# Rethinking Machine Learning Performance Metrics in Distributed Energy Resource Systems: A Concise Review


Qiushi Li
Cybersecurity and Intelligent Systems Engineering
Concordia University
Montreal, QC, Canada
qiushi.li@mail.concordia.ca

Chengming Hu
Computer Science
McGill University
Montreal, QC, Canada
chengming.hu@mail.mcgill.ca

Jun Yan*
Cybersecurity and Intelligent Systems Engineering
Concordia University
Montreal, QC, Canada
jun.yan@concordia.ca

Jie Gao
Transport and Planning
Delft University of Technology
Delft, Netherlands
j.gao-1@tudelft.nl

Hepeng Li
Electrical and Computer Engineering
University of Maine
Orono, ME, USA
hepeng.li@maine.edu

Yuhong Liu
Computer Science and Engineering
Santa Clara University
Santa Clara, CA, USA
yhliu@scu.edu

Xiao'ou Yang
Mechanical Engineering
Santa Clara University
Santa Clara, CA, USA
xyang13@scu.edu

Dafang Zhao
Institute of Science and Technology
University of Osaka
Osaka, Japan
zhao.dafang@ist.osakau.ac.jp



## Abstract

Machine learning (ML) techniques have been widely applied to distributed energy resource (DER) systems for a broad range of tasks, including load forecasting and operational control. However, evaluation practices remain inconsistent, conflating ML performance with DER system-level performance without a clear, systematic relationship, thereby limiting the interpretability and practical applicability of emerging ML applications in DER contexts. To this end, this paper introduces a hierarchical evaluation framework that systematically aligns ML performance with DER operational objectives. Specifically, this paper first discusses how ML-driven DER systems are currently evaluated, providing a comprehensive overview of performance evaluation metrics, their advantages and limitations, and practical insights. Given the challenges in evaluating ML-driven DER systems, we further organize performance evaluation metrics hierarchically, grouping them into six top-level categories that represent core DER objectives. Within each category, metrics are distinguished between ML-side and energy-side evaluation criteria. Consequently, our hierarchical DER-centric evaluation framework not only provides transparent interpretability of ML performance in the DER context but also clarifies recurring sources of confusion in current practice, particularly when ML performance is overgeneralized or implicitly assumed to reflect DER system-level performance. Finally, we highlight the challenges and identify future opportunities associated with the proposed hierarchical evaluation framework. This work serves as a valuable reference for researchers and engineers to better understand existing studies, as well as the potential, challenges, and opportunities in evaluating ML-driven DER solutions.



## CCS Concepts

• **General and reference → Evaluation**; **Metrics**; *Surveys and overviews*; • **Computing methodologies → Machine learning**; • **Hardware → Smart grid**.

## Keywords

Distributed Energy Resources, Machine Learning, Performance Evaluation, Hierarchical Framework, Metrics, Energy Management Systems

**ACM Reference Format:**
Qiushi Li, Chengming Hu, Jun Yan, Jie Gao, Hepeng Li, Yuhong Liu, Xiao'ou Yang, and Dafang Zhao. 2026. Rethinking Machine Learning Performance Metrics in Distributed Energy Resource Systems: A Concise Review. In *ACM Sustainability Week 2026 (ACM Sustainability Week Companion '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3765611.3815506


## 1 Introduction

Distributed energy resources (DERs), including rooftop photovoltaics, energy storage systems, electric vehicles, and flexible loads, are becoming an integral part of modern energy systems [21]. Beyond traditional residential and commercial settings, DER ecosystems are rapidly expanding to encompass emerging large-scale electricity consumers, such as data centers, electric transportation hubs, and flexible industrial loads, which increasingly participate in grid-edge operation and coordination [24]. In particular, data centers are transitioning from passive electricity consumers to grid-interactive entities through the deployment of microgrids and DERs, in which accurate load prediction, operational flexibility, and coordinated control are essential to ensure both energy efficiency and system resilience [50]. The growing penetration and diversification of DERs


*Corresponding Author




*ACM Sustainability Week Companion '26, Banff, AB, Canada*



ACM ISBN 979-8-4007-2199-1/26/06
https://doi.org/10.1145/3765611.3815506

introduce new challenges in coordination, forecasting, and operational decision-making, particularly at the grid edge, where uncertainty, heterogeneity, and local constraints play a central role [13]. Machine learning (ML) techniques have been widely adopted to address these challenges, enabling data-driven approaches to forecasting, control, scheduling, and decision support for DER applications. Compared with traditional rule-based or optimization-based methods, ML-based approaches provide superior adaptability and scalability across diverse DER tasks [2, 49].

However, evaluation practices for ML-assisted DER systems remain fragmented. Many studies rely on ML performance metrics such as accuracy, reward, or success rate, while others emphasize energy-system indicators such as cost, reliability, or constraint violations. These metrics are often reported independently, with limited discussion of how ML performance aligns with system-level outcomes in DER contexts, and reported performance improvements may not translate into comparable operational benefits [22, 60]. Furthermore, ML evaluation practices typically abstract away physical constraints, operational feasibility, and economic implications that are central to DER operation [49]. These challenges are further compounded by unresolved data infrastructure issues: the FAIR principles (Findable, Accessible, Interoperable, and Reusable) were established to ensure that research data and artifacts are discoverable, accessible, and reproducible across systems and disciplines [58], yet their adoption in the energy domain remains limited, with persistent gaps in metadata quality, standardized formats, and data provenance documentation [45, 57]. As a result, metric comparisons across studies frequently conflate differences in model capability with differences in data quality and availability, undermining reproducibility and cross-study comparability in ML-assisted DER research. This disconnect motivates the need for a more principled and transparent evaluation perspective.

In parallel, existing surveys have reviewed the application of AI and ML techniques in DER contexts [3, 5, 12, 31, 61]. As summarized in Table 1, these survey studies primarily focus on application scenarios and operational objectives in DER systems, while offering limited DER-level analysis of ML performance evaluation, particularly with respect to the alignment between ML evaluation metrics and DER operational objectives. Specifically, Lee et al. [31] analyze how integrating ML technologies enhances the security, scalability, reliability, and other operational constraints of energy management systems. Zaboli et al. [61] compare the capabilities of various ML models (such as convolutional neural networks and transformers) for load forecasting within integrated energy systems involving photovoltaic generation, battery energy storage, and electric vehicle charging infrastructures. Arevalo et al. [5] explore the specific roles that ML models play in optimizing operations, enhancing efficiency, and improving system resilience across energy system generation, transmission, and distribution. Bilal et al. [12] investigate AI-based hierarchical control mechanisms at the primary, secondary, and tertiary levels in microgrid energy systems, emphasizing their contributions to managing microgrid reliability and efficiency. Ali et al. [3] present a structured framework for evaluating key performance indicators (KPIs) for ML applications in the energy domain, with a particular emphasis on large language model (LLM) implementations in renewable energy management. On the other hand, within broader computer systems, Arp et al. [6] identify common pitfalls in the design, implementation, and evaluation of learning-based computer security systems and present actionable recommendations to mitigate them. Biderman et al. [11] identify common pitfalls throughout the ML model deployment pipeline, spanning algorithm design, data collection, and model evaluation. To date, however, comparable metric-centric analyses remain absent in the DER domain, where ML models are tightly coupled with physical processes, operational constraints, and system-level risk.

To this end, this review paper introduces a **hierarchical, DER-centric evaluation framework** that rethinks how ML performance is assessed in DER systems. Specifically, this paper first provides a comprehensive overview of ML-assisted DER systems and summarizes representative performance evaluation metrics, highlighting their advantages, limitations, and practical insights. Given the challenges in evaluating ML-assisted DER systems, we propose a hierarchical evaluation framework, in which performance evaluation metrics are grouped into **six top-level categories** representing core DER application categories, including: (1) state & knowledge consistency, (2) decision & constraint coherence, (3) operational & temporal robustness, (4) market & incentive fairness, (5) operational safety & risk exposure, and (6) human interpretability & action transparency. Within each category, metrics are explicitly distinguished between ML-side evaluation criteria, representing learning behavior and decision quality, and energy-side evaluation criteria, capturing system performance and operational outcomes. The hierarchical DER-centric evaluation framework provides transparent interpretability of ML performance while clarifying sources of confusion in DER systems, particularly in cases where ML performance is overgeneralized or implicitly assumed to reflect system performance. Finally, we highlight future directions associated with our hierarchical evaluation framework, including real-world deployment in non-stationary environments and reinforcement learning-based DER systems. To summarize, the **main contributions** of this paper are provided as follows:

- This paper provides a comprehensive overview of performance evaluation in ML-assisted DER systems, with a particular focus on analyzing the relationships between DER applications and ML performance metrics, as well as the advantages, limitations, and practical insights associated with these metrics.
- This paper introduces a hierarchical, DER-centric evaluation framework in which performance metrics are systematically categorized by DER use cases. Within each category, metrics are explicitly distinguished between ML-side and energy-side evaluation criteria, thereby enabling transparent interpretation of ML performance in DER systems.
- This paper outlines a series of actionable recommendations to enhance performance evaluation in ML-assisted DER systems, aimed at achieving more transparent, robust, and deployment-relevant evaluations.

The remainder of this paper is organized as follows. Section 2 analyzes commonly used ML evaluation metrics in ML-assisted DER systems, highlighting their application contexts and common pitfalls. Section 3 presents the proposed hierarchical DER-centric evaluation framework. Section 4 discusses limitations and recommendations for future research, and Section 5 concludes this paper.

**Table 1: Comparison of our paper with existing DER-related surveys.**

Metric Alignment: Represents the alignment between ML-side performance metrics and DER operational objectives.

| Reference | DER Systems | | ML Evaluation Metrics | Metric Alignment |
|---|---|---|---|---|
| | Applications | Operational Objectives | | |
| [31] | - energy supply management<br>- energy trading<br>- energy demand control | - security<br>- scalability<br>- interoperability<br>- reliability | ✗ | ✗ |
| [61] | - load forecasting | ✗ | - accuracy<br>- computational resource | ✗ |
| [5] | - demand forecasting<br>- energy flow optimization<br>- predictive maintenance | - scalability<br>- cost efficiency | - accuracy<br>- response time<br>- computational resource | ✗ |
| [12] | - energy allocation<br>- demand-side management<br>- hierarchical microgrid control | - reliability<br>- cost efficiency | ✗ | ✗ |
| [3] | - energy forecasting<br>- predictive maintenance<br>- resource optimization | - reliability<br>- operational efficiency<br>- energy efficient<br>- user experience | - accuracy<br>- inference time<br>- computational resource<br>- energy consumption | ✗ |
| **Our paper** | **- energy supply management**<br>**- energy trading**<br>**- demand-side management**<br>**- energy forecasting**<br>**- load forecasting**<br>**- resource optimization** | **- state consistency**<br>**- decision coherence**<br>**- operational robustness**<br>**- market fairness**<br>**- operational safety**<br>**- action transparency** | **- accuracy**<br>**- cumulative reward**<br>**- success rate**<br>**- robustness**<br>**- fairness**<br>**- interpretability** | **A hierarchical DER-centric evaluation framework** |

## 2 Evaluation in ML-Assisted DER Systems

This section reviews existing evaluation practices for ML–assisted DER systems and analyzes the challenges that arise when applying ML performance metrics in DER contexts. Rather than proposing new evaluation dimensions, it aims to clarify how evaluation is currently conducted and where conceptual gaps persist.

### 2.1 Relationships Between DER Applications and Performance Evaluation Metrics

Before examining evaluation metrics in detail, it is instructive to first consider how existing studies implicitly align DER application objectives with ML-side performance evaluation metrics. Figure 1 illustrates the high-level alignment between DER objectives and ML metrics synthesized from representative literature. The band widths reflect the relative frequency of each metric-objective pair, highlighting dominant evaluation patterns; each study may contribute to multiple bands.

From the insights of ML metrics, accuracy metrics, including both classification accuracy and regression error, pervade nearly all DER application categories. They dominate categories related to *State & Knowledge Consistency* [48], *Decision & Constraint Coherence* [26], and *Operational Safety & Risk Exposure* [55], where short-term predictive correctness or immediate response fidelity is often treated as a proxy for system performance. This widespread reuse suggests that accuracy metrics frequently serve as defaults, even when the underlying operational objectives differ substantially.

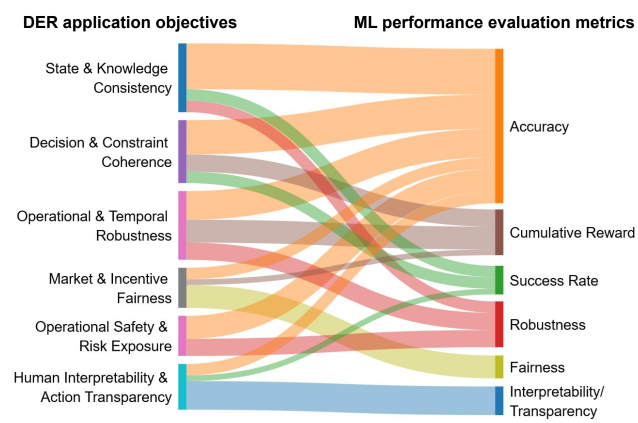


**Figure 1: Relationships between DER application objectives categories and ML performance evaluation metrics.**

Another insight is that cumulative reward metrics primarily align with sequential decision-making and adaptation, most notably *Decision & Constraint Coherence* [36] and *Operational & Temporal Robustness* [47]. This reflects the prevalence of reinforcement learning for scheduling and control, where long-term objective accumulation is explicitly modeled. However, the mapping also reveals that reward-based evaluation remains concentrated and is seldom applied to market fairness or interpretability. Moreover, success rate or task completion metrics are less frequent but span heterogeneous objectives, including *State & Knowledge Consistency* [41], *Decision & Constraint Coherence* [9], and *Human Interpretability & Action Transparency* [43]. In these contexts, success-oriented metrics are often used to indicate convergence, feasibility, or task-level completion, providing a coarse but intuitive signal of whether a system meets predefined operational criteria. Finally, the mapping highlights a notable concentration of fairness-, robustness-, and interpretability-oriented metrics within *Market & Incentive Fairness* [26] and *Human Interpretability & Action Transparency* [43]. Outside these categories, such metrics are comparatively under-represented, even in applications where long-term stability, risk exposure, or trustworthiness are central concerns. This imbalance suggests a systematic misalignment between what is operationally important and what is most commonly measured.

Consequently, the aggregated relationships underscore that identical ML-side metrics are often reused across fundamentally different DER objectives, while some objectives rely on narrowly scoped evaluation signals. This observation motivates a closer examination of how specific metric families are interpreted and applied in DER contexts. In the following, we therefore focus on three representative ML performance evaluation metrics, including accuracy, cumulative reward, and success rate, to illustrate their strengths, limitations, and potential misalignments when evaluated from a DER-centric system perspective.

## 2.2 ML Performance Evaluation Metrics

Building on the aggregated mapping between DER objectives and ML metrics, we now examine how these metrics are interpreted and applied in practice. The mapping reveals that a small set of ML-side metrics is repeatedly reused across heterogeneous DER objectives, often serving as proxies for different operational goals. To make these implicit assumptions explicit, this section analyzes three representative and widely reused metric families—accuracy, cumulative reward, and success rate/task completion rate—and analyzes their typical usage patterns, advantages, and limitations from a DER-centric perspective. Many of these metrics originate in the ML literature for characterizing algorithmic behavior, such as learning efficiency or prediction accuracy. While effective in controlled experiments, their interpretation is complex when transferred to real-world DER operations, where physical constraints, economic trade-offs, and reliability jointly shape system performance. Through structured analyses, including practical do's and don'ts, we illustrate recurring misalignments between ML performance indicators and the energy-side operational meaning, thereby motivating the need for more principled, DER-centric evaluation frameworks, as discussed in Section 3.

*2.2.1 Accuracy.* Accuracy-related metrics quantify how closely model outputs match ground-truth values, and are pervasively adopted in ML-driven DER systems. As illustrated in Figure 1, they evaluate diverse objectives from state estimation to safety monitoring. Depending on the task formulation, accuracy metrics can be broadly categorized into two classes: classification-oriented accuracy for discrete decision or event detection tasks, and regression-oriented error metrics for continuous-value forecasting. Although both are commonly referred to as "accuracy" in the literature, they capture fundamentally different aspects of model performance and entail distinct implications in DER contexts.

(1) **Classification Accuracy**
- *Description:* Classification accuracy measures the ratio of correctly predicted labels or events to total instances in tasks such as fault detection, anomaly detection, and event-based demand response, where the model output corresponds to discrete operational states or decisions.
- *What is Commonly Done:* Many studies report classification accuracy as a primary performance indicator due to its simplicity and interpretability, but often supplement it with precision, recall, and F1-score to better capture asymmetric error costs, particularly in safety-critical applications where false negatives and positives carry different operational consequences [14, 56].
- *Advantage:* Classification accuracy provides an intuitive first-order assessment of decision correctness, easily communicated to stakeholders.
- *Limitation:* Classification accuracy is highly sensitive to class imbalance. For instance, in anomaly detection, models may achieve deceptively high accuracy by predominantly predicting the majority class, thereby offering limited operational value and failing to distinguish between error types with varying system-level impacts. [52].

(2) **Regression Error**
- *Description:* Regression error measures deviations between predicted and observed values in tasks including load forecasting, renewable generation prediction, and state estimation, among others.
- *What is Commonly Done:* Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Squared Error (MSE) are routinely combined to evaluate prediction fidelity [30, 33]. MAE reflects average physical deviation, while RMSE penalizes large errors more heavily, and MSE is frequently used as a training loss function [29, 34, 53]. In renewable energy forecasting, RMSE and MAE dominate, highlighting extreme deviations and overall performance, respectively [19].
- *Advantage:* Regression error metrics are effective when error magnitude correlates with operational risk. RMSE is particularly suitable for grid-related forecasting, where rare but large errors may jeopardize system stability, whereas MAE supports routine monitoring [19, 54].
- *Limitation:* Dataset and task specifics strongly influence regression metrics, limiting cross-study comparability. Identical error values across studies do not guarantee similar operational performance, especially under different load profiles, temporal resolutions, or system constraints [52].

Consequently, we summarize the **Do & Don't** for applying accuracy-related metrics in ML-driven DER systems:

**Do:** Treat accuracy metrics as task-dependent indicators rather than universal measures. Combine it with precision, recall, and F1-score for classification, or with MAE and RMSE for forecasting, to balance interpretability and sensitivity to extreme errors [28]. Establish meaningful baselines and clearly report dataset characteristics to support fair comparison.

**Don't:** Avoid equating high accuracy or low error with practical effectiveness without considering data imbalance, error distribution, and system risk. Do not compare accuracy metrics across studies without clarifying task formulation, dataset properties, and operational context. Refrain from selecting accuracy metrics without accounting for response-time constraints, tolerance for large deviations, and the relative cost of different real-world failure modes.

#### 2.2.2 Cumulative Reward.

**Description:** Cumulative reward is the core objective in reinforcement learning (RL)-based DER management, as shown in Figure 1. It quantifies long-term returns by summing immediate rewards at each decision step, typically weighted by a discount factor to balance short-term gains against long-term benefits. In DER applications, the reward structure commonly comprises two components: revenue gains from optimization objectives (e.g., cost reduction, efficiency improvement) and penalty terms for constraint violations. High penalty coefficients are essential for enforcing operational bounds and guiding policies toward safe, high-performance states.

**What is Commonly Done:** In battery energy storage system (BESS) management, cumulative rewards are used to evaluate the optimization of charging and discharging decisions by balancing operational costs, battery aging, and economic performance [15, 59]. During training, agents continuously explore and learn charging strategies that maximize cumulative rewards over time [27]. Microgrid energy management systems use it to coordinate multiple DERs, including solar panels, wind turbines, fuel cells, and storage systems [15, 39, 40]. In demand response, cumulative rewards are employed to optimize load control strategies where agents learn to manage clusters of heterogeneous loads such as thermostatically controlled devices [17, 42]. Market participation strategies leverage cumulative reward to develop optimal energy trading and pricing policies based on historical price patterns [20, 42].

**Advantage:** Cumulative reward suits long-term DER optimization over myopic immediate reward maximization, especially in volatile renewable-powered microgrids where supply and demand change rapidly [38]. Unlike traditional model predictive control approaches, cumulative reward enables continuous strategy refinement as operational conditions evolve [51].

**Limitation:** Calibrating the revenue-penalty balance is a critical challenge. Insufficient penalties lead to safety violations [62], while multi-objective complexity can bias solutions toward certain goals at the expense of others. Furthermore, delayed consequences and reward sparsity hinder learning, particularly in contexts such as battery degradation or market dynamics, where immediate rewards are insufficient. Increasing system scale and DER counts further complicate the design of the reward function. Moreover, poor transferability across different configurations or market conditions requires extensive retraining in evolving operational environments.

Consequently, we summarize the **Do & Don't** for applying cumulative rewards in RL-based DER management:

**Do:** Carefully tune the discount factor to align with the application's temporal scale—smaller for immediate response, larger for long-term gains [1]. Implement a two-part structure that combines revenue gains with high penalties for constraint violations to ensure reliability [62]. Use progressive learning strategies that allow agents to explore the solution space while converging to optimal policies over multiple episodes [27]. For validation purposes, systematically examine both reward and loss curves over episodes. For example, rising total rewards can confirm effective balancing of multiple objectives such as photovoltaic energy utilization, grid dependence minimization, and battery management [8].

**Don't:** Avoid inadequate penalties for constraint violations, which directly compromise system safety [62]. Do not neglect the temporal scale of decision consequences; immediate rewards may fail to capture delayed effects such as battery degradation or long-term market impacts, leading to myopic policies over extended horizons. Refrain from assuming that reward functions optimized for one DER configuration will transfer directly to different systems without retraining or recalibration. Similarly, do not ignore the computational complexity that emerges as system scale increases; reward design should account for scalability to ensure practical deployability in large-scale DER management scenarios.

#### 2.2.3 Success Rate & Task Completion Rate.

**Description:** Success and task completion rates are outcome-oriented evaluation metrics that measure whether a system fulfills a predefined task under specified conditions, typically expressed as the ratio of successful executions to total attempts. In DER management, a task may involve a scheduling decision within a deadline, executing an offloading request within Quality of Service (QoS) constraints, or finishing a control procedure without violating operational rules. These metrics serve as high-level indicators of whether an ML-driven system is reliable and "works" in practice.

**What is Commonly Done:** In power generation scheduling, the success rate is commonly used to assess reliability by measuring the percentage of tasks completed within time limits. For example, recent studies show ML-based schedulers significantly outperform traditional heuristics in meeting dynamic deadlines [25]. In energy system modeling and optimization, success rate is often decomposed into multiple dimensions, such as whether objective functions are correctly specified (e.g., including fuel, start-up, and shut-down costs) and whether constraints are properly formulated and satisfied across all units and time periods [32]. In edge computing systems coupled with DERs, task completion rate is defined as the ratio of completed tasks to arrived tasks over a fixed interval, serving as a proxy for energy-aware execution capability under limited resources [7]. Similarly, task offloading scenarios evaluate success rate as the fraction of offloaded tasks completed within QoS constraints, reflecting system robustness under workload and energy fluctuations [4]. Some variants also quantify how sharply performance degrades under high renewable volatility compared to stable conditions [25].

**Advantage:** Success-oriented metrics are effective when objectives can be clearly operationalized as pass/fail conditions. In tasks like offloading or real-time scheduling, measuring whether tasks meet QoS or deadline constraints provides a direct and intuitive indicator of system reliability [4]. In these settings, the success rate enables straightforward comparison across algorithms and is easily interpretable by stakeholders, making it particularly suitable for benchmarking competing scheduling or control strategies [25].

**Limitation:** The primary limitation is their binary nature, which fails to capture partial progress or intermediate achievements, even when such progress may still deliver operational value. This lack of granularity is problematic in complex DER tasks composed of multiple interdependent sub-steps. Moreover, defining "success" for open-ended or long-horizon tasks is nontrivial, often requiring human judgment or task-specific heuristics [16, 23, 37]. Success rates also depend heavily on arbitrary threshold choices, such as fixed step limits or strict timing constraints, which may not align with real-world operational flexibility, consequently obscuring behavioral differences and masking performance variations [65].

Consequently, we summarize the **Do & Don't** for applying success-oriented metrics in ML-driven DER systems:

**Do:** Align success criteria with operational requirements, such as explicit deadline satisfaction or conflict avoidance, as demonstrated in power generation studies where tasks are counted as successful only if completed before assigned deadlines [25]. Complement success rate with constraint satisfaction or stability indicators to capture safety dimensions that binary outcomes alone cannot reveal [63]. For multi-stage tasks, adopt hierarchical or sub-goal success analysis to reflect partial completion [35].

**Don't:** Avoid relying on success rate as a standalone metric for complex problems where progress, quality, or trade-offs matter more than strict completion. Do not impose artificial thresholds without operational justification, as overly rigid definitions of success can misrepresent system capability and obscure meaningful performance differences. Finally, do not equate high success rates with overall optimality without verifying if efficiency and robustness requirements are met.

## 2.3 Challenges in Performance Evaluation

Having examined commonly used performance metrics in ML-assisted DER applications, we find that a more fundamental challenge remains: their interpretation and integration. Existing studies often report multiple metrics without a coherent perspective linking ML indicators to DER operational objectives. A key challenge stems from the disconnect between ML-centric evaluation and system validation. Metrics that effectively characterize model behavior, such as learning progress or task success, may not reflect whether decisions are safe and beneficial for DER operations, especially in distributed settings with heterogeneous ownership, local objectives, and physical constraints. Moreover, evaluation is often constrained by application-specific assumptions and simulation designs, making cross-study comparisons difficult even for similar use cases. The absence of a structured evaluation framework that jointly considers ML-side metrics and energy-side criteria limits interpretability, reproducibility, and the accumulation of cumulative knowledge.

**FAIR Data Challenges and Evaluation Reproducibility.** Beyond metric-level inconsistencies, evaluation outcomes in ML-assisted DER research are fundamentally shaped by underlying data challenges that remain largely unaddressed in the literature. In the DER context specifically, datasets are frequently proprietary, system-specific, or insufficiently documented in terms of provenance and operational context, making it difficult to reproduce reported ML evaluation results or assess their generalizability to real-world deployments. Ongoing efforts to address these barriers—including FAIR-compliant tools for time series dissemination [44], ontology-based metadata models for PV systems [46], FAIR maturity assessments for wind energy data sharing [10, 18], and broader community frameworks for low-carbon energy research [57]—highlight that achieving FAIR data practices in DER systems requires coordinated technical and cultural change. Until such standards are more widely adopted, the metric-level challenges identified above will persist, further motivating the structured evaluation perspective introduced in the next section.

# 3 Hierarchical DER-Centric Evaluation Framework

This section introduces a hierarchical, DER-centric evaluation framework to structure performance assessment of ML–assisted DER applications. Rather than proposing new metrics, the framework focuses on organizing existing evaluation practices to explicitly distinguish between ML-side and energy-side validation metrics. Table 2 provides an overview of the framework, summarizing the main categories and their interpretations in DER contexts.

## 3.1 Motivations and Design Principles

Several recent studies have explored structured ways of organizing evaluation metrics for ML-assisted power system applications. A representative example is the framework proposed in ElecBench [64], which categorizes performance metrics for large language models in power dispatch scenarios according to high-level system objectives. Such frameworks offer a useful organizational perspective, particularly for centralized decision-making tasks where evaluation is conducted primarily from a system operator's viewpoint.

In this work, we draw inspiration from the high-level categorization logic adopted in dispatch-oriented frameworks, while adapting the evaluation perspective to better align with DER applications. Our focus differs in scope and application context: whereas prior frameworks are primarily concerned with LLM-based dispatch tasks, this study targets a broader range of ML techniques and DER use cases, including decentralized control, local optimization, and market participation. To better reflect the characteristics of DER systems, we retain the top-level evaluation categories as an organizational starting point while reframing their interpretation in terms of DER-oriented use cases. For each category, we explicitly distinguish between ML-side and energy-side evaluation metrics. This separation is intended to clarify how commonly reported ML metrics relate to operational meaning in DER contexts, rather than to introduce an alternative or competing evaluation taxonomy. In addition, our proposed framework is guided by three design principles. First, it emphasizes interpretability by organizing evaluation dimensions around familiar high-level objectives. Second, it adopts a DER-centric perspective that accounts for distributed decision-making and heterogeneous stakeholder goals. Third, it makes the ML-side and energy-side evaluation dimensions explicit, facilitating clearer interpretation and more consistent reporting across studies.

## 3.2 Hierarchical Evaluation Categories

The proposed framework organizes evaluation dimensions hierarchically. At the top level, a set of high-level evaluation categories is used to capture core objectives commonly encountered in DER applications. These categories follow the general organizational logic of prior dispatch-oriented frameworks, while their interpretation is adjusted to align with the needs of DER-centric evaluation.

**Table 2: DER-Centric Evaluation Framework with ML-Side and Energy-Side Evaluation Metrics**

| Top-level Category | DER Use-case Interpretation | ML-side Evaluation Metrics (examples) | Energy-side Evaluation Metrics (examples) | Notes / Gaps |
|---|---|---|---|---|
| **State & Knowledge Consistency** | Whether ML-driven DER decisions rely on correct, up-to-date system states, forecasts, and market information | - state prediction accuracy<br>- forecasting error<br>- model loss<br>- data consistency checks | - state estimation error impact<br>- infeasible dispatch<br>- market signals misinterpretation<br>- operational degradation | Predictive accuracy does not guarantee correctness of DER operational states |
| **Decision & Constraint Coherence** | Whether ML-generated DER decisions respect physical limits, operational constraints, and market rules | - constraint satisfaction rate<br>- policy feasibility ratio<br>- success or task completion rate<br>- optimization convergence | - constraint violations<br>- infeasible dispatch actions<br>- corrective control interventions<br>- penalty costs | Task success or logical consistency does not ensure physical or market feasibility |
| **Operational & Temporal Robustness** | Robustness of DER decisions under uncertainty, variability, and long-horizon operation | - reward variance<br>- policy sensitivity<br>- convergence stability<br>- robustness test outcomes | - power quality deviations<br>- control oscillations<br>- curtailment frequency<br>- long-term cost variability | Learning stability does not imply stable DER or grid operation |
| **Market & Incentive Fairness** | Whether ML-driven DER decisions lead to incentive-rational and socially acceptable market outcomes | - reward distribution<br>- fairness metrics<br>- equilibrium convergence<br>- agent utility balance | - profit realization<br>- market efficiency<br>- price volatility<br>- risk of manipulation | AI-side fairness or equilibrium metrics may not reflect actual market efficiency |
| **Operational Safety & Risk Exposure** | Risks introduced by ML-driven decisions in DER operation under failures or adversarial conditions | - failure rate<br>- worst-case reward<br>- perturbation robustness<br>- stress test performance | - worst-case operational loss<br>- fault propagation<br>- safety margin erosion<br>- resilience degradation | Model robustness metrics do not fully capture physical and systemic risk |
| **Human Interpretability & Action Transparency** | Whether ML-assisted DER decisions are understandable, auditable, and actionable by stakeholders | - interpretability scores<br>- feature importance<br>- explanation consistency | - operational auditability<br>- operator trust<br>- regulatory compliance<br>- decision traceability | Model interpretability does not automatically translate to operational accountability |

At the second level, each top-level evaluation category is grouped into two complementary perspectives. One focuses on ML-side metrics that characterize model behavior, learning dynamics, or decision quality. The other focuses on energy-side metrics that reflect physical feasibility, operational outcomes, and economic or system-level impacts resulting from ML-driven decisions. The distinction between ML-side and energy-side metrics is not intended to impose a strict separation, as many metrics influence both learning behavior and system performance. Instead, it provides a lens to clarify the origin, interpretation, and appropriate role of different metrics when assessing ML-assisted DER systems.

**State & Knowledge Consistency:** This dimension concerns whether ML-driven DER decisions are based on correct, timely, and contextually appropriate information, including system states, forecasts, operational constraints, and market signals. In many DER applications, which underpin downstream decision-making, affecting scheduling, control actions, and economic outcomes. Errors or staleness in these inputs may propagate through learning-based decision processes, leading to suboptimal or infeasible DER behavior. ML-side metrics in this category typically assess accuracy, forecasting error, or internal data consistency. Energy-side metrics instead capture the operational consequences of incorrect information, such as infeasible dispatch actions, misinterpretation of price or demand signals, and overall degradation of system performance. Thus, high predictive accuracy at the model level could not guarantee correctness or reliability at the DER operational level.

**Decision & Constraint Coherence:** This dimension evaluates whether ML-generated DER decisions respect physical limits, operational constraints, and market rules inherent to the target use case. DER applications are governed by device-level capacity limits, network constraints, and regulatory or market participation rules. Decisions that appear valid from an optimization or learning perspective may nevertheless violate such constraints when deployed. Furthermore, ML-side evaluation metrics commonly include constraint satisfaction, policy feasibility, task success, or convergence behavior. Energy-side metrics focus on actual constraint violations, infeasible dispatch actions, corrective control interventions, and incurred penalty costs. Consequently, task-level success or logical consistency in learned policies does not ensure physical or market feasibility in real DER operations.

**Operational & Temporal Robustness:** This dimension characterizes the stability of ML-driven DER decisions under uncertainty, variability, and long-horizon operation. DER environments are inherently dynamic, with fluctuations in renewable generation, load demand, and market conditions. Robust decision-making requires both short-term performance and stable behavior over extended time horizons and under perturbations. ML-side metrics in this category often assess reward variance, policy sensitivity, convergence stability, or performance under robustness testing. Energy-side metrics capture system-level manifestations of instability, such as power quality deviations, oscillatory control actions, curtailment frequency, and long-term cost variability. Stable learning dynamics or convergence at the ML level, therefore, do not necessarily imply stable or reliable DER operations.

**Market & Incentive Fairness:** This dimension examines whether ML-driven DER decisions lead to economically rational, incentive-compatible, and socially acceptable outcomes in market-oriented applications. Many DER use cases involve participation in energy markets, peer-to-peer trading, or price-based demand response, where decisions directly affect profits, costs, and overall market efficiency. ML-side metrics typically include reward distribution, fairness metrics, equilibrium convergence, or agent utility balance in multi-agent settings. Energy-side metrics instead reflect realized profits, market efficiency, price volatility, and risks of manipulation or strategic behavior. As a result, fairness or equilibrium metrics evaluated at the ML level may not accurately represent economic efficiency or incentive fairness at the system level.

**Operational Safety & Risk Exposure:** This dimension focuses on the risks posed by ML-driven decisions in DER operations under failure, uncertainty, or adverse conditions. As DER systems increasingly rely on automated decision-making, erroneous or unstable policies can amplify faults, reduce safety margins, or degrade system resilience. In addition, ML-side evaluation metrics often quantify failure rates, worst-case rewards, robustness to perturbations, or stress-test performance. Energy-side evaluation metrics emphasize worst-case operational losses, fault propagation effects, erosion of safety margins, and degradation of system resilience. Model-level robustness metrics, therefore, provide only partial insight into the physical and systemic risks.

**Human Interpretability & Action Transparency:** This dimension assesses whether ML-assisted DER decisions are understandable, auditable, and actionable by stakeholders. Interpretability plays a critical role in operational trust, regulatory compliance, and effective human-in-the-loop decision-making, particularly in safety-critical or market-regulated environments. ML-side evaluation metrics typically measure interpretability scores, feature importance, or the model's explanation consistency. Energy-side evaluation metrics focus on operational auditability, operator trust, regulatory acceptance, and decision traceability. Consequently, improvements in model interpretability do not automatically translate into accountability or acceptability in real-world DER operations.

## 4 Discussion and Recommendations

This work proposes a DER-centric evaluation framework that emphasizes the interpretation and alignment of evaluation metrics across ML-side and energy-side perspectives. Rather than prescribing a fixed set of metrics or ranking their importance, the framework aims to make implicit evaluation assumptions explicit and to clarify how commonly used ML metrics relate to DER operational objectives. As such, it should be viewed as an interpretive and organizational tool that supports more transparent evaluation practices, rather than a normative benchmark.

Several limitations of the framework should be noted. First, although it is grounded in an extensive literature survey, the mapping between evaluation metrics and DER use cases is not exhaustive. Certain metrics or evaluation practices may be underrepresented due to reporting conventions or domain-specific focus. Second, applying the framework still requires domain expertise, as interpreting energy-side evaluation metrics depends on system context, operational constraints, and stakeholder objectives. Finally, the framework is not empirically validated through large-scale benchmarking, which remains an important direction for future work. Additionally, the framework does not explicitly address real-time feasibility constraints, such as inference latency and computational overhead, which are critical when ML models are deployed on resource-constrained DER edge devices operating under strict response-time requirements; this represents another important direction for future investigation. Taken together, these limitations do not undermine the value of the proposed framework; rather, they motivate concrete recommendations to improve evaluation practices in ML-assisted DER research. Building on the insights derived from the framework, we outline a series of actionable recommendations to guide future studies toward more transparent, robust, and deployment-relevant evaluation.

**Recommendation 1: Adopt Structured and Transparent Evaluation Protocols.** We recommend that future ML-assisted DER studies adopt structured evaluation protocols that explicitly distinguish between ML-side metrics and energy-side operational outcomes. Rather than reporting isolated performance indicators, researchers should clarify how selected metrics align with specific DER objectives, constraints, and stakeholder goals. Standardized reporting templates or evaluation checklists can clarify implicit assumptions and improve interpretability and reproducibility.

**Recommendation 2: Complement Dataset-Based Evaluation with Testbeds, Simulations, and Scenario-Driven Studies.** DER evaluation should move beyond static datasets, which often fail to capture real-world dynamics, feedback loops, and constraints. We recommend testbeds, high-fidelity simulations, and scenario-driven evaluations that reflect realistic conditions, such as fluctuating renewables, demand, and markets. This approach enables systematic stress testing and provides deeper insight into the robustness, safety, and long-horizon behavior of ML-driven systems.

**Recommendation 3: Emphasize Evaluations at Multiple Scales and Long Horizons.** We encourage the community to evaluate ML-assisted DER systems across multiple temporal and spatial scales. Short-term learning performance or convergence behavior does not necessarily translate into sustained operational benefits. Evaluation protocols should therefore examine how improvements in learning metrics relate to long-term system behavior, including stability, efficiency, and resilience under non-stationary conditions. Multi-scale evaluation is particularly important for adaptive DER systems that operate continuously in evolving environments.

**Recommendation 4: Address Robustness and Deployment-Related Risks Systematically.** As ML-driven controllers enter operational DER settings, evaluation must account for robustness under uncertainty, distribution shift, and adverse conditions. We recommend explicitly incorporating robustness-oriented evaluation dimensions, such as sensitivity to forecast errors, response to unexpected disturbances, and degradation under rare but high-impact events. These aspects are often underrepresented in conventional ML evaluation but are critical for safe and reliable DER operation.

**Recommendation 5: Strengthen Evaluation Practices for RL-Based DER Systems.** RL plays a central role in DER scheduling, control, and coordination, yet evaluation often relies heavily on task-specific rewards. We recommend making the relationship between reward design, policy stability, convergence behavior, and energy-side operational outcomes explicit. Beyond cumulative reward, future RL-based studies should also report constraint satisfaction and system-level impacts to improve cross-study comparability.

In summary, our framework provides a foundation for more coherent and interpretable evaluation practices in ML-assisted DER research. By shifting the discussion from abstract future directions to concrete recommendations, this work aims to guide the community toward evaluation paradigms that better align ML-side performance metrics with DER-specific operational objectives, thereby supporting both methodological rigor and practical relevance.

## 5 Conclusions

This paper presents a structured evaluation framework for ML-assisted DER applications, bridging the gap between ML-oriented metrics and energy-system-oriented objectives. By organizing evaluation dimensions and distinguishing ML-side and energy-side criteria, the framework clarifies how commonly used metrics should be interpreted and applied in DER contexts—including cases where ML-centric metrics are overgeneralized or implicitly assumed to reflect system-level performance. Beyond metric organization, the framework also highlights data-related challenges, particularly inconsistent reporting of dataset characteristics and limited reusability of evaluation results, which undermine reproducibility and hinder the FAIR use of energy data across studies. Overall, this work lays a foundation for more rigorous, transparent, and deployment-relevant evaluation practices, supporting clearer communication between ML researchers and power system engineers working on learning-integrated DER applications. Future work will expand this concise review into a comprehensive survey, incorporating a wider range of empirical validation and emerging ML techniques.

## Acknowledgments

This work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2025-05097, the U.S. National Science Foundation grant 2330504, and the Concordia University Research Chair grant.

## References

[1] Mohammed Al-Saffar and Petr Musilek. 2021. Distributed optimization for distribution grids with stochastic der using multi-agent deep reinforcement learning. *IEEE access* 9 (2021), 63059–63072.

[2] Talal Alazemi, Mohamed Darwish, and Mohammed Radi. 2024. Renewable energy sources integration via machine learning modelling: A systematic literature review. *Heliyon* 10 (2024). doi:10.1016/j.heliyon.2024.e26088

[3] Athar Javed Ali, Rohan Kumar, Singh, Rajkamal, Priya Sandip Karemore, Megha Paliwal, and Payal Pashine. 2025. Evaluation of Key Performance Indicators (KPIs) for Enhancing Efficiency, Sustainability, and Operational Optimization in Renewable Energy Management using Artificial Intelligence and Large Language Models. *Journal of Information Systems Engineering and Management* (2025). doi:10.52783/jisem.v10i11s.1594

[4] Endris Mohammed Ali, Jemal Abawajy, Frezewd Lemma, and Samira A Baho. 2025. Analysis of deep reinforcement learning algorithms for task offloading and resource allocation in fog computing environments. *Sensors* 25, 17 (2025), 5286.

[5] Paul Arévalo and Francisco Jurado. 2024. Impact of artificial intelligence on the planning and operation of distributed energy systems in smart grids. *Energies* 17, 17 (2024), 4501.

[6] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. 2022. Dos and don'ts of machine learning in computer security. In *31st USENIX Security Symposium (USENIX Security 22)*. 3971–3988.

[7] Mateen Ashraf, Shahab Jahanbazi, and Onel LA López. 2025. Evaluating Task Execution Performance Under Energy Measurement Overhead. *arXiv preprint arXiv:2508.08757* (2025).

[8] Mithul Raaj AT, Balaji B, Sai Arun Pravin RR, Rani Chinnappa Naidu, Rajesh Kumar M, Prakash Ramachandran, Sujatha Rajkumar, Vaegae Naveen Kumar, Geetika Aggarwal, and Arooj Mubashara Siddiqui. 2024. Intelligent energy management across smart grids deploying 6G IoT, AI, and blockchain in sustainable smart cities. *IoT* 5, 3 (2024), 560–591.

[9] Selen Avcı Azkeskin and Zerrin Aladağ. 2024. Evaluating regional sustainable energy potential through hierarchical clustering and machine learning. *Environmental Research Communications* 7 (2024). doi:10.1088/2515-7620/ada2e5

[10] Sarah Barber, Yuriy Marykovskiy, and Imad Abdallah. 2024. Improving data sharing in wind energy-structural health monitoring case study. In *Journal of Physics: Conference Series*, Vol. 2767. IOP Publishing, 032007.

[11] Stella Biderman and Walter J. Scheirer. 2020. Pitfalls in Machine Learning Research: Reexamining the Development Cycle. In *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops (Proceedings of Machine Learning Research, Vol. 137)*, Jessica Zosa Forde, Francisco Ruiz, Melanie F. Pradier, and Aaron Schein (Eds.). PMLR, 106–117. https://proceedings.mlr.press/v137/biderman20a.html

[12] Mohd Bilal, Abdullah A Algethami, Salman Hameed, et al. 2024. Review of computational intelligence approaches for microgrid energy management. *IEEE Access* (2024).

[13] Juan Caballero-Peña, Cristian Cadena-Zarate, Alejandro Parrado-Duque, and G. Osma-Pinto. 2022. Distributed energy resources on distribution networks: A systematic review of modelling, simulation, metrics, and impacts. *International Journal of Electrical Power & Energy Systems* (2022). doi:10.1016/j.ijepes.2021.107900

[14] Ioanna-M Chatzigeorgiou, Christos Diou, Kyriakos C Chatzidimitriou, and Georgios T Andreou. 2021. Demand Response Alert Service Based on Appliance Modeling. *Energies* 14, 10 (2021), 2953.

[15] Gaurav Chaudhary, Jacob J Lamb, Odne S Burheim, and Bjørn Austbø. 2021. Review of energy storage and energy management system control strategies in microgrids. *Energies* 14, 16 (2021), 4929.

[16] Jingxuan Chen, Derek Yuen, Bin Xie, Yuhao Yang, Gongwei Chen, Zhihao Wu, Li Yixing, Xurui Zhou, Weiwen Liu, Shuai Wang, Rui Shao, Liqiang Nie, Yasheng Wang, Jianye HAO, Jun Wang, and Kun Shao. 2024. SPA-BENCH: A COMPREHENSIVE BENCHMARK FOR SMARTPHONE AGENT EVALUATION. In *NeurIPS 2024 Workshop on Open-World Agents*. https://openreview.net/forum?id=ikXjMk8RUs

[17] Bert J. Claessens, Peter Vrancx, and Frederik Ruelens. 2018. Convolutional Neural Networks for Automatic State-Time Feature Extraction in Reinforcement Learning Applied to Residential Load Control. *IEEE Transactions on Smart Grid* 9, 4 (2018), 3259–3269. doi:10.1109/TSG.2016.2629450

[18] Andrew Clifton, Sarah Barber, Andrew Bray, Peter Enevoldsen, Jason Fields, Anna Maria Sempreviva, Lindy Williams, Julian Quick, Mike Purdue, Philip Totaro, et al. 2023. Grand challenges in the digitalisation of wind energy. *Wind Energy Science* 8, 6 (2023), 947–974.

[19] Xiaochao Dang, Xiaoling Shu, and Fenfang Li. 2025. Dynamic Graph Attention Meets Multi-Scale Temporal Memory: A Hybrid Framework for Photovoltaic Power Forecasting Under High Renewable Penetration. *Processes* 13, 3 (2025), 873.

[20] Luca Di Persio, Matteo Garbelli, and Luca Maria Giordano. 2025. Reinforcement learning for bidding strategy optimization in day-ahead energy market. *Energy Economics* 149 (2025), 108673. doi:10.1016/j.eneco.2025.108673

[21] J. Driesen and F. Katiraei. 2008. Design for distributed energy resources. *IEEE Power and Energy Magazine* 6, 3 (2008), 30–40. doi:10.1109/MPE.2008.918703

[22] Mohammad Mahdi Forootan, I. Larki, Rahim Zahedi, and A. Ahmadi. 2022. Machine Learning and Deep Learning in Energy Systems: A Review. *Sustainability* (2022). doi:10.3390/su14084832

[23] Linqiang Guo, Wei Liu, Yi Wen Heng, Tse-Hsun Peter Chen, and Yang Wang. 2026. Agent-SAMA: State-Aware Mobile Assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 29459–29467.

[24] J. Guzmán-Henao, R. Bolaños, O. Montoya, L. Grisales-Noreña, and Harold R. Chamorro. 2024. On Integrating and Operating Distributed Energy Resources in Distribution Networks: A Review of Current Solution Methods, Challenges, and Opportunities. *IEEE Access* 12 (2024), 55111–55133. doi:10.1109/access.2024.3387400

[25] Shaowei He, Wenchao Cui, Gang Li, Hairun Xu, Xiang Chen, and Yu Tai. 2025. Intelligent Scheduling of Virtual Power Plants Based on Deep Reinforcement Learning. *Computers, Materials and Continua* 84, 1 (2025), 861–886.

[26] J. Holzinger, A. Nagl, K. Bozem, C. Lecon, Andreas Ensinger, Jannik Roessler, and Christina Neufeld. 2025. Business Case for a Regional AI-Based Marketplace for Renewable Energies. *Sustainability* (2025). doi:10.3390/su17041739

[27] Jiaheng Jing, Meiheriayi Mutailipu, Qi Wang, Qiang Xiong, Mingyao Huang, and Jiuming Zhang. 2024. Research on Intelligent Scheduling Strategy for Electric Heavy Trucks Considering Photovoltaic Outputs. *Processes* 12, 7 (2024), 1493.

[28] Abdulgani Kahraman, Mehmed Kantardzic, and Muhammed Kotan. 2022. Dynamic modeling with integrated concept drift detection for predicting real-time energy consumption of industrial machines. *IEEE Access* 10 (2022), 104622–104635.

[29] Ömer Ali Karaman. 2023. Prediction of wind power with machine learning models. *Applied Sciences* 13, 20 (2023), 11455.

[30] Anam Nawaz Khan, Naeem Iqbal, Rashid Ahmad, and Do-Hyeun Kim. 2021. Ensemble prediction approach based on learning to statistical model for efficient building energy consumption management. *Symmetry* 13, 3 (2021), 405.

[31] Seongwoo Lee, Joonho Seon, Byungsun Hwang, Soohyun Kim, Youngghyu Sun, and Jinyoung Kim. 2024. Recent trends and issues of energy management systems using machine learning. *Energies* 17, 3 (2024), 624.

[32] Ran Li, Chuanqing Pu, Junyi Tao, Canbing Li, Feilong Fan, Yue Xiang, and Sijie Chen. 2023. LLM-based frameworks for power engineering from routine to novel tasks. *arXiv preprint arXiv:2305.11202* (2023).
[33] Songjiang Li, Wenxin Zhang, and Peng Wang. 2023. TS2ARCformer: A Multi-Dimensional Time Series Forecasting Framework for Short-Term Load Prediction. *Energies* 16, 15 (2023), 5825.
[34] Bojing Liu, Mengxiang Li, Zihui Ji, Hongming Li, and Ji Luo. 2024. Intelligent productivity transformation: corporate market demand forecasting with the aid of an AI virtual assistant. *Journal of Organizational and End User Computing (JOEUC)* 36, 1 (2024), 1–27.
[35] Guangyi Liu, Pengxiang Zhao, Yaozhen Liang, Liang Liu, Yaxuan Guo, Han Xiao, Weifeng Lin, Yuxiang Chai, Yue Han, Shuai Ren, Hao Wang, Xiaoyu Liang, WenHao Wang, Tianze Wu, Zhengxi Lu, Siheng Chen, LiLinghao, Hao Wang, Guanjing Xiong, Yong Liu, and Hongsheng Li. 2025. LLM-Powered GUI Agents in Phone Automation: Surveying Progress and Prospects. *Transactions on Machine Learning Research* (2025). https://openreview.net/forum?id=yWQqoi1G1K
[36] Renke Liu, R. Piplani, and C. Toro. 2022. Deep reinforcement learning for dynamic scheduling of a flexible job shop. *International Journal of Production Research* 60 (2022), 4049 – 4069. doi:10.1080/00207543.2022.2058432
[37] Yuxuan Liu, Hongda Sun, Wei Liu, Jian Luan, Bo Du, and Rui Yan. 2025. MobileSteward: Integrating Multiple App-Oriented Agents with Self-Evolution to Automate Cross-App Instructions. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1* (2025).
[38] Grace Muriithi and Sunetra Chowdhury. 2021. Optimal energy management of a grid-tied solar pv-battery microgrid: A reinforcement learning approach. *Energies* 14, 9 (2021), 2700.
[39] Nilam B. Panchal, Vaibhavi M. Parmar, Dharmistha Makwana, Mehulsinh G. Jadeja, and Rinal K. Ahir. 2024. Enhancing Energy Efficiency in Smart Grids through Reinforcement Learning-Based Control Strategies. *Journal of Electrical Systems* 20 (05 2024), 1659–1667. doi:10.52783/jes.3660
[40] Jeya Prakash R and Anju S Pillai. 2022. Enhancing Energy Efficiency of Intensive Computing Applications using Approximate Computing. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*. 1231–1236. doi:10.1109/ICESC54411.2022.9885429
[41] A. E. Labrador Rivas and T. Abrão. 2020. Faults in smart grid systems: Monitoring, detection and classification. *Electric Power Systems Research* 189 (2020), 106602. doi:10.1016/j.epsr.2020.106602
[42] Guangchun Ruan, Haiwang Zhong, Guanglun Zhang, Yiliu He, Xuan Wang, and Tianjiao Pu. 2020. Review of learning-assisted power system optimization. *CSEE Journal of Power and Energy Systems* 7, 2 (2020), 221–231.
[43] Ahmmod Musa Sazib, Joynul Arefin, Sabab Al Farabi, Fozlur Rayhan, Md Asif Karim, and Shamima Akhter. 2025. Advancing Renewable Energy Systems through Explainable Artificial Intelligence: A Comprehensive Review and Interdisciplinary Framework. *Journal of Computer Science and Technology Studies* (2025). doi:10.32996/jcsts.2025.7.2.5
[44] Andreas Schmidt, Mohamad Anis Koubaa, Jan Schweikert, Karl-Uwe Stucky, Wolfgang Süß, and Veit Hagenmeyer. 2023. ZeitGeist: A Generic Tool Supporting the Dissemination of Time Series Data Following FAIR Principles.. In *KMIS*. 303–310.
[45] Valeria Jana Schwanitz, August Wierling, Mehmet Efe Biresselioglu, Massimo Celino, Muhittin Hakan Demir, Maria Bałazińska, Mariusz Kruczek, Manfred Paier, and Demet Suna. 2022. Current state and call for action to accomplish findability, accessibility, interoperability, and reusability of low carbon energy data. *Scientific reports* 12, 1 (2022), 5208.
[46] Jan Schweikert, Karl-Uwe Stucky, Wolfgang Süß, and Veit Hagenmeyer. 2023. A photovoltaic system model integrating FAIR digital objects and ontologies. *Energies* 16, 3 (2023), 1444.
[47] C. Silva, P. Faria, Z. Vale, and J. Corchado. 2022. Demand response performance and uncertainty: A systematic literature review. *Energy Strategy Reviews* (2022). doi:10.1016/j.esr.2022.100857
[48] V. Simankov, P. Buchatskiy, A. Kazak, S. Teploukhov, S. Onishchenko, K. Kuzmin, and Petr Chetyrbok. 2024. A Solar and Wind Energy Evaluation Methodology Using Artificial Intelligence Technologies. *Energies* (2024). doi:10.3390/en17020416
[49] Arvind R. Singh, R. S. Kumar, Mohit Bajaj, Chetan B. Khadse, and Ievgen Zaitsev. 2024. Machine learning-based energy management and power forecasting in grid-connected microgrids with multiple distributed energy sources. *Scientific Reports* 14 (2024). doi:10.1038/s41598-024-70336-3
[50] Michael Stadler. 2025. *Digital Future-Proofing: Microgrids and DERs in Data Centers*. Xendee. https://eepower.com/industry-articles/digital-future-proofing-microgrids-and-ders-in-data-centers/ Industry Article.
[51] Phillip Stoffel, Laura Maier, Alexander Kümpel, Thomas Schreiber, and Dirk Müller. 2023. Evaluation of advanced control strategies for building energy systems. *Energy and Buildings* 280 (2023), 112709.
[52] Kyle A Thompson, Amos Branch, Tyler Nading, Thomas Dziura, Germano Salazar-Benites, Chris Wilson, Charles Bott, Andrew Salveson, and Eric RV Dickenson. 2022. Detecting industrial discharges at an advanced water reuse facility using online instrumentation and supervised machine learning binary classification. *Frontiers in Water* 4 (2022), 1014556.
[53] Wen-Chang Tsai, Chia-Sheng Tu, Chih-Ming Hong, and Whei-Min Lin. 2023. A review of state-of-the-art and short-term forecasting models for solar PV power generation. *Energies* 16, 14 (2023), 5436.
[54] Deenadayalan Venkatraman and Vaishnavi Pitchaipillai. 2024. Deep Learning-Based Auto-LSTM Approach for Renewable Energy Forecasting: A Hybrid Network Model. *Traitement du Signal* 41, 1 (2024).
[55] Jaya Verma, Laura Sandys, Allan Matthews, and Saurav Goel. 2024. Readiness of artificial intelligence technology for managing energy demands from renewable sources. *Eng. Appl. Artif. Intell.* 135 (2024), 108831. doi:10.1016/j.engappai.2024.108831
[56] Marc Wenninger, Jochen Schmidt, and Toni Goeller. 2017. Appliance usage prediction for the smart home with an application to energy demand side management-and why accuracy is not a good performance metric for this problem. In *International Conference on Smart Cities and Green ICT Systems*, Vol. 2. SciTePress, 143–150.
[57] August Wierling, Valeria Jana Schwanitz, Sebnem Altinci, Maria Bałazińska, Michael J Barber, Mehmet Efe Biresselioglu, Christopher Burger-Scheidlin, Massimo Celino, Muhittin Hakan Demir, Richard Dennis, et al. 2021. Fair metadata standards for low carbon energy research—A review of practices and how to advance. *Energies* 14, 20 (2021), 6692.
[58] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.
[59] Ziming Yan, Yan Xu, Yu Wang, and Xue Feng. 2020. Deep reinforcement learning-based optimal data-driven control of battery energy storage for power system frequency support. *IET Generation, Transmission & Distribution* 14, 25 (2020), 6071–6078.
[60] Zhenpeng Yao, Yanwei Lum, Andrew K. Johnston, L. M. Mejia-Mendoza, Xiaoxia Zhou, Yonggang Wen, Alán Aspuru-Guzik, E. Sargent, and Z. Seh. 2022. Machine learning for a sustainable energy future. *Nature Reviews. Materials* 8 (2022), 202 – 215. doi:10.1038/s41578-022-00490-5
[61] Aydin Zaboli, Swetha Rani Kasimalla, Kuchan Park, Younggi Hong, and Junho Hong. 2024. A Comprehensive Review of Behind-the-Meter Distributed Energy Resources Load Forecasting: Models, Challenges, and Emerging Technologies. *Energies* 17, 11 (2024), 2534.
[62] Xi Zhang, Qiong Wang, Jie Yu, Qinghe Sun, Heng Hu, and Ximu Liu. 2023. A multi-agent deep-reinforcement-learning-based strategy for safe distributed energy resource scheduling in energy hubs. *Electronics* 12, 23 (2023), 4763.
[63] Dongzhe Zheng and Wenjie Mei. 2025. Learning Dynamics under Environmental Constraints via Measurement-Induced Bundle Structures. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=mwSBIlNLdQ
[64] Xiyuan Zhou, Huan Zhao, Yuheng Cheng, Yuji Cao, Gaoqi Liang, Guolong Liu, Wenxuan Liu, Yan Xu, and Junhua Zhao. 2024. Elecbench: a power dispatch evaluation benchmark for large language models. *arXiv preprint arXiv:2407.05365* (2024).
[65] Yuxiang Zhou, Jichang Li, Yanhao Zhang, Haonan Lu, and Guanbin Li. 2026. Mobile-Agent-RAG: Driving Smart Multi-Agent Coordination with Contextual Knowledge Empowerment for Long-Horizon Mobile Automation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 29939–29947.