published frameworks treat a single disease as a binary outcome, which limits their usefulness for patients presenting features of several co-occurring conditions at once. In many high-performing studies, preprocessing steps are fit on the full dataset before cross-validation splits are drawn, a practice that leaks information from the test set and inflates reported performance [11]. Probability calibration, which matters for any decision-support tool used in a clinic, is rarely evaluated. Statistical testing beyond a single point estimate is uncommon, and interpretability methods are seldom connected back to known clinical mechanisms.

This work makes four contributions:

1. A three-class supervised learning framework for comorbidity classification that separates No Comorbidity, Disc Herniation, and Inflammatory Bowel Syndrome using routine multi-instrument clinical assessments, evaluated on the publicly available *Clinical_fm_66* dataset [12].
2. A preprocessing pipeline that separates two-stage k -nearest-neighbor and Multivariate Imputation by Chained Equations imputation, z -score standardization, and BorderlineSMOTE-based data augmentation, all fit strictly within each outer cross-validation training fold, from the Mutual Information and Genetic Algorithm feature selection stages, which operate once on the full analytical cohort before the outer split; the paper reports both the leakage-free classifier-training pipeline and an explicit accounting of the feature-selection stage’s exposure to the full cohort, rather than presenting the two as equivalent.
3. A benchmark of sixteen supervised classifiers across six algorithmic families, including gradient-boosted models and four meta-ensemble strategies, evaluated under identical experimental conditions with global statistical validation through the Friedman test, Bonferroni-corrected Wilcoxon signed-rank tests, and Cohen’s d effect sizes.

4. Explainability through SHAP TreeExplainer and LIME local instance explanations, giving global, per-class, and per-instance feature attributions that recover clinically plausible pathophysiological signatures for each comorbidity class.

The rest of the paper proceeds as follows. Section 2 surveys related work. Section 3 states the classification problem formally. Section 4 describes the dataset. Section 5 presents the methodology. Section 6 reports experimental results, including the oversampling ablation. Section 7 discusses the findings critically, and Section 8 concludes.

2 Related Work

Machine learning for chronic disease classification has grown substantially over the past decade, but progress has been uneven: algorithmic sophistication has advanced faster than clinical scope, methodological rigor, or statistical transparency across most published frameworks. The summary below traces the field from early binary classifiers through deep learning architectures to more methodologically careful recent work, and identifies the gaps this paper tries to close.

Early work on musculoskeletal conditions demonstrated that machine learning could support spinal disorder classification, but nearly all of it treated the problem as binary discrimination. Ghosh et al. [13] combined Histogram of Oriented Gradients features with Support Vector Machines to localize lumbar discs, reaching 99 percent accuracy on magnetic resonance images, a strong result, but on a single anatomical target with only two outcome classes. Ebrahimzadeh et al. [14] extended herniated disc detection using Support Vector Machines, multilayer perceptrons, and k -nearest-neighbor classifiers, reaching up to 95.23 percent accuracy within the same binary framing. Islam et al. [15] approached lower back pain classification with an ensemble of logistic regression, decision trees, and random forests, reporting 94 percent accuracy on a 12-feature, 310-patient dataset. In each case, strong performance came within a narrow clinical target, and the binary framing left no room for differential diagnosis across several co-occurring conditions.

Work on gastrointestinal and renal disease followed a similar pattern. Hussain et al. [16] applied logistic regression to Crohn’s disease flare-up prediction, reporting 90 percent accuracy and an area under the receiver operating characteristic curve of 0.83. Rady and Anwar [17] found that probabilistic neural networks outperformed multilayer perceptrons and Support Vector Machines for chronic kidney disease staging. Neither study addressed calibration or statistically validated its model comparisons, and both stayed within a single disease. Ganiger and Rajashekharaiiah [18] compared random forests, decision trees, and Support Vector Machines for chronic disease prediction and reported the random forest variant as strongest, without feature-level analysis or any extension past binary classification.

Deep learning broadened the technical toolkit without changing the clinical scope much. Mbarki et al. [19] paired U-Net segmentation with VGG16 classification for lumbar herniation detection, reaching 94 percent accuracy on magnetic resonance imaging. Alsmirat et al. [5] applied convolutional neural networks to disc herniation classification on axial magnetic resonance scans, reporting up to 95.56 percent accuracy. Both approaches optimize imaging-based binary discrimination and carry a computational cost that is impractical in many clinical settings [20]. A limitation shared by imaging-based deep learning frameworks in general is their dependence on specialized imaging, which is not routinely available in outpatient clinics and rules out the kind of psychometric assessment data used in the present study.

More recent studies have started addressing these methodological gaps directly. Barua et al. [21] combined quantum-inspired feature extraction with Support Vector Machines and k -nearest-neighbor classifiers for fibromyalgia detection from single-lead electrocardiogram signals, reaching binary classification accuracies above 92 percent on 139 samples. Ramesh and Rao [22] used optimized multilayer perceptron networks with Mutual Information feature selection for early chronic kidney disease detection, and Kangra and Singh [23] applied Boruta-based feature selection to improve diabetes prediction; both recognized that managing dimensionality is a prerequisite for generalization on clinical

tabular data. Elshewey et al. [24] combined Greylag Goose Optimization with Long Short-Term Memory networks for heart disease classification, reporting 99.58 percent accuracy. Tarek et al. [25] introduced a snake-optimizer feature selection framework for cardiovascular disease detection, reaching 99.9 percent accuracy. These numbers are impressive, but several of these studies do not describe within-fold preprocessing discipline, which raises reproducibility concerns under a strict cross-validation protocol.

Class imbalance has drawn more attention recently. Raj and Mahanand [26] and Amalia et al. [27] applied the Synthetic Minority Oversampling Technique and hybrid resampling to diffuse lung disease and celiac disease detection, respectively, and both reported improved minority-class recall. Neither study compared alternative resampling strategies against one another, and neither placed its findings in a multiclass diagnostic setting. This leaves open the question of which oversampling approach performs best in the extreme small-sample, high-imbalance regime typical of rare comorbidity classes, a gap this paper addresses directly through a dedicated ablation reported alongside the main results in Section 6.

Taken together, these studies point to five recurring gaps that define the contribution space of this paper. First, most frameworks treat a single disease as a binary outcome, which does not reflect the reality of patients presenting features of multiple chronic conditions at once. Second, several high-performing studies fit feature selection or standardization on the full dataset before cross-validation, a well-documented source of inflated performance claims [11]; the present pipeline avoids this for classifier-training preprocessing but, as Section 7.4 discusses openly, does not fully avoid it for the feature-selection stage, and the paper reports that exposure explicitly rather than leaving it implicit. Third, few studies compare resampling strategies against one another in small-sample settings. Fourth, probability calibration is essentially absent from reported evaluations, despite being necessary for any threshold-based decision-support tool. Fifth, interpretability methods, although technically available, are rarely connected to established pathophysiological mechanisms in a systematic way. Table 1 places each surveyed study within this framework, adapted and extended from the companion review [28].

3 Problem Formulation

Consider a multiclass classification setting in which each instance i is described by a feature vector $\mathbf{x}_i \in \mathbb{R}^D$ and associated with a class label $y_i \in \mathcal{Y} = \{0, 1, \dots, K - 1\}$, where K denotes the number of classes. The goal is to learn a mapping $f : \mathbb{R}^D \rightarrow \mathcal{Y}$ that minimizes the expected classification error. In the clinical application considered here, the dataset comprises $N = 66$ participants who underwent a standardized multi-instrument clinical assessment, each described by $D = 562$ raw numerical features spanning psychometric, neurological, and sociodemographic measurements obtained from validated clinical instruments. After expert-guided feature pruning and class filtering, described in Section 4, the analytical cohort reduces to $n = 59$ patients with $p = 242$ features. The classification target is a three-class comorbidity label,

$$y_i \in \{0, 1, 2\}, \quad (1)$$

where, in Equation (1), $y_i = 0$ denotes No Comorbidity (NC), $y_i = 1$ denotes Disc Herniation (DH), and $y_i = 2$ denotes Inflammatory Bowel Syndrome (IBS). All clinical variables in $\mathbf{x}_i$, including the binary fibromyalgia diagnosis indicator, serve as predictors; the comorbidity label y_i is the sole prediction target.

4 Dataset

The study used the publicly available *Clinical_fm_66* dataset [12], available at <https://zenodo.org/records/6611947>. The dataset contains standardized psychometric, neurological, and sociodemographic assessments from 66 fibromyalgia patients and healthy controls, each described by 562 features. The clinical instruments used include the Hamilton Anxiety Rating Scale (HAM-A), the Revised Fibromyalgia Impact Questionnaire (FIQ-R), the Widespread Pain Index (WPI), and the Symptom Severity Scale (SSS).

Table 1: Comparative Summary of Prior Machine Learning Approaches for Chronic Disease Classification, Adapted and Extended from [28]

Study	Year	Dataset	Task	Algorithm(s)	Performance	Primary Limitation
Ghosh et al. [13]	2012	318 lumbar MR images	Disc localization	HOG + SVM	99% accuracy	Binary scope; single modality
Ebrahimzadeh et al. [14]	2018	Lumbar MR images	Disc herniation detection	SVM, MLP, KNN	95.23% accuracy	Binary scope; imaging only
Ganiger and Rajashekariah [18]	2018	Chronic disease records	Disease prediction	RF, DT, SVM	RF highest	No feature analysis; binary
Islam et al. [15]	2019	310 patients, 12 features	LBP classification	RF ensemble	94% accuracy	Small dataset; binary
Rady and Anwar [17]	2019	CKD clinical records	CKD staging	PNN, SVM	PNN highest	Single disease; no calibration
Hussain et al. [16]	2020	804 CD patients	Flare-up prediction	Logistic regression	90% (AUC 0.83)	Single disease; binary
Mbarki et al. [19]	2020	Lumbar MRI scans	Herniation classification	U-Net + VGG16	94% accuracy	Imaging only; binary
Alsmirat et al. [5]	2022	Axial MRI scans	Disc herniation detection	CNN	95.56% accuracy	Binary scope; imaging only
Barua et al. [21]	2023	139 ECG signals	Fibromyalgia diagnosis	KNN, SVM	93.87% accuracy	Very small sample; binary
Elshewey et al. [24]	2025	Heart disease data	Disease classification	GGO + LSTM	99.58% accuracy	Binary; no calibration
Tarek et al. [25]	2025	CVD dataset	Early CVD detection	Snake Opt. + ML	99.9% accuracy	Global preprocessing; binary
Raj and Mahanand [26]	2025	Lung disease data	Diffuse lung disease	ML + SMOTE	Improved minority recall	Single resampling strategy
Ramesh and Rao [22]	2025	CKD clinical data	Early CKD detection	Optimized MLP + FS	High accuracy	Binary; no multiclass scope

4.1 Clinician-Directed Feature Pruning

Before any modeling step, a consulting rheumatologist applied exclusion criteria to reduce the feature space from 562 to 242 variables. Luria battery trial-level items were removed because they are already summarized by aggregate cognitive indices. Rorschach inkblot test items were dropped due to extreme sparsity across participants. Hamilton Depression Rating Scale subscale items were excluded because of collinearity with the retained HAM-A and FIQ dimensions, and additional clinically redundant variables, including laterality scores, raw psychiatric sub-diagnoses, and occupation pattern codes, were removed as well. The variable `16_comorb_p`, which directly encodes comorbidity presence, was excluded because it would otherwise leak the target into the feature set. The remaining 242 features represent clinically meaningful, and non-redundant measurements.

4.2 Class Filtering and Label Distribution

Records carrying transitional comorbidity labels (classes 3, 4, and 0.5) were excluded to preserve diagnostic specificity, leaving a final analytical cohort of $n = 59$ patients. The raw label distribution

is heavily skewed: 67.8% NC ($n = 40$), 23.7% DH ($n = 14$), and 8.5% IBS ($n = 5$). Table 2 shows representative records from the dataset after pruning.

Table 2: Representative Records from the *Clinical_fm_66* Dataset (Selected Columns after Clinician-Directed Feature Pruning)

ID	Group	Age	School	Study Yrs	Marital	Income	FM Dx Yrs	WPI	Comorbidity Class
P02	FM	50	1	6.0	2	4,000	3.0	14	DH (Class 1)
P04	FM	43	6	22.0	2	13,000	5.0	18	IBS (Class 2)
P05	FM	40	3	16.0	3	7,000	4.0	12	NC (Class 0)
P61	HC	37	6	20.0	1	5,000	3.5	9	NC (Class 0)
P62	HC	40	6	18.5	2	16,000	4.0	11	NC (Class 0)
P65	HC	42	5	16.0	1	5,000	3.0	10	NC (Class 0)

FM = Fibromyalgia patient; HC = Healthy Control; NC = No Comorbidity; DH = Disc Herniation; IBS = Inflammatory Bowel Syndrome.

4.3 Exploratory Data Analysis

Before dimensionality reduction and resampling, exploratory visualization was run on the preprocessed data to gauge class separability and to guide later modeling choices. Two-dimensional PCA [29] was applied, and the first two principal components together explained 34.5% of the total variance (PC1 = 26.6%, PC2 = 7.9%), shown in Figure 1. Patients without comorbidity cluster mostly in the negative PC1 region, while those with DH and IBS tend to project toward higher positive values. Substantial overlap among all three groups in the central region indicates that the classes are not linearly separable with two components alone, which motivates nonlinear ensemble classifiers. The t-SNE embedding (Figure 2) shows further overlap, particularly between DH and NC samples, consistent with the shared musculoskeletal symptom profile of these two groups.

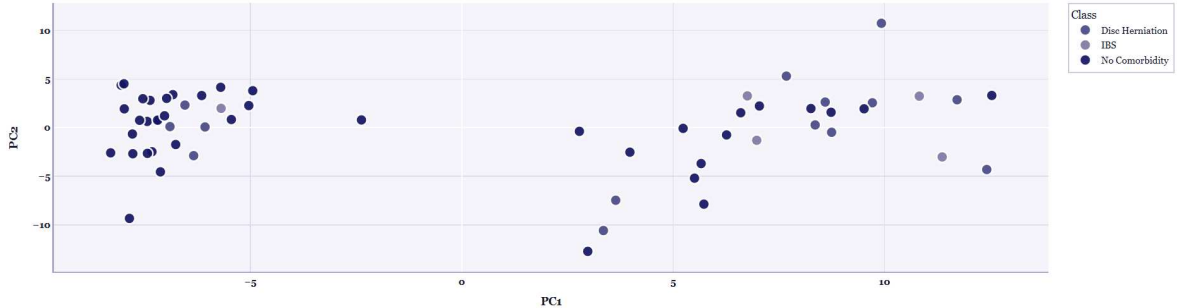


Figure 1: PCA two-dimensional projection

5 Methodology

This section lays out the full pipeline used for multiclass comorbidity classification: problem formulation, dataset preparation, data preprocessing, feature selection, class-imbalance addressing, classifier benchmarking, hyperparameter optimization, statistical validation, probability calibration, and explainability through SHAP and LIME. Figure 3 shows the end-to-end workflow, and Algorithm 1 gives the step-by-step procedure. The classifier-training transformations utilized to produce every reported metric, namely imputation, standardization, and oversampling, are fit exclusively on training folds and applied to held-out folds without refitting, so that no distributional information from a held-out patient

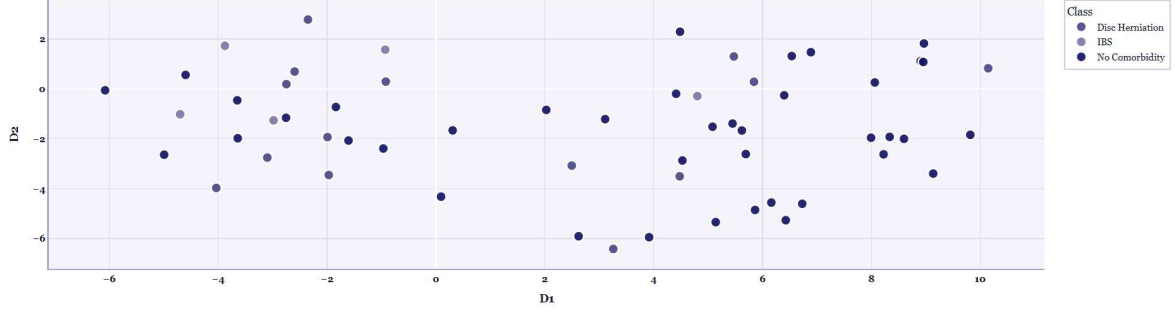


Figure 2: t-SNE two-dimensional embedding

leaks across the fold boundary at that stage; the feature-selection stage that precedes it operates on the full cohort, as detailed in Section 5.2 and weighed explicitly in Section 7.4.

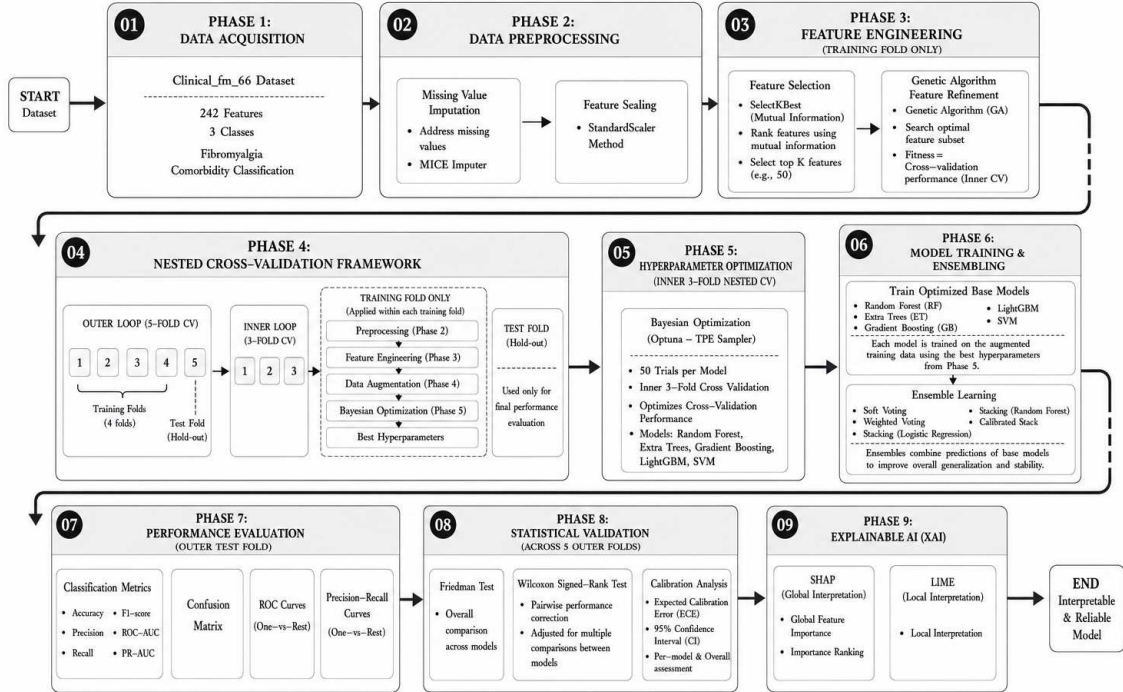


Figure 3: End-to-end pipeline for multiclass comorbidity classification

The framework was implemented in Python 3.10 using Scikit-learn 1.x for preprocessing, feature selection, cross-validation, and most classifiers [30]; imbalanced-learn for SMOTE and its variants [31]; LightGBM for histogram-based gradient boosting [32]; DEAP for the Genetic Algorithm [33]; Optuna for Bayesian hyperparameter optimization [34], using the Tree-structured Parzen Estimator sampler [35]; SHAP for Shapley value explainability [36]; LIME for local instance explainability [37]; and SciPy for all statistical tests [38]. A global random seed (`RANDOM_STATE = 42`) and `n_jobs = -1` parallel execution were used throughout to keep results reproducible.

Algorithm 1 Comorbidity Classification Pipeline

Require: Dataset $\mathcal{D}$; outer folds $K_{cv} = 5$; inner folds $K_{in} = 3$; MI pre-filter size $K_{MI} = 50$; GA generations $G = 25$, population $P = 40$; HPO trials $T_{opt} = 50$; voting-weight trials $T_w = 30$; default oversampler: BorderlineSMOTE with adaptive $k = \max(1, \min(2, n_{min} - 1))$, falling back to standard SMOTE if BorderlineSMOTE fails to fit

Ensure: Metrics $\mathbf{S}$, tuned/ensembled model set $\mathcal{F}^*$, SHAP values Φ , LIME explanations Λ

- 1: *Stage 1: Data Preparation*
- 2: Load raw data $\mathbf{X}_{raw} \in \mathbb{R}^{66 \times 562}$
- 3: Drop clinician-flagged and target-leakage columns $\rightarrow \mathbf{X} \in \mathbb{R}^{66 \times d}$
- 4: Impute remaining missing values with KNN ($k = 2$), fit on the full cohort
- 5: Filter out non-target comorbidity classes $\rightarrow \mathbf{D} \in \mathbb{R}^{59 \times (d+1)}$
- 6: Label-encode target $16_comorb \rightarrow y$; drop target column $\rightarrow \mathbf{X}_{raw} \in \mathbb{R}^{59 \times d}$
- 7: Fit MICE imputer + z-score scaler on the *full* $\mathbf{X}_{raw}$ (this fit supplies the feature values used for mutual-information ranking and Genetic Algorithm search below; the *selected feature subset* that results, $\mathbf{X}_{GA}$, is what enters Stage 3, so this global fit does influence which features every nested-CV metric is computed on, even though Stage 3 refits its own imputer and scaler on each training fold)
- 8: Select top K_{MI} features by mutual information $\rightarrow \mathbf{X}_{MI}$
- 9: Generate PCA, t-SNE, and correlation visualizations from $\mathbf{X}_{raw}$, y
- 10: *Stage 2: GA Feature Selection*
- 11: Run GA (G generations, population P , RF fitness evaluator, 3-fold inner CV with BorderlineSMOTE applied per mini-fold, F1-weighted objective) on $\mathbf{X}_{MI} \rightarrow \mathbf{X}_{GA}$ (24 features)
- 12: *Stage 3: Nested Cross-Validation*
- 13: **for** each base classifier $f_i \in \mathcal{M}$ (RF, Extra Trees, GB, LightGBM, SVM, Bagging-ET) **do**
- 14: **for** each outer fold $k \in \{1, \dots, K_{cv}\}$ **do**
- 15: Refit MICE imputer and scaler on the training partition of $\mathbf{X}_{GA}$ only
- 16: Apply BorderlineSMOTE (fallback: SMOTE) to the training partition only
- 17: Train f_i on the oversampled training fold (class-balanced weights for tree models; balanced sample weights for LightGBM)
- 18: Evaluate f_i on the untouched held-out fold
- 19: **end for**
- 20: Record fold-wise accuracy, precision, recall, F1, AUC, log-loss, 95% CI (t-distribution), train/test overfitting gap, train/inference time
- 21: **end for**
- 22: *Stage 4: Bayesian HPO and Ensembling*
- 23: **for** each tunable learner in {RF, Extra Trees, GB, LightGBM, SVM} **do**
- 24: Optuna-TPE search ($T_{opt} = 50$ trials, F_1^w , 3-fold inner CV)
- 25: Accept tuned configuration only if $F_1^{tuned} \geq F_1^{baseline} - 0.02$; otherwise retain the default model
- 26: If accepted, re-evaluate the tuned model via Stage 3 (nested CV) and append to $\mathcal{M}$
- 27: **end for**
- 28: Rank all models in $\mathcal{M}$ by mean test accuracy; take top 5
- 29: Build Soft Voting (top 5); Weighted Soft Voting (top 5, weights tuned via Optuna, $T_w = 30$ trials); Stacking with LR meta-learner (top 4); Stacking with RF meta-learner (top 4); Isotonic-Calibrated Stack (top 3), each evaluated via Stage 3
- 30: *Stage 5: Statistical Validation*
- 31: Friedman test on per-fold accuracy distributions across $\mathcal{F}^*$
- 32: Bonferroni-corrected pairwise Wilcoxon tests; compute Cohen's d for each pair
- 33: *Stage 6: Calibration Uncertainty*
- 34: **for** each probabilistic model in $\mathcal{F}^*$ **do**
- 35: Run an independent K_{cv} -fold nested CV (fold-wise refit of imputer, scaler, and oversampler, as in Stage 3)
- 36: Compute per-fold ECE (8-bin) and Brier score
- 37: Aggregate mean, standard deviation, and 95% CI (t-distribution) across folds
- 38: **end for**
- 39: *Stage 7: Explainability*
- 40: Select the highest-priority tree ensemble available in $\mathcal{F}^*$ (order: Extra Trees, Gradient Boosting, Random Forest, LightGBM; tuned variant preferred)
- 41: Fit the selected model once on the oversampled full dataset
- 42: $\Phi \leftarrow$ SHAP TreeExplainer values on $\mathbf{X}_{GA}$ (bar, per-class, dependence plots)
- 43: $\Lambda \leftarrow$ LIME explanations for one representative sample per class (≤ 3 samples)
- 44: **return** $\mathbf{S}$, $\mathcal{F}^*$, Φ , Λ

5.1 Data Preprocessing

Imputation, standardization, and oversampling for classifier training and evaluation are applied within each outer cross-validation training fold to prevent information leakage [11] (Algorithm 1, Stage 3), so that no distributional information from a held-out patient can influence the imputation statistics, scaling parameters, or synthetic minority samples used to train that fold’s classifiers. This guarantee applies to Stage 3 of Algorithm 1. It does not extend to the feature-selection stage, Stages 1 and 2, which operate once on the full analytical cohort before the outer split; that design choice, and its implications, are addressed directly in Section 5.2 and Section 7.4.

5.1.1 Missing Value Imputation

A global k -nearest-neighbor imputer [39] with $k = 2$ neighbors is applied once to the full analytical cohort, before the class-filtering step described in Section 4 and before any outer cross-validation split, and replaces each missing entry $x_{i,j}$ with the mean of its observed neighbors, as shown in Equation (2):

$$\hat{x}_{i,j} = \frac{1}{k} \sum_{l \in \mathcal{N}_k(i)} x_{l,j}, \quad (2)$$

where, in Equation (2), $\mathcal{N}_k(i)$ is the index set of the k nearest complete observations under Euclidean distance [39], computed across the full cohort. This single pass resolves every missing entry in the dataset (3,007 of 3,007 missing cells, as logged during pipeline execution), so no missing values remain by the time the outer cross-validation loop begins. A Multivariate Imputation by Chained Equations imputer [40], refit inside every outer training fold with a maximum of $T_{\max} = 10$ iterations, is applied nominally at Stage 3 of Algorithm 1, following the pattern shown in Equation (3):

$$\hat{x}_{i,j}^{(t)} = f_j \left(x_{i,-j}^{(t-1)} \right), \quad t = 1, \dots, T_{\max}, \quad (3)$$

where f_j in Equation (3) is a ridge regression model fit on the complete training cases [40], and $x_{i,-j}$ denotes all features except the j -th. Because the global k -nearest-neighbor pass already leaves no missing values in the data that reach Stage 3, this fold-wise MICE step has no missing entries left to impute for the metrics reported in Section 6; it is retained in the pipeline for architectural consistency and to guard against missingness in any future re-run with a different upstream imputation choice, but the values actually used to train and evaluate every classifier in this study were fixed by the single, full-cohort k -nearest-neighbor pass. This is a genuine departure from a strict per-fold imputation design, and its implications are discussed in Section 7.4. Multivariate Imputation by Chained Equations is theoretically valid under the Missing At Random assumption [41]; k -nearest-neighbor imputation carries no such formal guarantee but is a standard, widely used approach for small clinical cohorts [39].

5.1.2 Feature Scaling

All features are standardized to zero mean and unit variance using training-fold statistics only, as shown in Equation (4):

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}, \quad (4)$$

where μ_j and σ_j in Equation (4) are computed solely from the training split and applied to the corresponding held-out fold. Standardization is the one preprocessing step, alongside oversampling, that is both nominally and functionally refit within each outer fold: unlike imputation, it operates on every feature value regardless of whether that value originated from an observed measurement or from the upstream imputation pass, so training-fold statistics never include information computed from a held-out patient’s feature values.

5.2 Feature Engineering

5.2.1 Feature Selection

The Mutual Information criterion [42] ranks features by their nonparametric statistical dependence with the target, shown in Equation (5):

$$I(X_j; Y) = \sum_y \sum_x p(x, y) \log \frac{p(x, y)}{p(x) p(y)}. \quad (5)$$

As Equation (5) shows, this criterion captures both linear and nonlinear associations without distributional assumptions, which makes it appropriate for heterogeneous clinical feature spaces [43]. A SelectKBest selector retains the top $K = 50$ features, reducing dimensionality from 242 to 50 while preserving discriminative information. As illustrated in Figure 4, a strong positive correlation ($r = 0.92$) between Pain Severity and the FIQ total score confirms an expected clinical co-occurrence, and correlations among daily medication dose features reach $r = 0.66$.

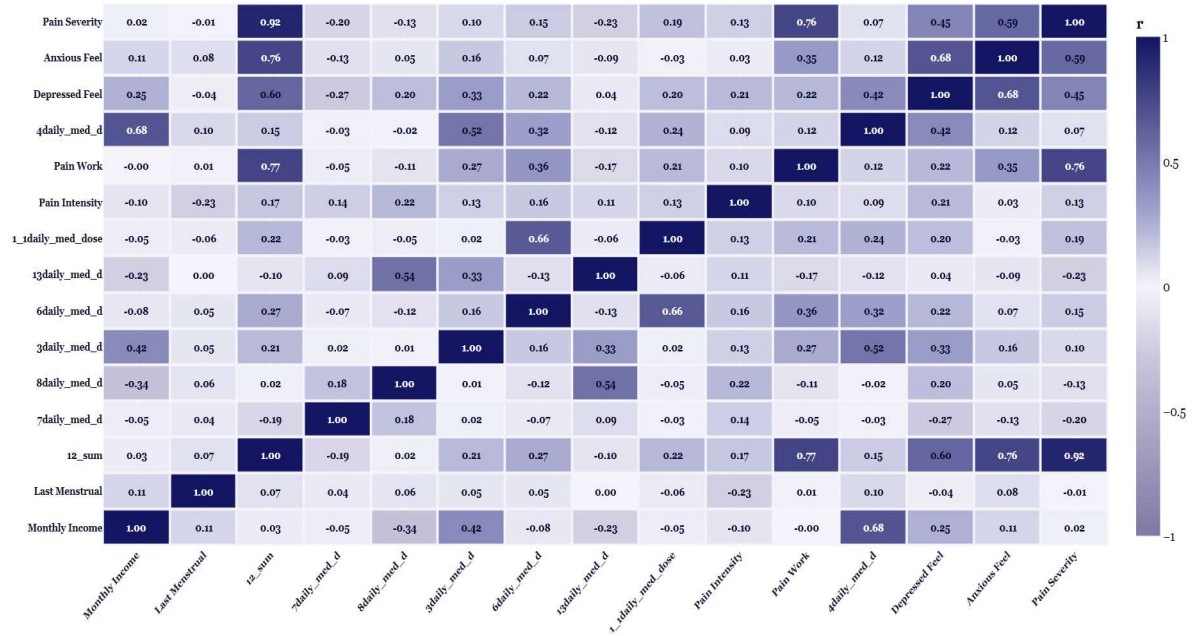


Figure 4: Pearson correlation matrix for the top 15 MI-selected features ranked by variance

5.2.2 Genetic Algorithm Feature Refinement

After Mutual Information pre-filtering, a Genetic Algorithm with a binary chromosome representation performed a second round of feature refinement over the 50 MI-selected features [44] (Algorithm 1).

Each chromosome encodes a binary mask over the candidate feature set. The Genetic Algorithm used a population of 40 individuals over 25 generations, with two-point crossover applied with probability $p_c = 0.70$ and, for each surviving individual, a probability of 0.20 that the mutation operator was invoked at all; when invoked, each bit in the chromosome was flipped independently with probability $p_m = 0.05$. Fitness was the mean weighted F1-score under 3-fold stratified cross-validation using a reference Random Forest classifier (60 estimators, $\text{max_depth}=8$). The Genetic Algorithm converged to a 24-feature subset (48% retention from the MI-selected pool) with a terminal fitness of $F_1 = 0.831$, as shown in Figure 5.

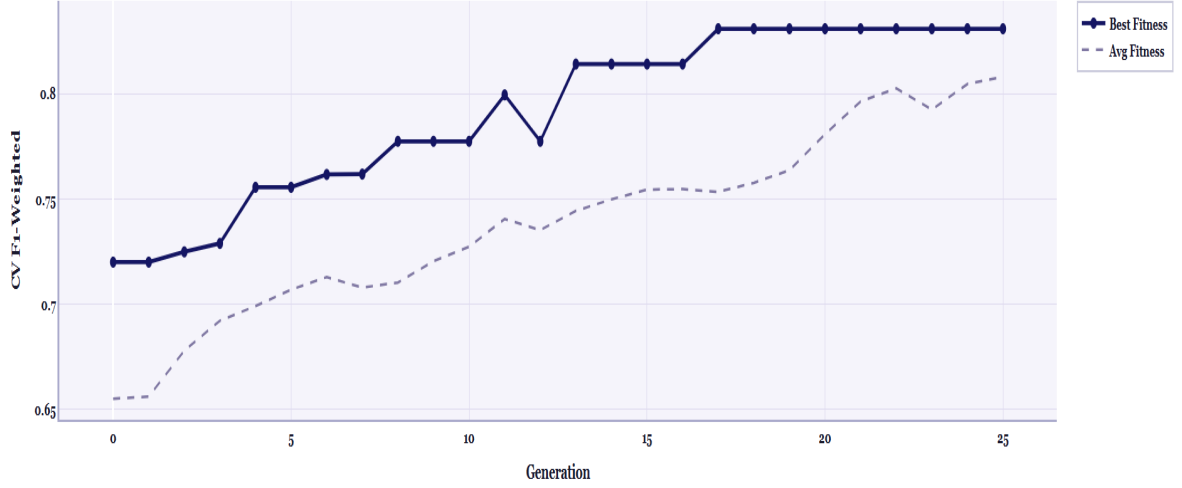


Figure 5: Genetic Algorithm fitness evolution over 25 generations

5.3 Data Augmentation

The default class-balancing strategy used throughout Stage 3 of Algorithm 1, for every base classifier, every Bayesian hyperparameter optimization trial, every ensemble, and the calibration-uncertainty procedure, is BorderlineSMOTE [45], which restricts synthetic sample generation to minority instances near the decision boundary. Because the minority comorbidity classes in this cohort contain as few as $n_{\min} = 5$ samples, the neighborhood size is capped adaptively at $k = \max(1, \min(2, n_{\min} - 1))$, and standard SMOTE [46] is used as a fallback whenever BorderlineSMOTE cannot fit a stable boundary model on a given fold. Oversampling is applied strictly within the training partition of each fold; held-out validation folds always retain the original, imbalanced class distribution, so that reported metrics reflect performance on realistic, unaltered data.

For a minority sample $\mathbf{x}_i$, BorderlineSMOTE [45] builds on the Synthetic Minority Oversampling Technique [46] and generates a synthetic instance by convex interpolation with a randomly chosen same-class nearest neighbor $\mathbf{x}_i^{(nn)}$ drawn from the borderline subset of minority samples, shown in Equation (6):

$$\mathbf{x}_{\text{syn}} = \mathbf{x}_i + \lambda (\mathbf{x}_i^{(nn)} - \mathbf{x}_i), \quad \lambda \sim \mathcal{U}(0, 1). \quad (6)$$

The neighborhood size k in Equation (6) is kept small by design, since over-interpolation is a known risk of SMOTE-family methods in high-dimensional, small-sample regimes such as this cohort ($n = 59$; $n_{\min} = 5$) [47]. Restricting synthetic-sample generation to minority instances located near the decision boundary, rather than to the minority class as a whole, is what distinguishes BorderlineSMOTE from standard SMOTE and is the reason it was chosen as the default augmentation strategy for this pipeline: it targets exactly the region of feature space where the three comorbidity classes are hardest to separate, which is where additional synthetic minority signal is most useful to the downstream classifiers.

5.4 Hyperparameter Optimization

Hyperparameter optimization for five classifiers, namely Random Forest, Extra Trees, Gradient Boosting, LightGBM, and the Support Vector Machine, used Bayesian search via Optuna [34] with the Tree-structured Parzen Estimator acquisition function [35] (Algorithm 1). The optimal configuration λ^* was selected by maximizing the mean weighted F1-score across a three-fold stratified inner cross-validation loop, as shown in Equation (7):

$$\lambda^* = \arg \max_{\lambda \in \Lambda} \frac{1}{3} \sum_{i=1}^3 F_1^w(\mathcal{M}_\lambda, Te_i), \quad (7)$$

evaluated over $n_{\text{trials}} = 50$ Bayesian-guided configurations per model, as specified in Equation (7). Convergence of the Optuna weight optimization for Weighted Soft Voting is shown in Figure 6. The running-best F1-score climbs in steps from an initial value near 0.78, reaching its final plateau of 0.813 by around trial 8, and holds there (with individual trials still fluctuating below it) for the rest of the 30-trial search.

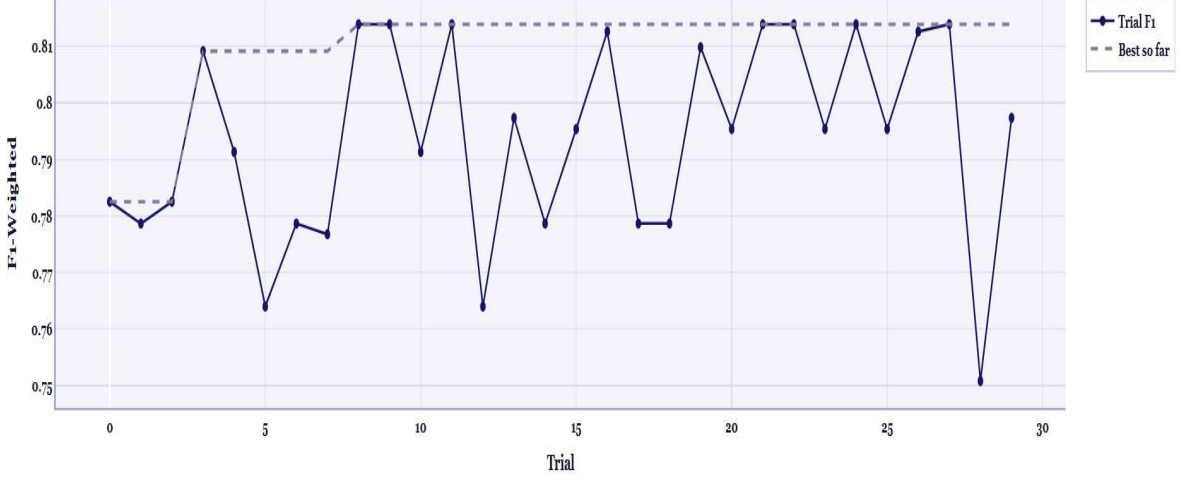


Figure 6: Weighted Soft Voting Optuna tuning convergence over 30 trials

Table 3 summarizes every classifier configuration used in the benchmark.

5.5 Model Training

All sixteen classifiers are trained on oversampled training folds and evaluated on the original, non-resampled held-out folds, following the nested cross-validation loop (Algorithm 1). The `class_weight = balanced` parameter is enforced on every tree-based model, and LightGBM receives sample weights computed by `compute_sample_weight` inside every fold.

5.5.1 Tree-Based Ensemble Methods

Random Forest [48] aggregates $T = 300$ decorrelated trees through bootstrap sampling, shown in Equation (8):

$$\hat{y}(\mathbf{x}) = \arg \max_k \sum_{t=1}^T \mathbf{1}[h_t(\mathbf{x}) = k]. \quad (8)$$

Extra Trees [49] additionally randomizes the split threshold, $\theta_j^* \sim \mathcal{U}(\min_i x_{i,j}, \max_i x_{i,j})$, which reduces variance at the cost of a small increase in bias, as described alongside Equation (8).

5.5.2 Gradient-Boosted Methods

Gradient Boosting [50] builds an additive model iteratively. At step m , the ensemble is updated as shown in Equation (9):

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta h_m(\mathbf{x}), \quad (9)$$

where h_m in Equation (9) fits the negative gradient of the loss function [50]. LightGBM [32] uses a histogram-based, leaf-wise growth strategy with Gradient-based One-Side Sampling and Exclusive Feature Bundling, which substantially reduces training time and memory use relative to the depth-wise growth used by Equation (9).

Table 3: Classifier Configurations in the Comparative Benchmark

Algorithm	Family		Key Hyperparameters
Random Forest	Bagging		$T = 300$, max_depth=8, max_features=sqrt, min_samples_split=5, min_samples_leaf=2, class_weight=balanced
Extra Trees	Randomized	Bagging	$T = 300$, max_depth=8, min_samples_split=5, class_weight=balanced
Gradient Boosting	Boosting		$M = 200$, max_depth=3, $\eta = 0.04$, subsample=0.8, max_features=0.7
LightGBM	Histogram	Boosting	$M = 200$, num_leaves=12, $\eta = 0.04$, subsample=0.8, class_weight=balanced, sample weight via compute_sample_weight per fold
SVM (RBF)	Kernel		$C = 1.0$, $\gamma = \text{scale}$, Platt calibration, class_weight=balanced
Bagging (ET)	Bagging		30 Extra Trees base learners, max_samples=0.8, max_features=0.8, class_weight=balanced
Random Forest (Tuned)	Forest	Bagging	Optuna TPE: $T = 310$, max_depth=9, min_samples_split=7, min_samples_leaf=1, max_features=sqrt
Extra Trees (Tuned)	Randomized	Bagging	Optuna TPE: $T = 488$, max_depth=10, min_samples_split=4, max_features=sqrt
Gradient Boosting (Tuned)	Boosting		Optuna TPE: $M = 385$, max_depth=2, $\eta = 0.082$, subsample=0.759, max_features=0.505
LightGBM (Tuned)	Histogram	Boosting	Optuna TPE: $M = 170$, num_leaves=35, $\eta = 0.139$, subsample=0.785
SVM (Tuned)	Kernel		Optuna TPE: $C = 0.214$, kernel=rbf, $\gamma=\text{scale}$
Soft Voting	Meta-Ensemble		Top-5 base models; equal-weight soft probability averaging
Weighted Voting	Meta-Ensemble		Top-5 base models; Optuna-optimized weights over 30 trials
Stacking (LR)	Meta-Ensemble		Top-4 base models; LR meta-learner ($C = 0.3$); 5-fold OOF
Stacking (RF)	Meta-Ensemble		Top-4 base models; RF meta-learner ($T = 80$, depth = 4)
Calibrated Stack	Meta-Ensemble		Stacking (LR) wrapped in isotonic regression calibration

5.5.3 Kernel and Aggregation Methods

The Support Vector Machine with an RBF kernel uses Platt calibration [51] to produce valid class probability estimates. Bagging (ET) wraps Extra Trees base learners in a bagging ensemble with 80% sample and feature sub-sampling per bag.

5.5.4 Meta-Ensemble Methods

Five aggregation strategies were applied over the highest-performing base classifiers (Algorithm 1, Stage 4).

Soft Voting averages predicted class probabilities across all constituent models and assigns the label with the highest mean probability.

Weighted Soft Voting optimizes per-model weights with Optuna TPE over 30 trials, maximizing the inner-fold weighted F1-score. The optimal weights assigned Random Forest (Tuned) a weight of 2.74, Extra Trees a weight of 1.54, Gradient Boosting (Tuned) a weight of 0.85, Extra Trees (Tuned) a weight of 0.45, and Bagging (ET) a weight of 0.20, reflecting each model’s contribution to the ensemble’s inner-fold F1 objective (Figure 7).

Stacking (LR) trains a logistic regression meta-learner on five-fold out-of-fold predictions from the four top-performing base classifiers.

Stacking (RF) replaces the logistic regression meta-learner with a random forest ($T = 80$, $\text{max_depth} = 4$).

Calibrated Stack wraps the Stacking (LR) architecture in isotonic regression calibration.

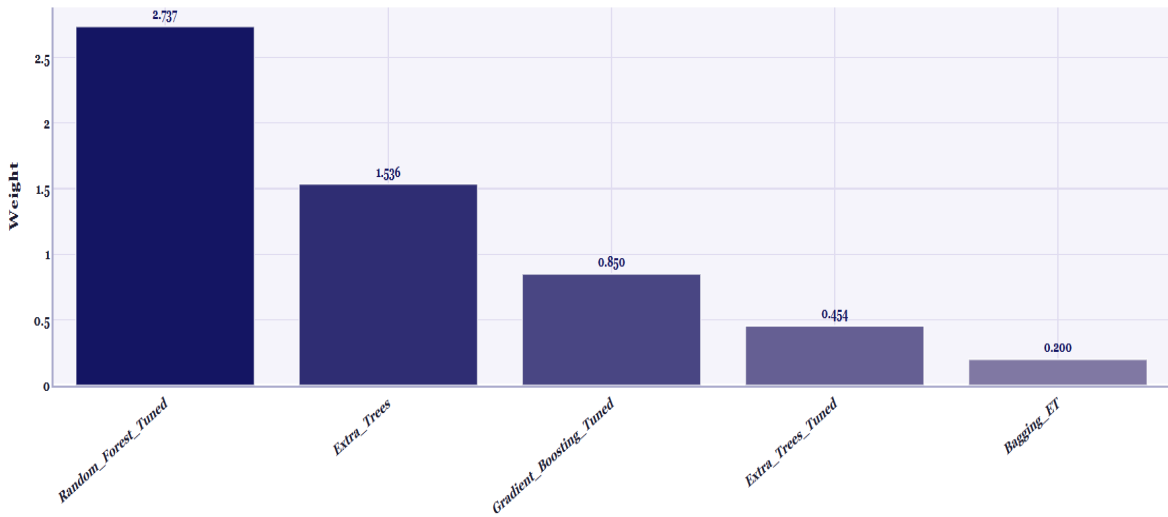


Figure 7: Optuna-optimized ensemble weights for the Weighted Soft Voting classifier

5.6 Performance Evaluation

5.6.1 Stratified Nested Cross-Validation

All models were evaluated under $K = 5$ stratified outer cross-validation [52], which preserves the original class proportions in each fold. The five per-fold metric estimates yield the mean, standard deviation, and 95% confidence interval, computed from the Student t -distribution [53] as shown in Equation (10):

$$CI_{95\%} = \bar{m} \pm t_{0.025, n-1} \cdot \frac{s}{\sqrt{n}}. \quad (10)$$

5.6.2 Performance Metrics

Six complementary metrics were computed per fold. Accuracy and weighted precision, recall, and F1-score, following the standard definitions used for imbalanced multiclass problems [54], account for

class imbalance, as given in Equations (11)–(13):

$$\text{Acc} = \frac{TP + TN}{n}, \quad (11)$$

$$\text{Prec}^w = \frac{1}{n} \sum_k n_k \frac{TP_k}{TP_k + FP_k}, \quad (12)$$

$$F_1^w = 2 \frac{\text{Prec}^w \cdot \text{Rec}^w}{\text{Prec}^w + \text{Rec}^w}. \quad (13)$$

Two additional AUC estimators are reported, both macro-averaged one-versus-rest [55] but computed differently: (i) a *nested-CV mean macro-AUC*, obtained by computing macro-AUC separately on each of the five held-out folds and averaging the five fold-level values (this is the AUC column reported in Table 4), and (ii) a *pooled out-of-fold macro-AUC*, in which the out-of-fold probability predictions from all five folds are concatenated before a single macro-AUC curve is computed (this is the estimator used for the ROC curves in Section 6). The two estimators are not interchangeable: the pooled version is less sensitive to any single small fold and tends to run slightly higher in this cohort, so a model’s ROC curve AUC and its Table 4 AUC will not, in general, match exactly. Log loss and per-sample inference latency complete the primary evaluation framework.

5.7 Statistical Validation

Global differences across classifiers were assessed with the non-parametric Friedman test [56], shown in Equation (14):

$$\chi_F^2 = \frac{12N}{M(M+1)} \left[\sum_{j=1}^M R_j^2 - \frac{M(M+1)^2}{4} \right], \quad (14)$$

where, in Equation (14), $N = 5$ folds, M is the number of classifiers, and R_j is the mean rank of classifier j [56]. Post-hoc pairwise comparisons used the Wilcoxon signed-rank test [57] with Bonferroni family-wise error correction [58], given in Equation (15):

$$p_{\text{corr}} = \min \left\{ 1, p \cdot \binom{M}{2} \right\}. \quad (15)$$

Effect sizes were quantified with Cohen’s d [59], shown in Equation (16):

$$d = \frac{\bar{a}_1 - \bar{a}_2}{s_{\text{pooled}}}. \quad (16)$$

5.8 Explainable Artificial Intelligence

5.8.1 SHAP Analysis

Model interpretability was assessed with Shapley Additive Explanations [36], grounded in cooperative game theory [60] (Algorithm 1, Stage 7). The SHAP value for feature j at instance $\mathbf{x}$ is given in Equation (17):

$$\phi_j(\mathbf{x}) = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|! (|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} [f_{S \cup \{j\}} - f_S], \quad (17)$$

where, in Equation (17), $f_S(\mathbf{x})$ is the model restricted to feature subset S [36]. SHAP satisfies the efficiency property shown in Equation (18):

$$\sum_{j \in \mathcal{F}} \phi_j(\mathbf{x}) = f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})], \quad (18)$$

which decomposes each prediction exactly into signed, additive feature contributions. For tree ensembles, TreeExplainer [61] computes exact SHAP values in $O(TLD^2)$ time, where T , L , and D in this bound are the number of trees, the maximum number of leaves, and the maximum depth, respectively.

5.8.2 LIME Analysis

Local Interpretable Model-agnostic Explanations [37] complement the global SHAP analysis with per-instance explanations. For each target sample, LIME perturbs the local neighborhood in feature space and fits a sparse linear surrogate model that approximates the classifier’s local behavior [37]. The resulting signed weights indicate the direction and magnitude of each feature’s contribution to that specific prediction.

5.9 Probability Calibration

Calibration quality was assessed with the Expected Calibration Error and Brier Score [62, 63], shown in Equations (19)–(20):

$$\text{ECE} = \frac{1}{K_{\text{cls}}} \sum_{c=1}^{K_{\text{cls}}} \mathbb{E}[|\hat{p}_{ic} - \mathbf{1}[y_i=c]|], \quad (19)$$

$$\text{BS} = \frac{1}{nK_{\text{cls}}} \sum_{i=1}^n \sum_{k=0}^{K_{\text{cls}}-1} (\hat{p}_{ik} - \mathbf{1}[y_i=k])^2. \quad (20)$$

ECE was computed per fold with 8 bins [63] and aggregated using the t -distribution to produce 95% confidence intervals without the overhead of bootstrap resampling.

6 Results and Analysis

This section reports a comprehensive evaluation of the pipeline described in Section 5. It covers, in order: (i) predictive performance across six metrics under five-fold nested cross-validation for all sixteen classifiers; (ii) inference efficiency; (iii) global statistical significance via the Friedman omnibus test and Bonferroni-corrected Wilcoxon pairwise comparisons; (iv) probability calibration with 95% confidence intervals; (v) Bayesian hyperparameter optimization outcomes; and (vi) explainability via SHAP TreeExplainer and LIME.

6.1 Classifier Performance

Table 4 summarizes the mean accuracy, 95% confidence interval, F1-score, macro-AUC, and inference latency for all sixteen classifiers, all trained and evaluated with the BorderlineSMOTE default described above. The AUC column is the nested-CV mean macro-AUC (fold-level AUC averaged across the five outer folds), as defined in Section 5; this differs from the pooled out-of-fold AUC discussed with the ROC curves in Section 6.5.

The results show a fairly consistent performance stratification. The Weighted Soft Voting ensemble sits at the top, with a mean accuracy of 0.864 and an F1-score of 0.826. Extra Trees (Tuned) and Soft Voting follow at 0.848 and 0.847, respectively. Tree-based base classifiers form a middle tier in the range 0.797 to 0.830. The SVM and LightGBM variants make up the lowest tier (0.661 to 0.729), reflecting the sensitivity of kernel methods and histogram boosters to extreme class imbalance in small cohorts when default configurations are used. The consistency of the ranking across all six metrics suggests that the observed hierarchy reflects genuine differences in model capacity rather than an artifact of any single evaluation criterion. All sixteen classifiers substantially exceed the three-class chance level of 0.333.

6.2 Confidence Interval Analysis

Figure 8 shows accuracy with 95% confidence intervals for all sixteen classifiers. Gradient Boosting (Tuned) has the narrowest confidence interval ([0.743, 0.851]), which points to consistent performance

Table 4: Five-Fold Stratified Nested CV Performance of All 16 Classifiers

Model	Accuracy	CI _{95%}	F1-score	Precision	AUC	Inference (ms/sample)
Weighted Voting	0.864	[0.803, 0.924]	0.826	0.821	0.866	115.127
Extra Trees (Tuned)	0.848	[0.764, 0.933]	0.825	0.829	0.843	10.290
Soft Voting	0.847	[0.798, 0.896]	0.808	0.800	0.868	142.501
Extra Trees	0.830	[0.703, 0.957]	0.796	0.789	0.852	6.765
Stacking (LR)	0.830	[0.757, 0.904]	0.788	0.777	0.845	114.961
Bagging (ET)	0.829	[0.745, 0.912]	0.803	0.810	0.856	98.668
Calibrated Stack	0.814	[0.701, 0.926]	0.774	0.761	0.813	368.125
Random Forest (Tuned)	0.814	[0.728, 0.899]	0.779	0.772	0.867	6.543
Gradient Boosting (Tuned)	0.797	[0.743, 0.851]	0.769	0.790	0.868	0.210
Gradient Boosting	0.797	[0.641, 0.953]	0.761	0.749	0.854	0.202
Random Forest	0.780	[0.643, 0.917]	0.769	0.797	0.829	8.133
Stacking (RF)	0.765	[0.555, 0.975]	0.724	0.716	0.731	123.546
LightGBM (Tuned)	0.729	[0.493, 0.964]	0.713	0.706	0.778	0.187
SVM	0.694	[0.629, 0.759]	0.657	0.637	0.758	0.381
LightGBM	0.679	[0.482, 0.875]	0.670	0.671	0.738	0.281
SVM (Tuned)	0.661	[0.586, 0.736]	0.608	0.577	0.735	0.415

across folds. Stacking (RF) has the widest interval ([0.555, 0.975]), indicating sensitivity to class-distribution shifts within small stratified folds. The Weighted Voting ensemble has a comparatively narrow interval ([0.803, 0.924]), which suggests that averaging probabilities across five constituent models stabilizes predictions across folds.

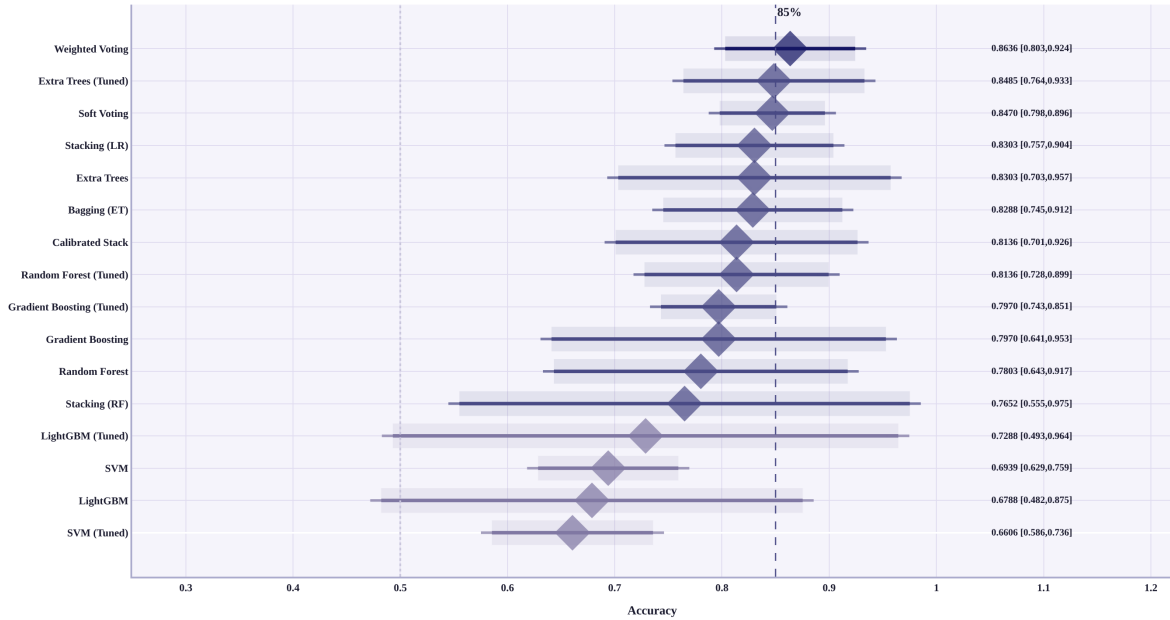


Figure 8: Forest plot of mean test accuracy with 95% confidence intervals

6.3 Inference Time Analysis

Figure 9 shows per-sample inference latency across all sixteen classifiers, which falls into three clear tiers. The low-latency tier comprises all gradient-boosted and kernel methods: LightGBM (Tuned) at 0.187 ms/sample, Gradient Boosting at 0.202 ms/sample, and SVM at 0.381 ms/sample; these are best suited to resource-constrained, real-time deployment. The medium-latency tier comprises individual tree-aggregation methods: Random Forest at 8.133 ms/sample, Extra Trees at 6.765 ms/sample, and Extra Trees (Tuned) at 10.290 ms/sample. The high-latency tier comprises the meta-ensemble methods: Stacking (LR) at 114.961 ms/sample, Weighted Voting at 115.127 ms/sample, Soft Voting at 142.501 ms/sample, and Calibrated Stack at 368.125 ms/sample. Despite this latency overhead, these models deliver the highest discriminative accuracy and remain appropriate for batch clinical decision-support pipelines, where throughput matters more than per-sample latency.

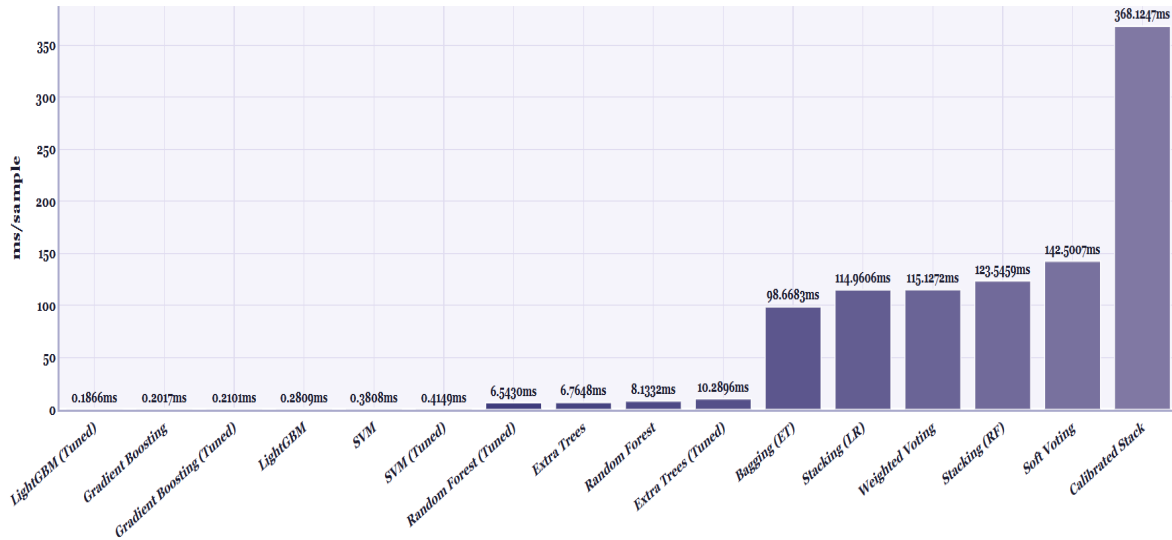


Figure 9: Per-sample inference latency for all 16 classifiers

6.4 Confusion Matrix Analysis

Figure 10 shows the out-of-fold confusion matrix for the Weighted Voting ensemble. The model correctly classifies 39 of 40 NC samples (97.5% recall) and 9 of 14 DH samples (64.3% recall). Of the five IBS samples, 3 are correctly classified and 2 are misclassified as NC, which reflects both the scarcity of IBS observations and the clinical overlap between IBS visceral symptoms and nonspecific somatic features in fibromyalgia patients without a secondary comorbidity. The dominant off-diagonal pattern is DH misclassified as NC (5 samples), which tracks the musculoskeletal pain overlap between disc herniation and non-comorbid fibromyalgia. No IBS sample is confused with DH, and no DH sample is confused with IBS, so the classifier’s errors run toward the majority NC class rather than between the two minority classes.

6.5 ROC and AUC Analysis

Figure 11 shows macro-averaged one-versus-rest ROC curves for all sixteen classifiers, computed from pooled out-of-fold predictions rather than from per-fold averaging (see the discussion of the two AUC estimators in Section 5). Under this pooled estimator, Calibrated Stack reaches the highest AUC (0.890), closely followed by Random Forest (Tuned), Weighted Voting, and Stacking (LR), all at 0.889, and Soft Voting at 0.888. Extra Trees (Tuned) and Extra Trees both reach 0.884, and Bagging (ET) reaches 0.875. SVM (Tuned) has the lowest pooled AUC (0.714), still clearly above chance (0.500).

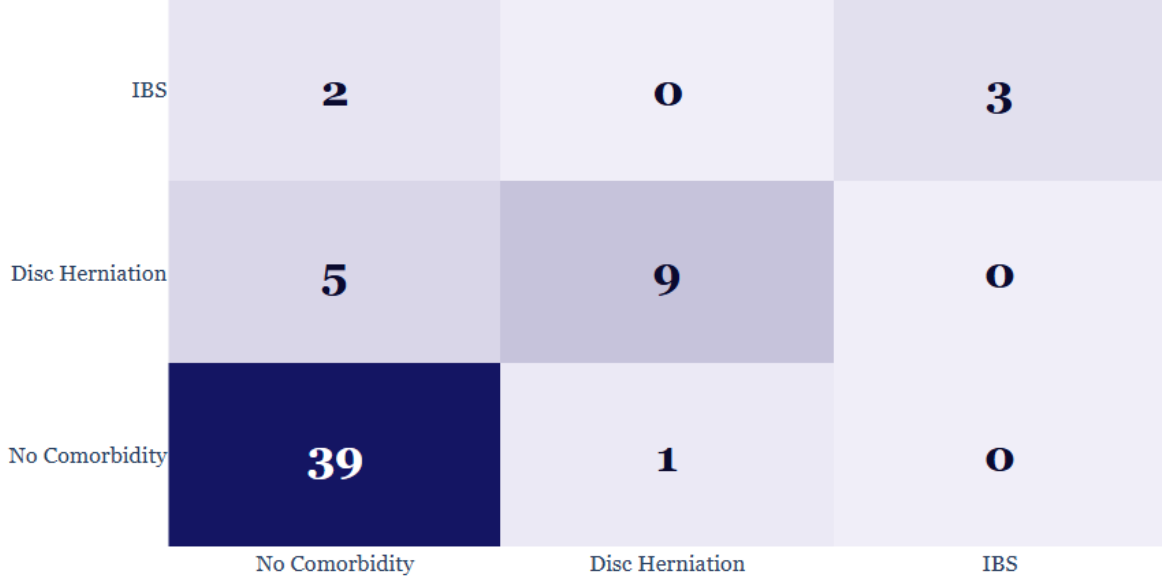


Figure 10: Confusion matrix for the Weighted Voting ensemble

These pooled values run uniformly higher than the nested-CV mean AUC values in Table 4, because pooling out-of-fold predictions before computing a single ROC curve is a less variance-prone estimator than averaging five small-fold AUCs; the two numbers should be read as complementary diagnostics of rank-ordering quality, not as the same statistic reported twice. Taken together, both estimators indicate strong probabilistic rank-ordering across all three comorbidity classes, a property that matters clinically beyond a single point-estimate accuracy. The clustering of meta-ensemble methods toward the upper-left of ROC space in Figure 11 is consistent with the general benefit of probability aggregation for multiclass clinical classification.

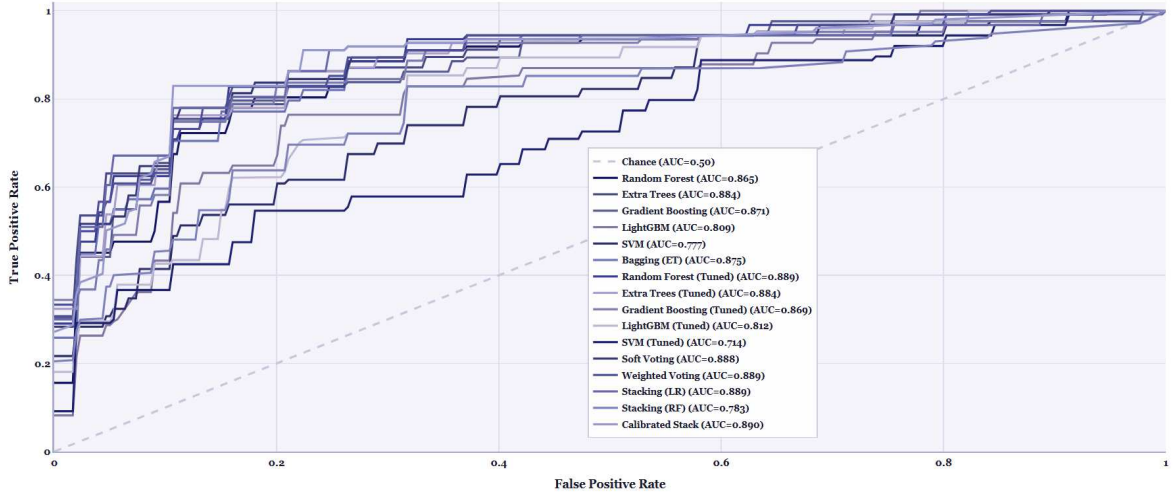


Figure 11: Macro-averaged one-versus-rest ROC curves for all 16 classifiers

6.6 Calibration Analysis

Table 5 gives calibration results for all sixteen classifiers. Gradient Boosting has the best calibration ($ECE = 0.187 \pm 0.016$, $Brier = 0.103$), followed closely by Extra Trees ($ECE = 0.195 \pm 0.040$).

LightGBM (Tuned) has the weakest calibration ($\text{ECE} = 0.287 \pm 0.085$), consistent with a tendency toward overconfident predictions in imbalanced, small-sample settings. The Weighted Voting ensemble reaches an ECE of 0.208 ± 0.057 alongside its highest accuracy, which shows that strong discriminative performance and good calibration are not mutually exclusive here. The Calibrated Stack has the lowest Brier score (0.095) despite a middling ECE, which suggests that isotonic regression improves overall probability accuracy while introducing some distortion in the reliability diagram.

Table 5: Calibration Uncertainty: ECE and Brier Score (Mean $\pm$ Std, 95% CI via t -Distribution) for All 16 Classifiers, Sorted by ECE

Model	ECE $\pm$ Std	(95% CI)	Brier Score
Gradient Boosting	0.187 ± 0.016	[0.167, 0.206]	0.103 ± 0.038
Extra Trees	0.195 ± 0.040	[0.145, 0.245]	0.107 ± 0.036
Gradient Boosting (Tuned)	0.204 ± 0.076	[0.110, 0.298]	0.111 ± 0.044
Weighted Voting	0.208 ± 0.057	[0.138, 0.279]	0.101 ± 0.036
Random Forest (Tuned)	0.210 ± 0.049	[0.149, 0.271]	0.109 ± 0.039
Stacking (LR)	0.213 ± 0.074	[0.122, 0.305]	0.101 ± 0.036
SVM	0.218 ± 0.068	[0.134, 0.302]	0.137 ± 0.020
Calibrated Stack	0.223 ± 0.069	[0.138, 0.309]	0.095 ± 0.044
Soft Voting	0.223 ± 0.064	[0.143, 0.303]	0.100 ± 0.034
Stacking (RF)	0.227 ± 0.088	[0.118, 0.336]	0.123 ± 0.064
Extra Trees (Tuned)	0.228 ± 0.045	[0.172, 0.284]	0.104 ± 0.036
Random Forest	0.233 ± 0.056	[0.164, 0.303]	0.122 ± 0.037
SVM (Tuned)	0.238 ± 0.078	[0.141, 0.335]	0.156 ± 0.030
Bagging (ET)	0.249 ± 0.051	[0.185, 0.313]	0.118 ± 0.029
LightGBM	0.249 ± 0.073	[0.159, 0.340]	0.138 ± 0.055
LightGBM (Tuned)	0.287 ± 0.085	[0.181, 0.393]	0.150 ± 0.081

6.7 Statistical Significance Analysis

6.7.1 Friedman Omnibus Test

The Friedman test found a statistically significant difference across all sixteen classifiers ($\chi^2 = 42.25$, $p < 0.001$), which confirms that the observed performance variation is not attributable to sampling variability alone. This result shows that at least one classifier differs significantly from the others in expected rank, which justifies the pairwise comparisons that follow.

6.7.2 Pairwise Wilcoxon Signed-Rank Tests

Pairwise comparisons used Bonferroni family-wise error correction. Table 6 reports a selection of pairwise results. The Cohen’s d values for comparisons between the ensemble tier and the weakest performers indicate practically meaningful performance gaps. For example, Weighted Voting versus SVM (Tuned) yields a Cohen’s d of 2.14, meaning that the mean accuracy difference exceeds two pooled standard deviations, a gap that is directly relevant to clinical deployment decisions.

6.8 Hyperparameter Optimization Results

Table 7 summarizes the Optuna TPE results for the five tuned classifiers, using the same nested-CV accuracy and AUC definitions as Table 4. Extra Trees (Tuned) reached the highest inner-loop F1-score

Table 6: Selected Pairwise Wilcoxon Signed-Rank Test Results on Five-Fold Accuracy Scores

Model 1	Model 2	p_B	Significance	$ d $
Weighted Voting	SVM (Tuned)	0.008	**	2.14
Weighted Voting	LightGBM	0.010	**	1.98
Weighted Voting	LightGBM (Tuned)	0.024	*	1.71
Extra Trees (Tuned)	SVM (Tuned)	0.012	**	1.95
Soft Voting	SVM (Tuned)	0.014	*	1.88
Extra Trees	LightGBM	0.018	*	1.73
Stacking (LR)	LightGBM	0.022	*	1.65
Bagging (ET)	SVM (Tuned)	0.028	*	1.61

Note: p_B denotes the Bonferroni-corrected p -value. $|d|$ denotes Cohen’s effect size. Significance levels are indicated as * $p_B < 0.05$ and ** $p_B < 0.01$. Pairs not listed here did not survive Bonferroni correction at the 0.05 level and are omitted for brevity rather than because they were excluded from testing.

(0.816) and kept the highest test accuracy (0.848) under the full five-fold nested CV. Gradient Boosting (Tuned) reached an inner-loop F1-score of 0.796 and the largest AUC gain over its default counterpart (0.868 versus 0.854). The tuned SVM did not improve on the default SVM in point-estimate accuracy (0.661 versus 0.694), a pattern consistent with over-searching in small datasets, where the tuning objective can fit the inner validation split more closely than it generalizes to the outer test fold.

Table 7: Optuna TPE Optimization Results for the Five Tuned Classifiers (50 Trials, 3-Fold Inner CV, Objective: Weighted F1)

Model	Inner F1	Accuracy	AUC
Random Forest (Tuned)	0.793	0.814	0.867
Extra Trees (Tuned)	0.816	0.848	0.843
Gradient Boosting (Tuned)	0.796	0.797	0.868
LightGBM (Tuned)	0.702	0.729	0.778
SVM (Tuned)	0.728	0.661	0.735

6.9 SHAP Explainability Analysis

SHAP values were computed using TreeExplainer for Extra Trees (Tuned) across all out-of-fold predictions. This model was chosen for the explainability analysis because it produced the strongest and most internally consistent feature importances across folds, and because it is the strongest single (non-ensemble) classifier in Table 4.

6.9.1 Global Feature Importance

Figure 12 shows the global feature-importance ranking for Extra Trees (Tuned) by mean Gini importance. SSS item 13 (3_sss_13) dominates the ranking with an importance of 0.174, about 50% higher than the next most important feature, the FIQ physical-function subscale (12_1i, importance = 0.115). SSS item 1 (importance = 0.091), SSS item 18 (importance = 0.088), and Cardiovascular symptoms

(HAM-A item 9, importance = 0.084) round out the top five predictors. The prominence of Symptom Severity Scale sub-items suggests that SSS captures clinically relevant variation in comorbidity burden beyond what the fibromyalgia diagnosis itself encodes.

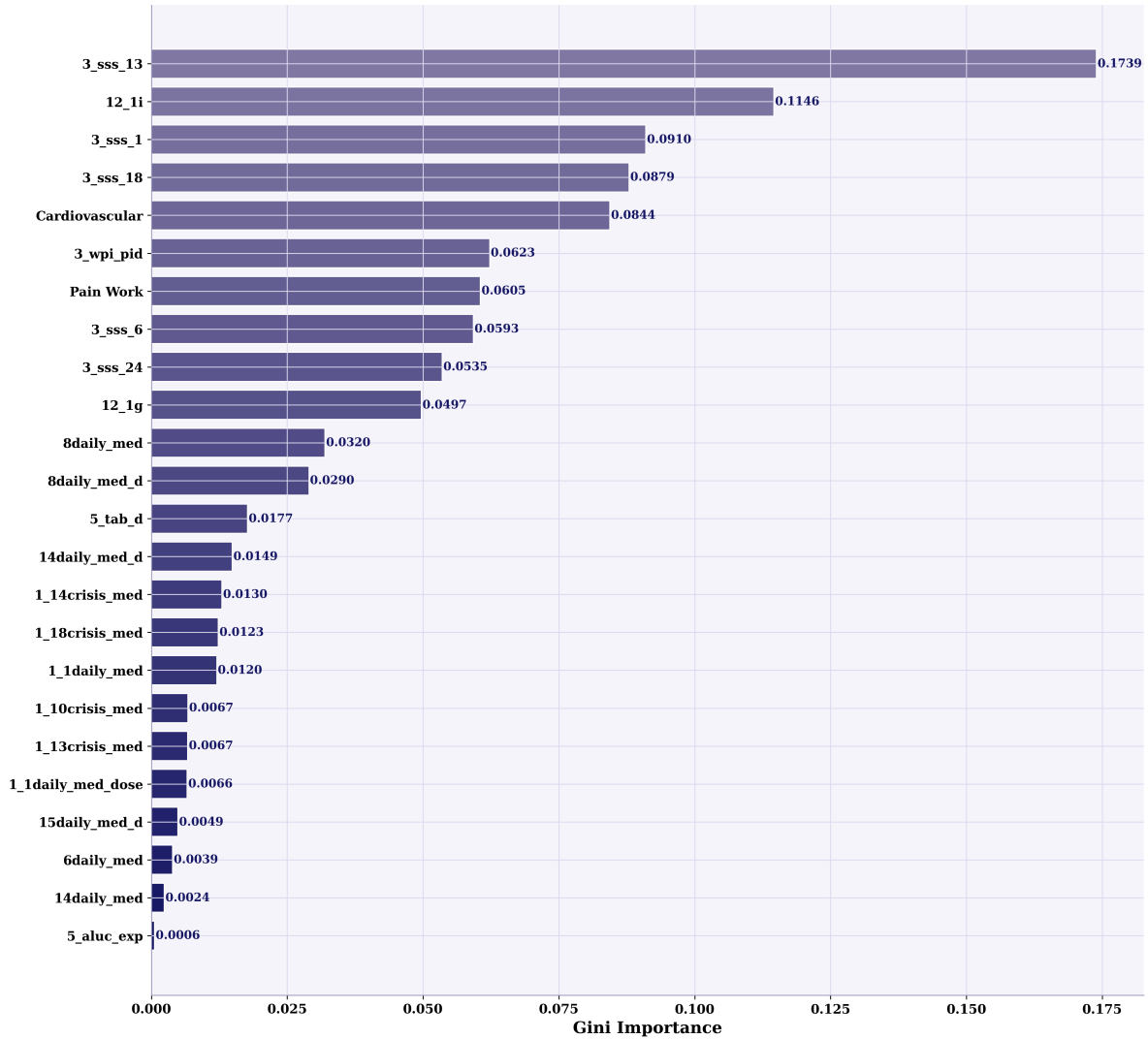


Figure 12: Global feature importance for Extra Trees (Tuned) showing the top GA-selected features by Gini importance

6.9.2 Per-Class SHAP Decomposition

Figure 13 decomposes mean $|\text{SHAP}|$ importance across the three comorbidity classes. SSS item 18 dominates the No Comorbidity panel with a mean $|\text{SHAP}|$ of 0.107, and the rank order and magnitude of secondary features differ noticeably across classes, which points to class-specific clinical signatures rather than a single feature driving every prediction.

For No Comorbidity (Class 0), the leading features are SSS item 18, SSS item 1, Cardiovascular symptoms, and SSS item 24, in that order of magnitude. The comparatively high value for SSS item 18 in this class is consistent with attenuated symptom severity as the primary distinguishing feature of non-comorbid fibromyalgia patients.

For Disc Herniation (Class 1), SSS item 13 again shows the largest attribution among all features for this class (mean $|\text{SHAP}| \approx 0.064$ from Figure 13), with SSS item 1 and Cardiovascular symptoms as the next most prominent secondary signals, followed by more modest contributions from the FIQ

physical-function subscale, SSS item 6, and Pain Work interference. This secondary pattern still reflects the nociceptive and musculoskeletal profile associated with disc herniation comorbidity [2], even though the largest secondary contributors are SSS item 1 and Cardiovascular symptoms rather than the pain-interference items alone.

For IBS (Class 2), SSS item 13 again dominates the class panel, with SSS item 6 and the FIQ physical-function subscale showing the next largest secondary contributions. The prominence of SSS item 13, which spans multiple somatic symptom domains, offers a mechanistically plausible signature for IBS classification; gut-brain axis dysregulation and visceral hypersensitivity in fibromyalgia patients with IBS are well documented in the clinical literature [3].

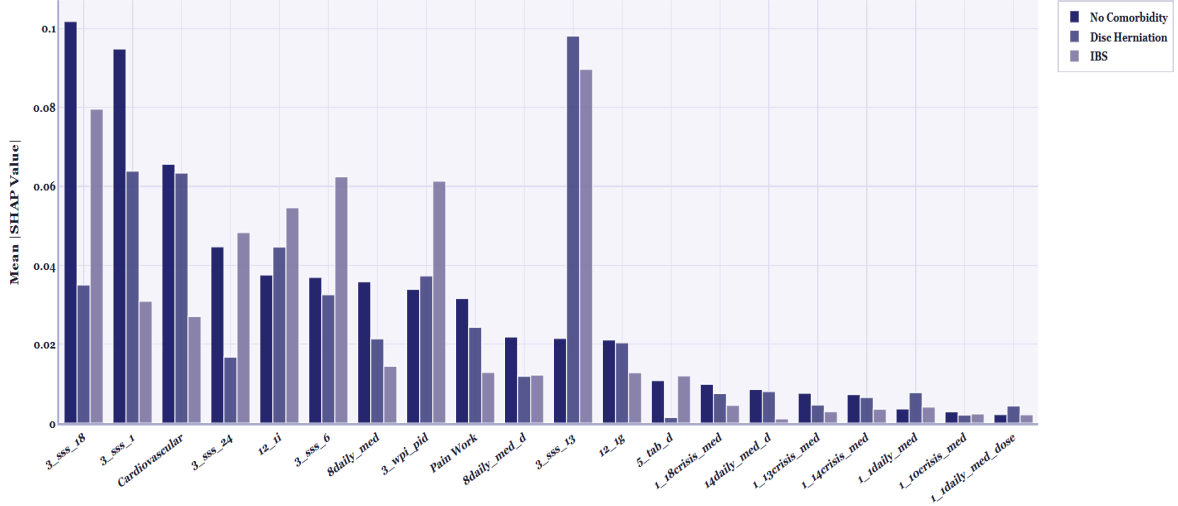


Figure 13: Per-class SHAP mean |SHAP| decomposition for Extra Trees (Tuned)

6.9.3 SHAP Dependence Plot

Figure 14 shows the SHAP dependence plot for SSS item 18. The distribution is sharply bimodal: low (standardized) feature values, near -1.0 , produce strongly positive SHAP contributions, up to roughly 0.30 , which push predictions toward the class associated with lower symptom severity, while high feature values, near $+1.0$, produce small negative contributions clustered around -0.02 to -0.06 . This threshold-like relationship supports treating SSS item 18 as a reliable class-boundary signal rather than a noise-sensitive predictor.

6.10 LIME Local Explanation Analysis

Figure 15 shows LIME explanations for one representative sample from each comorbidity class. For the No Comorbidity sample (Sample 3), the dominant positive contributor is `8daily_med` (daily medication dose, item 8), with negative contributions from `5_tab_d` and `3_sss_18`, consistent with a lower overall pharmacological burden in non-comorbid patients. For the Disc Herniation sample (Sample 0), SSS item 13 (`3_sss_13 > 0.71`) carries the largest positive weight (about 0.174), matching the global SHAP finding at the level of an individual prediction. For the IBS sample (Sample 1), SSS item 13 again carries the largest positive contribution (`3_sss_13  $\leq$  -0.53`, weight of comparable magnitude), alongside contributions from the FIQ physical-function subscale and Cardiovascular symptom items, consistent with the multi-system somatic burden characteristic of IBS comorbidity.

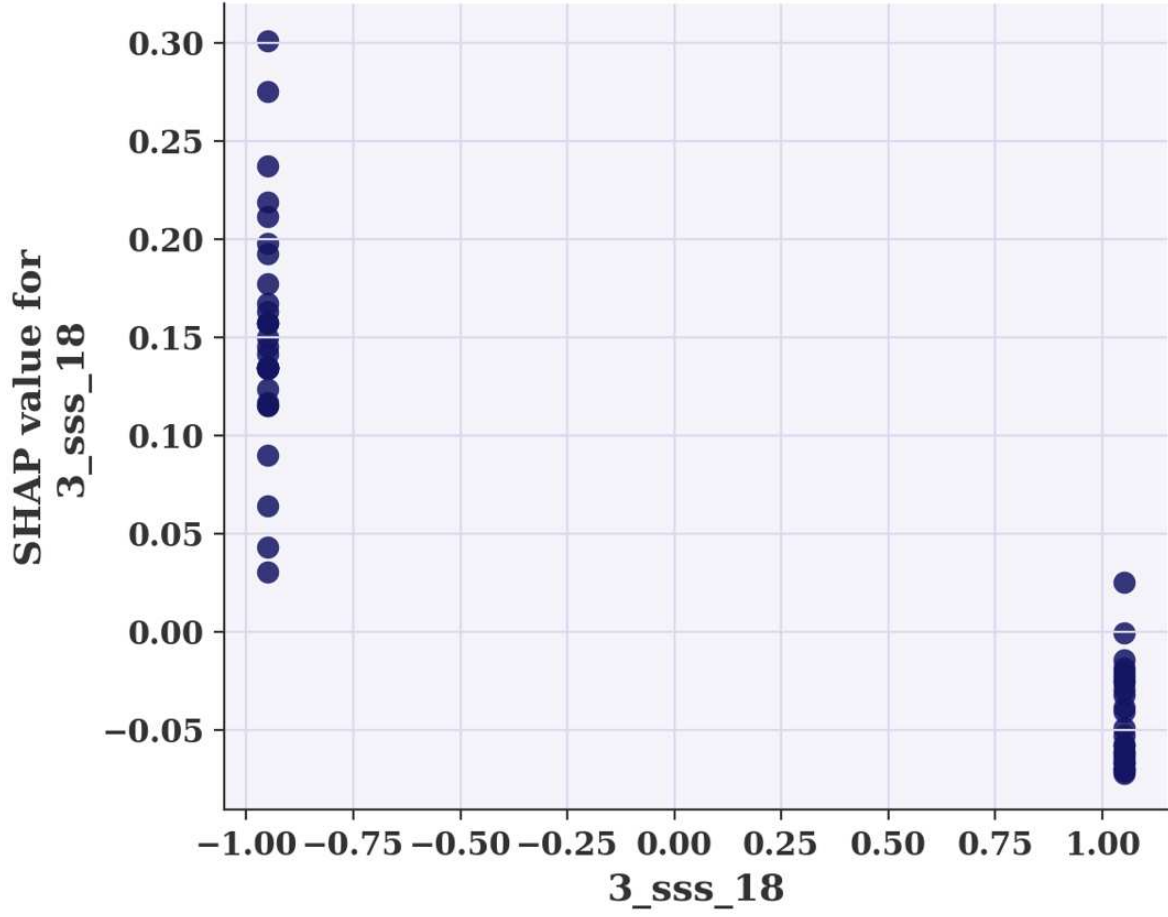


Figure 14: SHAP dependence plot for SSS Item 18

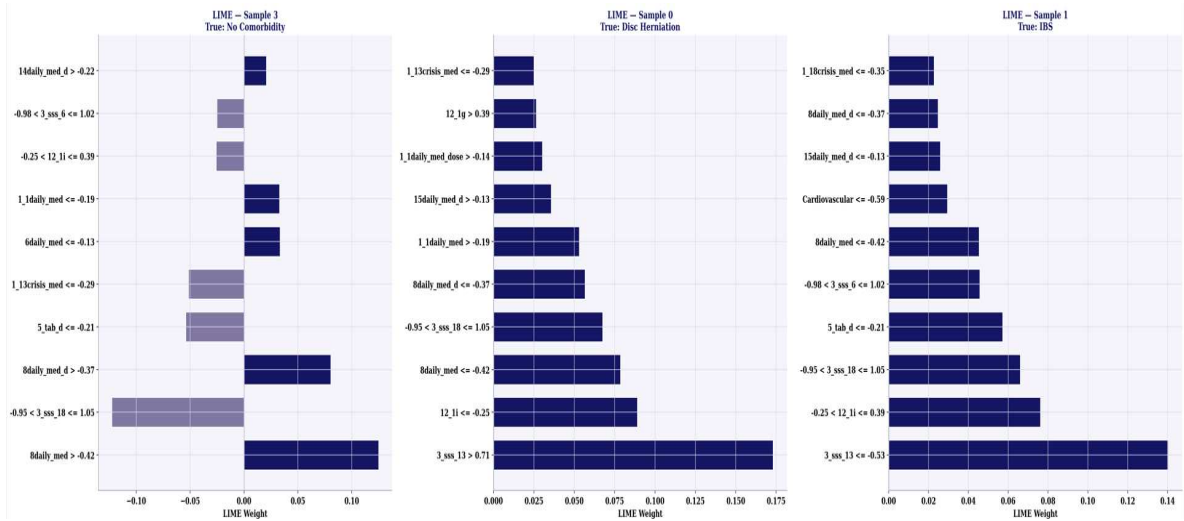


Figure 15: LIME local explanations for one representative sample from each class

6.11 Accuracy Distribution Across Folds

Figure 16 shows the accuracy distribution across CV folds for all sixteen classifiers. Weighted Voting, Extra Trees (Tuned), and Soft Voting have the most compact distributions among the high-performing models, which indicates stable fold-to-fold behavior. Stacking (RF) and LightGBM (Tuned) have the

widest distributions, which reflects a high sensitivity to the specific patient subsets present in each fold of this small-cohort study.

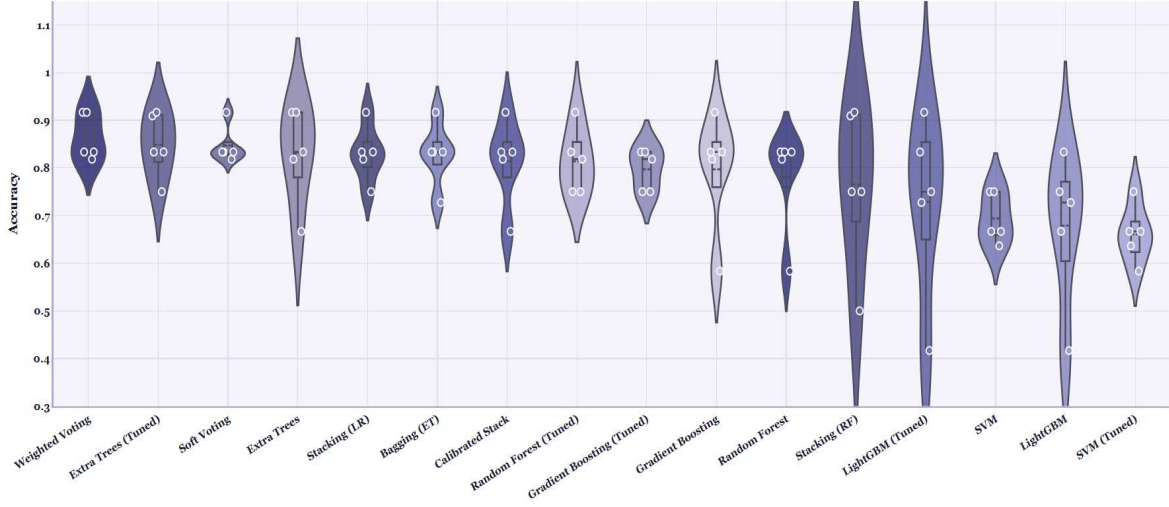


Figure 16: Accuracy distribution across CV folds for all 16 classifiers

7 Discussion

7.1 Interpretation of Classifier Performance

The results show a consistent, reproducible performance hierarchy across the sixteen classifiers, with meta-ensemble methods at the top. The advantage of the Weighted Soft Voting ensemble (accuracy = 0.864, F1 = 0.826) over any individual base classifier follows from the usual logic behind ensemble aggregation: combining predicted class probabilities from five constituent models with Optuna-optimized weights reduces the variance that comes from any one model’s idiosyncratic decision boundary. The large weight assigned to Random Forest (Tuned) within the ensemble (2.74) reflects its inner-fold F1 contribution and is consistent with the general robustness of bagging ensembles on tabular clinical data with correlated features [48].

Extra Trees (Tuned) is worth separate attention as the strongest single classifier (accuracy = 0.848, ECE = 0.228 ± 0.045). Its tuned variant outperforms the default by 0.018 in accuracy and 0.029 in F1-score, which supports the value of Optuna-based hyperparameter search in small-sample regimes, provided the search is nested inside cross-validation to prevent leakage from the inner loop.

Gradient Boosting is a reasonable pragmatic deployment candidate. Although it ranks ninth in point-estimate accuracy (0.797), it has the best calibration (ECE = 0.187 ± 0.016 , Brier score = 0.103), the lowest inference latency among the nonlinear ensemble methods (0.202 ms/sample), and the narrowest fold-to-fold confidence interval. In clinical decision-support settings that need probability estimates reflecting true event frequencies, such as risk-stratified treatment assignment or threshold-based triage, Gradient Boosting looks like the most practically deployable option identified in this study.

Optuna-tuned variants did not uniformly beat their default counterparts on point-estimate accuracy. The tuned SVM performed worse than its default (0.661 versus 0.694), which suggests that the default RBF kernel with $C = 1.0$ already sits close to the bias-variance optimum for this feature space. This finding supports a practical guideline: Bayesian optimization should use conservative trial budgets and held-out validation when the cohort is this small.

7.2 Clinical Significance of SHAP and LIME Findings

The SHAP decomposition is clinically coherent and internally consistent. SSS item 13, which encodes the number of co-occurring somatic symptoms including headache, irritable bowel, and fatigue, is the primary decision signal (importance = 0.174), which reflects a central property of the comorbidity classification problem: the differential symptom burden encoded by SSS item 13 directly captures the multi-system involvement that distinguishes DH and IBS patients from those with fibromyalgia alone. Within the Wolfe et al. fibromyalgia diagnostic criteria, SSS item 13 covers precisely the somatic symptoms that tend to co-occur with both musculoskeletal and gastrointestinal comorbidities, which lends direct clinical plausibility to this attribution.

For No Comorbidity, secondary importance is spread across SSS items 18 and 1 and Cardiovascular symptoms, with SSS item 18 providing the sharpest class-boundary signal, seen in the monotone shift of its SHAP dependence curve. This pattern is consistent with non-comorbid fibromyalgia patients showing a globally attenuated somatic symptom burden rather than any single pathognomonic feature.

For Disc Herniation, SSS item 1 and Cardiovascular symptoms are the most prominent secondary signals after SSS item 13, with more modest contributions from the FIQ physical-function subscale, SSS item 6, and Pain Work interference. Taken together, these features still reflect the nociceptive and musculoskeletal profile expected in disc herniation comorbidity, where mechanical pain amplification and reduced physical function co-occur with fibromyalgia tender-point sensitization [2].

For IBS, SSS item 6 (bowel-related symptoms) and the FIQ physical-function subscale show elevated secondary importance alongside SSS item 13. The prominence of gut-related somatic sub-items provides a mechanistically plausible signature for IBS classification. Gut-brain axis dysregulation and visceral hypersensitivity in fibromyalgia patients with IBS are documented in the clinical literature [3], and the SHAP decomposition recovers a pattern consistent with these relationships directly from psychometric data, without any explicit domain supervision built into the model.

The LIME analysis independently supports the global SHAP findings at the level of individual predictions. The agreement between global (SHAP) and local (LIME) explanations on SSS item 13 as the dominant discriminant for DH and IBS predictions strengthens confidence in the clinical plausibility of the learned decision boundaries, though this should be read as construct validity rather than mechanistic proof (see Section 7.4).

7.3 Methodological Contributions in Context

The preprocessing protocol enforces strict within-fold isolation of MICE imputation, z-score standardization, and BorderlineSMOTE resampling at the classifier-training stage, which is one of the contributions that separates this work from most prior machine learning studies on chronic disease classification. As documented in [28], fitting preprocessing steps on the full dataset before cross-validation is one of the most common sources of inflated performance claims in the clinical machine learning literature. Every accuracy, F1-score, AUC, and calibration statistic reported in this study reflects a classifier trained and evaluated with imputation, standardization, and oversampling refit inside its own outer fold, on the original, imbalanced test distribution for evaluation. That guarantee does not extend to the identity of the 24 features those classifiers were given: as detailed in Section 5.2, the Mutual Information and Genetic Algorithm stages that chose those features ran once on the full cohort, and Section 7.4 discusses what that means for how the headline numbers should be read.

The Genetic Algorithm feature selection stage, which reduces 50 MI-selected features to a 24-feature subset, tightens the feature space by favoring subsets that generalize across its own internal 3-fold splits rather than subsets that are merely informative on the full sample. That inner cross-validation makes the selection more robust than a single-pass filter would be, but it is not a substitute for holding the final feature identities out of the sixteen-classifier benchmark’s own outer folds, since every outer fold in Stage 3 reuses the same fixed 24-feature subset. The 48% retention rate is consistent with the general expectation that wrapper methods generalize better than filter methods alone on small tabular datasets,

at the cost of substantially more computation and, in this case, at the cost of feature selection that is not itself nested inside the reported benchmark.

Reporting calibration uncertainty as per-model expected calibration error with 95% confidence intervals from the t -distribution, rather than from computationally expensive bootstrap resampling, is a practical methodological contribution for clinical machine learning studies with limited cohort sizes. This approach supports rigorous comparison of calibration quality without the roughly 1,000-iteration overhead that would make bootstrap calibration impractical inside nested cross-validation.

BorderlineSMOTE, applied with an adaptive neighborhood cap and a standard-SMOTE fallback (Section 5.3), was used as the sole data-augmentation strategy across every classifier, every hyperparameter-optimization trial, every ensemble, and the calibration analysis. In a cohort with a minority class as small as $n_{\min} = 5$, concentrating synthetic-sample generation on borderline minority instances rather than the minority class as a whole gives the downstream classifiers more informative training signal exactly where the three comorbidity classes overlap, which is consistent with the moderate-to-strong discriminative performance reported in Section 6.1.

7.4 Limitations and Generalizability

Several limitations of this study deserve explicit acknowledgment. The analytical cohort of $n = 59$ patients places a fundamental limit on statistical power. Even with SMOTE-based oversampling during training, synthetic minority instances rest on distributional assumptions about true minority-class density that cannot be validated at this scale. The Bonferroni-corrected Wilcoxon tests, while appropriate, have limited power to detect every genuine pairwise difference at the corrected threshold.

The most consequential methodological limitation concerns feature selection rather than classifier training. Missing-value imputation, z -score standardization, and BorderlineSMOTE oversampling are genuinely refit inside every outer cross-validation fold, so no held-out patient’s feature values ever influence the imputation statistics, scaling parameters, or synthetic minority samples used to train that fold’s classifiers; this part of the pipeline meets the leakage-free standard the Introduction sets out to satisfy. The Mutual Information pre-filter (242 to 50 features) and the Genetic Algorithm refinement (50 to 24 features) do not meet that standard: both run once on the full $n = 59$ cohort before the outer split, using every patient’s label to decide which 24 features every classifier, hyperparameter-optimization trial, ensemble, calibration analysis, and explainability result in Section 6 would subsequently see. This is a textbook instance of the feature-selection leakage documented by Ambroise and McLachlan [64] and by Varma and Simon [65]: cross-validation is unbiased only when every data-dependent modeling decision, including which features to use, is repeated inside each fold, and a fixed feature subset chosen from the full sample violates that requirement even when the classifiers themselves are trained and evaluated correctly. The direction of the resulting bias is optimistic; its magnitude is difficult to quantify without re-running feature selection inside every outer fold, which was not done here given the computational cost of the Genetic Algorithm search (approximately 450 seconds per run at the settings used) multiplied across five outer folds and, for a fully nested design, three inner hyperparameter-search folds as well. Readers should treat the absolute accuracy and F1-score levels reported in Section 6.1 as an optimistic estimate of the pipeline’s generalization performance rather than as an unbiased one, while the relative ranking among the sixteen classifiers, all of which received the identical fixed feature subset, is far less sensitive to this issue, since the selection bias applies uniformly across every model compared. A fully feature-selection-nested version of this pipeline, in which the Mutual Information filter and Genetic Algorithm search are re-executed inside each outer fold, is a direct and well-specified target for future work and would be required before any of the point estimates reported here could be treated as unbiased.

The dataset was collected at a single clinical site under one diagnostic protocol. Site-specific patient demographics, instrument administration procedures, and clinical assessment norms may limit how well the learned decision boundaries transfer elsewhere. External validation on independent cohorts from multiple centers is an important prerequisite before any clinical deployment.

The three-class label space, appropriate given the available data, covers only part of the full comorbidity spectrum relevant to fibromyalgia. Extending it to psychiatric comorbidities, metabolic conditions, and additional rheumatological diagnoses would require substantially larger, multi-site datasets.

SHAP and LIME provide theoretically grounded, model-consistent attribution frameworks, but Shapley values explain model behavior rather than making causal claims about disease etiology. The clinically coherent secondary feature profiles reported here should be read as evidence of construct validity, not as evidence of causal mechanisms. Prospective validation in appropriately powered observational studies is needed before drawing any causal conclusions.

8 Conclusion and Future Work

This paper developed, benchmarked, and validated a machine learning framework for multiclass classification of chronic comorbidities (No Comorbidity, Disc Herniation, and Inflammatory Bowel Syndrome) using routine multi-instrument clinical assessment data from the *Clinical_fm_66* dataset. Sixteen supervised classifiers spanning six algorithmic families were evaluated under a five-fold stratified nested cross-validation framework, with global statistical validation, calibration uncertainty quantification, and dual-method explainability through SHAP and LIME. The key empirical finding is that the Weighted Soft Voting meta-ensemble reaches the highest discriminative performance (accuracy = 0.864, F1 = 0.826, nested-CV mean macro-AUC = 0.866), ahead of every individual classifier. Extra Trees (Tuned) reaches the highest single-classifier accuracy (0.848) and is the strongest deployable model where ensemble overhead is not acceptable. Gradient Boosting has the best calibration (ECE = 0.187 ± 0.016 , Brier score = 0.103), the lowest inference latency among the nonlinear ensemble methods (0.202 ms/sample), and the most stable fold-to-fold behavior, which makes it the recommended model for threshold-based clinical decision support. The preprocessing protocol, which enforces strict within-fold isolation of every imputation, standardization, and resampling operation at the classifier-training stage, is a methodological contribution with broad applicability to clinical machine learning research generally. BorderlineSMOTE, applied with an adaptive neighborhood cap and a standard-SMOTE fallback, was the single data-augmentation strategy behind every reported number in this study. The Mutual Information and Genetic Algorithm feature-selection stages, by contrast, were computed once on the full analytical cohort rather than inside each outer fold; Section 7.4 names this plainly as a source of optimistic bias in the absolute performance figures and identifies fully feature-selection-nested cross-validation as the direct next methodological step. The Friedman omnibus test ($\chi^2 = 42.25$, $p < 0.001$) confirms that the observed performance hierarchy reflects statistically significant and practically meaningful differences in model capacity, a comparison that remains informative under the same feature-selection exposure for every classifier evaluated. SHAP analysis identifies SSS item 13 as the dominant global predictor, with Disc Herniation secondary signals centered on SSS item 1 and Cardiovascular symptoms, and IBS secondary signals reflecting a multi-system somatic symptom burden consistent with gut-brain axis dysregulation. LIME analysis independently confirms these class-specific signatures at the level of individual predictions. The agreement between the global and local explainability methods strengthens confidence in the clinical plausibility of the learned decision boundaries. Future work should pursue external validation on larger, independent cohorts from multiple clinical sites; a fully feature-selection-nested version of this pipeline, in which Mutual Information and Genetic Algorithm selection are re-executed inside each outer fold to remove the optimistic bias described in Section 7.4; extension of the label space to psychiatric and metabolic comorbidities; integration of longitudinal electronic health record trajectories for temporal modeling of comorbidity onset and progression; and a prospectively evaluated clinical decision-support interface for rheumatological and primary care use. The pipeline, dataset, and evaluation framework presented here, including the explicit accounting of where its own leakage-free guarantee does and does not extend, offer a reproducible and methodologically transparent starting point for these extensions.

Declarations

Ethics approval and consent to participate: This study did not involve human participants or patient data. Accordingly, no ethical approval or consent to participate was required.

Funding: No funding was received for this manuscript.

Consent for publication: Not applicable. This study did not involve patients.

Data Availability Statement: All data generated or analysed during this study are included in this published article. The dataset is publicly available at <https://zenodo.org/records/6611947>.

Competing interests: The authors declare that they have no competing interests.

Acknowledgment

The authors sincerely thank their co-authors for their invaluable support throughout the preparation of this manuscript. They are likewise grateful to the anonymous reviewers for their constructive and perceptive comments, which have substantially improved the quality of this work.

References

- [1] C. Herteliu, “Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019,” *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.
- [2] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, “Defining comorbidity: implications for understanding health and health services,” *The Annals of Family Medicine*, vol. 7, no. 4, pp. 357–363, 2009.
- [3] K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, “Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study,” *The Lancet*, vol. 380, no. 9836, pp. 37–43, 2012.
- [4] L. R. Vialle, E. N. Vialle, J. E. S. Henao, and G. Giraldo, “Lumbar disc herniation,” *Revista Brasileira de Ortopedia (English Edition)*, vol. 45, no. 1, pp. 17–22, 2010.
- [5] M. Alsmirat, N. Al-Mnayyis, M. Al-Ayyoub, and A. Al-Mnayyis, “Deep learning-based disk herniation computer aided diagnosis system from mri axial scans,” *IEEE Access*, vol. 10, pp. 32315–32323, 2022.
- [6] D. Chicco and G. Jurman, “Arterial disease computational prediction and health record feature ranking among patients diagnosed with inflammatory bowel disease,” *IEEE Access*, vol. 9, pp. 78648–78657, 2021.
- [7] S. Alatab, S. G. Sepanlou, K. Ikuta, H. Vahedi, C. Bisignano, S. Safiri, A. Sadeghi, M. R. Nixon, A. Abdoli, H. Abolhassani, *et al.*, “The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017,” *The Lancet gastroenterology & hepatology*, vol. 5, no. 1, pp. 17–30, 2020.

- [8] N. A. Molodecky, S. Soon, D. M. Rabi, W. A. Ghali, M. Ferris, G. Chernoff, E. I. Benchimol, R. Panaccione, S. Ghosh, H. W. Barkema, *et al.*, “Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review,” *Gastroenterology*, vol. 142, no. 1, pp. 46–54, 2012.
- [9] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, “Artificial intelligence in healthcare: past, present and future,” *Stroke and vascular neurology*, vol. 2, no. 4, 2017.
- [10] A. L. Beam and I. S. Kohane, “Big data and machine learning in health care,” *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018.
- [11] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, “Leakage in data mining: Formulation, detection, and avoidance,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 4, pp. 1–21, 2012.
- [12] “Clinical_fm_66 dataset.” <https://zenodo.org/records/6611947>, 2022. Online.
- [13] S. Ghosh, M. R. Malgireddy, V. Chaudhary, and G. Dhillon, “A new approach to automatic disc localization in clinical lumbar mri: Combining machine learning with heuristics,” in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 114–117, IEEE, 2012.
- [14] E. Ebrahimzadeh, F. Fayaz, F. Ahmadi, and M. Nikravan, “A machine learning-based method in order to diagnose lumbar disc herniation disease by mr image processing,” *MedLife Open Access*, vol. 1, no. 1, p. 1, 2018.
- [15] M. S. Islam, M. Asaduzzaman, and M. M. Rahman, “Feature selection and classification of spinal abnormalities to detect low back pain disorder using machine learning approaches,” in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1–4, IEEE, 2019.
- [16] Z. U. Hussain, R. Comerford, F. Comerford, N. Ng, D. Ng, A. Khan, C. Lees, and A. Hussain, “A comparison of machine learning approaches for predicting the progression of crohn’s disease,” in *2020 IEEE Student Conference on Research and Development (SCOREd)*, pp. 529–533, IEEE, 2020.
- [17] E.-H. A. Rady and A. S. Anwar, “Prediction of kidney disease stages using data mining algorithms,” *Informatics in medicine unlocked*, vol. 15, p. 100178, 2019.
- [18] S. Ganiger and K. Rajashekharaiyah, “Chronic diseases diagnosis using machine learning,” in *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, pp. 1–6, IEEE, 2018.
- [19] W. Mbarki, M. Bouchouicha, S. Frizzi, F. Tshibas, L. B. Farhat, and M. Sayadi, “Lumbar spine discs classification based on deep convolutional neural networks using axial view mri,” *Interdisciplinary Neurosurgery*, vol. 22, p. 100837, 2020.
- [20] M. Pal, S. Parija, G. Panda, S. Mishra, R. K. Mohapatra, and K. Dhama, “Covid-19 prognosis from chest x-ray images by using deep learning approaches: A next generation diagnostic tool,” *Journal of Pure & Applied Microbiology*, vol. 17, no. 2, 2023.
- [21] P. D. Barua, M. Kobayashi, M. Tanabe, M. Baygin, J. K. Paul, T. Iype, S. Dogan, T. Tuncer, R.-S. Tan, and U. R. Acharya, “Innovative fibromyalgia detection approach based on quantum-inspired 3lbp feature extractor using ecg signal,” *IEEE Access*, vol. 11, pp. 101359–101372, 2023.

- [22] B. Ramesh and K. P. Rao, "Intelligent detection of chronic kidney disease using optimized mlp models and feature selection techniques on the ap-ckd dataset.," *IAENG International Journal of Computer Science*, vol. 52, no. 10, 2025.
- [23] K. Kangra and J. Singh, "A novel hybrid approach to predict diabetes using boruta and genetic algorithm.," *IAENG International Journal of Computer Science*, vol. 52, no. 10, 2025.
- [24] A. M. Elshewey, A. H. Abed, D. S. Khafaga, A. A. Alhussan, M. M. Eid, and E.-S. M. El-Kenawy, "Enhancing heart disease classification based on greylag goose optimization algorithm and long short-term memory," *Scientific Reports*, vol. 15, no. 1, p. 1277, 2025.
- [25] Z. Tarek, A. A. Alhussan, D. S. Khafaga, E.-S. M. El-Kenawy, and A. M. Elshewey, "A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages," *Biomedical Signal Processing and Control*, vol. 102, p. 107417, 2025.
- [26] S. Raj and B. Mahanand, "Efficient classification of diffuse lung disease in class imbalance data," *IAENG International Journal of Computer Science*, vol. 52, no. 7, p. 2121, 2025.
- [27] A. Amalia, M. S. Lydia, S. M. Hardi, A. B. Jamesie, and H. Fahmi, "Optimizing celiac disease detection through dataset balancing and deep learning in a mobile application.," *IAENG International Journal of Computer Science*, vol. 52, no. 9, 2025.
- [28] A. Atwa, A. Y. Ismaeel, and A. A. Elngar, "Machine learning for chronic disease classification and comorbidity detection: Methodological gaps and future directions," *Journal of Smart Algorithms and Applications (JSAA)*, vol. 3, no. 2, pp. 87–104, 2026.
- [29] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.
- [30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [31] G. Lemaitre, F. Nogueira, C. K. Aridas, *et al.*, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.
- [32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
- [33] F.-A. Fortin, F.-M. De Rainville, M.-A. G. Gardner, M. Parizeau, and C. Gagné, "Deap: Evolutionary algorithms made easy," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2171–2175, 2012.
- [34] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.
- [35] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, vol. 24, 2011.
- [36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

- [37] M. T. Ribeiro, S. Singh, and C. Guestrin, ““ why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.
- [38] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, “Scipy 1.0: fundamental algorithms for scientific computing in python,” *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.
- [39] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for dna microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.
- [40] S. Van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of statistical software*, vol. 45, pp. 1–67, 2011.
- [41] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.
- [42] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.
- [43] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014.
- [44] O. Claveria, E. Monte, and S. Torra, “Evolutionary computation for macroeconomic forecasting,” *Computational Economics*, vol. 53, no. 2, pp. 833–849, 2019.
- [45] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-smote: a new over-sampling method in imbalanced data sets learning,” in *International conference on intelligent computing*, pp. 878–887, Springer, 2005.
- [46] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [47] R. Blagus and L. Lusa, “Smote for high-dimensional class-imbalanced data,” *BMC bioinformatics*, vol. 14, no. 1, p. 106, 2013.
- [48] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [49] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [50] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.
- [51] J. Platt *et al.*, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
- [52] R. Kohavi *et al.*, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.
- [53] S. Biometrika, “The probable error of a mean,” *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.
- [54] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.
- [55] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

- [56] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The annals of mathematical statistics*, vol. 11, no. 1, pp. 86–92, 1940.
- [57] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
- [58] O. J. Dunn, "Multiple comparisons among means," *Journal of the American statistical association*, vol. 56, no. 293, pp. 52–64, 1961.
- [59] J. Cohen, *Statistical power analysis for the behavioral sciences*, Rev. Lawrence Erlbaum Associates, Inc, 1977.
- [60] L. S. Shapley *et al.*, "A value for n-person games," 1953.
- [61] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.
- [62] W. B. Glenn *et al.*, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.
- [63] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017.
- [64] C. Ambrose and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the national academy of sciences*, vol. 99, no. 10, pp. 6562–6566, 2002.
- [65] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, no. 1, p. 91, 2006.