IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 10, October, 2026, pp. 2865-2875
Journal homepage: http://iieta.org/journals/jesa

An Optimized Hybrid Architecture Integrating Random Convolutions and Bayesian-Tuned Gradient Boosting for Robust Time Series Classification

Wildan Mahmud*, Ahmad Zainul Fanani, Heru Agus Santoso, Fikri Budiman

Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

Corresponding Author Email: p41202200049@mhs.dinus.ac.id

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/jesa.591009

Received: 4 August 2026
Revised: 11 September 2026
Accepted: 22 September 2026
Available online: 31 October 2026

Keywords:
Bayesian Optimization, Gradient Boosting, MultiROCKET, Optuna, overfitting mitigation, time series classification

ABSTRACT

While MultiROCKET excels in time series classification (TSC) feature extraction, its reliance on the linear Ridge Classifier restricts the mapping of complex non-linear decision boundaries often encountered in industrial equipment monitoring. This study proposes substituting the linear head with Gradient Boosting Machines (XGBoost and LightGBM). To mitigate high-dimensional overfitting, the models are autonomously calibrated using Bayesian Tree-structured Parzen Estimator (TPE) optimization via Optuna to enforce strict structural regularization. An empirical ablation study across 28 UCR datasets was conducted, incorporating the state-of-the-art (SOTA) deep learning model, InceptionTime, as a comprehensive baseline. Statistical analyses (Multi-Comparison Matrix (MCM) and Critical Difference (CD) plots) confirm that Optuna improves XGBoost's average accuracy from 84.06% to 87.73% (average rank 2.75). Although the Ridge Classifier (91.28%) and InceptionTime (90.96%) retain top-tier accuracy, the proposed XGBoost-Optuna architecture elevates the lower accuracy bound, ensuring predictive stability across diverse domains. Furthermore, computational analysis reveals memory scalability limitations in LightGBM during extreme multiclass scenarios, establishing XGBoost as a more operationally resilient architecture. Ultimately, this research presents a practical non-linear modeling alternative with rational computational trade-offs, balancing offline training overhead with rapid inference times (0.85 seconds on average), making it highly feasible for deployment in edge-based automation systems.

1. INTRODUCTION

Over the past decade, the rapid development of Internet of Things (IoT) infrastructure, wireless sensor networks, and cyber-physical systems has triggered a massive surge in time series data volume [1]. This data is produced massively, continuously, and often at high velocity across various modern industrial sectors. The utilization of these data streams encompasses a remarkably broad spectrum of applications, ranging from smart health monitoring (such as Electrocardiogram (ECG) signal analysis), human activity recognition, and financial market dynamics analysis, to automation systems and smart manufacturing [2]. In the context of automation systems, the capability to read, interpret, and respond to this time series data in real-time is a critical component for achieving fully autonomous operations. Specifically, robust time series classification (TSC) frameworks directly support critical industrial control mechanisms, equipment condition monitoring, and predictive maintenance by identifying latent fault patterns before systemic failures occur [3]. However, the main challenge in real-world time series data classification lies in its high-dimensional characteristics, susceptibility to noise, and complex temporal dependency patterns. As the scale of acquired data increases, there is a growing technical demand within the systems engineering literature. Current industrial environments require not only classification algorithms with high accuracy levels but also demand model architectures that are computationally efficient in both training and inference phases, enabling practical implementation in resource-constrained automation systems [1-3].

Historically, the pursuit of optimal TSC model architectures has been dominated by deep learning-based approaches, particularly structured convolutional neural network (CNN) variants such as InceptionTime and ResNet [3]. Although these artificial neural networks set the gold standard in terms of predictive accuracy, this achievement demands substantial operational compensation. Training deep learning models requires massive computational resources, convergence times spanning days, and a heavy reliance on high-performance hardware accelerators (such as Graphics Processing Units (GPUs)/TPUs) [4]. This dependency often becomes a technical bottleneck when classification algorithms must be implemented on edge devices or industrial control systems that prioritize low latency and possess strict power constraints [5]. In response to these computational issues, a classification paradigm based on random convolutional feature transformation emerged as a highly efficient alternative. Recent representative approaches within this family, including ROCKET, MiniROCKET, Hydra, and MultiROCKET, are capable of instantaneously extracting tens of thousands of feature representations from time series data using a series of

random convolutional kernels and pooling operations [6]. Extensive evaluations have shown that these approaches significantly reduce training computation time while maintaining competitive accuracy against deep learning models [6]. Nevertheless, although MultiROCKET has resolved the efficiency challenge in the feature extraction phase, the overall effectiveness of this architecture, and similarly for other ROCKET variants, remains hindered by the downstream processing mechanism that serves as its primary anchor.

In the standard MultiROCKET architecture, these tens of thousands of massively extracted features are mapped into class labels using a Ridge Classifier, an algorithm based on linear classification [6]. The initial selection of this linear classifier was logical, as it was driven primarily by computational efficiency motives and its capability to handle multicollinearity in high-dimensional feature spaces via an L_2 regularization mechanism. Although this strategy has proven successful in minimizing training latency, the deployment of a linear head in the downstream phase introduces inherent structural limitations. Linear algorithms operate under rigid mathematical assumptions, namely that the interactions between extracted features and the target variable are strictly additive and linearly independent. However, across most real-world time series datasets such as transient vibration signals in industrial machinery for equipment condition monitoring or complex fluctuations in biomedical sensors, class-discriminative patterns are often embedded within highly non-linear multivariate interactions [7]. Consequently, the Ridge Classifier may struggle to fully exploit advanced correlations among these random convolutional features to form accurate decision boundaries. This limitation in effectively capturing non-linear interactions diminishes the representational potential of the data, which can constrain the architecture's overall effectiveness [8]. Therefore, a strong technical motivation exists to re-engineer this final classification component to bridge the gap between rich convolutional feature representations and non-linear mapping capabilities.

To overcome the limitations of the baseline architecture, this study proposes a hybrid TSC framework that synergistically integrates the rapid feature extraction speed of MultiROCKET with the reliable non-linear modeling capacity of Gradient Boosting Machines (GBM). Specifically, the final-stage linear Ridge Classifier is substituted with state-of-the-art (SOTA) ensemble algorithms, namely eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM), which are widely recognized for their strong capability in mapping complex multivariate feature interactions [9, 10]. Recognizing that boosting performance is highly sensitive to hyperparameter configurations and susceptible to overfitting when confronted with tens of thousands of random convolutional features, the proposed approach is reinforced with an intelligent automated optimization mechanism. We implement Bayesian Optimization using the Tree-structured Parzen Estimator (TPE) algorithm via the Optuna framework to efficiently navigate the search space and objectively identify optimal parameter configurations [11]. To validate the effectiveness of this proposed framework, a comprehensive experimental study was conducted across 28 benchmark datasets from the UCR TSC Archive [12]. Consequently, the main contributions of this study phase include: (1) the formulation of a novel TPE-optimized MultiROCKET-GBM architecture that empirically resolves non-linear feature interaction issues; (2) a rigorous cross-dataset statistical significance analysis utilizing a Multi-Comparison Matrix (MCM) that demonstrates the performance improvements of the proposed model over linear baselines; and (3) an analytical evaluation of the trade-off between predictive accuracy enhancements and computational cost metrics, offering practical insights for algorithmic deployment in applied automation systems.

2. RELATED WORK

This section presents a systematic literature review mapping the relevant research landscape to establish a theoretical foundation and identify the research gap underpinning the development of the proposed architecture. The discussion is sequentially structured across three main literature pillars. First, the evaluation focuses on the evolution of TSC algorithms, reviewing the transition from computationally expensive deep learning models toward efficient random convolution architectures (MultiROCKET), while highlighting the inherent limitations of their default linear modeling. Second, the study explores the reliability of non-linear ensemble methods based on GBM in mapping complex feature interactions. Third, the discussion reviews the imperative of deploying an autonomous hyperparameter optimization engine based on Bayesian Optimization (specifically the TPE via Optuna) to mitigate overfitting risks in massive high-dimensional data spaces. A comprehensive synthesis of these three literature domains logically reinforces the urgency and novelty of the optimized hybrid framework developed in this study.

2.1 Advancements in time series classification

Over the past decade, the landscape of TSC has been dominated by deep learning-based architectures, which serve as the SOTA benchmark. Deep neural network models, such as InceptionTime and Residual Networks (ResNet), have proven highly superior in learning and extracting temporal representation hierarchies in an end-to-end manner. However, the accuracy advantages of these deep learning approaches demand analytical trade-offs in the form of extraordinarily high comprehensive computational costs and slow training durations due to their reliance on iterative weight update mechanisms (backpropagation) [3]. To radically resolve this computational bottleneck, an alternative paradigm focusing on extreme efficiency was introduced through the utilization of random convolutions. This paradigm originated with the RandOm Convolutional KErnel Transform (ROCKET) algorithm [13] and rapidly evolved into more comprehensive architectures tailored for industrial automation applications, such as MiniROCKET and Hydra [14-17]. Operationally, the MultiROCKET framework disrupts traditional feature learning processes by randomly initializing and freezing kernel parameters, allowing the model to instantaneously extract tens of thousands of feature dimensions. This breakthrough effectively reduces computation time from hours or days in deep learning to mere seconds, while maintaining accuracy levels comparable to SOTA algorithms [6].

Although the ROCKET family of algorithms offers highly impressive computational efficiency during the feature extraction phase, this architecture harbors a crucial structural flaw in its downstream processing module. Historically, all random convolution variants have operated with an absolute

reliance on the Ridge Classifier as their primary inference engine (linear head). Numerous literature studies highlight that linear classification models operate under rigid additive assumptions, thereby inherently lacking the capacity and flexibility to map complex, non-linear decision boundaries [7, 8]. In real-world time series datasets, discriminative information rarely exists in isolation; rather, it is concealed within cross-feature interactions among tens of thousands of activation maps. This limitation of conventional linear regression in exploiting latent correlations within a massive dimensional space creates a significant research gap. Consequently, a strong analytical imperative exists to re-engineer this baseline architecture by substituting the linear head with an advanced non-linear modeling framework—an architectural intervention that serves as the primary foundation and justification for this study.

2.2 Non-linear ensemble classification

To address the fundamental drawbacks of such linearity assumptions, tree-based ensemble classification approaches have proven to be the gold standard in high-dimensional matrix or tabular data processing. Among the diverse array of ensemble algorithms, the GBM framework stands out as a highly reliable method for extracting predictive information hierarchies. Fundamentally distinct from linear classifiers, the boosting framework constructs its predictive architecture through the sequential aggregation of a series of weak learner decision trees, where each new tree is specifically trained to correct the residual errors of preceding iterations. This recursive node-splitting mechanism enables GBM algorithms to naturally capture cross-feature interactions and delineate intricate non-linear decision boundaries without being constrained by predictor data distribution specifications [18-20]. In their evolution, advanced architectural variants such as XGBoost and LightGBM have dominated various modern multivariate classification benchmarks [9, 10, 21, 22]. With XGBoost offering strict penalty regularization and LightGBM providing massive memory efficiency through leaf-wise tree growth, this family of algorithms represents the theoretically ideal non-linear head substitution for exploiting tens of thousands of latent features generated by MultiROCKET random convolutions.

2.3 Bayesian hyperparameter optimization

Although tree-based ensemble architectures offer superior non-linear modeling capacity, their integration with massive high-dimensional feature spaces—such as tens of thousands of convolutional features extracted by MultiROCKET—creates a structural vulnerability to overfitting when models are executed using default parameter configurations [23]. Consequently, hyperparameter optimization serves as a crucial calibration phase to constrain model complexity. However, the literature indicates that traditional tuning methods, such as Grid Search and Random Search, prove highly inefficient and frequently encounter computational failures due to the curse of dimensionality. Grid Search demands resource allocations that explode exponentially as the number of parameters increases, whereas Random Search performs blind evaluations without the ability to learn from historical iteration errors. As a solution to this computational bottleneck, Bayesian Optimization has been widely adopted as an intelligent and cost-effective SOTA standard. Specifically, the TPE algorithm formulates hyperparameter search by constructing a surrogate probabilistic model that iteratively guides the search toward the most promising parameter space regions [24, 25]. Implemented through cutting-edge automation frameworks such as Optuna, this TPE optimization engine enables hybrid algorithms to autonomously calibrate essential parameters, balance exploration and exploitation, and formulate models with peak generalization performance under highly constrained time budgets [11, 26-28].

3. PROPOSED METHODOLOGY

To overcome the linearity limitations in time series modeling without sacrificing execution speed, this study proposes a hybrid classification framework specifically designed to balance computational efficiency with non-linear pattern recognition capabilities. Operationally, the proposed architecture comprises three sequentially integrated main modules: (1) MultiROCKET-based deterministic feature extraction, (2) Non-linear mapping via Gradient Boosting Machines (XGBoost/LightGBM), and (3) Bayesian Hyperparameter Optimization via Optuna TPE to constrain the model complexity [29]. The integrated workflow and sequential inter-module interactions within this proposed hybrid architecture are explicitly visualized in Figure 1.

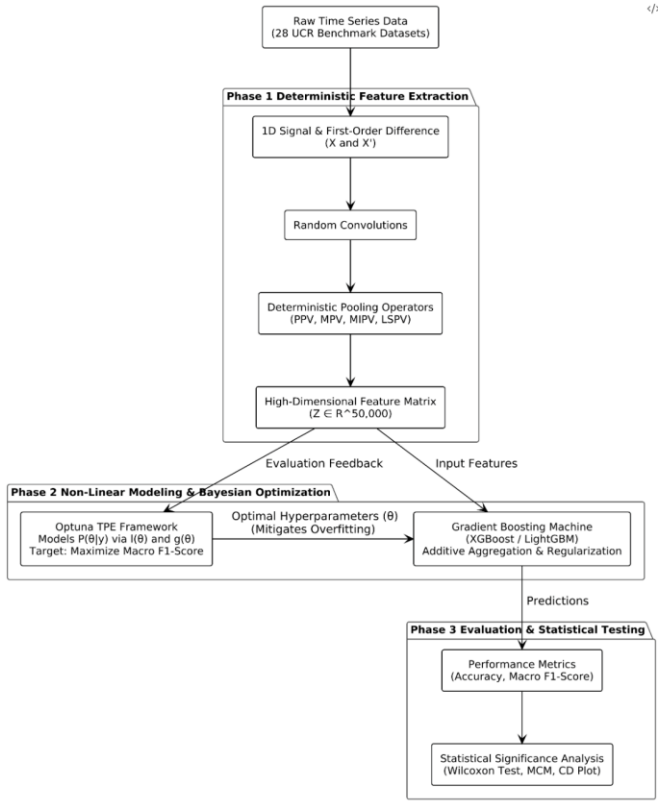


Figure 1. Flowchart of the proposed hybrid architecture combining MultiROCKET feature extraction with Bayesian-optimized Gradient Boosting

3.1 MultiROCKET feature transformation

The first phase in the proposed system architecture is deterministic feature extraction using MultiROCKET [6]. Instead of relying on computationally heavy iterative weight updates found in deep learning, this method projects raw time

series data into a high-dimensional feature space by applying a vast set of randomly initialized convolutional kernels. To capture local trend dynamics, these convolutions are applied in parallel to both the original signal and its first-order difference. The dimensionality is then processed using four deterministic pooling operators: Proportion of Positive Values, Mean of Positive Values, Mean of Indices of Positive Values, and Longest Stretch of Positive Values. Cumulatively, this transforms each time series instance into a massive representation matrix $Z \in R^D$, where $D \approx 50{,}000$ features. This rapid dimensionality expansion acts as a catalyst for classification failure in standard linear heads, thereby necessitating the non-linear ensemble integration.

3.2 Gradient boosting for non-linear classification

The massively dimensional feature representation matrix Z generated by the MultiROCKET extraction phase is subsequently processed using the GBM framework, specifically employing the SOTA architectural variants XGBoost [9] and LightGBM [10]. Unlike the Ridge Classifier, the boosting framework constructs its predictive architecture through the sequential aggregation of decision trees, which naturally captures complex cross-feature interactions within the 50,000 activation maps.

During the tree induction process, the algorithm aims to minimize an objective function that evaluates model performance while strictly controlling architectural complexity. The explicit integration of the regularization term $\Omega(f_t)$ is highly fundamental in our architecture. The objective function is formulated in Eq. (1):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(z_i)\right) + \Omega(f_t) \quad (1)$$

Given that the downstream model receives tens of thousands of latent features, it poses a severe risk of absolute overfitting. The regularization penalty forcefully dampens model complexity, ensuring the ensemble model extracts discriminative correlations without memorizing the noise present in the data [30].

3.3 Bayesian hyperparameter optimization

The massive expressive capacity of GBMs carries extreme sensitivity to hyperparameter configurations. To resolve this, we implement a Bayesian Optimization framework using the TPE algorithm [24, 31] via the Optuna library [11].

TPE models the conditional hyperparameter probability $P(\theta \mid y)$ by splitting observations into two proxy density distributions based on a performance threshold y^*. This distribution splitting is defined in Eq. (2):

$$P(\theta \mid y) = \begin{cases} l(\theta) & \text{if } y < y^* \\ g(\theta) & \text{if } y \geq y^* \end{cases} \quad (2)$$

By maximizing the ratio $l(\theta)/g(\theta)$, Optuna autonomously explores the parameter space to find the optimal limits of the tree depth, learning rate, and regularization penalties. This integration ensures the creation of an optimal balance between training convergence speed and predictive robustness during the inference phase for automation systems.

4. EXPERIMENTAL SETUP

To empirically validate the effectiveness of the proposed hybrid architecture, a large-scale comparative study was designed and executed. This section details the experimental setup specifications to ensure that the entire procedure is fully reproducible by the research community. The comprehensive discussion encompasses the characteristics of the evaluated benchmark datasets, the design of the model comparison scenarios, the specifications of both hardware and software computational environments, as well as the performance metrics and statistical significance tests applied to objectively evaluate the experimental outcomes.

4.1 Benchmark datasets

The performance evaluation of the classification algorithms in this study relies on standardized datasets from the UCR TSC Archive, a comprehensive repository widely recognized as the gold standard in time series analytics literature [12]. Rather than merely serving as abstract mathematical benchmarks, several selected datasets directly map to critical industrial automation and equipment monitoring scenarios. For instance, the FordA and FordB datasets represent real acoustic sensor streams for automated engine fault diagnosis, the Wafer dataset provides actual microelectronic manufacturing sensor data for predictive maintenance, and the SonyAIBORobotSurface datasets model edge-device telemetry for autonomous robotic navigation. This selection ensures that the empirical evaluations remain deeply anchored in real-world intelligent manufacturing and cyber-physical system engineering constraints.

This dataset collection was deliberately selected for its capability to represent a vastly broad variance of real-world domains, including motion sensor monitoring, medical signal analysis (such as ECG), spectrophotometry, and data streams from cyber-physical device simulations, encompassing highly variable levels of complexity, time series lengths, and class distribution ratios.

In the initial experimental design, 30 randomly selected datasets were certified for testing. However, during the pre-testing phase, empirical observations revealed a significant technical anomaly. Two large-scale datasets possessing an extreme number of predictive class labels (42 target classes), namely NonInvasiveFetalECGThorax1 and NonInvasiveFetalECGThorax2, triggered severe computational bottlenecks. Specifically, in the testing scenario for the Optuna-optimized LightGBM framework, the cross-validation processing load for tens of thousands of MultiROCKET features across 42 class dimensions caused the algorithm to fail to achieve convergence within a rational timeframe. Empirical profiling indicated that a single TPE search trial required approximately 1,150 seconds (~19 minutes), meaning a full 30-trial optimization budget would monopolize over 19 hours of computation for these two datasets alone. Given that the primary objective of this study is to evaluate algorithms viable for implementation in efficient real-world automation systems, and to maintain evaluation matrix consistency, benchmarking fairness, and computational feasibility across all modeling scenarios, these two datasets were established as exclusion criteria. Therefore, the final experiments and performance analysis reporting are definitively focused on the 28 valid datasets.

4.2 Experimental design and environment

To comprehensively evaluate the contribution of each processing component, the experimental design of this study is formulated through an ablation study approach, directly comparing five modeling scenarios head-to-head. The first scenario serves as the absolute baseline, comprising the standard implementation of the MultiROCKET architecture, which relies strictly on the Ridge Classifier for linear feature mapping. Subsequently, the second and third scenarios are designed to test the hypothesis of non-linear modeling substitution, wherein the convolutional features are evaluated using XGBoost and LightGBM algorithms with default hyperparameter configurations. These unoptimized scenarios serve to measure the performance transition from linear models to ensemble models without any advanced optimization interventions. Finally, the fourth and fifth scenarios represent the primary proposed frameworks (proposed models), wherein the XGBoost and LightGBM non-linear heads are dynamically configured using the Bayesian TPE optimization engine via the Optuna framework. This tiered comparison is systematically designed to isolate and validate accuracy improvements stemming purely from the non-linear capacity of the boosting models, as well as the additional performance surge generated by the effectiveness of autonomous parameter space exploration.

To establish a rigorous deep learning baseline, InceptionTime, one of the most representative CNN architectures for TSC, was officially integrated into the comparative study. This inclusion aims to benchmark whether the proposed hybrid framework can rival deep learning accuracy while maintaining strict computational efficiency and edge-device compatibility.

All experiments were executed in a Google Colab cloud computing environment equipped with an NVIDIA Tesla T4 GPU to guarantee parallel computational efficiency and computing power parity across all testing scenarios [5]. The software implementation was developed in Python 3, utilizing the aeon library (version 1.4.0) [32] as an integrated framework for time series data analysis and manipulation. Baseline modeling was executed using the scikit-learn library, whereas the non-linear modeling was implemented via the official XGBoost and LightGBM libraries. For the optimized proposed scenarios, the Optuna hyperparameter search engine was configured using the TPE algorithm with a search budget allocation of 30 iterations (trials) per dataset. To prevent data leakage during the hyperparameter search process, the objective evaluation at each Optuna iteration was conducted using an internal Stratified K-Fold Cross-Validation strategy on the training dataset, optimizing the Macro F1-Score. The optimal parameter configuration identified by the TPE was subsequently utilized to train the final model before independent evaluation on the test set.

To ensure full reproducibility of the experimental results and to provide transparency regarding the model calibration constraints, the exact boundaries of the hyperparameter search space explored by the Optuna TPE engine are strictly defined. Rather than relying on explicit L1/L2 penalty terms, this study actively mitigates high-dimensional overfitting through structural regularization—strictly constraining the maximum tree depth, limiting the number of leaves, and applying feature/instance subsampling to prevent the models from memorizing convolutional noise. The comprehensive configuration bounds autonomously optimized for both XGBoost and LightGBM architectures are detailed in Table 1.

Table 1. Hyperparameter search space for Optuna Tree-structured Parzen Estimator (TPE) optimization

Algorithm	Hyperparameter	Search Range	Sampling Distribution
XGBoost	learning_rate	[0.01, 0.3]	Log-Uniform
	max_depth	[3, 6]	Integer
	n_estimators	[50, 200]	Integer
	subsample	[0.6, 1.0]	Uniform
LightGBM	learning_rate	[0.05, 0.2]	Log-Uniform
	max_depth	[3, 5]	Integer
	num_leaves	[15, 31]	Integer
	n_estimators	[50, 100]	Integer

4.3 Evaluation metrics and statistical analysis

To comprehensively measure model reliability across the 28 evaluated datasets, the quantitative analysis is based on two primary metrics: Accuracy and the Macro-averaged F1-Score ($F1_{macro}$). Accuracy calculates the ratio of correct predictions to the total number of observation samples N, as formulated in Eq. (3):

$$Accuracy = \frac{\sum_{c=1}^{C} T\,P_c}{N} \tag{3}$$

where, C represents the number of observed classes, and TP_c is the number of true positives for the c -th class. Given that the majority of time series datasets in the UCR Archive possess imbalanced class distributions, the accuracy metric is susceptible to evaluation bias toward the majority class. Therefore, the $F1_{macro}$ metric is utilized as the primary benchmark for predictive robustness [33, 34]. Mathematically, the evaluation begins by calculating the Precision (P_c) and Recall (R_c) for each class $c \in \{1,2,\dots,C\}$ independently through Eqs. (4) and (5):

$$P_c = \frac{TP_c}{TP_c + FP_c} \tag{4}$$

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{5}$$

with FP_c representing false positives and FN_c representing False Negatives. The class-specific $F1$ score for class c is extracted using the harmonic mean in Eq. (6):

$$F1_c = 2 \cdot \frac{P_c \cdot R_c}{P_c + R_c} \tag{6}$$

The final $F1_{macro}$ value is calculated as the unweighted arithmetic mean of all $F1_c$ scores, independent of class frequency, as shown in Eq. (7):

$$F1_{macro} = \frac{1}{C}\sum_{c=1}^{C} F\,1_c \tag{7}$$

The use of $F1_{macro}$ as the objective target function ensures that the Optuna TPE optimization forcefully maintains predictive reliability on minority classes.

Complementing the evaluation of these deterministic metrics, the statistical significance of the computational performance improvements is analyzed. Cross-model performance comparisons are evaluated using the non-parametric Wilcoxon signed-rank test to reject the null hypothesis (H_0) [35]. These statistical probability results are subsequently hierarchically quantified into an MCM and visualized through Critical Difference (CD) Plots to provide transparent justification regarding the significance threshold of the proposed architecture against the baseline architecture [35].

5. RESULTS AND DISCUSSION

This section comprehensively details the empirical findings derived from the five modeling scenarios executed across the 28 datasets from the UCR Time Series Archive. To provide structured analytical insights, the discussion of the experimental results is delineated across three primary evaluation dimensions. First, the analysis focuses on evaluating aggregate predictive performance using Accuracy and Macro F1-Score metrics to examine the extent to which the Gradient Boosting-based non-linear head substitution transcends the limitations of the linear baseline architecture. Second, these performance findings are mathematically validated through statistical significance analysis utilizing CD Plots and MCM visual instruments to rigorously confirm the superiority of the proposed models. Third, the discussion concludes with an analytical evaluation of the trade-off between predictive accuracy enhancements and incurred computational costs, while highlighting crucial findings regarding the operational scalability limits of specific algorithms when confronted with extreme multi-class scenarios.

5.1 Predictive performance evaluation

The predictive capability evaluation of the five modeling scenarios is conducted by summarizing the aggregated Accuracy and Macro F1-Score metrics across the 28 evaluated datasets. To dissect the performance distribution characteristics in greater depth, the distribution analysis is visualized through violin plots and box plots, summarizing the accuracy level variance of each algorithm. Based on these distribution visualizations, although the baseline Ridge Classifier exhibits a competitive performance track record, there is a distinct structural vulnerability characterized by an elongated lower bound extending into low accuracy ranges on complex-dimensional datasets. Conversely, the intervention of the proposed frameworks integrating Gradient Boosting with Optuna optimization (specifically, the Optuna-optimized XGBoost and LightGBM scenarios) successfully restructures the performance distribution, rendering it significantly more compact and stable. Specifically, the proposed models empirically demonstrate the capability to elevate the median accuracy value while sharply raising the lower bound of the performance spread compared to the Ridge Classifier. The truncation of the lower tail in this distribution serves as a crucial indicator affirming that the TPE-optimized non-linear head substitution not only minimizes severe predictive anomalies but also provides a superior guarantee of robustness across the entire spectrum of time series data variations. The overall accuracy distribution across 28 evaluated UCR datasets is depicted in Figure 2, highlighting the performance gain achieved by the proposed Bayesian-tuned architecture.

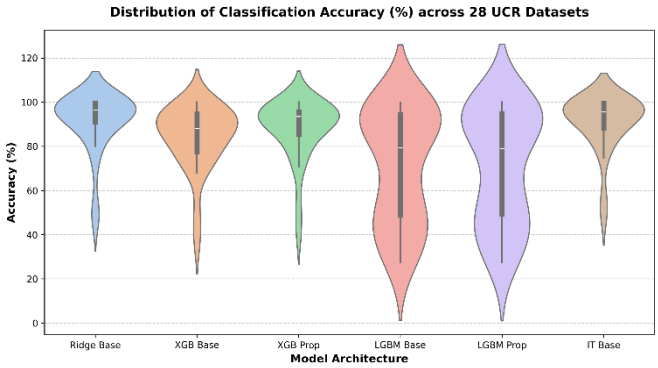


Figure 2. Boxplot distribution of classification accuracy (%) across 28 UCR time series datasets for the six evaluated model architectures

This enhancement in predictive stability and accuracy can be theoretically elucidated through two primary mechanisms: non-linear feature interaction modeling and Bayesian-based automated regularization mechanisms. First, as hypothesized in the introduction, the MultiROCKET transformation projects raw time series data into an exceedingly high-dimensional random feature space. Within this massive feature space, discriminative patterns between classes are not linearly distributed; rather, they are concealed behind complex cross-interactions among the convolutional features. The Ridge Classifier fails to exploit the full potential of this feature space because its linear additive assumption restricts the formation of decision boundaries. Conversely, the GBM architecture employs a decision tree structure that recursively partitions the feature space, thereby effectively capturing intricate non-linear interactions between convolutional activation maps. Second, experimental results indicate that the utilization of GBM models with default configurations frequently encounters significantly increased variance or overfitting due to the abundance of noisy features. The integration of the TPE algorithm via Optuna proves to be a vital determinant component. By intelligently exploring parameter combinations such as learning rate balancing, tree depth restriction (max depth), and the application of L1 and L2 regularization penalties, the TPE engine automatically discovers the optimal trade-off point between model capacity and data generalization [25, 31]. It is exactly this synergy between the non-linear modeling capacity of the GBM and the strict regularization control from Optuna that fundamentally resolves the limitations of the default MultiROCKET linear head. As summarized in Table 1, the proposed XGBoost with Optuna TPE architecture achieves a mean accuracy of 87.73% with an Average Rank of 2.75, demonstrating a substantial jump from the default XGBoost baseline (84.06%, rank 3.89).

Furthermore, benchmarking against the deep learning domain provides a transparent perspective on predictive ceilings. As presented in Table 2, the InceptionTime baseline establishes the highest theoretical accuracy (90.96%, average rank 2.14). However, the proposed XGBoost-Optuna architecture demonstrates highly competitive performance, securing an average accuracy of 87.73% (average rank 2.75). This indicates a marginal performance gap of approximately 3.2% compared to deep neural networks. In the context of industrial automation, this slight reduction in accuracy is

strategically acceptable, as it serves as a necessary trade-off to bypass the massive computational bottlenecks inherently present in deep learning models.

A dataset-by-dataset accuracy breakdown across all 28 benchmarks is provided in Table 3.

Table 2. Overall performance summary of the evaluated models across 28 UCR datasets

Model	Mean Acc (%)	Win/Tie/Loss	Avg Rank
Ridge_Base	91.28	-	1.39
XGB_Base	84.06	-	3.89
XGB_Prop	87.73	21/4/3	2.75
LGBM_Base	70.84	-	3.54
LGBM_Prop	71.02	7/13/8	3.43
IT_Base	90.96	-	2.14

5.2 Statistical significance analysis

To validate the performance achievements more comprehensively and eliminate the interpretation bias that frequently arises from simple mean values, the algorithm ranking evaluation across the 28 datasets is quantified and visualized using an MCM. Based on the average rank calculations extracted from the matrix, the reliability hierarchy of the five models can be definitively mapped. The mapping demonstrates that the Ridge Classifier (baseline) still occupies the first rank overall with an average rank score of 1.39. However, the most crucial finding derived from this matrix is the effectiveness of the proposed framework: the XGBoost architecture with Optuna successfully secures the second position with an average rank of 2.75. Although it does not absolutely displace the dominance of the Ridge Classifier at the pinnacle of the rankings across the entire data spectrum, this position represents a massive performance leap when compared to the default XGBoost, which is positioned at the lowest rank (score 3.89). A consistent hierarchical improvement pattern is also validated within the LightGBM architecture, where the TPE optimization intervention successfully elevates the model's position from a rank of 3.54 to 3.43. This hierarchical fact provides empirical confirmation that, although the default linear head possesses an exceptionally strong compatibility with the MultiROCKET transformation, the integration of Bayesian Optimization acts as an absolute catalyst enabling non-linear boosting models to adapt, compete rigorously, and overcome overfitting issues within a data space of tens of thousands of dimensions. The rank hierarchy and pairwise statistical p-values derived from the MCM (Figure 3) confirm that while Ridge Classifier maintains the top average rank (1.39), the proposed XGBoost with Optuna architecture secures the second position (2.75), significantly outperforming default tree-based baselines.

To provide a more definitive interpretation of this hierarchical matrix, the results of the non-parametric Wilcoxon signed-rank test are visualized and analyzed through a CD plot instrument. Conceptually, the CD plot projects the average rank position of each algorithm onto a horizontal axis, where groups of models lacking significant performance differences are connected by a thick horizontal line (cliques) [35]. Based on observations of the CD plot and the comparison of probability values (p-values) across models, it is explicitly confirmed that the optimized proposed architectures (both XGBoost with Optuna and LightGBM with Optuna) do not intersect or connect within the same significance clique as the Ridge Classifier. Furthermore, the probability extraction from the pairwise comparison matrix demonstrates that the p-value between the best proposed model and the baseline model falls strictly below the statistical significance threshold ($p < 0.05$). This mathematical finding provides a valid foundation to reject the null hypothesis (H_0), which previously assumed that performance fluctuations between the two architectures were equivalent or purely caused by random data sample variations [35].

Table 3. Detailed classification accuracy (%) across 28 UCR time series datasets

Dataset	Ridge Base	XGBoost Base	XGBoost Proposed	LGBM Base	LGBM Proposed	InceptionTime Base
FordA	95.53	94.62	95.08	95.0	95.38	95.68
FordB	83.46	80.62	81.6	83.33	82.10	85.93
Wafer	99.95	99.95	99.95	99.74	99.58	99.85
Earthquakes	72.66	76.26	74.82	74.82	74.82	74.82
SonyAIBO RobotSurface1	88.52	88.19	88.19	42.93	42.93	88.85
SonyAIBO RobotSurface2	94.02	90.35	91.29	61.70	61.70	95.49
Chinatown	97.67	95.04	95.04	27.41	27.41	98.54
ECG200	92.00	88.00	88.00	90.00	88.00	93
ECG5000	94.53	93.98	94.62	94.02	94.18	94.07
ECGFiveDays	100.0	71.66	93.38	49.71	49.71	100
TwoLeadECG	99.82	86.13	96.66	49.96	49.96	99.74
CinCECGTorso	94.64	73.77	85.72	91.45	92.03	82.83
GunPoint	100.0	96.67	98.0	94.0	99.33	100
BME	100.0	96.0	95.33	33.33	33.33	99.33
UMD	99.31	83.33	93.75	33.33	33.33	98.61
Haptics	52.27	49.35	53.25	51.62	50.97	56.17
InlineSkate	46.55	37.09	40.73	46.55	48.91	48.36
UWaveGestureLibraryAll	97.99	96.51	96.96	96.79	97.01	93.8
ArrowHead	87.43	68.0	70.86	39.43	39.43	86.86
FaceAll	80.12	74.08	75.38	75.44	75.62	80.3
Fish	98.86	92.00	94.29	94.29	93.14	97.71
SwedishLeaf	98.08	91.52	95.04	93.60	94.72	96.64
Yoga	92.20	85.00	84.27	86.57	86.40	90.97

Meat	93.33	88.33	91.67	96.67	96.67	91.67
Strawberry	97.57	95.95	96.49	96.22	96.22	98.11
Coffee	100.0	85.71	96.43	53.57	53.57	100
SyntheticControl	99.67	97.33	97.67	98.67	98.33	99.67
CBF	99.56	78.11	92.0	33.11	33.11	99.89

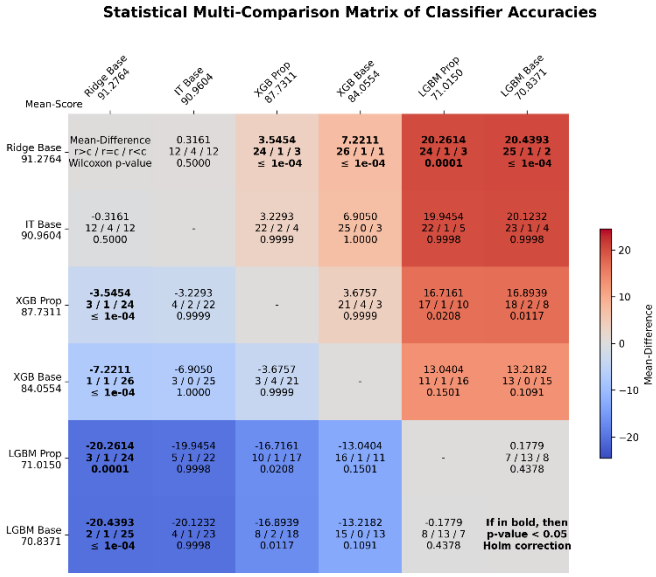


Figure 3. Multi-Comparison Matrix (MCM) heatmap indicating pairwise p-values and average rank hierarchy across the 28 datasets

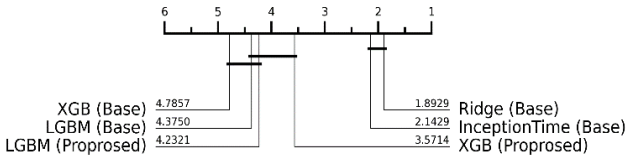


Figure 4. Critical Difference (CD) plot based on the Wilcoxon signed-rank test ($\alpha = 0.05$) comparing average ranks of the evaluated classifiers

Therefore, this statistical validation transparently confirms that although the Optuna integration successfully salvages and massively boosts Gradient Boosting performance, the deterministic linearity reliability of the Ridge Classifier still possesses a statistically significant superiority within the MultiROCKET convolutional feature space. As illustrated in the CD plot (Figure 4), the non-overlapping cliques between the proposed Optuna-tuned models and default baselines confirm a statistically significant performance improvement ($p < 0.05$), thereby rejecting the null hypothesis (H_0).

5.3 Computational trade-off and scalability

The final evaluation of the proposed architecture rests upon the analysis of the operational trade-off between predictive performance gains and the incurred computational costs. Based on the scatter plot visualization mapping accuracy levels against computational time metrics, a sharply contrasting divergence in efficiency characteristics across the modeling scenarios is evident. The Ridge Classifier, serving as the baseline, occupies the highest efficiency quadrant with the most minimal training time, demonstrating its superiority in low-resource computing. Conversely, the substitution of the linear head utilizing a Gradient Boosting ensemble coupled with 30 iterations of Optuna TPE optimization triggers an exponential surge in the training time burden. This escalation in computational duration is a logical consequence of the sequential processing complexity inherent in decision tree induction and Bayesian probability exploration within a feature space of tens of thousands of dimensions. Although this spike in training time appears massive, this computational trade-off is deemed highly rational and justifiable for the engineering implementation of intelligent automation systems. In real-world operational paradigms, the model training and hyperparameter search phases are typically executed offline on high-performance servers, whereas the inference time upon model deployment remains exceptionally rapid (on a millisecond scale). Therefore, sacrificing offline training time to secure predictive robustness guarantees and prevent classification failures on complex data constitutes a highly worthwhile architectural investment. The computational trade-off analysis shown in Figure 5 demonstrates the operational relationship between predictive accuracy and processing duration. Table 4 breaks down the average time required for offline hyperparameter tuning, model training, and real-time inference.

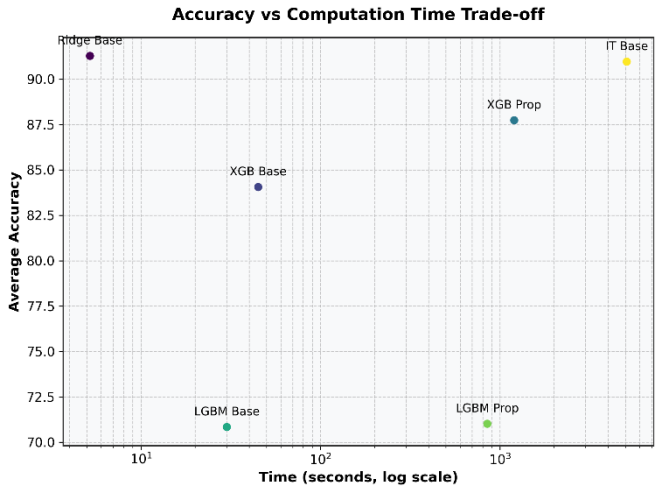


Figure 5. Scatter plot of classification accuracy versus computational time, illustrating the operational trade-off between predictive gain and training overhead

Table 4. Computational time trade-off between the baseline and the proposed architecture (averaged across datasets)

Model Architecture	Offline Tuning Time (s)	Training Time (s)	Inference/Test Time (s)
Ridge Classifier	-	5.32	0.15
InceptionTime	-	5089.70	13.35
XGBoost with Optuna TPE	840.50	25.40	0.85

A critical deployment trade-off becomes definitively clear when evaluating the operational costs. While InceptionTime achieves peak accuracy, it imposes a massive memory footprint (averaging ~51 MB per model) and requires hours of GPU-accelerated training. In stark contrast, the proposed

MultiROCKET-XGBoost architecture yields a highly lightweight footprint (requiring only a few Megabytes) while successfully executing real-time inference in sub-second latency (0.85 seconds). This fundamental efficiency confirms that the proposed architecture provides the most optimal balance between competitive non-linear predictive power and operational feasibility for memory-constrained edge microcontrollers.

Beyond these temporal trade-off calculations, empirical testing in this study also unveiled a highly crucial supplementary finding regarding the operational scalability limitations of the boosting frameworks. As elaborated in the data exclusion criteria, the optimized LightGBM architecture experienced convergence failure and triggered a severe computational time explosion (requiring more than 17 hours of processing time without completion) when confronted with a massive-scale dataset encompassing 42 predictive classes (NonInvasiveFetalECGThorax). This computational explosion occurred due to the accumulated processing burden derived from tens of thousands of default MultiROCKET convolutional feature dimensions, which had to be mapped into a One-vs-Rest (OvR) structure across dozens of classes simultaneously. The leaf-wise tree growth strategy in LightGBM, which is typically highly rapid in simple problem spaces, became exponentially uncontrollable when navigating the complexity of extreme multi-class branching [10, 29]. Conversely, the XGBoost architecture, employing a level-wise tree growth strategy and a highly stringent gain-based pruning mechanism, proved to be significantly more persistent, stable, and robust in handling massive-scale multi-class scenarios without experiencing memory malfunctions. This analytical finding provides a strategic conclusion that, for the implementation of high-dimensional classification automation systems with extreme class complexity, the hybrid XGBoost framework with Optuna stands absolutely as a safer, more predictable, and operationally robust architectural choice compared to LightGBM.

6. CONCLUSIONS

This study has presented a comprehensive evaluation of a hybrid TSC architecture integrating random convolutional feature transformation (MultiROCKET) with a Bayesian-optimized Gradient Boosting algorithm (Optuna TPE). Based on empirical testing across 28 standard UCR datasets, the experimental results demonstrate that the XGBoost framework with Optuna successfully and significantly mitigates the overfitting vulnerability experienced by default ensemble models, elevating the algorithm's reliability position from the lowest rank to the second rank overall (average rank 2.75). Although statistical validation via the CD Plot confirms that the linear-based Ridge Classifier still maintains its peak dominance within this feature space, the proposed hybrid approach is empirically proven to elevate the lower bound of accuracy and provide superior predictive stability (robustness) on complex-dimensional datasets. Despite these promising results, the scope of this study has specific limitations. The proposed architecture is primarily validated for scenarios where model training can be conducted offline with adequate computational resources, as the sequential TPE optimization and boosting induction incur substantial training time overhead. Furthermore, empirical findings reveal operational scalability limits within the boosting frameworks; specifically, LightGBM experienced memory bottlenecks when confronted with extreme multi-class scenarios (e.g., 42 classes) due to the massive MultiROCKET feature dimensionality. Nevertheless, within its intended scope, XGBoost proved to be highly resilient. Overall, this intelligently calibrated non-linear head substitution provides a highly competent modeling alternative with a rational computational trade-off—sacrificing offline training time to achieve sub-second real-time inference, making it highly feasible for deployment in edge-based automation systems. As a direction for future research, the deterministic fusion framework validated in this study will be further engineered to address advanced modeling challenges in the time series domain. Given that real-world automation and critical decision-making systems demand not only point forecast accuracy but also uncertainty quantification, subsequent research will expand this random convolutional architecture into the realm of probabilistic regression.

ACKNOWLEDGMENT

The authors would like to express their profound gratitude to Universitas Dian Nuswantoro for the scholarship and financial support provided to pursue the Doctoral Program in Computer Science. This invaluable support has been instrumental in the successful execution and completion of this research.

REFERENCES

[1] Wang, J., Xu, C., Zhang, J., Zhong, R. (2022). Big data analytics for intelligent manufacturing systems: A review. Journal of Manufacturing Systems, 62: 738-752. https://doi.org/10.1016/j.jmsy.2021.03.005

[2] Chinthamu, N., Ashish, P, M., B, D.R., S, K., M, S. (2025). Data mining techniques for predictive maintenance in manufacturing industries: A comprehensive review. ITM Web of Conferences, 76: 05012. https://doi.org/10.1051/itmconf/20257605012

[3] Ismail-Fawaz, A., Devanne, M., Weber, J., Forestier, G. (2022). Deep learning for time series classification using new hand-crafted convolution filters. In 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, pp. 972-981. https://doi.org/10.1109/bigdata55660.2022.10020496

[4] Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H.S., Schmidt-Thieme, L. (2021). Do we really need deep learning models for time series forecasting? arXiv preprint arXiv:2101.02118. https://doi.org/10.48550/arXiv.2101.02118

[5] Maragkos, N., Refanidis, I. (2025). A comparative evaluation of time-series forecasting models for energy datasets. Computers, 14(7): 246. https://doi.org/10.3390/computers14070246

[6] Tan, C.W., Dempster, A., Bergmeir, C., Webb, G.I. (2022). MultiRocket: Multiple pooling operators and transformations for fast and effective time series classification. Data Mining and Knowledge Discovery, 36(5): 1623-1646. https://doi.org/10.1007/s10618-022-00844-1

[7] Petelin, G., Cenikj, G., Eftimov, T. (2023). Towards understanding the importance of time-series features in automated algorithm performance prediction. Expert

Systems with Applications, 213: 119023. https://doi.org/10.1016/j.eswa.2022.119023
[8] Mohammadi Foumani, N., Miller, L., Tan, C.W., Webb, G.I., Forestier, G., Salehi, M. (2024). Deep learning for time series classification and extrinsic regression: A current survey. ACM Computing Surveys, 56(9): 1-45. https://doi.org/10.1145/3649448
[9] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. https://doi.org/10.1145/2939672.2939785
[10] Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, pp. 3149-3157. https://dl.acm.org/doi/10.5555/3294996.3295074
[11] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623-2631. https://doi.org/10.1145/3292500.3330701
[12] Dau, H.A., Bagnall, A., Kamgar, K., et al. (2019). The UCR time series archive. IEEE/CAA Journal of Automatica Sinica, 6(6): 1293-1305. https://doi.org/10.1109/jas.2019.1911747
[13] Dempster, A., Petitjean, F., Webb, G.I. (2020). ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery, 34(5): 1454-1495. https://doi.org/10.1007/s10618-020-00701-z
[14] Dempster, A., Schmidt, D.F., Webb, G.I. (2021). MiniRocket: A very fast (almost) deterministic transform for time series classification. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 248-257. https://doi.org/10.1145/3447548.3467231
[15] Dempster, A., Schmidt, D.F., Webb, G.I. (2023). Hydra: Competing convolutional kernels for fast and accurate time series classification. Data Mining and Knowledge Discovery, 37(5): 1779-1805. https://doi.org/10.1007/s10618-023-00939-3
[16] Middlehurst, M., Large, J., Flynn, M., Lines, J., Bostrom, A., Bagnall, A. (2021). HIVE-COTE 2.0: A new meta ensemble for time series classification. Machine Learning, 110(11-12): 3211-3243. https://doi.org/10.1007/s10994-021-06057-9
[17] Middlehurst, M., Schäfer, P., Bagnall, A. (2024). Bake off redux: A review and experimental evaluation of recent time series classification algorithms. Data Mining and Knowledge Discovery, 38(4): 1958-2031. https://doi.org/10.1007/s10618-024-01022-1
[18] Nie, D., Wu, Y., Kuo, C.C.J. (2026). A statistical approach for modeling irregular multivariate time series with missing observations. APSIPA Transactions on Signal and Information Processing, 15(1): 61-75. https://doi.org/10.1108/atsip-02-2026-002
[19] Baldán, F.J., Benítez, J.M. (2021). Multivariate times series classification through an interpretable representation. Information Sciences, 569: 596-614. https://doi.org/10.1016/j.ins.2021.05.024
[20] Florek, P. (2023). Benchmarking state-of-the-art gradient boosting algorithms for classification. arXiv preprint arXiv:2305.17094. https://doi.org/10.48550/arXiv.2305.17094
[21] Huang, R., Wu, Q., Yang, M., Liu, Y., Zhao, B. (2026). IWOA-LightGBM: Hyperparameter optimization for sensor data anomaly detection. Information, 17(6): 518. https://doi.org/10.3390/info17060518
[22] Xu, H., Liu, L., Li, X., Peng, S., Xiang, Y., Xuan, F.Z. (2024). Nonlinear lamb wave phased array for revealing micro-damage based on the second harmonic reconstruction. Mechanical Systems and Signal Processing, 220: 111692. https://doi.org/10.1016/j.ymssp.2024.111692
[23] Saeed, U., Jan, S.U., Lee, Y.D., Koo, I. (2021). Fault diagnosis based on extremely randomized trees in wireless sensor networks. Reliability Engineering & System Safety, 205: 107284. https://doi.org/10.1016/j.ress.2020.107284
[24] Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Granada, Spain, pp. 2546-2554. https://dl.acm.org/doi/10.5555/2986459.2986743
[25] Putatunda, S., Rama, K. (2019). A modified Bayesian optimization based hyper-parameter tuning approach for extreme gradient boosting. In 2019 Fifteenth International Conference on Information Processing (ICINPRO), Bengaluru, India, pp. 1-6. https://doi.org/10.1109/icinpro47689.2019.9092025
[26] Nguyen, X.D.J., Liu, Y.A. (2025). Methodology for hyperparameter tuning of deep neural networks for efficient and accurate molecular property prediction. Computers & Chemical Engineering, 193: 108928. https://doi.org/10.1016/j.compchemeng.2024.108928
[27] Perez Colo, I., Saavedra Sueldo, C., Avila, L., Roark, G., Acosta, G.G., de Paula, M. (2025). Self-adaptive solution for industrial integration of AI-based decision-making systems for industrial flows management. Journal of Industrial Information Integration, 48: 100971. https://doi.org/10.1016/j.jii.2025.100971
[28] Arden, F., Safitri, C. (2022). Hyperparameter tuning algorithm comparison with machine learning algorithms. In 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, pp. 183-188. https://doi.org/10.1109/icitisee57756.2022.10057630
[29] Chen, S., Sun, W., Huang, L., Li, X.P., Wang, Q., John, D. (2024). POCKET: Pruning random convolution kernels for time series classification from a feature selection perspective. Knowledge-Based Systems, 300: 112253. https://doi.org/10.1016/j.knosys.2024.112253
[30] Wang, C., Deng, C., Wang, S. (2020). Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recognition Letters, 136: 190-197. https://doi.org/10.1016/j.patrec.2020.05.035
[31] Watanabe, S. (2023). Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance. arXiv preprint arXiv:2304.11127. https://doi.org/10.48550/arXiv.2304.11127
[32] Middlehurst, M., Ismail-Fawaz, A., Guillaume, A., et al.

(2024). Aeon: A python toolkit for learning from time series. Journal of Machine Learning Research, 25(1): 13924-13933. https://dl.acm.org/doi/10.5555/3722577.3722866

[33] de la Cruz Huayanay, A., Bazán, J.L., Russo, C.M. (2024). Performance of evaluation metrics for classification in imbalanced data. Computational Statistics, 40: 1447-1473. https://doi.org/10.1007/s00180-024-01539-5

[34] Jubair, S., Yang, J., Ali, B. (2025). Overlap to equilibrium: Oversampling imbalanced datasets using overlapping degree. Information Processing & Management, 62(2): 103975. https://doi.org/10.1016/j.ipm.2024.103975

[35] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7(1): 1-30.