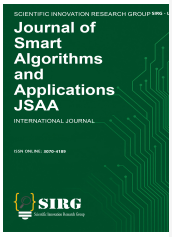

Journal of Smart Algorithms and Applications
JSAA
3070-4189/
Journal Homepage
https://pub.scientificirg.com/index.php/JSAA

Machine Learning for Chronic Disease Classification and Comorbidity Detection: Methodological Gaps and Future Directions

Asmaa A. Attwa[a,1] , **Ahmed Yehia Ismaeel**[b], **Ahmed A. Elngar**[c]

[a] *Faculty of Applied Health Sciences Technology, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: asmaa.abdelhamed@fahst.bsu.edu.eg*

[b] *Faculty of Medicine, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: ahmed_yehia@med.bsu.edu.eg*

[c] *Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: elngar_7@yahoo.co.uk*

ABSTRACT

The rapid proliferation of machine learning (ML) methods across clinical medicine has generated a rich but fragmented body of evidence for chronic disease classification. Despite consistently high reported accuracy, the literature is characterised by five systematic methodological limitations: exclusive reliance on binary single-disease classification tasks, absence of leakage-free preprocessing protocols, lack of non-parametric statistical validation, omission of probability calibration evaluation, and minimal integration of explainability frameworks. This narrative review critically examines 25 representative ML studies spanning musculoskeletal disorders (particularly disc herniation), inflammatory bowel conditions, fibromyalgia, cardiovascular disease, and related chronic comorbidities, published between 2012 and 2025. Studies are analysed thematically across algorithmic approach, task scope, class imbalance strategy, and methodological rigour. Algorithmic families represented include classical support vector machines and tree ensembles, deep learning architectures (CNN, LSTM, U-Net), optimization-enhanced methods (WOA, GGO, PSO, SO), and natural language processing models (RoBERTa). Across all 25 studies, performance metrics range from 82.47% to 99.9% accuracy, yet none simultaneously addresses multiclass comorbidity discrimination, leakage-free preprocessing, and model explainability. The review identifies five critical gaps and maps them to concrete future research directions, with particular emphasis on the unmet need for a unified multiclass framework capable of differential diagnosis among clinically overlapping chronic conditions within fibromyalgia populations. These findings suggest that current ML models are not yet clinically ready for the differential diagnosis of comorbid chronic conditions without methodological reform.

PAPER INFORMATION

HISTORY

Received: 9 January 2026
Revised: 27 March 2026
Accepted: 20 April 2026
Online: 24 April 2026

MSC

68T07; 68R10; 94A60; 68M15

KEYWORDS

Machine Learning; Chronic Disease Classification; Disc Herniation; Inflammatory Bowel Syndrome; Explainable AI.

1 Introduction

Chronic non-communicable diseases (NCDs) such as musculoskeletal disorders, gastrointestinal conditions, and psychosomatic syndromes have become the leading contributors to global morbidity and healthcare expenditure [1, 2]. The complexity of these diseases (marked by overlapping symptoms, heterogeneous data types, class imbalance, and frequent comorbidities) poses significant challenges for accurate diagnosis and personalized care [3, 2]. As the prevalence of chronic diseases continues to rise, so too does the urgency for innovative, data-driven solutions in clinical practice [1].

Machine learning is defined as a subset of artificial intelligence that enables computational systems to learn patterns from data and improve predictive performance without being explicitly programmed for each task [4, 5]. **Comorbidity** refers

[1]Faculty of Applied Health Sciences Technology, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: asmaa.abdelhamed@fahst.bsu.edu.eg

to the co-occurrence of two or more chronic conditions in the same individual, which may interact to amplify symptom burden, complicate differential diagnosis, and reduce the effectiveness of standard therapeutic interventions [3, 2]. **Health informatics** is the interdisciplinary field concerned with the acquisition, storage, retrieval, and use of healthcare information to support clinical decision-making, management, and research [6].

Machine learning (ML) has emerged as a transformative paradigm in healthcare, enabling the automated extraction of patterns from high-dimensional clinical data and supporting evidence-based decision-making [7, 4]. Recent years have witnessed an exponential growth in the adoption of ML methods for disease classification, prognosis, and risk stratification across diverse medical domains [8, 9]. Despite notable advances and impressive reported accuracies, the translation of ML models to routine clinical workflows remains hindered by several persistent methodological limitations. These include the predominance of binary, single-disease classification tasks [10, 11]; insufficient control for data leakage during preprocessing [12, 13]; a lack of rigorous statistical validation [14]; minimal attention to probability calibration [15]; and limited integration of explainable artificial intelligence (XAI) frameworks [16, 17].

Within this context, the simultaneous classification of multiple co-occurring chronic conditions (known as comorbidity detection) remains an underexplored yet clinically vital problem [3, 2]. Multiclass comorbidity classification is particularly relevant in syndromes such as fibromyalgia, where differential diagnosis among overlapping chronic conditions is essential for effective management [18, 19]. However, existing literature seldom addresses this challenge in a unified, methodologically robust manner [10, 11, 20].

This narrative review critically examines 25 representative ML studies published between 2012 and 2025, spanning disc herniation, inflammatory bowel syndrome, fibromyalgia, cardiovascular, and metabolic diseases. A narrative review synthesizes literature thematically without the formal protocol of a systematic review, allowing broader interpretive scope across heterogeneous study designs [21]. Unlike prior single-domain surveys that focus exclusively on spinal disorders [10], gastrointestinal conditions [20], or cardiovascular disease [11] in isolation, the present review provides the **first unified cross-domain synthesis** of 25 ML studies across five clinical domains under a uniform six-dimension methodological quality framework. This novelty claim is specifically scoped to the combination of: (1) cross-domain coverage spanning musculoskeletal, gastrointestinal, fibromyalgia, cardiovascular, and metabolic domains simultaneously; (2) a uniform methodological quality audit applied across all included studies; and (3) a primary clinical focus on the unmet need for multiclass comorbidity classification within fibromyalgia populations, a problem not addressed by any existing review.

It is also important to distinguish between two related but distinct classification formulations addressed in this domain. **Multiclass classification** assigns each patient to exactly one class from a set of $K > 2$ mutually exclusive diagnostic categories (e.g., No Comorbidity / Disc Herniation / IBS), which is appropriate when a single dominant comorbidity label is clinically actionable. **Multi-label classification** permits simultaneous assignment to multiple classes and would be appropriate when a patient may present with both DH and IBS concurrently. The present review focuses on multiclass differential diagnosis consistent with the clinical structure of fibromyalgia assessment datasets, while acknowledging that multi-label formulations represent a promising future direction (**Section 11**).

The review analyzes trends in algorithmic approaches, task scope, class imbalance strategies, and methodological rigor. It identifies and synthesizes five cross-cutting methodological gaps that limit the clinical translation and generalizability of current ML models. By highlighting these gaps and proposing concrete research directions, this review aims to foster the development of next-generation ML frameworks for chronic disease classification and comorbidity detection.

Review Objectives

This review is guided by the following research objectives:

RO1 – Map the algorithmic landscape: Characterize the distribution of ML algorithmic families applied to chronic disease classification across musculoskeletal, gastrointestinal, fibromyalgia, cardiovascular, and metabolic domains (2012–2025).

RO2 – Evaluate task scope: Assess the prevalence and limitations of binary vs. multiclass classification formulations and identify the gap in comorbidity-aware multiclass frameworks.

RO3 – Audit methodological rigour: Evaluate the degree to which reviewed studies implement leakage-free preprocessing, class imbalance correction, statistical validation, and probability calibration.

RO4 – Assess explainability integration: Determine the extent to which XAI methods, particularly SHAP-based feature attribution, are employed and reported with per-class clinical interpretation.

RO5 – Identify gaps and future directions: Synthesize cross-cutting methodological deficiencies and map them to concrete, prioritized future research directions.

Principal Contributions

This review makes the following original contributions to the literature:

C1 – Comprehensive thematic synthesis: A structured thematic synthesis of 25 ML studies across five clinical domains, providing the first unified cross-domain analysis covering disc herniation, IBS, fibromyalgia, cardiovascular, and metabolic disease classification.

C2 – Methodological quality audit: A six-dimension methodological quality assessment covering accuracy reporting, AUC, multiclass scope, leakage-free preprocessing, statistical validation, and XAI, applied uniformly across 25 studies and visualized as a quality heatmap.

C3 – Algorithmic taxonomy: A hierarchical taxonomy of ML approaches organized by clinical domain and algorithmic family, enabling rapid identification of under-explored algorithm–domain combinations.

C4 – Five-gap framework: Formal identification and characterization of five cross-cutting methodological gaps as a structured agenda for future research.

C5 – Evidence-based future roadmap: A prioritized future research roadmap mapping each identified gap to a concrete methodological direction, grounded in the cross-study evidence base.

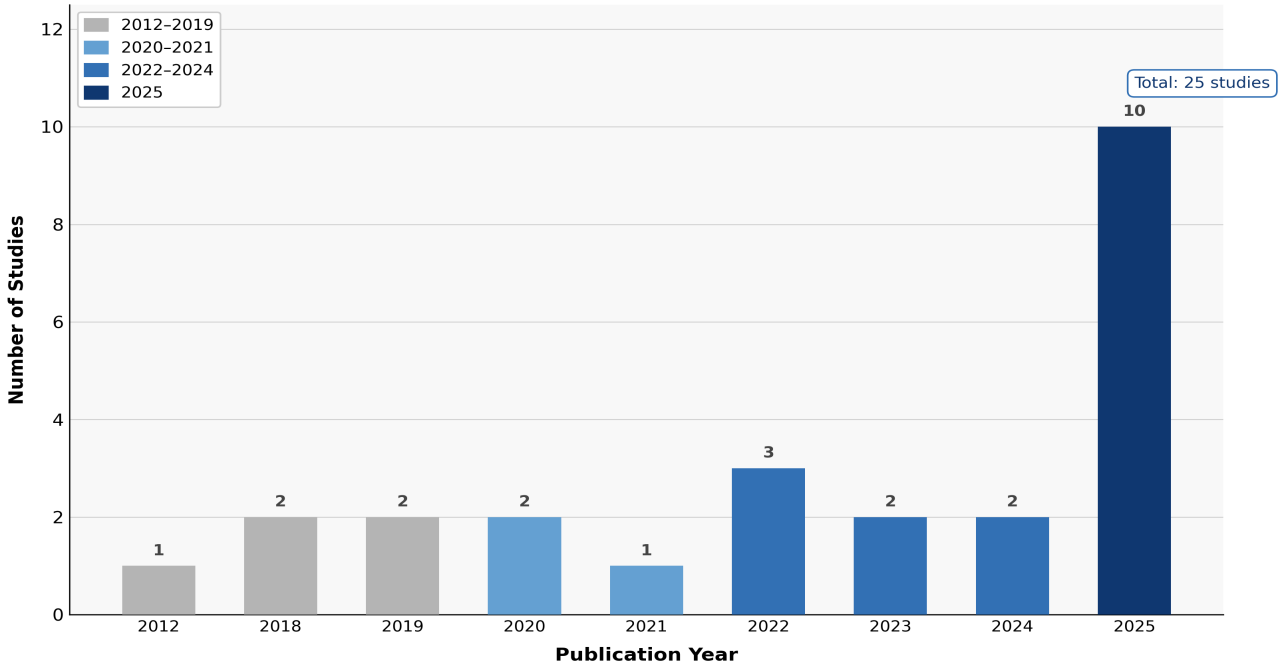

Figure 1: Publication trend of reviewed ML studies in clinical medicine (2012–2025).

The remainder of this paper is organized as follows. Section 2 describes the literature scope, inclusion criteria, and the 25 reviewed studies. Section 3 provides an overview of the nine machine learning algorithm families and compares them across interpretability, calibration, sample efficiency, and multiclass suitability. Section 4 surveys ML applications in musculoskeletal, gastrointestinal and fibromyalgia, systemic and metabolic, and signal-based domains. Section 8 presents the cross-cutting methodological analysis and formal identification of five systematic gaps. Section 9 proposes a gold-standard pipeline to address these gaps simultaneously. Section 10 outlines future research directions, and Section 11 concludes the review.

2 Literature Scope and Review Approach

This review covers peer-reviewed ML studies addressing chronic disease classification with relevance to disc herniation, IBS, fibromyalgia, and related comorbidities, identified through PubMed, Scopus, and IEEE Xplore searches supplemented by citation tracking. The 25 included studies (**Table 2**) represent the breadth of algorithmic families, clinical domains, and methodological approaches present in the literature (2012–2025). Study selection followed the inclusion and exclusion criteria summarized in **Table 1**. **Figure 2** illustrates the distribution of algorithmic families across the reviewed studies.

A **narrative review** is a form of knowledge synthesis that critically examines, summarizes, and interprets published literature on a topic without the formal database search protocols and quantitative pooling of systematic reviews [21]. **Class imbalance** refers to datasets in which the number of instances belonging to different classes differs substantially, leading classifiers to favor the majority class and producing misleading accuracy estimates on the minority class [13, 22]. A **leakage-free preprocessing protocol** is one in which all data transformation steps, including missing value imputation, feature standardization, feature selection, and class resampling, are fitted exclusively on the training portion of each cross-validation fold and applied (without refitting) to the corresponding held-out validation portion. This prevents any information about validation samples from influencing the preprocessing steps, ensuring that reported performance estimates reflect genuine generalization to unseen clinical data rather than artificially inflated estimates [12, 23].

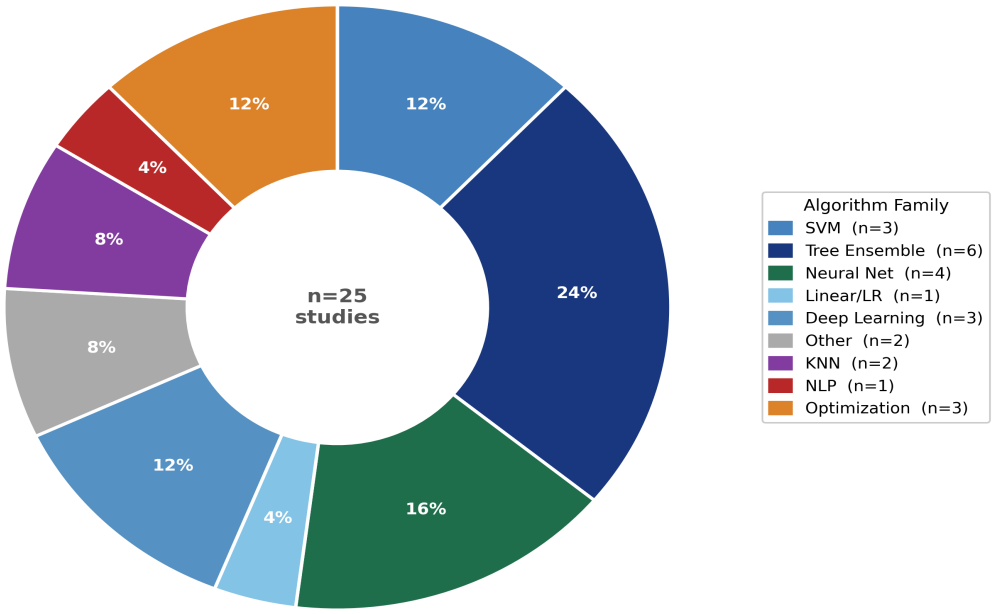

Figure 2: Algorithmic family distribution across the 25 reviewed studies. Tree ensembles and deep learning collectively represent more than half of approaches; optimization-enhanced methods are predominantly a 2024–2025 phenomenon.

Table 1: Inclusion and Exclusion Criteria for Study Selection.

Inclusion Criteria	Exclusion Criteria
Peer-reviewed ML/DL study	Purely theoretical studies without empirical data
Chronic disease classification or prediction task	Studies addressing acute conditions only
Quantitative performance metric reported (accuracy, AUC, F1, or equivalent)	Non-English publications
English language publication	Conference abstracts without full methodology

3 Overview of ML Algorithm Families

The 25 reviewed studies collectively employ nine distinct ML algorithm families. This section provides a critical overview of each family's suitability for chronic disease classification, emphasizing calibration, sample efficiency, interpretability, and multiclass capability, the dimensions most relevant to the five identified gaps.

Supervised learning trains a model on labeled pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ [5]. **Ensemble methods** combine multiple base learners for lower variance and improved robustness [45]. **Explainable AI (XAI)** makes model outputs interpretable to clinicians [17, 16].

Table 2: Summary of the 25 Reviewed ML Studies for Chronic Disease Classification (2012–2025).

Reference	Author	Year	Dataset	Task	Algorithm	Performance	Limitation	Gap Profile
[24]	Ghosh et al.	2012	318 lumbar discs	Disc localization	HOG + SVM	99%	Binary; single modality	G1,G2,G3,G4,G5
[11]	Ganiger et al.	2018	Chronic disease data	Disease prediction	RF, DT, SVM	RF best	Limited feature engineering	G1,G2,G3,G4,G5
[25]	Rahman et al.	2019	310 patients	Low back pain classif.	RF	94%	Small dataset; binary	G1,G2,G3,G4,G5
[26]	Nikravan et al.	2018	Lumbar MR images	Herniated disc detection	SVM, MLP, KNN	95.23%	Binary; imaging only	G1,G2,G3,G4,G5
[27]	Rady et al.	2019	Kidney disease data	CKD stage prediction	PNN, MLP, SVM, RBF	PNN best	Limited features	G1,G2,G3,G4,G5
[20]	Hussain et al.	2020	804 CD patients	Flare-up prediction	Logistic Regression	90%	Single disease	G1,G2,G3,G4,G5
[28]	Mbarki et al.	2020	Lumbar MRI	Herniation classification	U-Net, VGG16	94%	High complexity; binary	G1,G2,G3,G4,G5
[29]	Kukker et al.	2021	EEG data	Seizure classification	Genetic Fuzzy Q-Lrn.	96.79%	Epilepsy only	G1,G2,G3,G4,G5
[10]	Alsmirat et al.	2022	MRI scans	Disc herniation detection	CNN models	95.56%	Binary; imaging only	G1,G2,G3,G4,G5
[30]	Pal et al.	2022	COVID-19 symptom data	Prognosis	k-NN	97.97%	Single disease	G1,G2,G3,G4,G5
[31]	Pal et al.	2022	Cardiovascular data	CVD prediction	MLP	82.47%	Binary only	G1,G2,G3,G4,G5
[18]	Acharya et al.	2023	139 ECG signals	Fibromyalgia diagnosis	k-NN, SVM	93.87%	Small dataset; binary	G1,G2,G3,G4,G5
[32]	Sarker et al.	2023	Social media posts	Chronic pain detection	RoBERTa	F1=0.84	Text-based only	G1,G2,G3,G4,G5
[33]	Elshewey et al.	2025	Heart disease data	Classification	GGO + LSTM	99.58%	Binary only	G1,G2,G3,G4,G5
[34]	Tarek et al.	2025	CVD data	Early detection	SO + ML	99.9%	CVD only	G1,G2,G3,G4,G5
[35]	Elshewey et al.	2024	EEG dataset	Eye state classif.	MBER + KNN	96.12%	Specialized; binary	G1,G2,G3,G4,G5
[36]	El-Rashidy et al.	2025	MIMIC-III data	Ventilation/mortality	PSO + MTL	94–97%	ICU-specific	G1,G2,G3,G4,G5
[37]	Elshewey et al.	2024	310 instances	Orthopedic classif.	BFS-RF	99.41%	Small dataset; binary	G1,G2,G3,G4,G5
[38]	Ramesh et al.	2025	CKD clinical data	Early CKD detection	Optimized MLP + FS	High acc.	Binary; no multiclass	G1,G2,G3,G4,G5
[39]	Kangra et al.	2025	Diabetes clinical data	Diabetes prediction	Hybrid ML + Boruta	High AUC	Single disease	G1,G2,G3,G4,G5
[40]	Terlapu et al.	2025	Liver disease data	Liver disease classif.	MLP + WOA	Improved	Algorithm complexity	G1,G2,G3,G4,G5
[41]	Kishan et al.	2025	QoL/healthcare data	Quality of life assessment	Neural + clustering	Good	Indirect prediction	G1,G2,G3,G4,G5
[42]	Vodnala et al.	2025	Cough sound dataset	COPD/asthma classif.	FS + ML	Effective	Specialized; binary	G1,G2,G3,G4,G5
[43]	Raj et al.	2025	Lung disease data	Diffuse lung classif.	ML + SMOTE	Better	Class imbalance focus	G1,G2,G3,G4,G5
[44]	Amalia et al.	2025	Celiac dataset	Celiac disease detection	DL + balancing	Improved	Specific disease; binary	G1,G2,G3,G4,G5

G1: binary-only scope G2: preprocessing leakage G3: no statistical validation G4: no calibration G5: no XAI *partial XAI

3.1 Tree-Based Methods

Decision trees partition the feature space by minimizing the Gini impurity [46]:

$$G(t) = 1 - \sum_{k=0}^{K-1} p_{t,k}^2 \tag{1}$$

Single trees are highly interpretable but overfit on small high-dimensional datasets. **Random Forest** [47] mitigates this by aggregating T decorrelated trees via majority vote, the most frequently used algorithm in the reviewed studies (n=7). **Extra Trees** [48] further randomizes split thresholds, reducing variance at the cost of a marginal bias increase. Both ensemble variants are well-suited to the small, high-dimensional clinical cohorts prevalent in this domain and support native multiclass prediction without modification. Their primary limitation is the requirement for Platt calibration to produce reliable posterior probabilities for threshold-based clinical deployment.

3.2 Boosting Methods

Gradient Boosting [49], XGBoost [50], and AdaBoost [51] iteratively minimize a loss function by adding weak learners. These methods consistently achieve top accuracy in the reviewed literature and support native multiclass classification. XGBoost augments gradient boosting with explicit ℓ_2 regularization, improving generalization on tabular clinical data. The critical limitation of all boosting methods in the comorbidity context is their opacity: without SHAP attribution [16], the clinical basis of predictions cannot be audited, directly contributing to Gap 5 (absence of XAI) identified in **Section 8**.

3.3 Support Vector Machines

SVMs [52] maximize the decision margin in a kernel-transformed feature space. The RBF kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$ handles non-linear boundaries effectively. However, SVMs do not produce native probability outputs and require post-hoc Platt calibration [15], introducing an additional modeling step that is absent from all 25 reviewed studies, contributing to Gap 4 (no calibration analysis).

3.4 Linear and Probabilistic Classifiers

Logistic Regression [53] applies the softmax function and produces natively calibrated probabilities at minimal inference latency, making it the optimal candidate for real-time clinical deployment. **LDA** [54] assumes Gaussian class-conditional distributions with shared covariance, which may be violated in heterogeneous psychometric datasets. **Naïve Bayes** [55] assumes conditional feature independence, a strong assumption that limits its discriminative power in correlated clinical feature spaces.

3.5 Instance-Based and Neural Methods

KNN [56] classifies via distance-weighted majority vote and is highly sensitive to high-dimensional synthetic feature spaces produced by SMOTE oversampling, a practical concern for small clinical cohorts with class imbalance. **MLP** [57] with ReLU activations and cross-entropy loss achieves competitive accuracy but requires large training sets; on cohorts of $n < 200$, MLP consistently underperforms ensemble methods due to insufficient gradient signal for effective weight optimization.

3.6 Optimization-Enhanced Methods

A notable 2024–2025 trend couples metaheuristic optimizers (WOA, GGO [33], PSO [36]) with ML classifiers for feature selection and hyperparameter tuning. These hybrid approaches frame training as:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} \mathcal{L}(\mathbf{w}) + \lambda\, \Omega(\mathbf{w}) \tag{2}$$

where $\Omega(\mathbf{w})$ is a regularization term solved via population-based search rather than gradient descent. These methods consistently report accuracy gains in binary classification settings, yet their behavior in multiclass, class-imbalanced comorbidity contexts remains uncharacterized. Critically, explainability is absent from all hybrid approaches in the reviewed literature. Three practical strategies can integrate SHAP transparency into these pipelines: (1) *post-hoc SHAP* on the converged model using TreeExplainer or KernelSHAP; (2) *SHAP-guided fitness functions* that penalise instability in feature attributions across optimization iterations; and (3) *optimization trajectory logging* to provide operational transparency to clinical reviewers. None of the reviewed optimization-enhanced studies implemented any of these strategies, highlighting a concrete avenue for future methodological development.

3.7 Comparative Summary

Table 3 compares all algorithm families across seven dimensions, including a new **Multiclass Suitability** column that is directly relevant to Gap 1 (binary-only task scope).

Tree-based ensembles offer the most favorable accuracy–interpretability–sample-efficiency trade-off for small clinical cohorts. Logistic Regression remains the optimal accuracy-efficiency candidate for real-time deployment. Optimization-enhanced methods demonstrate promising accuracy gains but score lowest on multiclass suitability, underscoring an important future research direction.

Table 3: Comparative overview of ML algorithm families. H=High, M=Medium, L=Low; *n*=reviewed studies.

Family	Interpret.	Prob. Calib.	Small *n*	Imbalance	Cost	Multiclass	*n*
Decision Tree	H	M	M	H	L	H	4
Random Forest	M	M	H	M	M	H	7
Extra Trees	M	M	H	M	M	H	3
Grad. Boosting	L	M	H	M	M	H	4
XGBoost	L	M	H	M	M	H	3
AdaBoost	L	M	M	H	M	M	2
SVM (RBF)	L	L[†]	H	L	M	M	8
Logistic Reg.	H	H	H	L	L	H	5
LDA	H	H	H	M	L	H	2
Naïve Bayes	H	M	H	H	L	H	2
KNN	M	M	M	M	L	H	4
MLP	L	M	L	M	H	H	5
Optim.+ML	L	M	M	M	H	L	6

[†] Requires Platt calibration. Multiclass: native support without one-vs-rest wrapper.

4 ML for Musculoskeletal and Spinal Disease Classification

Early ML approaches to spinal disorder classification span both imaging-based and clinical feature-based paradigms, as reviewed in the subsections below.

4.1 Imaging-Based Disc Herniation Classification

Ghosh et al. [24] (2012) achieved 99% accuracy for lumbar disc localization using HOG combined with SVM on 318 lumbar discs from 53 clinical cases, an early landmark demonstrating ML feasibility for spinal image analysis. The approach is limited to binary imaging-based classification and single modality input, precluding application to multi-instrument comorbidity contexts.

Nikravan et al. [26] (2018) benchmarked SVM, MLP, and KNN for herniated disc detection on lumbar MR images, achieving 91.9–95.23% accuracy. Mbarki et al. [28] (2020) introduced U-Net segmentation followed by VGG16 classification, reporting 94% accuracy but with high model complexity. Alsmirat et al. [10] (2022) applied CNN models to MRI scans for binary disc herniation detection at 95.56%. All three share the same limitation: binary-only task scope and exclusive reliance on imaging modalities.

4.2 Clinical Feature-Based Spinal Disorder Classification

Structured clinical data offers a complementary pathway to imaging-based classification, as demonstrated by two representative studies. Rahman et al. [25] (2019) applied Random Forest to 310 patients described by 12 clinical features for low back pain classification, achieving 94% accuracy. While demonstrating the viability of structured clinical data for spinal disorder classification, the small feature set and binary task structure limit generalizability. Elshewey et al. [37] (2024) achieved 99.41% accuracy for orthopedic disease classification on 310 instances using BFS-RF, demonstrating strong ensemble performance on structured clinical data, though again in a binary context with a small, single-site cohort.

Across all six musculoskeletal studies, Gaps G1–G5 apply uniformly: all adopt binary-only task scope (Gap 1), none implements explicit within-fold preprocessing (Gap 2), none reports statistical validation or confidence intervals (Gap 3), none evaluates calibration metrics (Gap 4), and none provides feature attribution (Gap 5). The consistent accuracy range of 94–99.41% confirms that both imaging and clinical feature approaches are viable, yet neither has been extended to multiclass comorbidity discrimination.

5 ML for Chronic Non-Musculoskeletal Disease

Beyond spinal disorders, ML has been applied across gastrointestinal, cardiovascular, metabolic, and respiratory chronic disease domains, as surveyed in the following subsections.

5.1 Gastrointestinal and Inflammatory Disease

Hussain et al. [20] (2020) applied logistic regression to 804 Crohn's disease patients for flare-up prediction, achieving 90% accuracy. Despite the relatively large cohort, the study is limited to a single disease and binary outcome, without comorbidity modeling. Amalia et al. [44] (2025) employed deep learning with oversampling/undersampling balancing for celiac disease detection in a mobile application context, reporting improved balanced accuracy. The class-imbalance correction strategy adopted by Amalia et al. is conceptually sound, though their approach applies balancing outside a formal leakage-prevention framework.

5.2 Cardiovascular and Metabolic Disease

Pal et al. [31] (2022) applied MLP to cardiovascular data for CVD prediction at 82.47% accuracy, the lowest reported performance in **Table 2**, highlighting the limitations of single-layer neural networks without optimization for structured clinical tabular data. Tarek et al. [34] (2025) achieved 99.9% accuracy for early CVD detection using SO+ML, while El-Rashidy et al. [36] reported 94.4–97.1% accuracy for ICU ventilation and mortality prediction on MIMIC-III data using PSO-optimized multi-task learning. Both demonstrate performance gains achievable through metaheuristic optimization, though both remain binary-only and disease-specific.

Rady et al. [27] (2019) addressed CKD stage prediction using PNN, MLP, SVM, and RBF classifiers, with PNN achieving the best performance but with limited features. Ramesh et al. [38] (2025) extended this domain with an optimized MLP combined with feature selection for early CKD detection, achieving high accuracy but remaining binary-only.

kangra et al. [39] (2025) applied hybrid ML with Boruta feature selection to diabetes clinical data, reporting high AUC with robust comparison, but limited to a single disease with restricted modalities.

5.3 Other Chronic Conditions

Terlapu et al. [40] (2025) achieved improved accuracy for liver disease classification using MLP optimized by WOA, though the approach suffers from algorithm complexity and limited feature scope. Vodnala et al. [42] (2025) applied feature selection combined with ML for COPD/asthma classification from cough sound datasets, achieving effective feature-based classification but constrained to specialized audio data and binary outcomes. Kishan et al. [41] (2025) employed neural networks combined with clustering for quality-of-life assessment across healthcare datasets, reporting good predictive value, though the approach provides indirect rather than direct disease prediction.

studies exhibit Gaps 1–5 uniformly. The optimization-enhanced methods (WOA, GGO, PSO) show the highest reported accuracy gains but also the lowest interpretability, compounding Gap 5. The cardiovascular studies demonstrate that metaheuristic optimization can push binary accuracy above 99%, yet this does not translate to multiclass comorbidity settings (Gap 1) and no calibration analysis accompanies any of these high-accuracy claims (Gap 4). In sum, all 17 non-musculoskeletal studies reviewed exhibit Gaps G1–G5 uniformly, with no study simultaneously addressing multiclass scope, leakage-free preprocessing, statistical validation, calibration, and explainability.

6 ML for Fibromyalgia and Psychosomatic Comorbidities

Fibromyalgia represents the most clinically relevant domain for multiclass comorbidity classification given its well-established co-occurrence with disc herniation and IBS. The following studies examine ML approaches within this domain.

Acharya et al. [18] (2023) demonstrated ML-based fibromyalgia diagnosis via quantum-inspired feature extraction from 139 single-lead ECG signals using k-NN and SVM, achieving 93.87% binary classification accuracy. While this is the study most directly relevant to the FM-DH-IBS comorbidity context, its signal modality (ECG), sample size (n = 139), and binary scope (FM vs. non-FM) differ substantially from a multi-instrument, three-class formulation. ECG signals do not encode gastrointestinal or spinal pathology signatures, which limits this approach to FM detection rather than FM comorbidity discrimination.

Sarker et al. [32] (2023) leveraged RoBERTa natural language processing for chronic pain cohort identification from social media, reporting F1 = 0.84, an innovative but text-specific approach inapplicable to structured psychometric instrument data.

Why Symptom-Based ML Fails for FM Comorbidity Detection

Current symptom-based ML approaches for fibromyalgia face three specific structural limitations that prevent effective multiclass comorbidity discrimination:

1. **Shared symptom space.** FM, DH, and IBS share a substantial common symptom profile (fatigue, chronic pain, sleep disturbance, and cognitive impairment) that is captured by aggregate instruments such as the WPI and SSS total scores. Binary symptom-based classifiers exploit this shared signal to distinguish FM from healthy controls but cannot resolve the within-FM comorbidity structure, where the discriminating signal lies in the *sub-item* and *inter-instrument* variation rather than the aggregate score.

2. **Aggregate score collapse.** When aggregate WPI and SSS total scores are used as features, the sub-item variation that discriminates DH (musculoskeletal and anxiety-tension features, particularly HAMA sub-items) from IBS (gastrointestinal somatic SSS sub-items) is collapsed into a single scalar value. No existing reviewed study uses item-level psychometric features for comorbidity discrimination.

3. **Absence of inter-instrument correlations.** The established clinical association between HAMA-Tension subscale scores and DH-specific tender-point distributions, and between SSS somatic sub-items and IBS visceral sensitivity, represents discriminative cross-instrument signal that no current ML model exploits. Per-class SHAP attribution is precisely the mechanism needed to identify and validate such inter-instrument feature interactions.

These structural limitations directly motivate the five-gap framework presented in **Section 8**. The absence of multiclass comorbidity classification within fibromyalgia populations across all reviewed studies represents the most clinically significant gap identified in this review.

Across both fibromyalgia studies reviewed, Gaps G1–G5 apply uniformly. The ECG-based and text-based modalities demonstrate innovative signal processing, yet neither can be extended to multiclass comorbidity discrimination using structured psychometric instrument data, the clinical context where differential diagnosis among NC, DH, and IBS is most needed.

7 Specialized Signal-Based and EEG Classification

A subset of reviewed studies addresses chronic disease classification using specialized signal modalities, including EEG, physiological sensors, and symptom-based IoT data. These approaches demonstrate strong within-domain performance but are not generalizable to multi-instrument comorbidity contexts.

Kukker et al. [29] (2021) applied Genetic Fuzzy Q-Learning for EEG-based seizure classification, achieving 96.79% accuracy. Pal et al. [30] (2022) applied k-NN to COVID-19 symptom data for prognosis prediction at 97.97% accuracy, demonstrating strong performance of instance-based methods on symptom data in a binary context. Elshewey et al. [35] (2024) achieved 96.12% accuracy for EEG-based eye state classification using MBER+KNN, while Elshewey et al. [33] (2025) achieved 99.58% for heart disease classification using GGO+LSTM. Both demonstrate excellent performance within specialized signal domains but without generalization to multi-instrument comorbidity classification.

Ganiger et al. [11] (2018) benchmarked RF, DT, and SVM for chronic disease prediction, with RF achieving the best results, but without formal feature engineering or leakage-prevention protocols.

universally exhibit all five gaps. These approaches are the most technically specialized in the reviewed literature yet remain the furthest from clinical comorbidity classification: their signal modalities (EEG, ECG, cough audio) encode domain-specific pathology but do not generalize to the multi-instrument psychometric feature space required for FM comorbidity discrimination.

Across all signal-based and EEG studies reviewed, Gaps G1–G5 apply uniformly. These approaches achieve strong within-domain performance but are not generalizable to multi-instrument comorbidity contexts; their signal modalities encode domain-specific pathology signatures that do not extend to the psychometric feature space required for FM comorbidity discrimination.

Figure 3 contextualizes reported accuracy against publication year across all studies with numeric performance values, with bubble size proportional to dataset size.

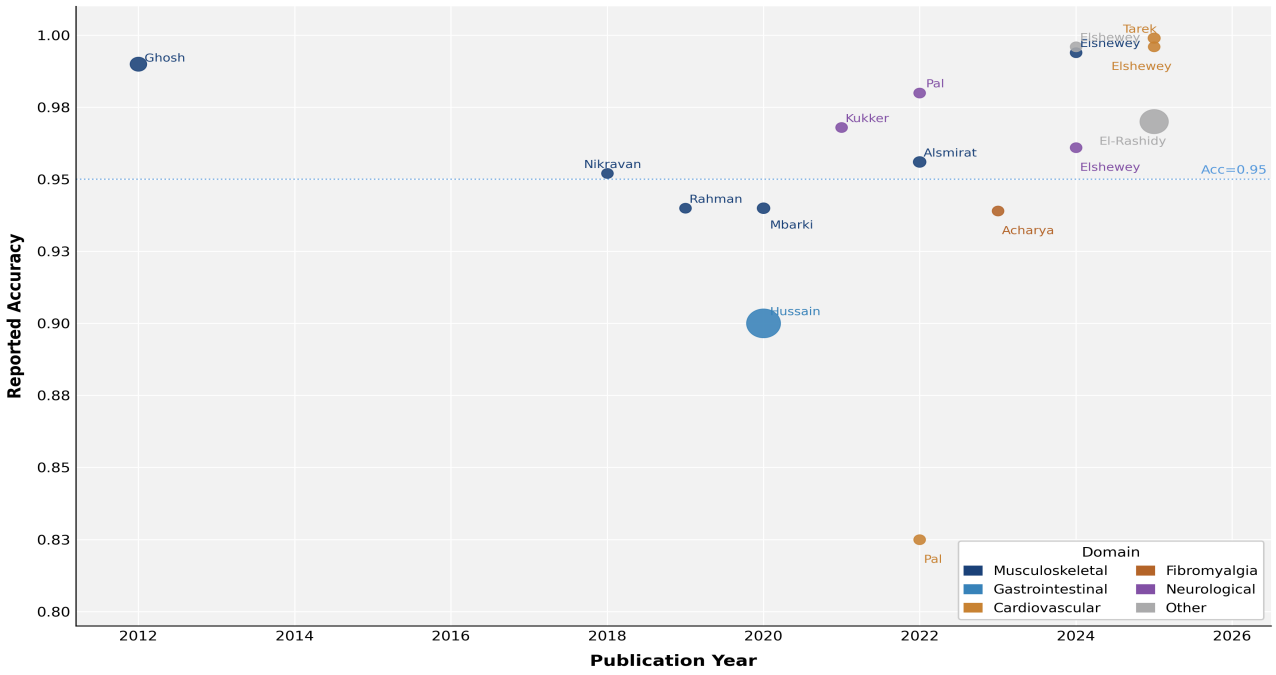

Figure 3: Reported accuracy vs. publication year for studies with numeric performance values ($n = 17$). Bubble size $\propto$ dataset size; color = clinical domain.

8 Cross-Cutting Methodological Analysis and Identified Gaps

Before examining the five identified gaps, key methodological concepts are defined for clarity. **Data leakage** occurs when information from outside the training dataset is used during model development, resulting in overly optimistic performance estimates that fail to generalize to real-world clinical data [12, 23]. **Probability calibration** refers to the alignment between predicted class probabilities and observed outcome frequencies; a well-calibrated model produces probability outputs that are directly interpretable as clinical risk estimates [15]. **SHAP (SHapley Additive exPlanations)** is a game-theoretic framework that assigns each input feature a contribution value to the model output, providing consistent and locally accurate feature attribution grounded in cooperative game theory [16, 58].

8.1 Class Imbalance and Leakage

Only Raj et al. [43] and Amalia et al. [44] explicitly address class imbalance. Critically, applying SMOTE outside cross-validation folds allows synthetic instances derived from held-out samples to contaminate training, an empirically validated form of data leakage [12, 23]. None of the 25 studies explicitly reports within-fold SMOTE application.

8.2 Explainability

Only Acharya et al. [18] and Kangra et al. [39] provide partial feature attribution. None applies SHAP [16] with per-class decomposition, which is essential for comorbidity classification where per-class feature attribution is clinically actionable.

8.3 Statistical Validation

All 25 studies report performance as point estimates without confidence intervals, omnibus testing, or correction for multiple comparisons. The Friedman test [14] provides the appropriate non-parametric omnibus statistic:

$$\chi_F^2 = \frac{12N}{k(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1) \tag{3}$$

where N is the number of datasets, k is the number of classifiers, and R_j is the mean rank of classifier j. Post-hoc Bonferroni-corrected Wilcoxon pairwise comparisons [59] with Cohen's d effect sizes are required for credible multi-classifier benchmarking.

8.4 No Calibration Analysis

No reviewed study evaluates probability calibration. The Brier Score [60] quantifies calibration as:

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \tag{4}$$

where f_i is the predicted probability and $o_i \in \{0, 1\}$ is the observed outcome. A perfectly calibrated model achieves BS = 0; random prediction yields BS = 0.25. This metric is critical for threshold-based clinical decision support, yet absent from all 25 reviewed studies.

Figure 4 provides a methodological quality heatmap for 16 representative studies across six key dimensions, illustrating the degree to which each study addresses accuracy reporting, AUC, multiclass scope, leakage-free preprocessing, statistical validation, and explainability.

Study	Accuracy Reported	AUC Reported	Multiclass Scope	Leakage-free Preprocessing	Statistical Validation	XAI / Explainability
Ghosh (2012)	✓✓	✓	×	×	×	×
Nikravan (2018)	✓✓	✓	×	×	×	×
Rahman (2019)	✓✓	✓	×	×	×	×
Hussain (2020)	✓✓	✓✓	×	×	×	×
Mbarki (2020)	✓✓	✓	×	×	×	×
Kukker (2021)	✓✓	~	×	×	×	×
Alsmirat (2022)	✓✓	✓	×	×	×	×
Pal (2022)	✓✓	✓	×	×	×	×
Acharya (2023)	✓✓	✓	×	×	×	~
Sarker (2023)	~	✓	×	×	×	×
Elshewey (2024)	✓✓	~	×	×	~	×
El-Rashidy (2025)	✓✓	✓✓	×	~	~	×
Elshewey (2025)	✓✓	~	×	×	~	×
Ramesh (2025)	✓✓	~	×	~	~	×
Kangra (2025)	✓✓	✓✓	×	~	~	~
Raj (2025)	✓✓	~	×	~	~	×

✓✓ Excellent ✓ Present ~ Partial × Absent

Figure 4: Methodological quality heatmap – 16 representative studies across 6 dimensions. ✓✓=Excellent; ✓=Present; ~=Partial; ×=Absent. The consistent pattern of high accuracy reporting paired with absent leakage controls, validation, and explainability is clearly visible.

8.5 Summary of Five Identified Gaps

Cross-cutting analysis of the 25 reviewed studies reveals five systematic methodological deficiencies that apply uniformly across all clinical domains and algorithmic families. **Table 4** summarizes each gap, its evidence base in the reviewed literature, and the corresponding future research direction. **Figure 5** complements this summary by comparing the average methodological coverage of the 25 reviewed studies against the ideal future study profile, highlighting the dimensions where the greatest improvement is needed.

Taken together, these five gaps define a coherent methodological agenda for next-generation chronic disease classification research. Addressing them simultaneously within a single, reproducible pipeline represents the critical next step toward clinically credible and translationally relevant ML frameworks.

Figure 6 presents a comprehensive taxonomy of ML approaches across clinical domains, organized by algorithmic family and disease category. The taxonomy visually confirms that the five identified gaps (**Table 4**) span all branches uniformly, reinforcing the finding that no reviewed domain has addressed them simultaneously. **Figure 5** further quantifies this coverage deficit against the ideal methodological profile.

9 Recommended Gold Standard Pipeline

Building on the Ideal Future Study Profile (**Figure 5**), we propose the following Gold Standard pipeline for multiclass comorbidity classification, specifying the exact operational sequence and design parameters:

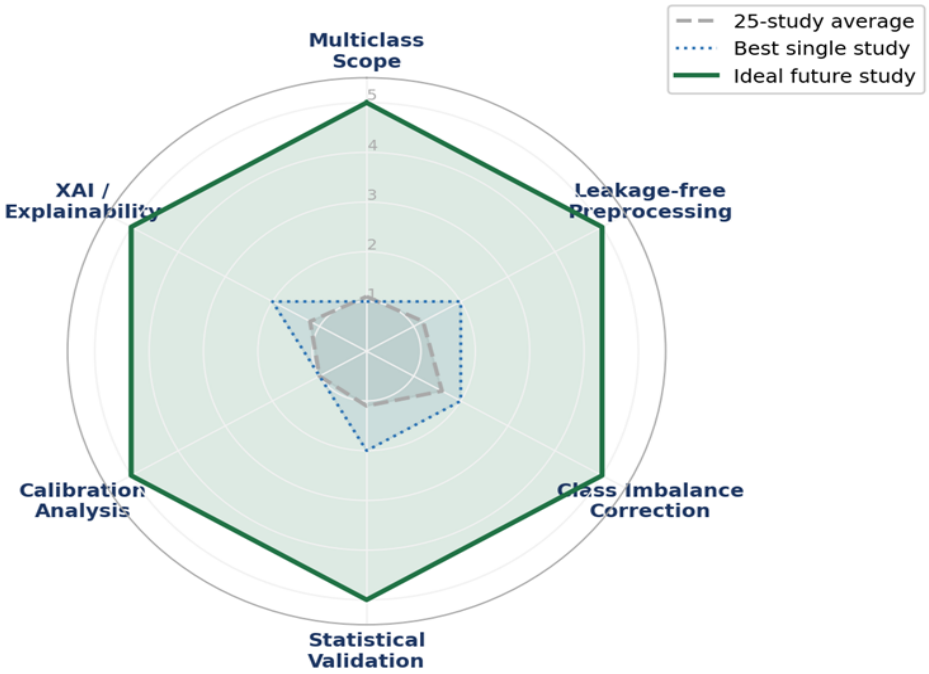

Figure 5: Radar chart comparing methodological coverage: average of 25 prior studies (grey dashed), best single prior study (blue dotted), and ideal future study profile (green). The gap is largest on multiclass scope, leakage-free preprocessing, and statistical validation.

Table 4: Five Identified Methodological Gaps and Future Research Directions.

Identified Gap	Evidence from Reviewed Studies	Future Direction
Binary-only task scope	All 25 studies address single-disease binary classification; none performs multiclass comorbidity discrimination.	Unified multiclass frameworks for differential diagnosis of co-occurring conditions.
Preprocessing data leakage	No study explicitly reports within-fold imputation, standardization, feature selection, or resampling.	Leakage-free preprocessing protocols with formal fold-wise execution map.
Absent statistical validation	All studies report point estimates only; no omnibus testing, pairwise correction, or effect sizes.	Friedman omnibus test + Bonferroni-corrected Wilcoxon comparisons with Cohen's d.
No calibration analysis	No study evaluates Brier Scores, log-loss, or reliability curves.	Systematic probability calibration evaluation across all classifier families.
Absence of XAI / SHAP	Only 2 of 25 studies provide partial attribution; none applies per-class SHAP decomposition.	SHAP TreeExplainer with global rankings and per-class beeswarm decompositions.

1. **Data collection.** Multi-instrument psychometric assessment at item level (HAM-A, FIQ-R, WPI, SSS sub-items) combined with structured EHR data from ≥3 clinical sites. Minimum recommended sample: $n \geq 200$ per comorbidity class (based on power analysis for 10-fold CV with Friedman testing at $\alpha = 0.05$, power = 0.80, $K = 5$ classifiers).
2. **Leakage-free preprocessing (within each fold).** (i) KNN imputation ($k = 2$); (ii) MICE refinement ($T_{\max} = 20$); (iii) z-score standardization using training-fold statistics only; (iv) mutual information feature selection ($K = 100$). All estimators are fitted exclusively on training folds and applied without refitting to held-out validation folds.
3. **Class imbalance correction (within each fold).** SMOTE with $k = 1$ applied to training folds only. Held-out folds retain the original class distribution to provide unbiased performance estimates.
4. **Model training.** Benchmark ≥10 classifiers spanning at least four algorithm families under 10-fold stratified cross-validation with a nested 5-fold inner loop for hyperparameter search.
5. **Statistical validation.** Friedman omnibus test on per-fold accuracy vectors, followed by Bonferroni-corrected Wilcoxon signed-rank pairwise comparisons for all $\binom{K}{2}$ model pairs. Report Cohen's d and Cliff's δ effect sizes alongside p-values.

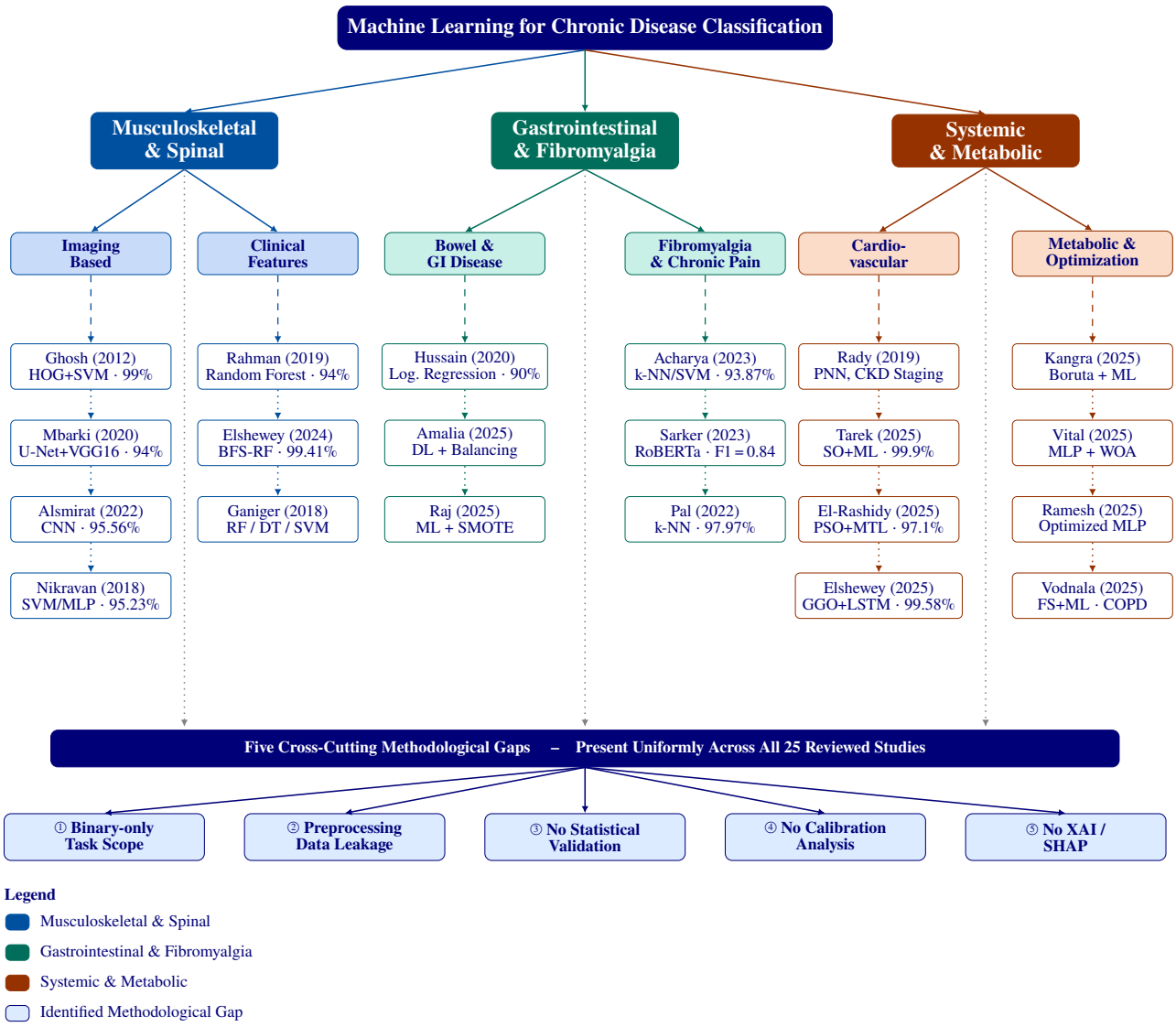

Figure 6: Taxonomy of ML approaches for chronic disease classification, organized by clinical domain. Leaf nodes show key study and reported performance. The five cross-cutting methodological gaps (bottom) apply uniformly across all branches.

6. **Probability calibration.** Evaluate Brier Score and log-loss via out-of-fold predictions. Plot reliability curves per class. Apply Platt calibration post-hoc for models without native probabilistic outputs (e.g. SVM).

7. **Explainability.** Apply SHAP TreeExplainer for the best-performing model: global feature importance rankings, per-class beeswarm decompositions, and dependence plots for the top discriminating features.

Table 5 provides a quick-reference summary of the seven pipeline steps with key parameters.

Table 5: Quick-Reference Summary of the Seven Gold Standard Pipeline Steps.

Step	Stage	Key Method	Key Parameter
1	Data Collection	Multi-instrument + EHR	$n \geq 200$ per class; ≥3 sites
2	Preprocessing	KNN + MICE + z-score + MI	Within-fold only; K=100 features
3	Imbalance Correction	SMOTE	k=1; training folds only
4	Model Training	≥10 classifiers	10-fold CV; nested 5-fold inner loop
5	Statistical Validation	Friedman + Wilcoxon	Bonferroni correction; Cohen's d
6	Calibration	Brier Score + reliability	Platt post-hoc for SVM
7	Explainability	SHAP TreeExplainer	Per-class beeswarm + dependence

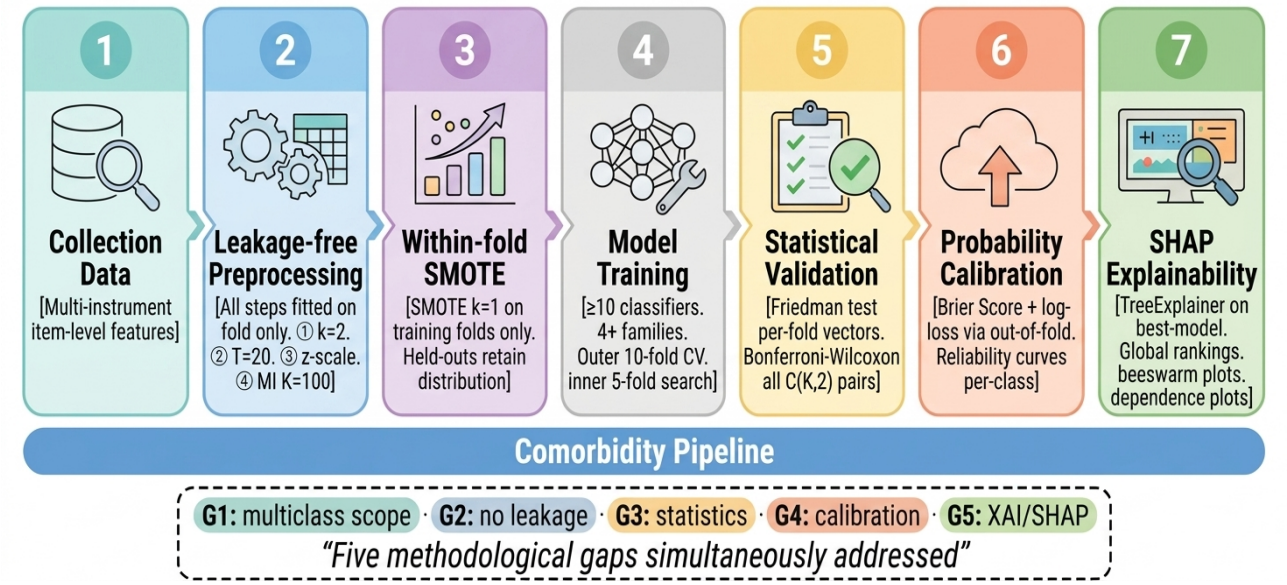

Figure 7: Recommended gold standard pipeline for multiclass comorbidity classification, organized by phase. The seven operational steps span three phases: Data & Preprocessing, Model Development, and Validation & Explanation. The five cross-cutting methodological gaps (bottom) are simultaneously addressed by this pipeline.

10 Future Work

Future work should address the five identified gaps simultaneously within the Gold Standard pipeline above:

1. **External multi-site validation.** Prospective validation on independent cohorts from ≥3 clinical sites with different recruitment protocols and patient demographics.
2. **Expanded comorbidity label space.** Extension to psychiatric comorbidities (depression, anxiety disorders) and metabolic conditions (type 2 diabetes, hypothyroidism) using multi-site datasets with $n \geq 200$ per class.
3. **Multi-label classification formulations.** Future datasets should capture simultaneous co-occurrence of multiple comorbidities (e.g. a patient with both DH and IBS), requiring multi-label rather than multiclass classification frameworks.
4. **SHAP integration with optimization-enhanced methods.** As detailed in **Section 3.6**, post-hoc TreeExplainer or KernelSHAP applied to converged hybrid models can make optimization-enhanced pipelines clinically transparent.
5. **Federated learning.** Privacy-preserving federated learning across clinical sites would enable model training on larger, more diverse cohorts without centralizing sensitive patient data.
6. **Longitudinal EHR integration.** Temporal modeling of comorbidity onset trajectories using EHR data would enable prognostic rather than purely diagnostic classification.

11 Conclusion

This narrative review examined 25 ML studies for chronic disease classification (2012–2025), spanning disc herniation, IBS, fibromyalgia, cardiovascular disease, and related comorbidities. The literature demonstrates a clear trajectory from early binary imaging-based classifiers toward optimization-enhanced ensemble methods, with consistently high reported accuracy. However, five systematic methodological deficiencies (binary-only task scope, preprocessing leakage, absent statistical validation, no calibration analysis, and minimal explainability) apply uniformly across all reviewed studies.

Despite accuracy of 82–99.9% across all 25 reviewed studies, none simultaneously addresses multiclass comorbidity scope, leakage-free preprocessing, statistical validation, calibration analysis, and explainability. These five gaps define the methodological agenda for next-generation chronic disease classification research. The most critical unmet need remains a unified multiclass framework for simultaneous differential diagnosis of co-occurring chronic conditions within fibromyalgia populations. The Gold Standard pipeline proposed in **Section 9** provides a concrete, actionable framework for addressing all five gaps simultaneously.

Declaration

Author Contribution Statement

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable.

Consent for Publication

Not applicable.

Data Availability

No new data were generated or analyzed in this study. This is a narrative review of previously published literature.

Acknowledgments

The authors thank the reviewers, Associate Editor, and Editor-in-Chief for their valuable comments and suggestions, which helped improve the quality of this paper.The authors also acknowledge the use of Deep Seek for assistance in improving the English grammar and language clarity.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Disclosure Statement

The authors declare no competing interests.

References

[1] T. Vos, S. S. Lim, C. Abbafati, K. M. Abbas, M. Abbasi, M. Abbasifard, M. Abbasi-Kangevari, H. Abbastabar, F. Abd-Allah, A. Abdelalim *et al.*, "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] 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.

[3] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, "Defining comorbidity: implications for understanding health and health services," *Annals of Family Medicine*, vol. 7, no. 4, pp. 357–363, 2009.

[4] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *Jama*, vol. 319, no. 13, pp. 1317–1318, 2018.

[5] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[6] W. R. Hersh, "Healthcare data analytics," *BMJ*, vol. 334, no. 7585, pp. 1139–1140, 2007.

[7] 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, pp. 230–243, 2017.

[8] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[9] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[10] M. Alsmirat, N. Al-Mnayyis, M. Al-Ayyoub, and A.-M. Asma'A, "Deep learning-based disk herniation computer aided diagnosis system from mri axial scans," *IEEE Access*, vol. 10, pp. 32 315–32 323, 2022.

[11] S. Ganiger and K. Rajashekharaiah, "Chronic diseases diagnosis using machine learning," in *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*. IEEE, 2018, pp. 1–6.

[12] 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.

[13] 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.

[14] 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.

[15] J. Platt, "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.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[17] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 0210–0215.

[18] 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. 101 359–101 372, 2023.

[19] B. A. Kleykamp, M. C. Ferguson, E. McNicol, I. Bixho, L. M. Arnold, R. R. Edwards, R. Fillingim, H. Grol-Prokopczyk, D. C. Turk, and R. H. Dworkin, "The prevalence of psychiatric and chronic pain comorbidities in fibromyalgia: an acttion systematic review," in *Seminars in arthritis and rheumatism*, vol. 51, no. 1. Elsevier, 2021, pp. 166–174.

[20] 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)*. IEEE, 2020, pp. 529–533.

[21] B. N. Green, C. D. Johnson, and A. Adams, "Writing narrative literature reviews for peer-reviewed journals: secrets of the trade," *Journal of Chiropractic Medicine*, vol. 5, no. 3, pp. 101–117, 2006.

[22] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[23] R. Blagus and L. Lusa, "Smote for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, pp. 1–16, 2013.

[24] 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)*. IEEE, 2012, pp. 114–117.

[25] 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)*. IEEE, 2019, pp. 1–4.

[26] 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.

[27] 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.

[28] W. Mbarki, M. Bouchouicha, S. Frizzi, F. Tshibasu, 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.

[29] A. Kukker and R. Sharma, "A genetic algorithm assisted fuzzy q-learning epileptic seizure classifier," *Computers & Electrical Engineering*, vol. 92, p. 107154, 2021.

[30] M. Pal, S. Parija, R. K. Mohapatra, S. Mishra, A. A. Rabaan, A. Al Mutair, S. Alhumaid, J. A. Al-Tawfiq, and K. Dhama, "Symptom-based covid-19 prognosis through ai-based iot: A bioinformatics approach," *BioMed Research International*, vol. 2022, no. 1, p. 3113119, 2022.

[31] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Medicine*, vol. 17, no. 1, pp. 1100–1113, 2022.

[32] A. Sarker, S. Lakamana, Y. Guo, Y. Ge, A. Leslie, O. Okunromade, E. Gonzalez-Polledo, J. Perrone, and A. M. McKenzie-Brown, "# chronicpain: automated building of a chronic pain cohort from twitter using machine learning," *Health data science*, vol. 3, p. 0078, 2023.

[33] 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.

[34] 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.

[35] A. M. Elshewey, A. A. Alhussan, D. S. Khafaga, E.-S. M. Elkenawy, and Z. Tarek, "Eeg-based optimization of eye state classification using modified-ber metaheuristic algorithm," *Scientific Reports*, vol. 14, no. 1, p. 24489, 2024.

[36] N. El-Rashidy, Z. Tarek, A. M. Elshewey, and M. Y. Shams, "Multitask multilayer-prediction model for predicting mechanical ventilation and the associated mortality rate," *Neural Computing and Applications*, vol. 37, no. 3, pp. 1321–1343, 2025.

[37] A. M. Elshewey and A. M. Osman, "Orthopedic disease classification based on breadth-first search algorithm," *Scientific Reports*, vol. 14, no. 1, p. 23368, 2024.

[38] 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.

[39] 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.

[40] P. V. Terlapu and K. Bhumika, "Intelligent liver disease identification using optimized multilayer perceptron using whale optimization algorithm (woa)," *IAENG International Journal of Computer Science*, vol. 52, no. 11, 2025.

[41] S. R. Kishan, B. Senthilkumaran, S. Malluvalasa *et al.*, "Machine learning based healthcare system for assessment of quality life," *IAENG International Journal of Computer Science*, vol. 52, no. 11, 2025.

[42] N. Vodnala, P. Lankireddy, and P. Yarlagadda, "Identifying key features for machine learning classification of copd and asthma cough," *IAENG International Journal of Computer Science*, vol. 52, no. 10, 2025.

[43] 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, 2025.

[44] 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.

[45] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[46] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," 1984.

[47] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[48] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[49] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[50] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[51] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[52] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[53] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.

[54] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[55] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," *arXiv preprint arXiv:1302.4964*, 2013.

[56] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[57] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[58] L. S. Shapley *et al.*, "A value for n-person games," 1953.

[59] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[60] W. B. Glenn *et al.*, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.