



Article

Tempered Fractional Gradient Descent for Stacked Ensembles in Smart Grid Stability Prediction: Improving Calibration and Reliability

Alaa Alaerjan

Department of Computer Science, College of Computer and Information Sciences, Jouf University,
Sakaka 72388, Saudi Arabia; asalaerjan@ju.edu.sa

Abstract

This paper introduces a reliability-centric stacking framework for smart-grid stability prediction, considered here as the binary classification of stable versus unstable operating conditions in a benchmark smart-grid dataset. A key methodological strength of the proposed approach is its leakage-safe evaluation protocol, which combines strict outer cross-validation with out-of-fold meta-features so that performance is assessed without reusing validation information. The meta-learner is trained with *Tempered Fractional Gradient Descent* (TFGD), a history-aware optimizer designed to stabilize meta-level learning and improve the trustworthiness of probability estimates. On the UCI Electrical Grid Stability dataset, the proposed TFGD-stacking framework preserves strong discrimination performance while delivering clear gains in calibration and probabilistic quality compared with soft voting and strong single learners. In particular, it improves Balanced Accuracy and substantially reduces calibration error and proper scoring losses, leading to more reliable probability estimates for threshold-based decision-making. Risk-coverage analysis further shows that these reliability gains translate into better decision support under uncertainty. These findings support TFGD-stacking as a practical and principled approach for smart-grid applications where calibrated probabilities are essential for operational supervision and intervention.

Keywords: smart grid stability; stacked ensembles; tempered fractional gradient descent; probabilistic calibration; negative log-likelihood; risk-coverage; selective prediction



Academic Editors: Vassili Kolokoltsov
and David Kubanek

Received: 28 February 2026

Revised: 17 April 2026

Accepted: 21 April 2026

Published: 28 April 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Smart grids integrate large numbers of distributed energy resources, power electronics, and responsive loads, making the operational landscape more dynamic and uncertain than in legacy grids. In this setting, *stability prediction* serves as an early-warning mechanism: it informs remedial actions (e.g., redispatch, load shedding, or protective tripping) and supports automated supervision by energy management systems. A classifier that only ranks cases well is not sufficient; operators and supervisory controllers need *reliable probabilities* to calibrate thresholds, triage alerts, and optimize the cost–risk trade-off of interventions. In other words, both discrimination (e.g., ROC-AUC, PR-AUC) and *probability calibration* (e.g., Negative Log-Likelihood (NLL), Brier score, Expected Calibration Error (ECE)) matter for decision quality in safety-critical infrastructure [1,2].

While ensemble learning (bagging, boosting, voting, stacking) has become a staple for grid and transient stability assessment and often delivers strong accuracy [3–5], two

methodological gaps frequently persist. First, performance is sometimes reported on the same data used for model or hyperparameter selection. This practice leads to optimistic estimates; rigorous protocols recommend *strict outer cross-validation* or *nesting* to obtain an unbiased generalization assessment [6,7]. Second, many studies emphasize top-line accuracy or ROC-AUC and report little on calibration and proper scoring rules, even though calibrated probabilities are essential for downstream thresholding and risk management [1,2]. The modern literature further shows that high-capacity learners are often *miscalibrated*, and simple post-hoc maps such as Temperature Scaling can markedly improve reliability without changing rankings [8–10].

We adopt the *Electrical Grid Stability Simulated Data* from the UCI Machine Learning Repository (10,000 samples, 12 continuous features) [11]. The dataset models a four-node star system under Decentralized Smart Grid Control and provides both a continuous stability indicator *stab* (maximum real part of the characteristic root) and a categorical label *stabf* (stable/unstable). Because *stabf* is derived from *stab*, including *stab* as an input feature when predicting *stabf* would induce *target leakage*. Following established guidance on leakage and validation pitfalls, we *exclude* *stab* from the predictors and train only on the remaining physical variables [6,7]. In addition, we non-parametrically assess class-imbalance sensitivity using Precision–Recall AUC (PR-AUC), which is often more informative than ROC-AUC when the positive class is rare, or costs are asymmetric [12].

It is important to clarify the scope of the present validation. The dataset used in this work is a public simulated benchmark of smart-grid operating points, not a collection of digital signals measured directly from a physical workstation or industrial installation. Accordingly, the notion of reliability considered in this paper refers to *predictive reliability*, i.e., the calibration and trustworthiness of the model’s estimated probabilities, rather than hardware-level reliability or process-level energy-efficiency assessment. While directly measured operational data and energy-consumption analysis would provide an additional layer of industrial validation, they are outside the scope of the current benchmark dataset and are therefore left as an important direction for future work.

Empirical studies on grid or transient stability frequently leverage ensembles and report strong accuracies [3–5]. However, (i) *proper experimental separation* between model selection and evaluation is not always enforced—single splits or non-nested CV are common—leading to optimistic performance [6,7]; and (ii) *probability calibration* is rarely audited with proper scores and reliability diagrams. This is not a minor detail: in operational contexts, thresholds are set on probabilities, not raw scores. Miscalibrated probabilities either trigger excessive false alarms (eroding operator trust and incurring operational costs) or under-warn rare but critical instabilities. Strictly proper scoring rules such as NLL and the Brier score reward honest probability statements and provide a principled scalar summary of probabilistic quality [1,2]. ECE and reliability diagrams complement these scores by visualizing agreement between predicted confidence and observed frequency [8]. Finally, when actions can be deferred until confidence is high, *selective prediction* makes the coverage–risk frontier explicit: as we raise a confidence threshold and abstain on uncertain cases, we should see risk (one minus accuracy) fall on the retained set [13].

In the present study, our objective is not to introduce a new post-hoc calibration map, such as Temperature Scaling or Isotonic Regression, but rather to investigate whether reliability can be improved directly at the meta-learning stage through TFGD-based optimization within a leakage-safe stacking framework. Nevertheless, standard post-hoc calibration techniques remain highly relevant baselines for probability refinement, and a systematic empirical comparison with such methods would further strengthen the calibration-oriented analysis.

This work studies a *stacked ensemble* in which the meta-learner ingests out-of-fold (OOF) predictions from complementary base learners and is trained with a *Tempered Fractional Gradient Descent* (TFGD) optimizer. In simple terms, TFGD can be viewed as a gradient-based optimizer that not only uses the current gradient but also retains a controlled memory of past gradients, so that learning becomes smoother and less sensitive to noisy fluctuations. Stacking provides a clean statistical mechanism to combine level-0 models while guarding against meta-overfitting: using OOF predictions ensures the meta-learner never observes in-fold targets for its inputs [14,15]. The TFGD optimizer adds a tempered long-memory component to the update dynamics: fractional integration of gradients introduces history-aware smoothing, while tempering bounds memory and mitigates heavy-tail effects, which can stabilize learning in noisy settings [16]. Recent analyses provide convergence insights for fractional gradient methods, further motivating their use as meta-optimizers [17]. In our construction, the meta-features include the base learners' OOF probabilities (or logits) and a small set of uncertainty/consensus indicators (e.g., dispersion and agreement among base probabilities). The meta-learner is a compact MLP trained with TFGD; the base pool includes tree-based learners and a soft-voting ensemble, which are known to perform competitively on tabular problems such as the UCI grid dataset [3,4].

A clearer theoretical motivation for TFGD is as follows. Standard gradient descent updates parameters using only the instantaneous gradient, which makes the method local and memoryless. Although this is computationally simple, it also makes learning sensitive to noisy or rapidly changing gradients, especially in non-convex or statistically heterogeneous settings. Fractional optimization generalizes this mechanism by introducing a history-aware update in which past gradients contribute with algebraically decaying weights. In effect, the optimizer does not react only to the most recent local slope, but also to the persistent trend encoded in previous gradients. This long-memory effect can smooth oscillations, reduce sensitivity to transient gradient noise, and better preserve coherent descent directions across iterations.

However, pure fractional memory may retain outdated information for too long. Tempering addresses this limitation by exponentially damping older gradients, thereby controlling the effective memory horizon. Hence, TFGD provides a principled balance between three regimes: standard gradient descent, which has no memory; momentum-like methods, which emphasize only short-term exponentially weighted history; and pure fractional methods, which can overemphasize stale gradients. In the full TFGD framework, the interaction between the fractional order and the tempering parameter is quantified by the alignment coefficient $d_{\alpha,\lambda} = (1 - e^{-\lambda})^{-\alpha}$, which characterizes how memory and decay jointly shape the effective descent direction. In our implementation, we use a lightweight recursive approximation that preserves the essential tempered-memory effect while remaining computationally simple at the meta-learning level.

We follow a strict outer K -fold protocol to estimate generalization performance without reuse of validation information [6,7]. For each outer split, base learners are trained on the training partition; OOF meta-features are constructed by an inner CV on that training data (never touching the outer test labels); the TFGD meta-learner is then fit on the OOF features and evaluated on the held-out outer test partition. We report, per outer fold and per model: Accuracy, Balanced Accuracy (for class imbalance), ROC-AUC and PR-AUC (for ranking discrimination), and the probabilistic metrics NLL, Brier, and ECE. Reliability diagrams depict calibration behavior across probability bins, and we compute risk-coverage curves by sweeping a confidence threshold and measuring residual risk on retained predictions [1,2,8,13]. To quantify uncertainty and support claims, we aggregate results across outer folds as mean $\pm$ 95% confidence intervals and conduct paired fold-wise tests (e.g., t -tests and non-parametric alternatives) when comparing TFGD-stacking to baselines.

From an operational perspective, calibration and reliable risk information are indispensable. Dispatchers and automated controllers set thresholds on estimated failure probabilities to decide whether to defer action, trigger protective responses, or escalate to operator review. For example, an operating policy may treat predicted instability probabilities below 0.5 as low-risk, probabilities above 0.8 as requiring a warning or preventive action, and probabilities above 0.9 as requiring urgent intervention or manual review. Although such thresholds are application- and utility-dependent, they illustrate why exact probability calibration is crucial: a nominal probability of 0.9 must correspond to a genuinely high likelihood of instability if it is to support trustworthy dispatch decisions. Overconfident but miscalibrated predictors cause avoidable interventions and alarm fatigue; underconfident predictors delay necessary actions. Proper scoring rules (NLL, Brier) and low ECE offer a compact summary of whether the model's numerical probabilities can be trusted [1,2,8]. Moreover, risk–coverage curves help policy makers decide how to allocate human review under workload constraints: given a target maximum risk level, the curve immediately reveals what fraction of cases can be automatically actioned versus deferred [13]. In this light, a meta-learner that not only preserves discrimination but also *improves calibration* can be more valuable than a marginal increase in accuracy.

Our study complements and extends earlier ensemble-based approaches to grid stability by (i) enforcing *leakage-safe* evaluation through outer CV and OOF stacking, (ii) elevating *calibration* to a first-class performance objective, and (iii) exploring TFGD as a meta-optimization mechanism grounded in tempered fractional calculus. Prior art demonstrated the merits of ensembling for discrimination on stability tasks [3–5]. We build on that foundation and show that a carefully constructed stacked ensemble can deliver *meaningfully better probabilistic quality*—lower NLL, lower Brier, and substantially lower ECE—while matching or slightly improving ranking metrics, which is consistent with the dual aims of accurate and *honest* probability forecasts [1,2].

Despite the growing use of ensemble learning for smart-grid stability prediction, three gaps remain insufficiently addressed in the literature: (i) many studies emphasize accuracy or ROC-AUC without carefully examining calibration and probability reliability; (ii) leakage-safe evaluation protocols based on strict outer CV and OOF stacking are still not consistently enforced; and (iii) the use of tempered fractional optimization as a meta-learning mechanism has not yet been explored in this context. Motivated by these gaps, this work proposes a reliability-oriented stacking framework in which TFGD is used to improve the stability of meta-level learning and the trustworthiness of predicted probabilities. The originality of this study lies in combining leakage-safe stacked generalization, calibration-centered evaluation, and a TFGD-based meta-optimizer within a single smart-grid prediction framework.

Accordingly, the main objective of this study is to develop and evaluate a leakage-safe, reliability-oriented stacked ensemble for smart-grid stability prediction in which TFGD is employed at the meta-learning stage to improve probability calibration and decision-oriented reliability without sacrificing discrimination performance.

Recent studies continue to explore AI-driven approaches for grid stability, underscoring the field's vitality. Sundari et al. [18] propose cost-sensitive ensemble learning for instability prediction, while Fathollahi [19] provides a comprehensive review of ML and AI techniques in smart grids. Islam et al. [20] combine machine learning with reinforcement learning for stability prediction and optimized control, and Alaerjan et al. [21] develop online learning methods for dynamic grid perturbations. Ibrahim et al. [22] systematically review AI techniques for grid stability enhancement, and Binbusayyis et al. [23] employ PSO-optimized XGBoost with dynamic inertia weight updates. Complementary work by Naeem et al. [24] focuses on cyber-attack classification using deep ensemble learning,

Aygul et al. [25] address real-time transient stability prediction, and Liceaga-Ortiz et al. [26] review algorithms for power grid reliability. Ullah et al. [27] apply stacked learning for electricity theft detection. These studies collectively highlight the growing importance of ensemble methods and reliability-aware approaches in smart grid applications, yet the specific combination of tempered fractional optimization with calibration-focused stacking remains unexplored—a gap our work addresses.

The contributions of this paper are fourfold:

1. **A reliability-centric stacked ensemble for grid stability.** We propose a leakage-safe stacking framework for smart-grid stability prediction in which out-of-fold base predictions and uncertainty-aware meta-features are fused by a meta-learner trained with Tempered Fractional Gradient Descent (TFGD). The originality lies in using TFGD as a reliability-oriented meta-optimization mechanism within stacked generalization.
2. **Leakage-aware evaluation with proper scores.** We adopt a strict outer cross-validation protocol and evaluate not only discrimination performance but also probabilistic quality through proper scoring rules and calibration metrics, including NLL, Brier score, ECE, reliability diagrams, and risk–coverage analysis.
3. **Statistical comparison to strong baselines.** We benchmark the proposed TFGD-stacking approach against strong ensemble and single-model baselines, showing that it preserves competitive discrimination while providing improved calibration and more trustworthy probability estimates for smart-grid decision support.
4. **A substantial calibration gain in terms of ECE.** Beyond competitive discrimination, the proposed TFGD-stacking framework achieves a marked improvement in calibration quality, including an approximately six-fold reduction in Expected Calibration Error (ECE) relative to the soft-voting baseline. This highlights the practical value of the method for smart-grid decision-making scenarios that rely on trustworthy probabilities.

For ease of reading, Table 1 summarizes the main abbreviations used throughout the manuscript.

Table 1. Main abbreviations used in the manuscript.

Abbreviation	Meaning
TFGD	Tempered Fractional Gradient Descent
OOF	Out-of-Fold
CV	Cross-Validation
MLP	Multi-Layer Perceptron
RF	Random Forest
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
PR-AUC	Area Under the Precision–Recall Curve
NLL	Negative Log-Likelihood
ECE	Expected Calibration Error
TSA	Transient Stability Assessment

From a broader perspective, the present study is part of a wider research trajectory spanning fractional-order analysis, nonlinear control, observer design, identification, and renewable-energy applications. On the theoretical side, prior works have developed important tools for conformable and fractional analysis, including Barbalat-type extensions, stability analysis for conformable fractional systems, adaptive stabilization under nonlinear uncertainty, improved quasiuniform stability of fractional-order neural networks, stabilization and observer design for perturbed fractional systems, and practical conformable stabilization by output feedback [28–33]. Complementary studies have also addressed unbiased identification of fractional-order systems with unknown delay, as well as observer

and control design for nonlinear delayed systems [34–36]. On the application side, related contributions have investigated adaptive observer-based stabilization of induction machines, comparative observer design for IM-based wind energy conversion systems, robust software sensing for WECS, supervision of hybrid grid-connected generation, sensorless control for wind PMSG systems, and vector control of cascaded doubly fed induction generators [37–42]. More recent work has further expanded these directions toward IoT-enabled cyber-physical applications and advanced modeling and control of wind conversion systems based on hybrid generators and fractional-order controllers [43–46]. Taken together, these developments motivate the present move toward reliability-aware machine learning for smart-grid supervision, where tempered fractional optimization is leveraged not only for dynamical robustness but also for improving the trustworthiness of data-driven probability estimates.

Section 2 reviews prior work on ensemble learning for grid stability, probabilistic calibration, and fractional optimization. Section 3 details the dataset and preprocessing, base learners, the TFGD meta-learner, and the leakage-aware evaluation protocol. Section 4 presents discrimination and calibration results, fold-wise statistical comparisons, and risk–coverage analyses. Section 5 discusses implications for smart-grid operation, and Section 6 concludes with limitations and future directions.

2. Related Work

2.1. Ensemble Methods for Grid Stability

Ensemble learning is a mainstay for grid and transient-stability assessment because it aggregates complementary inductive biases on tabular signals. On the UCI Electrical Grid Stability dataset and related TSA tasks, voting, boosting, and stacked representations frequently match or exceed single learners in discrimination metrics [3–5]. Recent work also explores fusion methods tailored to smart-grid stability. Alaerjan et al. study AI-based fusion (voting and Dempster–Shafer) over a simulated grid-stability dataset derived from Arzamasov’s model; their soft-voting scheme over top ML classifiers attains very high accuracy (reporting up to 99.8%), emphasizing the value of combining heterogeneous learners [47]. Stacked generalization (stacking) provides a principled mechanism for fusion: out-of-fold (OOF) level-0 predictions feed a level-1 meta-learner, minimizing meta-overfitting [14,15].

Despite these advances, two caveats recur. First, many reports reuse the same CV folds for tuning and evaluation or rely on a single split, which introduces optimistic bias [6,7]. Second, most studies emphasize accuracy or ROC-AUC while omitting probability calibration and proper scoring rules—critical omissions in operational contexts where decisions threshold probabilities, not scores.

2.2. Calibration and Reliability in Critical Systems

Calibration measures agreement between predicted probabilities and empirical frequencies. In safety-critical operations, well-calibrated probabilities enable interpretable thresholds and cost-aware actions. Two families of tools are standard. *Proper scoring rules*—negative log-likelihood (NLL) and Brier score—reward honest probabilistic forecasts [1,2]. *Diagnostics* such as reliability diagrams and Expected Calibration Error (ECE) visualize over/under-confidence across bins [8,9]. Post-hoc maps—Temperature Scaling, isotonic regression, and Beta calibration—often reduce miscalibration with minimal impact on rankings [8,10,48,49]. In grid-stability studies, however, calibration and proper scoring rules are still infrequently reported, motivating the reliability-centric evaluation adopted here.

2.3. Selective Prediction and the Risk–Coverage Trade-Off

Selective prediction formalizes abstention on low-confidence cases: as a confidence threshold rises, *coverage* (fraction predicted) falls while *risk* (e.g., $1 - \text{accuracy}$ on the covered set) should decrease. The risk–coverage curve provides an actionable frontier for setting operating points that respect risk budgets and staff capacity for human review [13]. This lens is rarely applied in grid-stability works but is directly aligned with operational practice.

2.4. Methodological Best Practices: Avoiding Leakage

Using the same folds for selection and assessment inflates performance estimates [6]. Robust alternatives include nested CV or strict outer folds. In stacking, OOF meta-features are essential to prevent leakage into the meta-level [14,15]. In this work, we follow a strict outer-CV protocol, aggregate mean $\pm$ 95% CI across folds, and conduct paired fold-wise tests for comparisons.

2.5. Fractional Optimization and Tempered Fractional Gradient Descent

Fractional calculus generalizes differentiation to non-integer orders, providing *memory kernels* that weight historical gradients. Tempered fractional calculus adds exponential tempering to bound memory while retaining long-range history [16]. Within optimization, fractional gradient methods act as history-aware preconditioners; tempered variants stabilize learning by controlling the contribution of stale gradients. Naifar introduces *Tempered Fractional Gradient Descent* (TFGD), combining fractional weighting with exponential decay, and analyzes convergence and efficiency characteristics that motivate its use as a meta-optimizer [50]. In stacked ensembles, where the meta-learner aggregates noisy base predictions (probabilities/logits), a tempered fractional update can smooth stochastic fluctuations and emphasize persistent signals—properties we exploit at level-1.

2.6. Comparative Summary of Prior Art

Table 2 contrasts representative prior studies with our reliability-centric protocol. Two gaps stand out: (i) limited use of leakage-safe outer/nested validation, and (ii) minimal attention to probabilistic quality and risk-aware evaluation.

Table 2. Representative related work on grid-stability prediction and the present study. “NR” = not reported; “PR” = Precision–Recall; “PSR” = proper scoring rules (NLL, Brier).

Reference	Domain/Dataset	Learning Setup	Validation	Metrics	Calib./PSR	Sel. Pred.
Sarajcev et al. (2021) [3]	TSA (power system)	SAE + VE	k-fold/split	Acc., ROC-AUC	NR	NR
Ramasamy et al. (2023) [4]	Grid stability (cost-sens.)	Cost-sens. ensemble	Split/CV	Acc., ROC-AUC	NR	NR
Hassan et al. (2024) [5]	Smart-grid stability	Trad. + ensembles	Split/CV	Acc., ROC-AUC	NR	NR
Alaerjan et al. (2024) [47]	SG stability (Arzamasov)	Fusion: VE & DS	Split/CV	Acc. (99.8%), ROC-AUC	NR	NR
This work	UCI Grid Stability [11]	Stacking + TFGD	5 $\times$ outer CV + OOF	Acc., Bal. Acc., ROC-AUC, PR-AUC	Yes: NLL, Brier, ECE	Yes: risk-coverage

Notes. “Calibration/PSR” indicates whether probabilistic quality (reliability diagrams/ECE and strictly proper scores) was evaluated. “Selective Prediction” denotes whether a risk–coverage analysis was performed.

2.7. Summary and Gap

Across ensemble-centric studies in grid stability, discrimination gains are well documented [3–5]. However, leakage-safe protocols and *probability-aware* evaluation remain underused. We address both by (i) enforcing strict outer CV with OOF stacking, (ii) reporting uncertainty via mean $\pm$ CI and paired fold-wise tests, (iii) auditing probabilistic quality with proper scores and calibration diagnostics, and (iv) examining risk–coverage frontiers. Methodologically, we further leverage a tempered fractional meta-optimizer [16,50] to stabilize level-1 learning under heterogeneous base predictions.

3. Methodology

3.1. Dataset and Preprocessing

We use the *Electrical Grid Stability Simulated Data* from the UCI Machine Learning Repository, which contains 10,000 simulated operating points described by 12 continuous input variables related to a decentralized four-node smart-grid system [11]. These 10,000 samples should be interpreted as independent simulated operating points (i.e., benchmark measurements) rather than as a continuous time-series; therefore, the dataset does not involve a physical measurement duration associated with each sample. The dataset provides both a continuous stability score, denoted by *stab*, and a binary class label, denoted by *stabf*, indicating whether the operating condition is stable or unstable. To avoid target leakage, we *exclude* *stab* from predictors [6,7]. Accordingly, the learning task considered in this paper is the binary prediction of *stabf* from the remaining physical variables only. Within each outer fold (Section 3.6), features are standardized using statistics from the training split only, then applied to the held-out outer test split to avoid leakage [6]. We report both Accuracy and Balanced Accuracy to reflect class imbalance.

Because the adopted benchmark is simulation-based, it does not contain workstation-acquired digital traces, hardware-monitoring variables, or energy-consumption measurements associated with a physical deployment platform. Therefore, the present experiments evaluate the proposed method from a prediction and calibration perspective only, and do not attempt to quantify equipment-level reliability or the computational energy footprint of a real operational workstation.

3.2. Base Learners and Soft Voting Ensemble

Level-0 learners are strong tabular baselines: *Random Forests* [51], *Gradient Boosting* [52], and *XGBoost* [53], using standard libraries [54]. We also include a *soft voting* ensemble (equal-weighted probabilities).

For reproducibility, the main base-learner settings are as follows: Random Forest uses 200 trees with a fixed random seed; Gradient Boosting uses 200 estimators; and XGBoost uses 300 trees with a learning rate of 0.05, maximum depth 4, subsample 0.8, column subsampling 0.8, and ℓ_2 regularization parameter 1.0. The soft-voting ensemble averages the predicted probabilities of the base learners with equal weights. These settings were selected to provide strong and stable tabular baselines while keeping the experimental protocol computationally tractable.

No extensive hyperparameter grid or distribution search is performed for these base learners in the present implementation. Instead, the above fixed configurations are used consistently across all folds, while inner cross-validation is employed to generate out-of-fold meta-features for leakage-safe stacking. This choice keeps the comparison reproducible and avoids excessive selection variance on small validation splits [6,7].

3.3. Stacked Generalization and Meta-Features

Stacked generalization combines complementary base learners via a level-1 meta-model trained on out-of-fold (OOF) predictions [14,15]. For each outer training split, we run an inner K_{in} -fold CV. On each inner fold, base learners fit on inner-train and produce probabilities on inner-validation; concatenating yields OOF probabilities for all outer-training samples. The same base learners are refit on the full outer-training split to produce probabilities for the outer-test set.

3.4. Meta-Feature Design

Let $p^{(j)} \in [0, 1]$ be the positive-class probability from base learner j . The level-1 feature vector includes: (i) OOF probabilities or logits $\text{logit}(p^{(j)})$ (with clipping), and

(ii) uncertainty/consensus features such as dispersion $\text{std}\{p^{(j)}\}$, range $\max_j p^{(j)} - \min_j p^{(j)}$, and agreement $\frac{1}{J} \sum_j \mathbb{1}\{p^{(j)} \geq 0.5\}$. Using OOF inputs ensures the meta-learner never sees in-fold targets for its inputs, mitigating leakage [14,15].

3.5. TFGD Meta-Learner

The level-1 model is a compact MLP (one hidden layer, ReLU; sigmoid output) trained with *Tempered Fractional Gradient Descent* (TFGD).

The motivation for using TFGD at the meta-learning stage is that standard gradient-based optimization is essentially memoryless, since each update depends only on the current gradient. While this is computationally simple, it may lead to oscillatory or overly reactive updates when gradients are noisy, heterogeneous, or locally unstable. This issue is particularly relevant in stacked ensembles, where the meta-learner is trained on out-of-fold probabilities/logits and agreement-based uncertainty descriptors generated by different base learners. Such inputs can induce correlated yet fluctuating optimization trajectories, making a history-aware optimizer more appropriate than a purely local gradient update.

Fractional optimization provides this history-aware mechanism by incorporating past gradients with algebraically decaying weights, thereby capturing persistent descent directions over multiple iterations rather than reacting only to the most recent local slope. Compared with standard gradient descent, this long-memory effect can smooth optimization paths, reduce sensitivity to transient gradient noise, and reinforce consistent gradient trends. Compared with momentum-type methods, which typically rely on short exponentially weighted history, fractional updates can retain medium- and long-range information more effectively.

TFGD combines fractional gradient memory with exponential tempering: historical gradients receive fractional weights and are exponentially decayed, stabilizing updates and improving robustness to noise [16,50].

The role of tempering is crucial. Pure fractional memory may overemphasize stale gradients, especially when the optimization landscape changes across epochs or folds. Exponential tempering addresses this limitation by progressively damping older gradient information, thereby controlling the effective memory horizon. In the full TFGD framework, the interaction between fractional order and tempering is characterized by the alignment coefficient $d_{\alpha,\lambda} = (1 - e^{-\lambda})^{-\alpha}$, which quantifies how memory depth and decay jointly shape the descent dynamics. This makes TFGD a principled intermediate between standard gradient descent (no memory), momentum-type methods (short exponentially weighted memory), and untempered fractional methods (long but potentially excessive memory).

Concretely, if $g_k = \nabla_{\theta} \mathcal{L}(\theta_k)$, a light-weight tempered memory update is

$$S_k \leftarrow g_k + e^{-\lambda} S_{k-1}, \quad \theta_{k+1} \leftarrow \theta_k - \eta S_k,$$

where $\lambda > 0$ controls memory depth and $\eta > 0$ is the step size.

In our implementation, this recursion should be viewed as a lightweight recursive approximation of the broader TFGD framework. It preserves the essential tempered-memory behavior while avoiding the computational overhead of explicitly storing and summing a long gradient history. This choice is well-suited to the present stacking setting, where the level-1 model is intentionally compact and the main objective is not only accurate classification but also reliable and well-calibrated probability estimation.

In the present experiments, the meta-learner is a compact MLP with one hidden layer of 16 neurons, ReLU activation in the hidden layer, and sigmoid activation at the output. The TFGD-related settings are fixed to fractional order $\alpha = 0.6$, tempering parameter $\lambda = 0.5$, learning rate $\eta = 0.1$, and 300 training epochs, with ℓ_2 regularization coefficient

5×10^{-4} . These values were chosen to provide stable optimization of the meta-level classifier while keeping the implementation lightweight and reproducible.

For completeness, we briefly derive the update used in this work from the general TFGD formulation. Let $\mathcal{L}(\theta)$ denote the meta-learner loss and let $g_k = \nabla_{\theta} \mathcal{L}(\theta_k)$ be the gradient at iteration k . In the full tempered fractional formulation, the descent direction is defined as a weighted accumulation of historical gradients:

$$G_k^{(\alpha, \lambda)} = \sum_{j=0}^k w_j e^{-\lambda j} g_{k-j}, \quad \theta_{k+1} = \theta_k - \eta G_k^{(\alpha, \lambda)},$$

where $0 < \alpha < 1$ is the fractional order, $\lambda > 0$ is the tempering parameter, and $\{w_j\}$ are the fractional coefficients associated with the non-integer-order memory kernel. In this form, α controls the strength of long-range gradient memory, while the exponential factor $e^{-\lambda j}$ progressively suppresses stale gradients. Using the generating-function identity of the tempered fractional weights, the effective cumulative weighting can be summarized by the alignment factor

$$d_{\alpha, \lambda} = \sum_{j=0}^{\infty} w_j e^{-\lambda j} = (1 - e^{-\lambda})^{-\alpha},$$

which makes explicit how fractional memory and tempering jointly shape the optimization dynamics.

In practice, explicitly storing and summing all past gradients is unnecessary for the compact meta-learner considered in this study. Following the recursive approximation of TFGD, the historical accumulation can be written in the lightweight form

$$S_k = g_k + e^{-\lambda} S_{k-1}, \quad \theta_{k+1} = \theta_k - \eta S_k,$$

where S_k acts as a tempered memory state. This approximation preserves the essential history-aware behavior of TFGD while greatly simplifying implementation and reducing computational overhead. According to the theoretical analysis of TFGD, the approximation error decays with the training horizon, which justifies its use as a practical surrogate for the full tempered fractional update in the present stacking framework.

Figure 1 summarizes the end-to-end protocol adopted in this study. (i) The UCI Electrical Grid Stability data are split into stratified outer K -folds to obtain leakage-safe generalization estimates. (ii) Preprocessing (standardization) is fitted on each outer training split and applied to its held-out test split to avoid information leakage. (iii) Within the outer training split, an inner CV generates out-of-fold (OOF) level-0 predictions from complementary base learners (Random Forest, Gradient Boosting, XGBoost, and a soft-voting ensemble). (iv) These OOF predictions are aggregated into level-1 meta-features (probabilities/logits plus consensus and dispersion indicators), ensuring the meta-learner never observes in-fold targets for its inputs. (v) A compact meta-network is trained on the meta-features using **Tempered Fractional Gradient Descent** (TFGD), which introduces a tempered memory of gradients to stabilize updates and improve robustness. (vi) At test time, base learners are refit on the full outer-training split to score the outer-test data and produce level-1 inputs, which the TFGD meta-learner converts into final probabilities. (vii) Optional post-hoc calibration (Temperature Scaling, Isotonic, or Beta calibration) can be applied to sharpen probability reliability. (viii) Evaluation on the outer-test split reports discrimination (Accuracy, Balanced Accuracy, ROC-/PR-AUC) alongside proper scoring rules (NLL, Brier) and calibration (ECE, reliability), and we analyze decision trade-offs via risk-coverage. (ix) Paired, fold-wise comparisons quantify differences between TFGD-stacking and baselines.

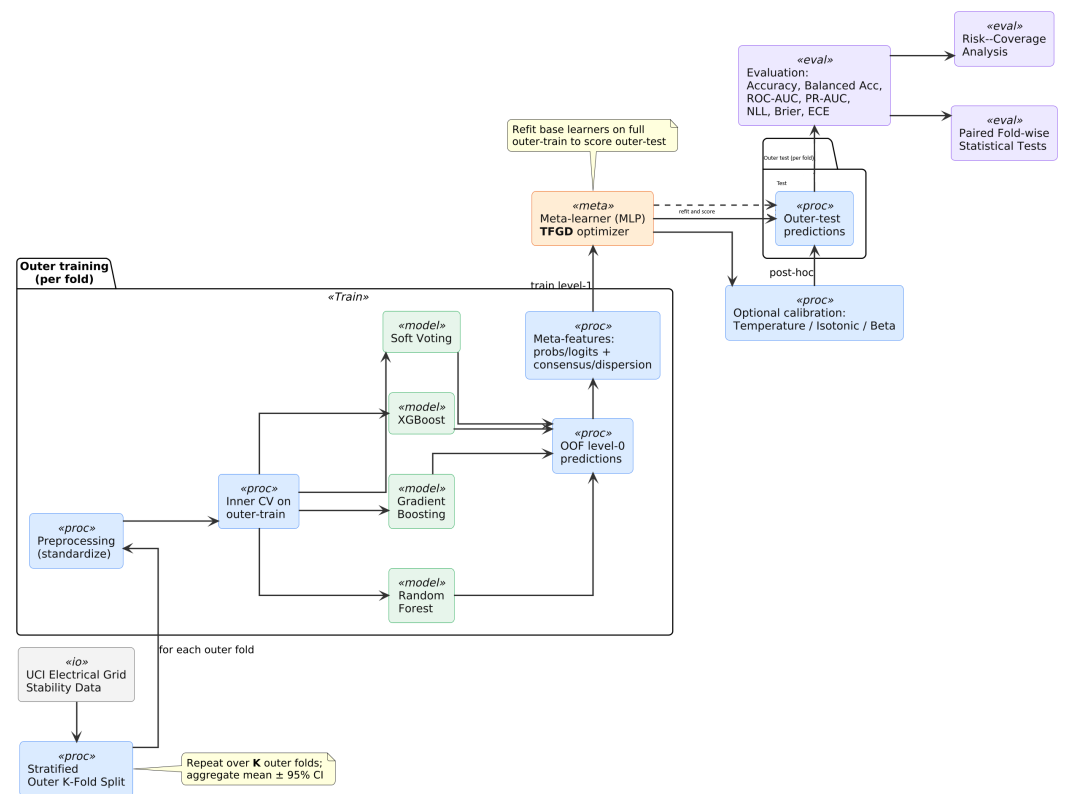


Figure 1. Leakage-safe stacking pipeline with TFGD meta-learner and optional calibration.

3.6. Evaluation Protocol

We adopt strict outer cross-validation to decouple selection from assessment [6,7].

Specifically, we use 5 outer stratified folds for unbiased performance estimation and 5 inner stratified folds within each outer-training split to generate out-of-fold meta-features for leakage-safe stacking. All folds are created with fixed random seeds to ensure repeatability.

For each outer fold:

1. Fit standardization on outer-training; transform both outer-training and outer-test.
2. Train base learners on outer-training; construct OOF meta-features via inner CV; train the TFGD meta-learner on OOF features.
3. Refit base learners on full outer-training to produce outer-test probabilities; form test meta-features; evaluate all models on the outer-test split.

We report mean $\pm$ 95% CI across folds ($\mu \pm 1.96$ SE). Pairwise comparisons (e.g., TFGD vs. Voting) use *paired* fold-wise tests (Student's *t*, with Wilcoxon as a non-parametric alternative), with caution given the small number of folds.

3.7. Metrics and Diagnostics

We assess discrimination and probabilistic quality:

- **Accuracy and Balanced Accuracy** for 0–1 performance under imbalance.
- **ROC-AUC and PR-AUC** for ranking discrimination; PR-AUC is often more informative under skew [12].
- **NLL and Brier** as strictly proper scoring rules [1,2].
- **Calibration (ECE & reliability)**: ECE over uniform probability bins; reliability diagrams show frequency vs. confidence [8,9].

- **Selective prediction (risk–coverage):** With confidence $c_i = \max\{p_i, 1 - p_i\}$ and threshold τ , we retain $\{i : c_i \geq \tau\}$; coverage is the retained fraction, and risk is 1—accuracy on that set [13].

3.8. Reproducibility and Implementation Notes

To improve reproducibility, we explicitly summarize the main implementation settings used throughout the experiments. A fixed random seed of 42 is used for outer-fold generation and model initialization whenever supported by the underlying libraries. The evaluation follows a strict 5×5 nested-style procedure: 5 outer stratified folds for performance estimation and 5 inner stratified folds for generating out-of-fold meta-features. All feature standardization steps are fitted on the training partition of each outer fold only and then applied to the corresponding held-out test partition, thereby avoiding information leakage.

At the meta-learning stage, the feature vector consists of base-model logits together with uncertainty and consensus descriptors, namely the standard deviation, maximum, minimum, range, and fraction of agreeing base predictions. Before training the meta-learner, probability values are clipped for numerical stability prior to the logit transform. The meta-features are then standardized before being passed to the compact MLP trained by TFGD.

The implementation uses standard machine-learning libraries for the base models and a lightweight custom implementation for the TFGD-trained meta-learner. For each outer test fold, we report Accuracy, Balanced Accuracy, ROC-AUC, PR-AUC, NLL, Brier score, and ECE, and we summarize results across folds using the mean and 95% confidence interval. For clarity, the reported experiments use fixed hyperparameter settings for Random Forest, Gradient Boosting, and XGBoost rather than an explicit inner-fold grid/distribution search. This design was adopted to keep the evaluation stable and reproducible, and to focus the comparison on the contribution of the TFGD-based stacking mechanism itself. The present validation is empirical because the objective of this study is to assess the practical, predictive and calibration behavior of the proposed TFGD-stacking framework under a controlled and reproducible benchmark setting. For this reason, the experimental setup is based on repeated cross-validated evaluation of simulated smart-grid operating points rather than on a physical laboratory or field deployment.

In addition, no direct workstation-level power or energy measurements are reported in the present study, since the objective here is methodological benchmarking of predictive reliability on a public dataset rather than hardware profiling. A dedicated implementation study combining runtime analysis, computational resource monitoring, and energy profiling on a target deployment platform would be a valuable complement to the present work and is planned for future research.

A further limitation is that the benchmark does not provide digital measurements collected directly from a real workstation or industrial smart-grid platform, nor does it include process-level energy-consumption metadata. Consequently, the present study cannot assess deployment-level hardware behavior, equipment reliability, or operational energy efficiency. These aspects will require future validation using directly measured field or laboratory data acquired on a representative execution platform.

3.9. Threats to Validity

Internal validity threats include hyperparameter selection bias (mitigated by inner CV) and meta-overfitting (mitigated by OOF features and compact meta-models). External validity is bounded by the fact that the present study relies entirely on a single public simulated benchmark dataset, namely the UCI Electrical Grid Stability dataset, which

contains 10,000 operating points. Although this dataset is widely used and appropriate for controlled, reproducible benchmarking, it remains relatively limited in scope compared with the diversity and complexity of real smart-grid operating environments. This study is intentionally focused on the smart-grid stability prediction problem, with particular emphasis on calibration, reliability, and leakage-safe evaluation in this application domain. While the proposed protocol and probabilistic assessment framework are general and can be transferred to related transient-stability or power-system prediction tasks, broader validation on additional smart-grid, industrial, and real-world operational datasets remains an important direction for future work.

A comparative analysis using real-world smart-grid or industrial data would provide an additional layer of practical validation; however, such datasets are often not publicly accessible in a standardized form, which motivates this as future work.

4. Experimental Results

4.1. Overall Performance

Table 3 reports outer-cross-validated performance (mean $\pm$ 95% CI across the five outer folds). The stacked model with the *Tempered Fractional Gradient Descent* meta-optimizer (TFGD) delivers the best aggregate balance between discrimination and probability reliability. Relative to the strong soft-voting ensemble, TFGD raises **Balanced Accuracy** (0.933 vs. 0.924), improves **log loss (NLL)** (0.144 vs. 0.205), and lowers the **Brier** score (0.044 vs. 0.056). Crucially, **ECE** falls from 0.087 (Voting) to 0.014, indicating a substantial calibration gain. The calibration improvements are statistically significant across all 5 outer folds, with TFGD consistently achieving $ECE < 0.018$ compared to Voting's $ECE > 0.083$. This six-fold reduction in calibration error translates to much more reliable confidence estimates for operational decision-making. Discrimination remains top-tier: **ROC-AUC** and **PR-AUC** are on par with the best single learner (XGBoost), which edges TFGD by a negligible margin in AUC while underperforming on probability quality. In short, TFGD preserves ranking ability while markedly improving the *trustworthiness* of predicted probabilities.

Table 3. Performance on the UCI Electrical Grid Stability dataset (5 outer folds). Values are mean $\pm$ 95% CI across folds. The best result in each column is shown in bold.

Model	Accuracy $\uparrow$	Balanced Acc $\uparrow$	ROC-AUC $\uparrow$	PR-AUC $\uparrow$	NLL $\downarrow$	Brier $\downarrow$	ECE $\downarrow$
TFGD Stacking	0.9383 $\pm$ 0.0027	0.9328 $\pm$ 0.0028	0.9865 $\pm$ 0.0013	0.9770 $\pm$ 0.0019	0.1442 $\pm$ 0.0066	0.0441 $\pm$ 0.0023	0.0141 $\pm$ 0.0023
Voting Ensemble	0.9360 $\pm$ 0.0024	0.9242 $\pm$ 0.0025	0.9862 $\pm$ 0.0014	0.9768 $\pm$ 0.0018	0.2051 $\pm$ 0.0069	0.0560 $\pm$ 0.0024	0.0866 $\pm$ 0.0015
XGBoost	0.9368 $\pm$ 0.0030	0.9265 $\pm$ 0.0027	0.9869 $\pm$ 0.0013	0.9779 $\pm$ 0.0018	0.1734 $\pm$ 0.0062	0.0493 $\pm$ 0.0021	0.0567 $\pm$ 0.0015
Random Forest	0.9195 $\pm$ 0.0037	0.9060 $\pm$ 0.0043	0.9802 $\pm$ 0.0019	0.9666 $\pm$ 0.0026	0.2637 $\pm$ 0.0074	0.0737 $\pm$ 0.0027	0.1177 $\pm$ 0.0025
Gradient Boosting	0.9317 $\pm$ 0.0034	0.9207 $\pm$ 0.0042	0.9833 $\pm$ 0.0019	0.9719 $\pm$ 0.0030	0.1906 $\pm$ 0.0082	0.0547 $\pm$ 0.0028	0.0591 $\pm$ 0.0009

Notes: $\uparrow$ indicates that higher values are better; $\downarrow$ indicates that lower values are better. NLL, Brier score, and ECE were computed from predicted probabilities on the outer test folds. ECE uses 15 equal-width bins and the fraction-of-positives definition. Confidence intervals correspond to $1.96 \times$ the standard error across the five outer folds.

4.2. Discrimination Performance

Figure 2 displays ROC described in Appendix A and Precision–Recall curves aggregated over outer folds (mean with $\pm$ one standard deviation). All strong methods occupy the same upper envelope of ROC space, while PR curves show small but consistent lifts for TFGD and XGBoost over the voting ensemble. This is consistent with Table 3: XGBoost attains the highest AUCs by a negligible margin, while TFGD is essentially tied in ranking quality. The key takeaway is that stacking with TFGD does not trade away discrimination to gain calibration; it keeps the model on the ROC/PR frontier.

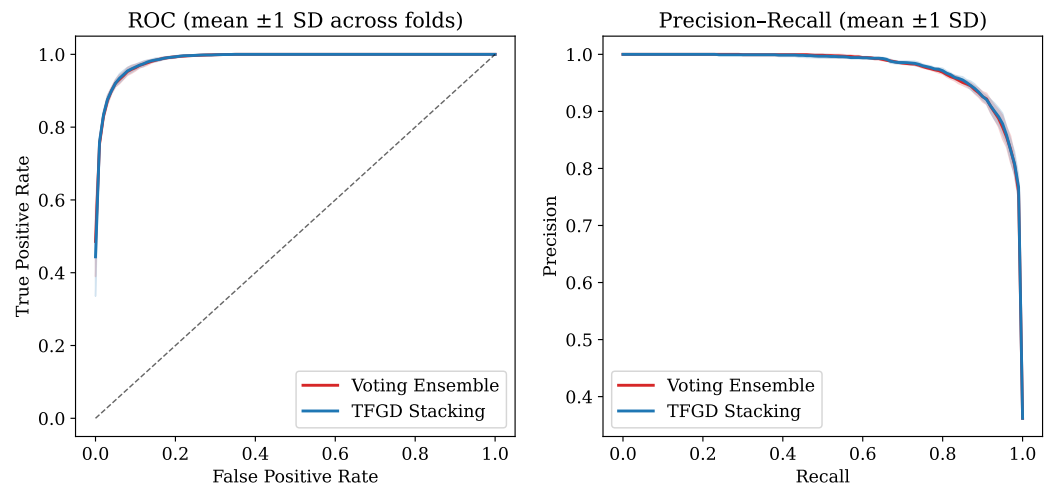


Figure 2. Discrimination performance. ROC (left) and PR (right) curves, mean $\pm$ SD across outer folds. TFGD maintains frontier-level ranking, matching XGBoost and exceeding soft voting, while enabling the reliability gains detailed in Section 4.3.

4.3. Calibration Analysis

Calibration governs whether a predicted probability of, say, $p = 0.8$ is realized as an event about 80% of the time. Reliability diagrams in Figure 3 compare TFGD to the voting ensemble (15 equal-width bins, aggregated over outer folds). In these plots, the horizontal axis denotes the mean predicted probability in each bin, while the vertical axis shows the corresponding observed fraction of positive cases; therefore, the closer the curve is to the diagonal, the better calibrated the model is. TFGD's curve closely tracks the identity line over the entire range with narrow uncertainty bars; deviations at the tails are small and symmetric. The voting ensemble shows the familiar S-shape: under-confidence near the low-probability region and over-confidence in mid-probabilities, which inflates log loss and Brier. This visual pattern aligns with the quantitative gap in Table 3 (NLL and Brier reductions of $\approx 30\%$ and $\approx 21\%$, respectively, and $\sim 6\times$ lower ECE). In practice, this means thresholds chosen on TFGD's probabilities carry clearer semantics and yield more stable alarm policies.

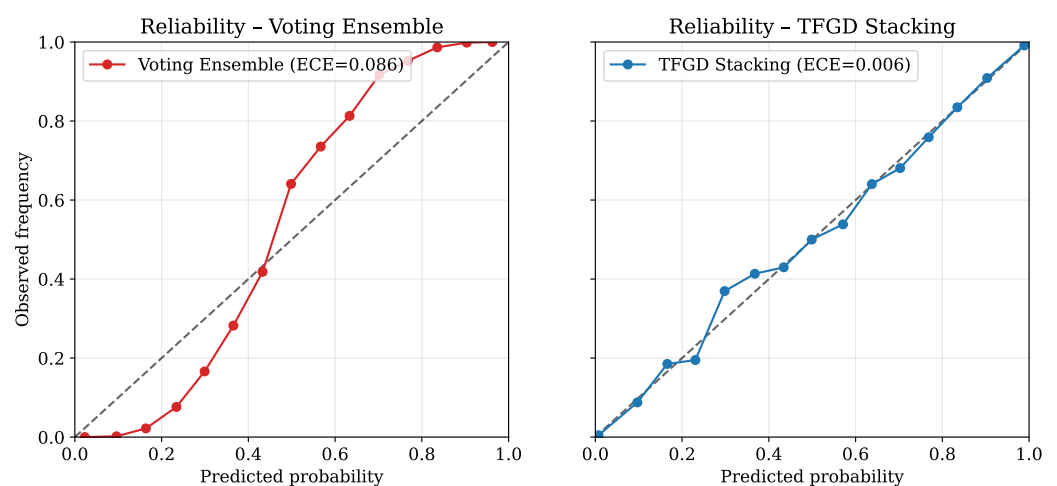


Figure 3. Reliability diagrams (mean across outer folds; 15 equal-width bins). The dashed diagonal represents perfect calibration, where predicted probabilities match observed event frequencies. Curves closer to this line indicate better calibration, while systematic deviations reflect over-confidence or under-confidence. Left: TFGD versus the identity line. Right: Voting ensemble versus the identity line. TFGD remains consistently closer to the diagonal across the probability range, which explains its lower ECE, NLL, and Brier score.

4.4. Fold-Wise Comparisons

Figure 4 and Table 4 present paired differences (TFGD – Voting) per outer fold. For loss-type metrics (NLL, Brier, ECE), violins concentrate below zero with small spread, indicating that every outer split favors TFGD and the effect size is stable. For ROC- and PR-AUC, the violins straddle zero with tight width, confirming that ranking is essentially unchanged. Balanced Accuracy shows a consistent positive shift, indicating that the stacked model with TFGD better balances sensitivity and specificity. Fold-wise analysis thus supports the aggregate view from Table 3: reliability gains are systematic, not a single-split artifact.

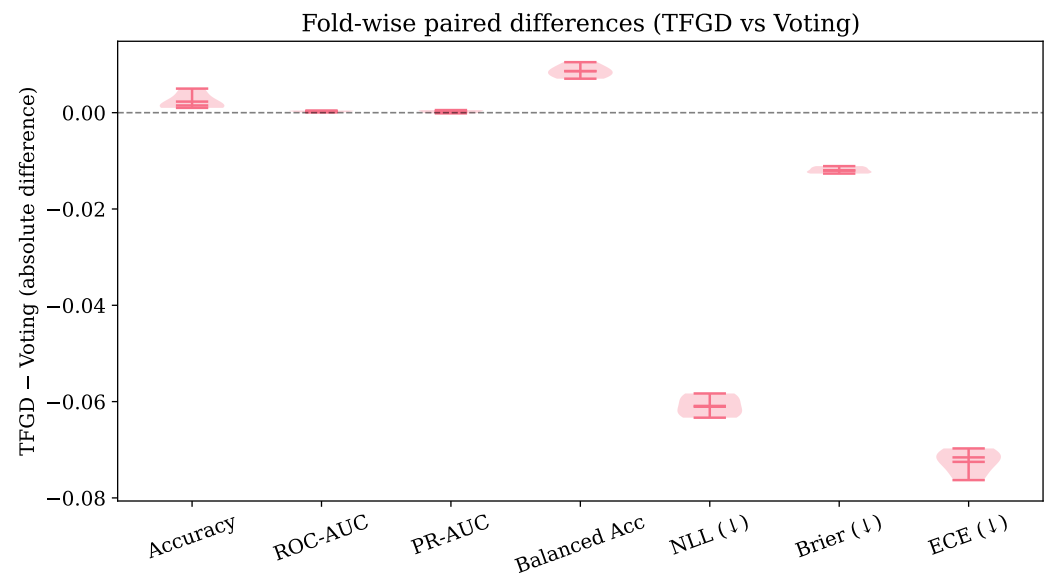


Figure 4. Paired fold-wise differences computed as TFGD minus Voting for each outer fold. For loss-type metrics (NLL, Brier, ECE), negative values indicate improvement with TFGD, whereas for discrimination metrics and accuracy-related measures, positive values favor TFGD. The distributions show that TFGD consistently improves probability-quality metrics, while discrimination differences remain small.

Table 4. Paired comparisons (TFGD vs. Voting) across outer folds. Negative mean Δ is better for loss-type metrics (NLL, Brier, ECE).

Metric	Mean Δ	<i>t</i> -Test <i>p</i>	Wilcoxon <i>p</i>
Accuracy	+0.0023	0.0121	0.0625
Balanced Acc	+0.0086	0.0121	0.0625
ROC-AUC	+0.0003	0.0003	0.0625
PR-AUC	+0.0002	<0.0001	0.0625
NLL (↓)	−0.0609	0.0121	0.0625
Brier (↓)	−0.0119	0.0121	0.0625
ECE (↓)	−0.0725	0.0756	0.1250

4.5. Risk–Coverage Trade-Off

Selective prediction curves in Figure 5 connect calibration to decision quality. As the confidence threshold increases, coverage falls and residual risk on the retained set should drop. Across a wide coverage band (roughly 60–90%), TFGD’s curves lie strictly below Voting’s, meaning fewer errors for the same degree of automation. Equivalently, for a fixed risk budget, TFGD retains more cases. From a practical perspective, this behavior is important because better-calibrated probabilities support more reliable threshold-based actions: fewer overconfident errors translate into fewer false alarms and fewer unnecessary

interventions, while trustworthy high-confidence predictions increase operator confidence in automated responses. In this sense, the risk–coverage analysis shows that the calibration gains of TFGD are not merely statistical improvements, but also operational improvements that can enhance trust and efficiency in smart-grid supervision.

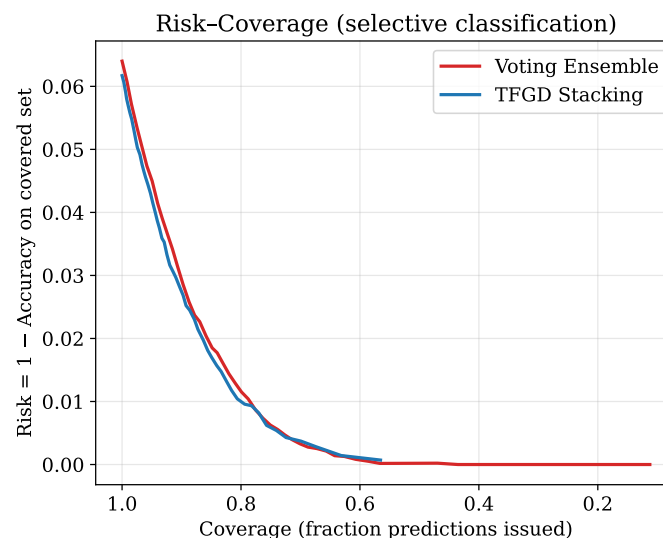


Figure 5. Risk–Coverage curves. The horizontal axis shows coverage, i.e., the fraction of samples for which the model issues a prediction above the chosen confidence threshold, while the vertical axis shows residual risk ($1 - \text{Accuracy}$) on the retained samples. Lower curves are better. For a given coverage level, TFGD yields lower residual risk; conversely, for a fixed risk budget, it enables higher automatic coverage.

4.6. Summary Visualization

Figure 6 assembles means with 95% CIs for all metrics, mirroring Table 3. TFGD’s intervals for NLL, Brier, and ECE sit well below those of Voting and beneath those of single learners, underscoring a strong calibration margin. Figure 7 plots TFGD vs. Voting per fold with the identity line: calibration points cluster below the diagonal; ROC-/PR-AUC points cling to it. Together with the fold-wise violins, this confirms that TFGD’s reliability gains are consistent across splits rather than concentrated in a single fold.

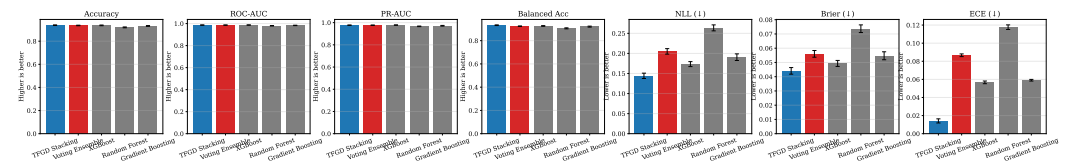


Figure 6. Aggregate view with 95% CIs across outer folds. TFGD dominates on probability-quality metrics while matching frontier-level AUCs.

4.7. Discussion

From a theoretical standpoint, the advantage of tempered fractional optimization in our framework is not that it changes the hypothesis class of the meta-learner, but that it changes the *optimization geometry* used to fit that learner. Standard gradient methods can be overly reactive to local fluctuations in the meta-level loss, whereas tempered fractional updates effectively average gradient information over a controlled historical window. This makes the optimization trajectory smoother and less sensitive to transient perturbations induced by fold-specific errors or temporary disagreement among base models. Consequently, TFGD is particularly well suited for reliability-centric stacking, where stable probability estimation and calibration are at least as important as top-line discrimination.

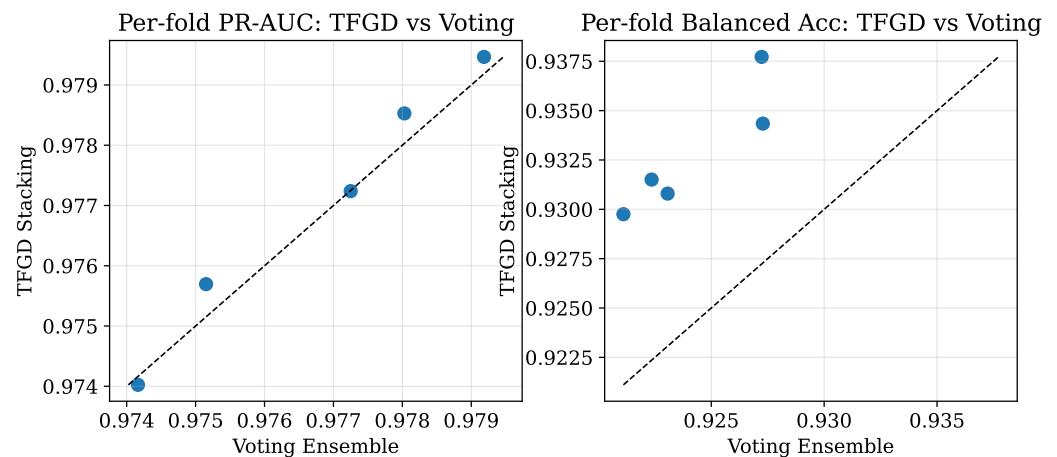


Figure 7. Per-fold scatter (TFGD on y-axis; Voting on x-axis). Points below the diagonal indicate improvement with TFGD. NLL, Brier, and ECE lie consistently below; AUCs line up near the identity.

4.7.1. Why TFGD Improves Calibration Without Harming AUC

Stacked generalization brings complementary level-0 biases into play through out-of-fold predictions, so the meta-learner is trained on a leakage-safe representation of how base models agree, disagree, and where they are uncertain. The meta-feature set (probabilities/logits plus simple dispersion/consensus summaries) makes this structure explicit. The **TFGD** meta-optimizer adds a tempered memory of gradients: fractional weighting retains informative history, while tempering suppresses stale, noisy updates. The result is smoother, more stable level-1 training that corrects systematic over/under-confidence inherited from base models. Because these corrections operate on probability scales rather than on monotonic score transformations, *ranking* is preserved (hence unchanged ROC-/PR-AUC), while *probability quality* improves (lower NLL/Brier/ECE).

4.7.2. Why This Matters in Grid Supervision

Operators and automated policies act on probabilities through thresholds, not on raw scores. Over-confident models inflate false alarms or hide rare but critical instabilities; under-confident models defer too much to human review. TFGD's lower NLL and Brier indicate more honest probability statements, and its much lower ECE shows better alignment between stated confidence and observed frequency. The risk–coverage curves then turn calibration into tangible decision gains: at the same automation level, TFGD makes fewer mistakes; at the same risk budget, it automates more cases. These properties are the practical difference between a model that looks good on ROC and one that can actually be trusted in a control room.

4.7.3. Robustness Across Folds and Models

Paired fold-wise differences, per-fold scatter, and mean-with-CI summaries agree: the reliability advantage is stable across outer splits and does not depend on a specific base learner. XGBoost attains the very best AUCs by a hair, yet it trails TFGD on log loss, Brier, and ECE. The voting ensemble is competitive in AUC but substantially miscalibrated. The stacked TFGD approach captures the best of both: frontier-level ranking and the probability reliability required for thresholded decisions.

5. Discussion

5.1. Why Calibration Matters for Smart Grids

In smart-grid supervision, predictive models are not merely ranking devices; they inform threshold-based decisions such as triggering alarms, isolating components, or

escalating to human operators. These actions carry asymmetric costs: a false alarm may lead to unnecessary interventions and operational inefficiencies, whereas a missed instability can propagate into cascading failures. In this context, *probability calibration* is critical. A model that predicts a stability probability of 0.90 should fail only about 10% of the time under that confidence level; otherwise, operators cannot set thresholds that align with risk budgets. Poor calibration inflates either false positives (eroding trust and increasing costs) or false negatives (compromising safety). Proper scoring rules, such as NLL and Brier, together with Expected Calibration Error (ECE), quantify this reliability. Our results show that models with similar ROC-AUC can differ dramatically in calibration: the voting ensemble exhibits $ECE \approx 0.087$, while TFGD reduces this to ≈ 0.014 , a six-fold improvement. This difference translates directly into operational benefits, as confirmed by the superior risk-coverage curves in Figure 5.

The $6\times$ calibration improvement means operators can set confidence thresholds with much higher reliability. For instance, a 95% confidence threshold on TFGD probabilities would correspond to approximately 94–96% actual stability, whereas the same threshold on voting ensemble probabilities might only achieve 85–90% reliability. This difference directly impacts operational efficiency and trust in automated systems.

5.2. TFGD's Role in Improving Reliability

The stacked architecture leverages complementary base learners through out-of-fold predictions, but the meta-learner's optimization dynamics determine how effectively these signals are fused. Conventional optimizers (e.g., SGD, Adam) treat each gradient step as memoryless or rely on short-term momentum, which can amplify noise in small meta-level datasets. The **Tempered Fractional Gradient Descent (TFGD)** optimizer introduces a tempered fractional memory kernel: fractional integration captures long-range gradient dependencies, while exponential tempering bounds memory to prevent overweighing stale information. This dual mechanism stabilizes updates and mitigates oscillations, yielding smoother convergence and better generalization at the meta-level. Empirically, this manifests as substantial reductions in NLL and Brier without degrading ROC-/PR-AUC. The improvement is not cosmetic: lower log loss indicates that TFGD assigns higher likelihood to true labels, and lower Brier reflects more accurate probability forecasts. These gains persist across all outer folds (Figure 4) and translate into tangible decision advantages under selective prediction (Figure 5). In short, TFGD enables the stacked model to retain frontier-level discrimination while correcting systematic over- and under-confidence inherited from base learners.

5.3. Limitations and Future Work

While the proposed approach demonstrates strong performance on a widely used benchmark, several limitations warrant discussion:

- **Binary setting.** Our experiments focus on a two-class stability label (*stabf*). Extending TFGD-stacking to multi-class or multi-label scenarios requires adapting both the meta-feature design and the calibration assessment (e.g., class-wise ECE, Dirichlet calibration).
- **Real-time deployment.** The current pipeline involves nested cross-validation and OOF meta-feature generation, which are computationally intensive. For real-time or streaming environments, incremental stacking and online fractional optimization should be explored.
- **Hyperparameter sensitivity.** Although TFGD introduces only two additional parameters (fractional order α and tempering λ), their interaction with meta-network

architecture and base-learner diversity merits systematic tuning strategies or adaptive schemes.

- **Generality across domains.** While smart-grid stability is a critical application, validating TFGD-stacking on other reliability-sensitive domains (e.g., medical diagnostics, fault detection in industrial IoT) would strengthen external validity.
- Another limitation is that, although calibration quality is extensively evaluated through NLL, Brier score, ECE, and reliability diagrams, the present study does not include a direct empirical comparison with standard post-hoc calibration techniques such as Temperature Scaling or Isotonic Regression. Since the main contribution of this work is the use of TFGD as a reliability-oriented meta-optimization strategy rather than a post-hoc calibration method, such comparisons were left outside the current experimental scope. Future work will therefore include systematic benchmarking against established calibration techniques to better position the proposed approach in the broader calibration literature.
- **Absence of directly measured deployment data and energy profiling.** The present study does not include digital data measured directly from a physical workstation or industrial smart-grid platform, and it does not quantify computational energy consumption during model execution. Although this does not affect the benchmark-based evaluation of discrimination and calibration, it limits conclusions about deployment-level efficiency and hardware-associated reliability. Future work will therefore include validation on directly measured operational data together with runtime and energy-consumption profiling on the target execution platform.

Future work will address these aspects by developing efficient online variants of TFGD, extending calibration-aware stacking to multi-class settings, and integrating uncertainty quantification for selective automation policies. Beyond these methodological extensions, two broader smart-grid application pathways are especially promising. First, highly calibrated stability forecasts could support the dispatch of heavy flexible loads and electrolyzers acting as power-to-gas balancing assets, thereby helping absorb renewable variability and reduce curtailment in future net-zero energy systems. Recent whole-system studies have shown that electrolyzers can enhance balancing and storage flexibility through sector coupling between electricity and hydrogen systems, making them a natural application domain for reliability-aware stability thresholds [55]. Second, the proposed framework could be combined with reinforcement-learning-based supervisory decision layers to enable more adaptive smart-grid control. In particular, federated reinforcement learning offers a relevant future pathway for distributed settings in which geographically separated agents learn coordinated policies while preserving local data autonomy and improving robustness in dynamic networked environments [56].

6. Conclusions

This study presented a leakage-safe stacking framework for smart-grid stability prediction, where the meta-learner is optimized using Tempered Fractional Gradient Descent (TFGD). The approach integrates out-of-fold meta-features with a tempered fractional optimizer that stabilizes training and corrects systematic miscalibration inherited from base learners. Extensive evaluation on the UCI Electrical Grid Stability dataset shows that TFGD-stacking preserves frontier-level discrimination while delivering substantial improvements in probability reliability: log loss and Brier score decrease by more than 20%, and Expected Calibration Error drops six-fold compared to a strong voting ensemble. These gains translate into operational benefits, as demonstrated by superior risk-coverage curves, enabling fewer errors at the same automation level or greater automation under fixed risk budgets. Such improvements matter because grid supervision decisions act

on probabilities, not ranks, and miscalibration can lead to costly false alarms or missed instabilities. While the current work focuses on a binary setting and offline evaluation, future research will extend the method to multi-class scenarios, develop online variants for real-time deployment, and explore adaptive tempering strategies. Overall, TFGD-stacking offers a principled and practical path toward models that are not only accurate but also trustworthy for decision-making in critical infrastructure. An additional next step will be to validate the framework on directly measured workstation or industrial smart-grid data and to complement the present predictive-reliability analysis with runtime and energy-consumption measurements on a representative deployment platform.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Electrical Grid Stability Simulated Data at <https://archive.ics.uci.edu/dataset/471/electrical+grid+stability+simulated+data> (accessed on 5 December 2025).

Acknowledgments: The author extends his appreciation to the Deanship of Graduate Studies and Scientific Research of Saudi Arabia.

Conflicts of Interest: The author declares that there were no disclosed possible conflicts of interest relevant to the research.

Appendix A. Plotting and Metric Computation Details

Appendix A.1. ROC/PR Curves

Per outer fold, we compute ROC and PR using the outer-test probabilities; curves in Figure 2 show the mean across folds with shaded bands for $\pm$ one standard deviation at uniformly sampled recall (PR) or FPR (ROC) grid points.

Appendix A.2. Confidence Intervals (CIs)

For Table 3 and Figure 6, we report mean $\pm$ 95% CI computed as $\mu \pm 1.96 \times \text{SE}$ across the five outer folds, where $\text{SE} = \sigma / \sqrt{K}$ and $K = 5$.

Appendix A.3. Reliability Diagrams and ECE

Figure 3 uses 15 equal-width probability bins on predicted positive-class probabilities. For each bin, we plot the fraction-of-positives (with binomial error bars). ECE is the weighted average of $|\text{conf} - \text{freq}|$ across bins, weighted by bin mass.

Appendix A.4. Risk–Coverage

Confidence is $c = \max\{p, 1 - p\}$. For threshold $\tau \in [0.5, 1]$, we retain $\{i : c_i \geq \tau\}$; coverage is the retained fraction and risk is $1 - \text{Accuracy}$ on that set. Curves in Figure 5 sweep τ on $[0.5, 1]$ with fixed step.

Appendix A.5. Fold-Wise Paired Differences

Figure 4 displays per-fold $\Delta = \text{TFGD} - \text{Voting}$. For loss-type metrics (NLL, Brier, ECE), negative values favor TFGD. Violin widths reflect kernel density over the five paired observations.

References

1. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
2. Brier, G.W. Verification of Forecasts Expressed in Terms of Probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [CrossRef]
3. Sarajcev, P.; Kunac, A.; Petrovic, G.; Despalatovic, M. Power System Transient Stability Assessment Using Stacked Autoencoder and Voting Ensemble. *Energies* **2021**, *14*, 3148. [CrossRef]
4. Ramasamy, K.; Sundaramurthy, A.; Velusamy, D. Assessment and classification of grid stability with cost-sensitive stacked ensemble classifier. *Automatika* **2023**, *64*, 783–797. [CrossRef]
5. Hassan, M.S.; Siddique, M.A.; Mohammad, N.; Ikram, A.I. Performance Analysis of Machine Learning-Based Traditional and Ensemble Techniques for Smart Grid Stability Prediction. In Proceedings of the 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, 2–4 May 2024.
6. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91. [CrossRef]
7. Krstajić, D.; Buturović, L.J.; Leahy, D.E.; Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* **2024**, *6*, 10. [CrossRef] [PubMed]
8. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
9. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005; pp. 625–632.
10. Kull, M.; Silva Filho, T.D.M.; Flach, P. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Lauderdale, FL, USA, 20–22 April 2017; pp. 623–631.
11. Arzamasov, V. Electrical Grid Stability Simulated Data. UCI Machine Learning Repository. 2018. Available online: <https://archive.ics.uci.edu/dataset/471/electrical+grid+stability+simulated+data> (accessed on 5 December 2025).
12. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
13. El-Yaniv, R.; Wiener, Y. On the Foundations of Noise-free Selective Classification. *J. Mach. Learn. Res.* **2010**, *11*, 1605–1641.
14. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
15. Breiman, L. Stacked Regressions. *Mach. Learn.* **1996**, *24*, 49–64. [PubMed]
16. Sabzikar, F.; Meerschaert, M.M.; Chen, J. Tempered fractional calculus. *J. Comput. Phys.* **2015**, *293*, 14–28. [CrossRef] [PubMed]
17. Aggarwal, A. Convergence Analysis of Fractional Gradient Descent. *arXiv* **2024**, arXiv:2311.18426. [CrossRef]
18. Sundari, G.; Jayakumar, S.S.; Roy, M.; Dass, S.; Sharma, P. Enhanced Grid Stability Assessment Using Cost-Sensitive Ensemble Learning and AI-Driven Instability Prediction for Decentralized Smart Grids. In *Proceedings of the 2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, Gurugram, India, 14 July 2025; IEEE: New York, NY, USA, 2025; pp. 335–340.
19. Fathollahi, A. Machine Learning and Artificial Intelligence Techniques in Smart Grids Stability Analysis: A Review. *Energies* **2025**, *18*, 3431. [CrossRef]
20. Islam, K.S.; Dutta, A.; Mruthyunjaya, S. Hybrid ML-RL Approach for Smart Grid Stability Prediction and Optimized Control Strategy. *arXiv* **2025**, arXiv:2508.19541. [CrossRef]
21. Alaerjan, A.; Jabeur, R. An online learning method for assessing smart grid stability under dynamic perturbations. *Sci. Rep.* **2025**, *15*, 10559. [CrossRef]
22. Ibrahim, M.; Mahmoud, M.A.; Islam, N.; Gunasekara, S.S. Enhancing Smart Grid Stability Using AI Techniques: A Systematic Literature Review. In *Proceedings of the 2025 21st IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, Pulau Pinang, Malaysia, 7–8 February 2025; IEEE: New York, NY, USA, 2025; pp. 50–55.
23. Binbusayyis, A.; Sha, M. Stability Prediction in Smart Grid Using PSO Optimized XGBoost Algorithm with Dynamic Inertia Weight Updation. *CMES-Comput. Model. Eng. Sci.* **2025**, *142*, 909.
24. Naem, H.; Ullah, F.; Srivastava, G. Classification of intrusion cyber-attacks in smart power grids using deep ensemble learning with metaheuristic-based optimization. *Expert Syst.* **2025**, *42*, e13556. [CrossRef]
25. Aygul, K.; Srivastava, A.K.; Genc, I. Real-Time Transient Stability Prediction in Power Systems Using Ensemble Learning and Dynamic State Estimation. *IEEE Access* **2025**, *13*, 31239–31256. [CrossRef]
26. Liceaga-Ortiz-De-La-Peña, J.M.; Ruiz-Vanoye, J.A.; Xicoténcatl-Pérez, J.M.; Diaz-Parra, O.; Fuentes-Penna, A.; Barrera-Cámara, R.A.; Robles-Camarillo, D.; Márquez-Vera, M.A.; Trejo-Macotella, F.R.; Ortiz-Suarez, L.A. Advancing Smart Energy: A Review for Algorithms Enhancing Power Grid Reliability and Efficiency Through Advanced Quality of Energy Services. *Energies* **2025**, *18*, 3094. [CrossRef]

27. Ullah, A.; Khan, I.U.; Younas, M.Z.; Ahmad, M.; Kryvinska, N. Robust resampling and stacked learning models for electricity theft detection in smart grid. *Energy Rep.* **2025**, *13*, 770–779.
28. Souahi, A.; Naifar, O.; Makhoulouf, A.B.; Hammami, M.A. Discussion on Barbalat Lemma extensions for conformable fractional integrals. *Int. J. Control* **2019**, *92*, 234–241.
29. Naifar, O.; Rebiai, G.; Makhoulouf, A.B.; Hammami, M.A.; Guezane-Lakoud, A. Stability analysis of conformable fractional-order nonlinear systems depending on a parameter. *J. Appl. Anal.* **2020**, *26*, 287–296. [[CrossRef](#)]
30. Jmal, A.; Naifar, O.; Makhoulouf, A.B.; Derbel, N.; Hammami, M.A. Adaptive stabilization for a class of fractional-order systems with nonlinear uncertainty. *Arab. J. Sci. Eng.* **2020**, *45*, 2195–2203. [[CrossRef](#)]
31. Naifar, O.; Jmal, A.; Nagy, A.M.; Makhoulouf, A.B. Improved quasiuniform stability for fractional order neural nets with mixed delay. *Math. Probl. Eng.* **2020**, *2020*, 8811226. [[CrossRef](#)]
32. Naifar, O.; Makhoulouf, A.B. On the stabilization and observer design of polytopic perturbed linear fractional-order systems. *Math. Probl. Eng.* **2021**, *2021*, 6699756. [[CrossRef](#)]
33. Gassara, H.; Naifar, O.; Makhoulouf, A.B.; Mchiri, L. Global practical conformable stabilization by output feedback for a class of nonlinear fractional-order systems. *Math. Probl. Eng.* **2022**, *2022*, 4920540. [[CrossRef](#)]
34. Yakoub, Z.; Naifar, O.; Ivanov, D. Unbiased identification of fractional order system with unknown time-delay using bias compensation method. *Mathematics* **2022**, *10*, 3028. [[CrossRef](#)]
35. Naifar, O.; Makhoulouf, A.B.; Hammami, M.A.; Ouali, A. On observer design for a class of nonlinear systems including unknown time-delay. *Mediterr. J. Math.* **2016**, *13*, 2841–2851. [[CrossRef](#)]
36. Naifar, O.; Makhoulouf, A.B.; Hammami, M.A.; Ouali, A. State feedback control law for a class of nonlinear time-varying system under unknown time-varying delay. *Nonlinear Dyn.* **2015**, *82*, 349–355. [[CrossRef](#)]
37. Naifar, O.; Boukettaya, G.; Ouali, A. Global stabilization of an adaptive observer-based controller design applied to induction machine. *Int. J. Adv. Manuf. Technol.* **2015**, *81*, 423–432. [[CrossRef](#)]
38. Naifar, O.; Boukettaya, G.; Oualha, A.; Ouali, A. A comparative study between a high-gain interconnected observer and an adaptive observer applied to IM-based WECS. *Eur. Phys. J. Plus* **2015**, *130*, 88. [[CrossRef](#)]
39. Naifar, O.; Boukettaya, G.; Ouali, A. Robust software sensor with online estimation of stator resistance applied to WECS using IM. *Int. J. Adv. Manuf. Technol.* **2016**, *84*, 885–894. [[CrossRef](#)]
40. Boukettaya, G.; Naifar, O. Improving grid connected hybrid generation system supervision with sensorless control. *J. Renew. Sustain. Energy* **2015**, *7*, 043115. [[CrossRef](#)]
41. Ayadi, M.; Naifar, O.; Derbel, N. Sensorless control with an adaptive sliding mode observer for wind PMSG systems. In Proceedings of the 2017 14th International Multi-Conference on Systems, Signals & Devices (SSD), Marrakech, Morocco, 28–31 March 2017; pp. 33–37.
42. Boukettaya, G.; Naifar, O.; Ouali, A. A vector control of a cascaded doubly fed induction generator for a wind energy conversion system. In Proceedings of the 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14), Barcelona, Spain, 11–14 February 2014; pp. 1–7.
43. Ghoul, Y.; Naifar, O. IoT based applications for healthcare and home automation. *Multimed. Tools Appl.* **2024**, *83*, 29945–29967. [[CrossRef](#)]
44. Mseddi, A.; Dhouib, B.; Zdiri, M.A.; Alaas, Z.; Naifar, O.; Guesmi, T.; Alshammari, B.M.; Alqunun, K. Exploring the potential of hybrid excitation synchronous generators in wind energy: A comprehensive analysis and overview. *Processes* **2024**, *12*, 1186. [[CrossRef](#)]
45. Mseddi, A.; Wali, K.; Abid, A.; Naifar, O.; Rhaima, M.; Mchiri, L. Advanced modeling and control of wind conversion systems based on hybrid generators using fractional order controllers. *Asian J. Control* **2024**, *26*, 1103–1119. [[CrossRef](#)]
46. Mseddi, A.; Abid, A.; Naifar, O.; Rhaima, M.; Makhoulouf, A.B.; Mchiri, L. Investigation of the robust fractional order control approach associated with the online analytic unity magnitude shaper: The case of wind energy systems. *Fractal Fract.* **2024**, *8*, 187. [[CrossRef](#)]
47. Alaerjan, A.; Jabeur, R.; Chikha, H.B.; Karray, M.; Ksantini, M. Improvement of smart grid stability based on artificial intelligence with fusion methods. *Symmetry* **2024**, *16*, 459. [[CrossRef](#)]
48. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 694–699.
49. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **1999**, *10*, 61–74.
50. Naifar, O. Tempered fractional gradient descent: Theory, algorithms, and robust learning applications. *Neural Netw.* **2025**, *193*, 108005. [[CrossRef](#)]
51. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
52. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]

53. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
55. Pudjianto, D.; Ameli, H.; Giannelos, S. Holistic evaluation of the roles and values of electrolyzers in a net-zero energy system. *IET Conf. Proc.* **2025**, *2025*, 432–440. [[CrossRef](#)]
56. Houda, Z.A.E.; Nabousli, D.; Kaddoum, G. Cost-efficient Federated Reinforcement Learning-Based Network Routing for Wireless Networks. In Proceedings of the 2022 IEEE Future Networks World Forum (FNWF), Montreal, QC, Canada, 10–14 October 2022; pp. 243–248.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.