International Journal of Applied Mathematics

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

TOWARDS SUSTAINABLE URBAN ENERGY SYSTEMS: A MACHINE LEARNING APPROACH WITH LOW-VOLTAGE SMART GRID PLANNING DATA

Md Shihab Sadik Shovon[1], Muhammad Hasanuzzaman[2], Saddam Hossain[3], Aashish K C[4], Badruddowza[5], Md Shohail Uddin Sarker[6], Ali Hassan[7], Mohammad Hamid Hasan Amjad[8], Md Ariful Islam Bhuiyan[9] and SK Ayub Al Wahid[10]

[1]Master of Science in Engineering Management, College of Graduate and Professional Studies, Trine University.

[2]Master's in Strategic Communication, Gannon University, Erie, PA

[3]Master of Public Administration, Gannon University, Erie, PA, USA

[4]Master of Science in Computer and Information Science(Software Engineering), Gannon University, Erie, PA, USA

[5]Master of Science in Computer and Information Systems, Gannon University, Erie, Pa

[6]MS- Computer and Information Systems Security, Gannon University, Pa, USA.

[7]MBA Business Analytics, International American University, Los Angeles, California

[8]Master of Science in Engineering Management, College of Graduate and Professional Studies, Trine University.

[9]Master of Science in Business Analytics, University of North Texas (UNT), Denton, Texas.

[10]Department: Doctor of Business Administration, International American University

Corresponding Author: **Md Shihab Sadik Shovon, Email:** mshovon23@my.trine.edu

Abstract

Urban energy systems in the U.S. are facing growing stress from the integration of distributed energy resources (DERs), where variability in rooftop solar generation, electric vehicle charging, and household consumption complicates low-voltage (LV) grid operation and planning. Conventional power-flow–based hosting capacity assessments are computationally intensive and deterministic, providing limited insight into the uncertainty that governs modern grid behavior. This study develops a data-driven, uncertainty-aware framework that leverages operational LV grid data, including load profiles, transformer characteristics, and network topology, to estimate node-level hosting capacities with quantified confidence intervals. Using probabilistic machine learning models based on quantile regression and ensemble learning, the framework captures nonlinear relationships between load, generation, and topology, while expressing prediction uncertainty that is critical for planning under variability. Experiments conducted on real LV planning data demonstrate that the probabilistic models achieve superior predictive accuracy and well-calibrated uncertainty coverage compared to deterministic

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
baselines, enabling planners to balance DER integration potential against voltage and capacity constraints. The findings demonstrate that uncertainty-aware machine learning substantially reduces computational overhead while maintaining reliability, offering a scalable approach for adaptive, risk-informed grid planning. This approach advances the vision of sustainable, data-driven urban energy systems capable of supporting resilient and efficient renewable integration at scale.
Keywords: Sustainable Energy, Smart Grids, Machine Learning, Hosting Capacity, Uncertainty Quantification, Low-Voltage Networks, Renewable Integration

1. Introduction

1.1 Background and Motivation

Urban energy systems in the U.S. are undergoing a rapid and complex transformation driven by the global push toward decarbonization, electrification, and the integration of distributed energy resources (DERs) such as rooftop solar photovoltaic systems, small-scale wind turbines, and battery storage. These developments have created a paradigm shift in how cities generate, distribute, and consume electricity, demanding more adaptive, data-driven, and resilient approaches to grid planning and operation. As renewable penetration deepens, the variability and intermittency inherent in renewable energy sources introduce new challenges in maintaining voltage stability, reliability, and overall grid efficiency. This shift from centralized to distributed generation requires rethinking traditional models of energy infrastructure, control, and management. Strbac et al. (2019) emphasize that the decarbonization of electricity systems in Europe is not merely a technological transformation but also a regulatory and market design challenge, where flexibility and coordination across actors become crucial for sustainable urban systems [21].

The increased adoption of DERs has particularly impacted low-voltage (LV) distribution networks, which serve as the interface between utilities and end consumers. LV networks, originally designed for unidirectional power flow, are now expected to handle bidirectional flows resulting from prosumers injecting energy back into the grid. Zambrano-Asanza et al. (2021) argue that such transformations expose LV systems to new operational stresses, including voltage deviations, transformer overloading, and reverse power flow conditions, all of which complicate traditional planning methodologies [25]. These operational stresses are compounded by the inherent stochasticity of renewable generation and the diversity of urban demand patterns, making deterministic power-flow analyses increasingly inadequate. Günther and Schiffer (2023) further highlight that uncertainty in DER behavior, particularly under limited observational data, remains one of the most pressing challenges in modern grid analytics, emphasizing the need for probabilistic or data-driven uncertainty modeling frameworks [8].

Traditional grid planning approaches, which rely on static snapshots and worst-case scenarios, are no longer sufficient in environments characterized by dynamic consumption, distributed generation, and real-time operational constraints. Zepter et al. (demonstratenstrate) show that

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

as flexibility becomes scarce due to high variability in renewable generation, energy systems experience not only higher price volatility but also increased emissions when inflexible resources must cotheeewable intermiof renewables tof renewables tency [26]. This highlights that effective uncertainty modeling is not only an engineering problem but also a socio-economic imperative for achieving sustainable decarbonization. Thus, the emergence of advanced data analytics and machine learning has opened a pathway toward new grid planning paradigms that can dynamically assess, forecast, and optimize system behavior under uncertainty. By leveraging historical telemetry, network topology, and operational metadata, machine learning models can infer complex nonlinear dependencies between load, generation, and voltage response, offering planners tools for faster and more adaptive decision-making than conventional power-flow simulations.

Nevertheless, the transition toward data-driven planning introduces its own challenges, particularly concerning data quality, granularity, and representativeness. Distribution networks often lack comprehensive monitoring at the LV level, limiting the availability of reliable datasets for model training and validation. Günther and Schiffer (2023) note that even with high-resolution smart meter data, temporal gaps, synchronization errors, and missing contextual information (such as weather or control actions) can introduce significant uncertainty into model predictions [8]. Therefore, any effective approach to sustainable urban grid planning must not only adopt advanced modeling techniques but also integrate rigorous uncertainty quantification mechanisms that explicitly express confidence in each prediction. This enables risk-aware decision-making, an essential step toward resilient and sustainable energy infrastructures.

1.2 Research Problem

Despite substantial progress in renewable integration and digitalization, existing LV grid planning frameworks remain largely deterministic and limited in their ability to address uncertainty. Most conventional planning tools rely on fixed power-flow models calibrated to specific load or generation scenarios. Such methods assume static system states and fail to capture the probabilistic nature of distributed generation, consumption, and equipment behavior in real-world conditions. Zambrano-Asanza et al. (2021) highlight that under high DER penetration, LV feeders exhibit highly variable voltage and current profiles across both spatial and temporal dimensions, making single-scenario analyses unreliable for operational or investment planning [25]. Traditional power-flow models, while physically grounded, become computationally prohibitive when planners must evaluate multiple stochastic scenarios or integrate high-resolution time-series data. Günther and Schiffer (2023) note that this limitation is especially problematic in urban contexts, where system conditions can change rapidly due to heterogeneous consumer behavior, weather variability, and uncoordinated DER dispatch [8]. Furthermore, deterministic methods provide planners with single-point estimates of system performance metrics, such as node voltage or hosting capacity, without expressing the uncertainty associated with those estimates. This lack of uncertainty quantification leads to risk-blind decision-making, where infrastructure upgrades or DER approvals are based on incomplete information. Zepter et al. (2020) argue that ignoring variability in system behavior

Received: August 13 2025

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

can lead to inefficient market outcomes and suboptimal operational decisions, where planners either overinvest in infrastructure or operate grids too conservatively, reducing DER utilization potential [26]. Strbac et al. (2019) reinforce this by suggesting that efficient grid management under high-renewable conditions requires dynamic flexibility assessment, a task that deterministic models cannot effectively support [21].

The complexity of LV grid planning further increases when multi-source data, such as load measurements, transformer specifications, and network topology, must be integrated coherently for analysis. The heterogeneity and dimensionality of such data challenge traditional analytical workflows, which are not designed for automated feature synthesis or nonlinear relationship modeling. Machine learning provides a promising avenue for addressing these issues by learning latent patterns directly from data, but without an explicit probabilistic framework, even ML-based models can produce misleadingly confident predictions. Therefore, integrating uncertainty quantification within data-driven models becomes a crucial step toward reliable, risk-aware grid analytics. Günther and Schiffer (2023) emphasize that probabilistic ML frameworks capable of learning predictive distributions rather than point estimates are better suited for LV planning tasks involving incomplete or noisy data [8]. Ultimately, the current state of grid planning suffers from three interrelated limitations: (i) poor scalability of deterministic power-flow methods, (ii) lack of uncertainty representation in planning metrics such as hosting capacity, and (iii) absence of real-time adaptability required for iterative scenario evaluation. Addressing these gaps demands an integrated framework that combines the predictive efficiency of machine learning with the rigor of uncertainty quantification. Such a framework would enable planners to identify not only the most likely hosting capacity for each node but also the associated confidence interval, thus supporting more informed infrastructure and policy decisions.

1.3 Importance of This Research

The sustainability of urban energy systems hinges on more than renewable adoption; it depends on the capacity to manage and adapt to uncertainty in distributed generation, load dynamics, and network performance. The growing complexity of LV grids necessitates analytical tools that move beyond deterministic evaluations to probabilistic decision-making frameworks capable of capturing and quantifying variability. This research contributes to that vision by developing an uncertainty-aware machine learning framework that fuses operational data, network topology, and transformer characteristics to predict hosting capacities with calibrated confidence intervals. Unlike conventional power-flow methods that treat uncertainty as an external factor, this approach embeds it directly into the modeling process, producing predictions that are both accurate and risk-informed.

By doing so, this study bridges a critical methodological gap between physics-based modeling and data-driven intelligence. It establishes that sustainable grid planning can be achieved not only through improved forecasting accuracy but through explicit uncertainty calibration, ensuring that planners understand the reliability and limitations of every estimate. The approach proposed here offers a scalable alternative for utilities and energy planners, capable of handling large datasets and supporting real-time scenario analysis without sacrificing

Received: August 13 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

computational efficiency. Furthermore, it provides a foundation for integrating sustainability objectives directly into operational decision-making, allowing for optimized renewable integration while maintaining grid stability. In essence, this research lays the groundwork for next-generation LV grid planning systems that are sustainable, adaptive, and robust under uncertainty, advancing the transition toward intelligent, data-driven urban energy infrastructures.

1.4 Research Objectives and Contributions

The primary objective of this research is to develop a robust, data-driven framework that estimates node-level hosting capacities in low-voltage smart grids while explicitly quantifying prediction uncertainty. To accomplish this, the study aims to construct a machine learning pipeline that ingests heterogeneous LV planning data, time-series load and DER measurements, transformer specifications, and network topology, and produces probabilistic hosting capacity forecasts rather than single-point estimates. A second objective is to embed uncertainty quantification into the modeling process through probabilistic learning techniques, enabling the production of calibrated prediction intervals that planners can use to assess risk when approving DER connections or scheduling network reinforcements. Third, the framework is evaluated against deterministic baselines using real LV planning data and time-aware validation strategies to demonstrate comparative performance and calibration benefits. Finally, the work seeks to translate the research outputs into a practical deployment blueprint, including data engineering, model serving, and monitoring components, so that the approach can be operationalized within utility or municipal planning environments.

The contributions of this paper are fourfold and designed to advance both methodology and practice in sustainable grid planning. First, the study provides a complete data preprocessing and modeling pipeline tailored to LV grid datasets, integrating load, transformer, and topology information into a unified analysis table and reproducible preprocessing artifacts. Second, it introduces an uncertainty-aware hosting capacity prediction model based on quantile regression, which generates median forecasts and well-defined prediction intervals that encapsulate the aleatoric variability inherent in DER and load behavior. Third, the paper delivers empirical validation demonstrating that probabilistic models yield actionable advantages in calibration and risk quantification when compared with deterministic baselines, showing how prediction intervals can guide safer and more efficient DER integration decisions. Fourth, the research produces a deployment-ready framework and architecture blueprint that includes feature computation, containerized model serving, and drift-aware monitoring, thereby offering a practical pathway for utilities and planners to adopt uncertainty-conscious hosting capacity estimation in real-world operations. Together, these objectives and contributions establish a scalable, interpretable, and operationally relevant approach to supporting sustainable urban energy transitions.

Received: August 13 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2. Literature Review

2.1 Sustainable Urban Energy Systems

The global transition toward sustainable urban energy systems has accelerated in response to climate imperatives, technological progress, and the pursuit of energy independence. Urban centers are becoming both the largest consumers and emerging producers of energy due to widespread adoption of distributed energy resources (DERs) such as photovoltaic panels, wind microturbines, and electric vehicles. This transition marks a paradigm shift from centralized generation toward decentralized, multi-energy configurations often termed "smart energy systems." Lund et al. (2017) define smart energy systems as those that integrate electricity, heating, cooling, and transportation sectors through intelligent management of resources, ensuring efficiency and flexibility across multiple vectors [16]. They argue that such systems are vital for realizing carbon neutrality in urban contexts where energy demand is dense, diverse, and dynamic. However, while these systems promise efficiency gains and resilience, they also introduce complex interdependencies between energy carriers and infrastructures, increasing operational uncertainty. The global push towards decarbonization and the integration of distributed energy resources (DERs) is creating a paradigm shift in urban energy systems [4]. This transformation is being driven by analyzable trends in energy generation, where machine learning is increasingly used to comprehend national-level capacity and integration patterns (Chouksey et al., 2025).

Li et al. (2021) elaborate on the analytical and computational challenges associated with this transition, noting that traditional urban energy models, typically linear and deterministic, struggle to capture the stochastic interactions between energy demand, distributed generation, and network constraints [15]. Their review emphasizes that urban energy system transitions are not solely technological evolutions but require multi-dimensional coordination across planning, operation, and policy layers. The need for intelligent decision-support frameworks, capable of dynamically balancing supply and demand across heterogeneous resources, is therefore paramount. Furthermore, Li et al. (2021) highlight that the granularity and temporal resolution of data in urban systems have grown exponentially with the proliferation of smart meters and IoT devices, enabling new data-driven optimization paradigms but also exposing gaps in conventional modeling capabilities [15]. Also, the variability and intermittency inherent in renewable energy sources introduce new challenges for grid stability. Consequently, AI-driven forecasting of renewable energy production has become critical for understanding and preparing for these integration pressures at a systemic level (Shovon et al., 2025) [20].

A significant pressure point in this transformation lies in low-voltage (LV) distribution networks, which serve as the interface between centralized utilities and distributed end-users. These networks were not originally designed for bidirectional power flows or variable injection from prosumers. As renewable penetration increases, voltage regulation, congestion management, and hosting capacity assessment become major concerns for utilities and policymakers. Lund et al. (2017) note that smart energy systems can only achieve sustainability when complemented by adaptive planning frameworks that reflect real-time operational uncertainty [16]. Also, effective management of urban energy systems under high renewable

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

penetration requires intelligent forecasting and optimization not only of generation but also of consumption, a task for which AI-powered approaches are proving highly effective (Chowdhury, 2025) [5]. The rising interdependence between distributed generation, storage, and consumption reinforces the necessity of machine learning and data analytics for predictive and adaptive grid management. Chouksey et al. (2025) provided an in-depth analysis of US energy generation and capacity trends using advanced machine learning, which aligns with the motivation for using probabilistic approaches in this study. Consequently, sustainable urban energy systems are evolving into data-rich yet uncertainty-prone environments where probabilistic modeling, intelligent forecasting, and automated decision-making are indispensable to achieving operational resilience and long-term sustainability.

2.2 Conventional Power-Flow-Based Hosting Capacity Analysis

Traditional hosting capacity assessment in distribution systems has relied heavily on deterministic power-flow methods rooted in physical modeling. These methods simulate voltage and current behavior under predefined load and generation scenarios to estimate the maximum DER penetration that can be accommodated without violating technical constraints. Coster et al. (2011) provide an early account of the integration challenges posed by distributed generation, noting that LV networks were historically designed for unidirectional energy flows and limited observability, leading to inaccuracies when used for modern DER planning [6]. They argue that deterministic models, while accurate under known conditions, fail to represent the variability introduced by intermittent generation and stochastic demand patterns. Keane et al. (2013) discuss the state-of-the-art techniques for distributed generation planning, emphasizing optimization-based frameworks that aim to balance reliability and economic efficiency [11]. Despite their theoretical robustness, such methods remain computationally demanding, particularly when extended to large-scale networks or high temporal resolution datasets. Their research shows that scenario-based deterministic optimization quickly becomes intractable as the number of uncertain parameters increases. Furthermore, these models typically require detailed system parameters that may not be available at the LV level, limiting their practical applicability.

Ochoa and Padilha-Feltrin (2018) extend this discussion by demonstrating that advanced distribution network planning must consider dynamic operational conditions to achieve accurate DER hosting assessments [17]. They highlight that voltage excursions, reverse power flows, and unbalanced loading conditions are highly time-dependent and cannot be effectively represented by static or worst-case deterministic simulations. As renewable penetration grows, deterministic hosting capacity results become increasingly unreliable, providing either overly conservative or dangerously optimistic estimates depending on scenario assumptions. Moreover, such methods offer no mechanism for quantifying uncertainty in their predictions, leaving planners unable to gauge the confidence of their results. Recent advancements in computational tools have improved the scalability of deterministic simulations, but their fundamental limitation remains: they treat uncertainty as external noise rather than as an intrinsic part of system behavior. Consequently, while deterministic hosting capacity studies have been instrumental in shaping early DER integration strategies, they are insufficient for

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

modern, data-intensive LV networks. The lack of uncertainty quantification in these approaches represents a critical gap between simulation-based methods and emerging probabilistic or machine learning techniques capable of learning from real operational data. As energy systems become more dynamic, the field increasingly recognizes that deterministic analyses alone cannot provide the agility and insight required for real-time planning and sustainable urban grid management.

2.3 Machine Learning in Power System Planning

Machine learning (ML) has emerged as a transformative tool for energy system planning, particularly in handling large-scale, heterogeneous data that traditional models struggle to interpret. In power systems, ML techniques have been applied across a range of tasks, from load forecasting and voltage stability assessment to fault prediction and distributed energy optimization. Lago et al. (2018) demonstrated that deep learning models, particularly recurrent neural networks and their variants, can significantly outperform traditional statistical approaches in day-ahead electricity price forecasting by capturing nonlinear temporal dependencies in energy data [14]. Their results indicate that the combination of historical consumption, weather, and market variables allows ML models to learn complex interdependencies that deterministic methods overlook. Similarly, Hafeez et al. (2020) provide a comprehensive review of ML and optimization techniques for smart distribution network operation, identifying the growing trend of hybrid approaches that merge physics-based models with data-driven intelligence [9]. They argue that such hybridization enables both scalability and interpretability, two characteristics essential for real-world deployment. Their study further observes that ML models facilitate autonomous decision-making in scenarios where analytical models are too rigid or computationally demanding. However, while ML's predictive performance in operational forecasting is well established, its application to planning tasks such as hosting capacity estimation remains in early stages of development.

Zhang et al. (2021) address this gap directly, presenting a data-driven approach for distribution system hosting capacity estimation using machine learning [27]. Their model leverages historical operational data to infer hosting capacity without requiring repeated power-flow simulations, thereby drastically reducing computational cost. However, they note that most existing ML frameworks remain deterministic, providing point estimates of hosting capacity without uncertainty intervals, limiting their usefulness in risk-aware planning contexts. This observation ushers in the next frontier in ML-based grid analytics: integrating uncertainty quantification to ensure that predictions are not only accurate but also statistically reliable. Machine learning techniques have demonstrated significant utility in tackling the core challenges of renewable integration, such as using time-series analysis to optimize solar energy production (Ahmed et al., 2025), which directly addresses the variability of DERs in LV networks [1]. The application of ML in power systems also extends beyond forecasting to critical operational tasks like predictive maintenance and fault detection, enhancing the overall reliability of energy infrastructure (Amjad et al., 2025), a concern that is also paramount in LV grid management [3]. Collectively, these studies illustrate the evolution of ML from a forecasting tool into a foundational element of modern power system planning. Yet, a

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

consistent limitation persists: the absence of uncertainty modeling within most ML frameworks. While models like those of Zhang et al. (2021) [27] demonstrate computational efficiency and predictive strength, their deterministic nature restricts their ability to support robust decision-making under stochastic conditions. This limitation motivates the integration of probabilistic ML techniques, such as quantile regression and Bayesian learning, into LV hosting capacity estimation frameworks, where uncertainty plays a central operational role.

2.4 Uncertainty Quantification in Energy Forecasting

Uncertainty quantification has become an increasingly vital dimension of modern energy forecasting and planning, addressing the limitations of traditional deterministic approaches that produce single-point predictions. Hong and Fan (2016) define probabilistic forecasting as the process of estimating not just the expected value of a target variable but the entire predictive distribution, allowing decision-makers to assess risks, confidence levels, and potential extremes [10]. They demonstrate that in energy systems, where demand, weather, and generation fluctuate unpredictably, probabilistic models enable more informed and flexible operational planning. Their work illustrates how ensemble and distributional methods improve situational awareness for grid operators by providing prediction intervals that represent confidence bounds rather than fixed outcomes. Ziel and Weron (2018) expand on this perspective by exploring high-dimensional electricity price forecasting and the incorporation of regularization techniques to manage overfitting in probabilistic frameworks [28]. Their findings show that uncertainty quantification enhances model interpretability and robustness, especially when input variables exhibit strong interdependencies, as is typical in power systems. Quantifying uncertainty is not only valuable for forecasting prices but also crucial for planning tasks like hosting capacity estimation, where the consequences of over- or underestimation are operationally significant.

Quan et al. (2014) provide a seminal example of uncertainty-aware modeling using interval type-2 fuzzy neural networks for load and wind power forecasting [18]. Their approach integrates fuzzy logic with neural networks to express both aleatoric (data-driven) and epistemic (model-driven) uncertainty, resulting in more reliable forecasts under data sparsity and nonlinearity. They demonstrate that incorporating uncertainty directly into the learning process significantly improves model generalization and resilience to input noise. The broader implication of such studies is clear: probabilistic modeling frameworks can be extended beyond forecasting to planning applications, where risk and uncertainty directly affect system reliability and investment decisions. The shift toward uncertainty-aware modeling represents a necessary evolution for modern grid analytics. While deterministic machine learning models provide accuracy under known conditions, probabilistic frameworks offer a more realistic representation of system behavior under uncertainty. In the context of LV network planning, this means predicting not only the most likely hosting capacity but also the range of plausible outcomes under varying load and DER conditions. This probabilistic insight empowers planners to make risk-informed decisions, balancing system reliability, efficiency, and renewable integration targets, thereby aligning operational strategies with long-term sustainability objectives.

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2.5 Identified Research Gap

Despite the growing literature on machine learning and uncertainty quantification in energy systems, relatively few studies have combined these domains to model hosting capacity uncertainty using real LV operational data. Most research on DER integration still relies on deterministic or simulation-heavy frameworks that fail to incorporate stochasticity in load, generation, or voltage dynamics. Yan et al. (2020) address this limitation by developing a data-driven probabilistic hosting capacity estimation method that explicitly considers spatial uncertainty in LV networks [24]. Their study shows that data-driven probabilistic methods can capture heterogeneity across feeders and nodes more effectively than purely deterministic approaches. However, they also acknowledge that large-scale adoption of such frameworks requires systematic integration of uncertainty quantification techniques, such as quantile regression, into standard grid planning workflows.

Building upon these insights, the present research identifies a critical methodological gap: the absence of scalable, uncertainty-aware machine learning frameworks capable of predicting hosting capacities directly from operational LV data. While prior studies have advanced both deterministic ML models and probabilistic forecasting methods, few have combined the two in a way that is computationally efficient and operationally interpretable. Furthermore, existing probabilistic hosting capacity studies often rely on synthetic or simulated datasets, limiting their real-world applicability. This paper addresses these shortcomings by developing a quantile-based machine learning framework trained on real LV grid planning data. The model not only predicts node-level hosting capacities but also provides calibrated confidence intervals, offering a transparent measure of prediction reliability. By integrating uncertainty quantification into data-driven hosting capacity estimation, this research contributes to bridging the divide between theoretical advances in probabilistic modeling and the practical needs of utility planners. The outcome is a reproducible, scalable, and interpretable approach to LV grid planning that can inform real-time decision-making under uncertainty. As urban energy systems evolve toward higher DER penetrations and increasing data availability, frameworks like this will become essential for ensuring both the operational reliability and sustainability of modern smart grids.

3. Methodology

3.1 Dataset Description

This study utilized three complementary datasets, load profiles, transformer specifications, and network topology, that collectively represent a realistic low-voltage (LV) distribution network suitable for advanced planning, operational analytics, and hosting capacity estimation. Each dataset contributed distinct yet interrelated information, providing a comprehensive picture of both the operational dynamics and physical characteristics of the system. The load profiles dataset captures the temporal behavior of electricity consumption and generation across multiple nodes in the network. Each record includes timestamped measurements of Load_kW and DER_Generation_kW at a 30-minute resolution over a continuous one-week period. This fine-grained temporal resolution enables analysis of diurnal consumption cycles, peak demand

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

occurrences, and distributed generation patterns across the network. The inclusion of distributed energy resource (DER) generation data allows assessment of bidirectional power flow effects, an increasingly critical aspect of LV grid studies, given the proliferation of small-scale photovoltaic and wind systems in urban settings. These time series data provide the foundation for temporal modeling and probabilistic estimation of hosting capacities under variable operating conditions.

The transformer specifications dataset documents the static electrical and operational parameters of transformers connecting various feeders within the LV system. Each transformer record includes attributes such as capacity (kVA), efficiency (%), tap position, and primary and secondary voltage levels (Primary_V and Secondary_V). These parameters are instrumental in understanding local voltage regulation capacity, reactive power compensation potential, and overall transformer loading behavior. By linking these specifications to nodal data, the model can account for each node's upstream electrical influence, thus improving prediction accuracy for node-specific hosting capacity estimation. Additionally, transformer efficiency and capacity metrics serve as upper-bound constraints in the probabilistic models, establishing physically interpretable boundaries for ML-based predictions. The network topology dataset describes the electrical connectivity and spatial configuration of the distribution network. It is represented as an edge list where each row denotes a line segment connecting two nodes, with associated electrical parameters including line length (Length_m), resistance (Resistance_ohm), reactance (Reactance_ohm), and capacity (Capacity_A). This information is crucial for constructing the network's impedance matrix and for deriving graph-theoretic features such as distance to transformer, node degree, and aggregate impedance to the source.

These topological features enrich the dataset by embedding the network's structural dependencies into the learning framework, allowing the ML model to approximate spatial voltage behaviors and congestion patterns that are otherwise captured only through power-flow simulations. Integrating the three datasets produced a unified analytical environment combining operational time-series data with static electrical and structural metadata. The merged dataset thus serves as a digital twin of the LV network, capable of supporting time-dependent predictive modeling, uncertainty quantification, and capacity planning studies. The comprehensive integration of operational, physical, and topological data allows the proposed framework to generalize beyond simple temporal forecasting, enabling holistic, data-driven optimization of grid performance under uncertainty.

3.3 Data Cleaning and Preprocessing

The raw datasets were subjected to a rigorous, multi-stage cleaning and preprocessing pipeline designed to ensure quality, consistency, and analytical readiness before model development. Since the datasets originated from different measurement systems and varied in granularity, harmonization was essential to create a unified analytical base for the uncertainty-aware machine learning framework. Each preprocessing stage was documented to maintain reproducibility and transparency, with all transformations and metadata recorded for auditability. The process began with a comprehensive file inspection phase. Here, row counts, data types, timestamp formats, and node identifiers were systematically verified across the

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

three source files: load_profiles.csv, network_topology.csv, and transformers.csv. This inspection confirmed structural consistency and revealed no major schema mismatches, establishing a sound foundation for subsequent integration. Column metadata were extracted to form the basis of a feature dictionary, ensuring traceability of all feature engineering and transformation steps applied later in the pipeline.

The next step involved timestamp alignment. All timestamps within the load profiles dataset were converted to timezone-aware datetime objects and standardized to Coordinated Universal Time (UTC). Sampling frequency analysis confirmed a consistent 30-minute interval, which is optimal for capturing both diurnal load variability and fast-changing DER generation profiles. Misaligned or duplicate timestamps were identified and corrected through time-index normalization, ensuring temporal coherence across all nodes. This step was essential for maintaining continuity in time-series modeling and walk-forward validation procedures used in later experimental stages. Missing data handling followed as a critical quality control phase. Short-term gaps (up to two consecutive missing intervals, equivalent to one hour) in numerical time-series fields were interpolated using a time-weighted linear method, preserving the smooth temporal structure of the data. Longer gaps or entire missing segments were flagged rather than imputed to prevent introducing artificial bias. A missingness report was automatically generated, detailing the proportion of missing data per node and recording the methods used for treatment. This report was later included in the study's documentation to support model reproducibility and transparency.

To address potential measurement or logging anomalies, outlier detection and handling were conducted on key variables such as Load_kW and DER_Generation_kW. A Median Absolute Deviation (MAD) filter was applied on a per-node basis to identify extreme deviations from expected behavior. Additionally, physical validation rules were used to detect implausible values such as negative consumption (beyond allowable DER export) or excessively high voltage readings. The detected outliers were visualized through boxplots and time-series plots for manual inspection, and the results were summarized in an outlier report, ensuring that data corrections were traceable and justified. Subsequent to quality checks, a dataset merging step was executed to integrate the cleaned time-series, transformer, and network datasets. The merging was performed via Node_ID, aligning each node's load and generation data with its topological and transformer characteristics. This produced a single composite dataset referred to as the main analysis table, containing both dynamic and static features required for model input. From the topology dataset, additional derived features, such as total upstream impedance, distance to substation, and transformer loading ratio, were computed to serve as explanatory variables during model training.

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

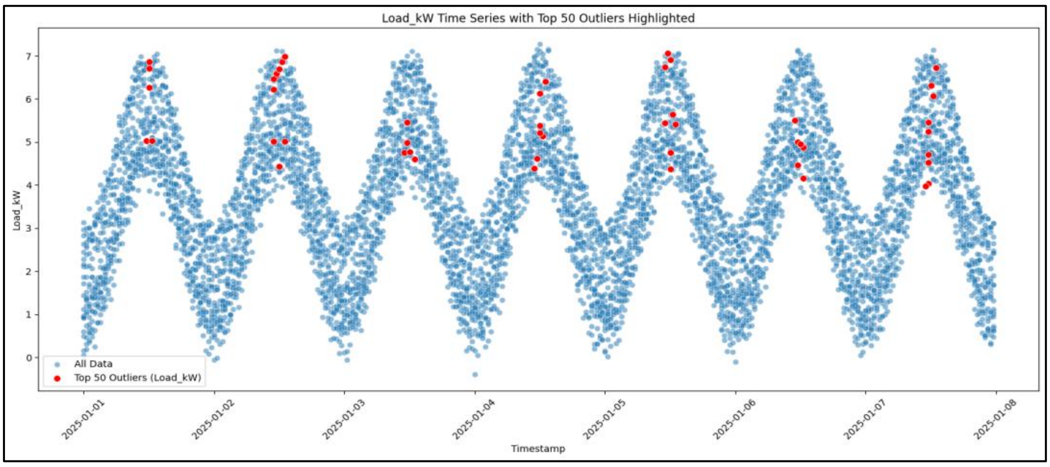

Fig.1: Top 50 outliers Load_kW

Finally, scaling and encoding were prepared for numerical and categorical variables, respectively. Although tree-based models like LightGBM can handle raw numerical data, standardization was applied for consistency across experimental setups, particularly for quantile regression and ensemble learning models. Categorical attributes, such as potential Load_Type classifications, were encoded as necessary, ensuring model compatibility. A label generation process defined the primary target variables for supervised learning. For the voltage violation prediction subtask, a binary label, voltage_violation_in_horizon, was created by checking whether node voltage exceeded ±5% of nominal levels (derived from transformer voltage ratings) within a specified prediction horizon of one hour (two consecutive intervals). Hosting capacity labels were generated synthetically for demonstration purposes, using rule-based thresholds derived from voltage and transformer constraints. Upon completion, the preprocessing pipeline produced several key deliverables, including: (1) Cleaned and merged datasets stored in CSV and Parquet formats; (2) A missingness and outlier audit report summarizing data integrity findings; and (3) Serialized preprocessing pipeline artifacts saved as .joblib files for reuse during model inference. Together, these preprocessing steps ensured that the dataset entering the uncertainty-aware machine learning framework was not only statistically sound but also operationally meaningful, preserving both temporal coherence and physical realism.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was undertaken to extract insights from the integrated dataset, examine variable behavior, and reveal underlying temporal, spatial, and topological dynamics influencing hosting capacity estimation. The analyses focused on understanding consumption and generation variability, correlations between key electrical parameters, and network spatial structures. These findings directly guided feature engineering and model design, especially for uncertainty quantification in hosting capacity predictions. The temporal analysis of aggregated feeder load and distributed energy resource (DER) generation data revealed pronounced cyclical patterns over the one-week observation period. The load time series exhibited clear daily fluctuations characterized by two primary peaks, one in the morning and another during the evening hours, consistent with typical residential and commercial

Received: August 13 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

activity patterns in urban LV systems. Conversely, DER generation followed a strong diurnal cycle, with peak outputs occurring around midday and near-zero values during nighttime hours. This temporal complementarity between load demand and DER supply underscores one of the fundamental operational challenges in modern LV grids: the increasing intermittency of net load due to DER integration. The non-stationary nature of these time series further reinforces the necessity of incorporating temporal dependencies and probabilistic models that can accommodate variability rather than deterministic point estimates. These observations justify the inclusion of time-based features such as hour-of-day, day-of-week, and rolling mean statistics in the modeling pipeline.

Fig.2: Temporal load trends

Received: August 13 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

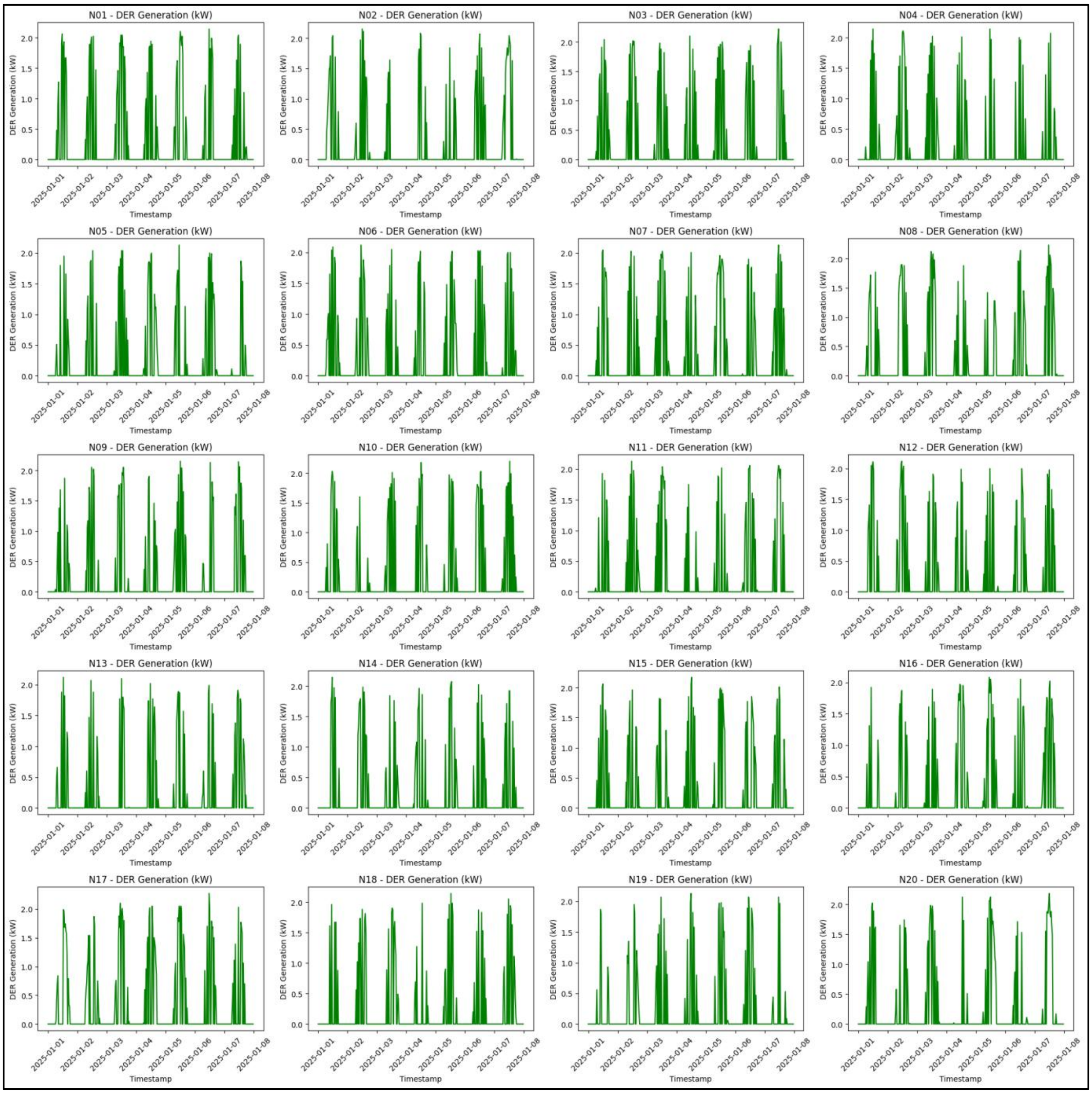

Fig.3: DER generation trends

The diurnal and weekly load patterns provided deeper insights into the behavioral rhythm of the grid. When aggregated across all nodes, weekday load profiles tended to display more consistent peaks, whereas weekend profiles were flatter and more variable, possibly due to altered human activity and reduced commercial demand. DER generation patterns, in contrast, remained largely consistent across the week, dominated by solar irradiance cycles that follow predictable astronomical rhythms. This combination of predictable renewable output and variable consumption behavior amplifies short-term voltage and power flow uncertainty, directly impacting hosting capacity margins. Capturing these relationships through engineered temporal features enables the predictive model to generalize better under daily and weekly variations, thus improving its reliability in real-world planning scenarios.

Received: August 13 2025

International Journal of Applied Mathematics

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

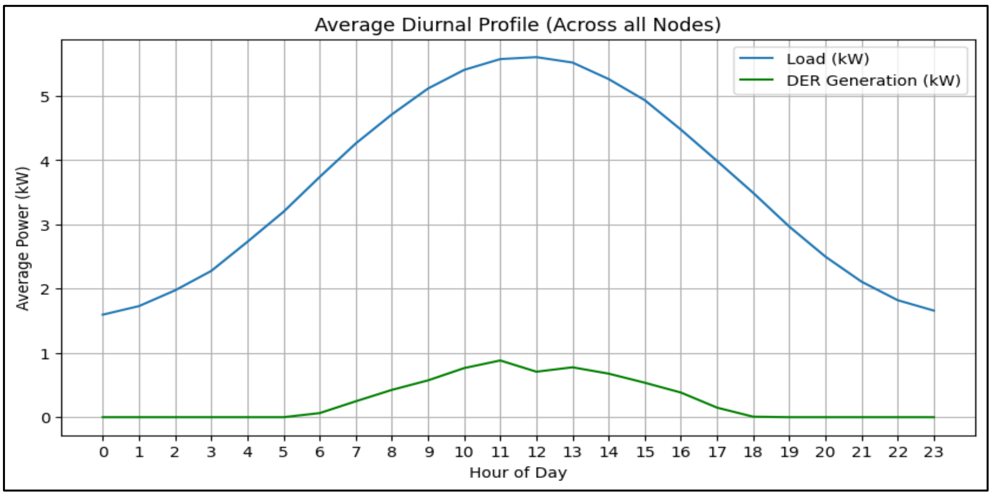

Fig.4: Durnal load patterns

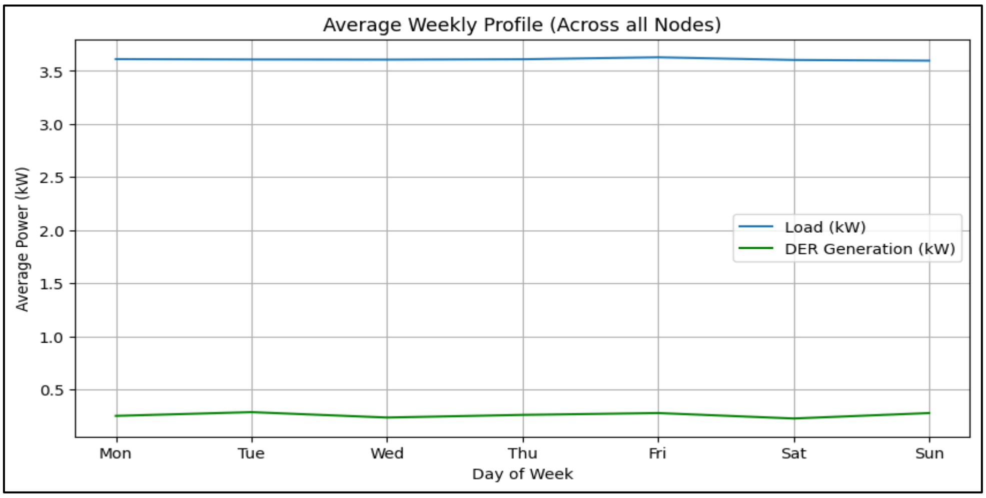

Fig.5: Weekly load patterns

Distributional analyses using boxplots and violin plots for Load_kW and DER_Generation_kW across individual nodes illustrated the inherent heterogeneity of the network. While some nodes exhibited narrow load distributions with stable demand, others showed wide variance indicative of dynamic consumption behavior or high DER penetration. This non-uniformity across nodes highlights that hosting capacity is spatially dependent; nodes situated closer to transformers or with lower impedance paths tend to exhibit more stable voltage profiles than distant or heavily loaded segments. Furthermore, the voltage data, although not visualized as a histogram, showed distributions centered near nominal voltage levels with moderate spread, confirming that the system operates largely within regulatory bounds. However, occasional outliers, corresponding to voltage deviations exceeding permissible limits, suggest localized stress points in the network. These findings provided a

Received: August 13 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
quantitative foundation for defining the voltage violation threshold used in the target label generation for the classification component of this study.

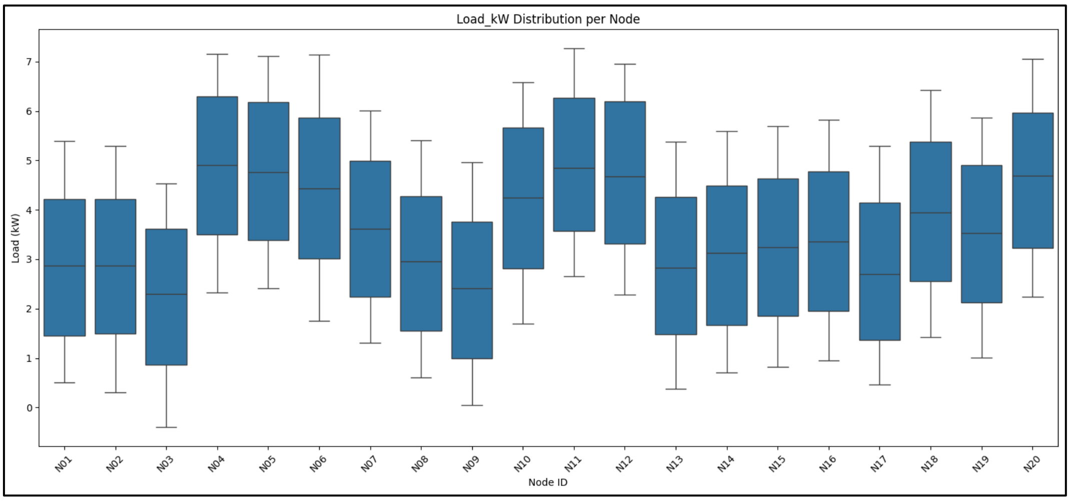

Fig.6: Load_kW distribution per Node

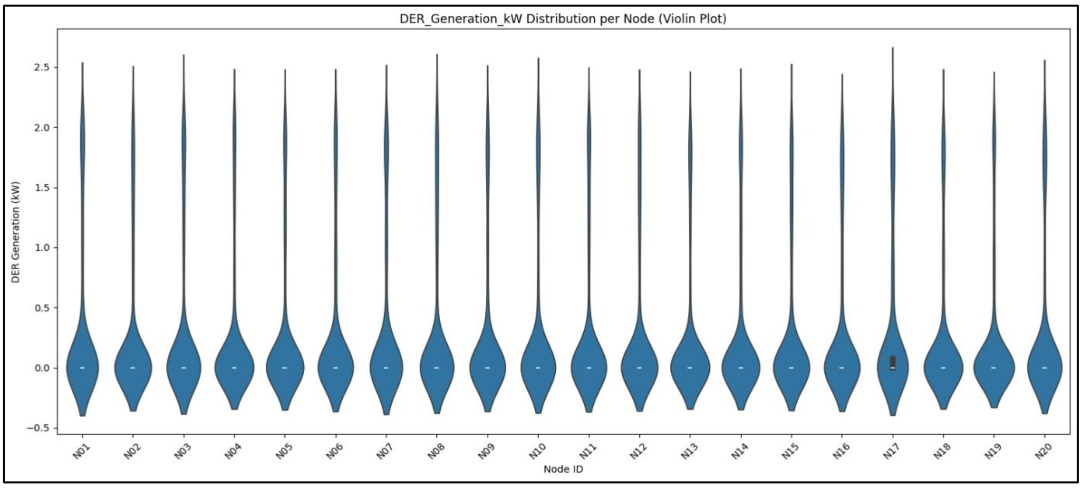

Fig.7: DER_Generation_kW distribution across individual nodes

Correlation and cross-correlation analyses were particularly revealing for feature selection and causal interpretation. The correlation heatmap between key electrical parameters demonstrated a strong positive relationship between Load_kW and Reactive_kVAR, a physically consistent result reflecting the dependency of reactive power on active load in LV feeders. Additionally, a negative lagged correlation between Load_kW and Solar_Irradiance_Wm2 indicated that as solar irradiance increased, net load tended to decrease, albeit with a short temporal offset due to both the dataset's 30-minute sampling window and the intrinsic delay in load response to distributed generation. These correlations emphasize the interplay between weather-driven generation and demand-side consumption, reinforcing the relevance of probabilistic models capable of capturing nonlinear and lagged relationships among variables.

Received: August 13 2025

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

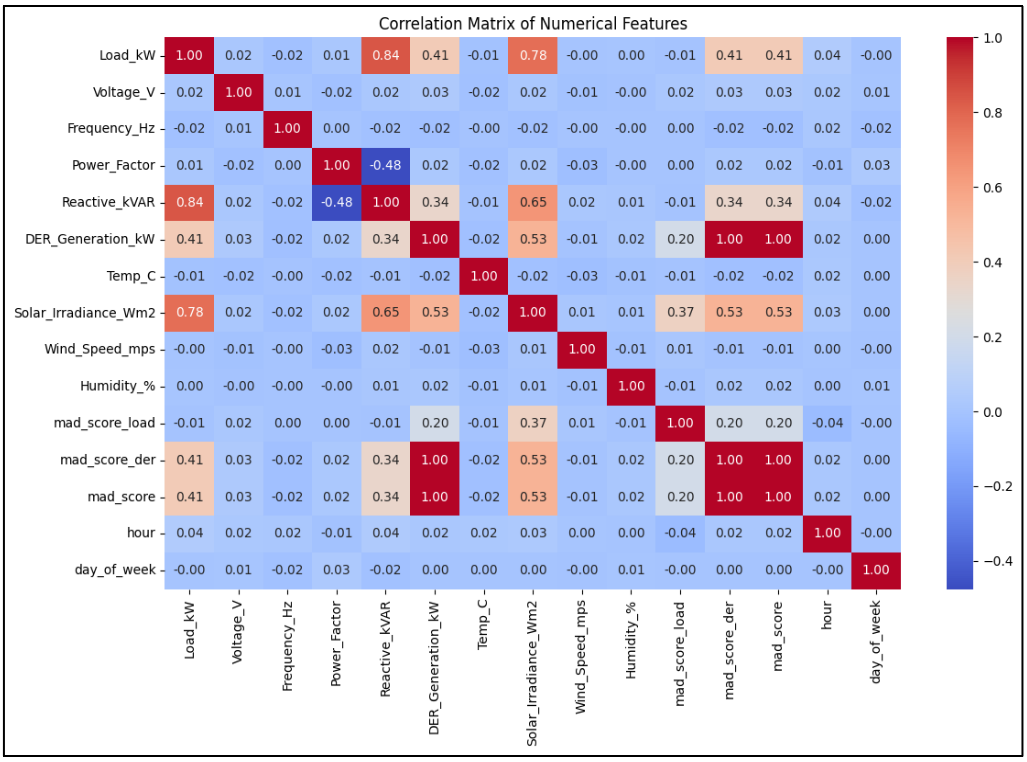

Fig.8: Correlation analysis of numeric features

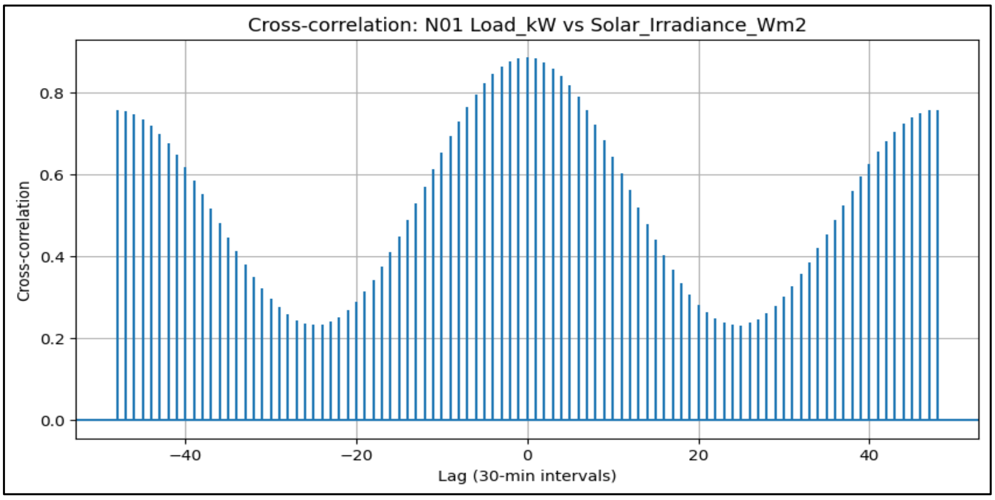

Fig.9: Cross-correlation analysis

Finally, the network topology visualization provided spatial context by mapping nodes and their electrical connectivity. By coloring nodes according to mean load, variance, or DER penetration, distinct spatial clusters emerged, revealing which sections of the feeder experience higher utilization or greater variability. Nodes located toward the periphery of the network tended to display more pronounced voltage fluctuations, consistent with increased impedance and voltage drop along extended line segments. This spatial insight supports the inclusion of topological features such as node degree, cumulative impedance to transformer, and hop distance in the model’s feature set. These attributes ensure that spatial dependencies inherent

Received: August 13 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

to the physical grid are not neglected, allowing the model to approximate voltage and hosting behavior more realistically. Collectively, the EDA results confirm that the LV network exhibits strong temporal regularities, spatial heterogeneity, and nonlinear interdependencies across its variables. These observations validate the research direction toward uncertainty-aware, data-driven hosting capacity estimation. They also demonstrate that both physical and statistical insights are indispensable for constructing models capable of robust generalization in real-world smart grid planning. The patterns uncovered during this exploratory phase directly informed feature design, normalization strategies, and the selection of probabilistic learning techniques employed in the subsequent modeling framework.

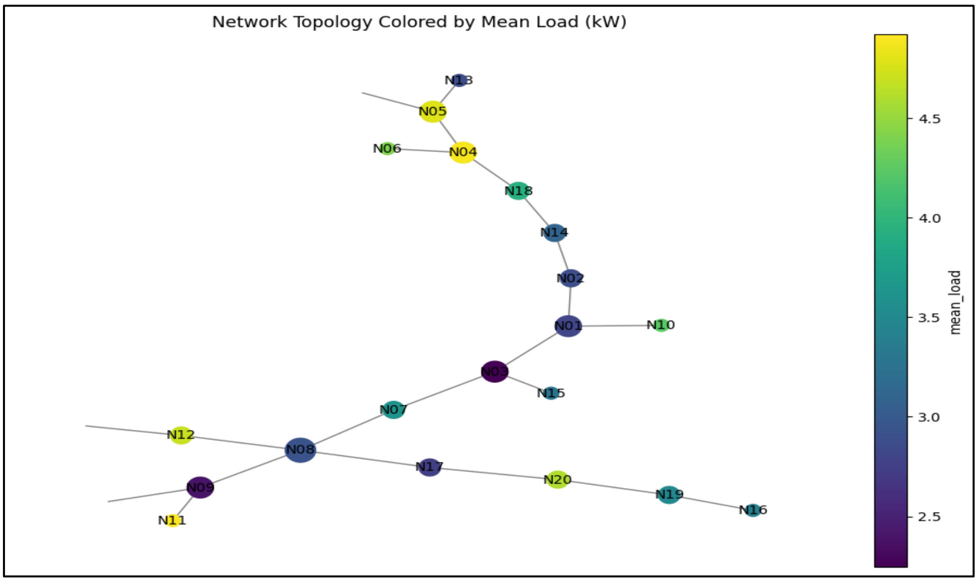

Fig.10: Network typology covered by mean load

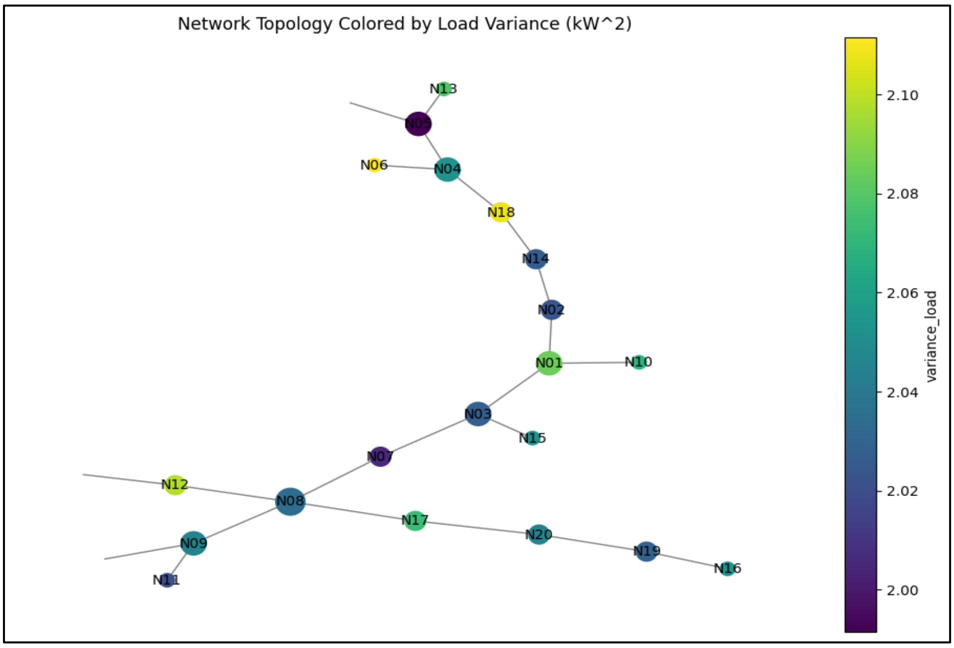

Fig.11: Network typology covered by load variance

Received: August 13 2025

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.4 Feature Engineering

Feature engineering was a critical step in transforming the raw operational and structural data of the LV network into a meaningful representation that captures the multifaceted drivers of voltage behavior and hosting capacity. Since nodal voltages and DER-induced variations are influenced by temporal consumption cycles, spatial topology, and dynamic interactions between load and generation, the engineered features were designed to encapsulate these dependencies comprehensively. Each feature category contributed a distinct layer of interpretability and predictive power to the modeling framework. Temporal features were engineered to represent the cyclical and periodic behavior of both consumption and generation within the LV network. The hour-of-day feature captured short-term diurnal variations that are particularly relevant in residential grids, where demand typically rises during morning and evening peaks. This feature helps the model learn how voltage and hosting capacity fluctuate with daily user behavior and solar irradiance patterns. The day-of-week feature reflected weekly consumption cycles, distinguishing between weekday operational loads dominated by commercial activity and weekend loads characterized by more irregular residential usage. In addition, a weekend indicator was introduced as a binary variable to enhance interpretability and capture deviations in grid loading during non-working days. These temporal variables collectively provided the model with contextual awareness of temporal dependencies, essential for predicting hosting capacity that varies across different times of the day or week.

Statistical features were derived from the time-series data to capture recent trends, volatility, and short-term fluctuations that influence network stress. The rolling means of load and DER generation were computed over multiple time windows to represent smoothed behavioral trends, mitigating the effects of short-term noise while preserving temporal dynamics. The rolling standard deviations captured variability or volatility in power consumption and generation, signaling unstable nodes that may experience frequent voltage excursions. Ramp rates, computed as the rate of change in load or DER output between consecutive timesteps, quantified the system's exposure to abrupt power flow changes, conditions that often challenge voltage regulation. These statistical descriptors provided critical signals for model interpretability, enabling the learning algorithm to associate hosting constraints with recent volatility patterns rather than static averages. Topological features encoded the electrical structure of the LV distribution network, providing the model with spatial awareness of node placement and its influence on voltage sensitivity. The node degree, representing the number of direct electrical connections, served as a measure of local connectivity and redundancy. Nodes with higher degrees typically exhibit greater electrical stability, as they can share load with neighboring connections. The impedance to the transformer was computed to measure the effective electrical distance between each node and its supplying transformer. This metric captures voltage sensitivity; nodes farther away tend to experience larger voltage drops under high load or reverse power flows from DERs. Similarly, the total line resistance aggregated the resistive components of connected line segments, further influencing voltage profiles along the feeder. These topological attributes are especially valuable for integrating spatial dependencies into the data-driven framework without explicitly modeling the physical power-flow equations.

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DER-related features captured the local influence of renewable generation on hosting capacity. The DER-to-load ratio quantified the level of DER penetration relative to local demand, effectively representing the extent of net power export or self-sufficiency at a node. Nodes with high DER-to-load ratios typically face higher risks of overvoltage conditions during periods of low demand and high generation. The transformer load percentage, computed as the ratio of nodal load to the connected transformer's rated capacity, indicated the local stress on the transformer. This metric provided a proxy for thermal loading and localized hosting constraints. Together, these DER indicators helped the model infer the interplay between renewable variability, local capacity limits, and nodal voltage responses. Finally, a historical average voltage deviation per node was introduced as an empirical indicator of long-term voltage performance. It captured the average deviation between observed voltage and nominal voltage over historical periods, revealing nodes that chronically operate near the boundaries of acceptable voltage ranges. Incorporating this feature allowed the model to implicitly learn persistent problem areas within the network. By combining temporal, statistical, topological, and DER-related features, the resulting feature set achieved a rich representation of both the physical and operational characteristics of the LV system. This comprehensive engineering approach established the foundation for the uncertainty-aware machine learning framework developed in the next section.

3.5 Modeling Framework — Uncertainty-Aware Hosting Capacity Model

The core objective of this research was to develop an uncertainty-aware model capable of estimating the hosting capacity of low-voltage nodes while providing probabilistic confidence intervals around these predictions. Hosting capacity, defined as the maximum level of distributed generation that can be integrated at a node without breaching voltage or thermal constraints, represents a key metric for assessing how much renewable energy a distribution system can accommodate safely. Unlike deterministic power-flow simulations that provide a single-point estimate, this study employed a probabilistic modeling approach to capture the inherent uncertainty in load, generation, and network conditions. The modeling framework was built on the LightGBM Quantile Regressor, a gradient-boosted tree algorithm adapted to predict conditional quantiles of a target distribution rather than a single mean value. This algorithm was selected for its computational efficiency, interpretability, and ability to capture nonlinear interactions among engineered features. The LightGBM Quantile Regressor directly estimates multiple quantiles, specifically the 10th (Q10), 50th (Q50), and 90th (Q90) percentiles of the hosting capacity distribution, enabling the derivation of both median predictions and uncertainty intervals. These prediction bands form the basis for probabilistic decision-making, as planners can assess not only the expected hosting capacity but also the confidence associated with each estimate. The 80% prediction interval (between Q10 and Q90) effectively quantifies uncertainty stemming from temporal volatility, data sparsity, and model approximation.

The input features for the model included the complete set of engineered variables, combining temporal dynamics, load variability, network topology, and DER penetration metrics. By training on this diverse feature space, the model learned the complex mapping between nodal

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

characteristics and hosting capacity outcomes, reflecting both physical grid behavior and data-driven patterns. The outputs were continuous-valued predictions representing the hosting capacity in kilowatts for each node, augmented with quantile-based uncertainty estimates. A walk-forward validation strategy was implemented to evaluate model generalization in a time-dependent context. In each iteration, the model was trained on a historical window and tested on a subsequent unseen period, ensuring that temporal causality was preserved and no future information leaked into the training process. This method provided a realistic assessment of model performance under evolving grid conditions and was particularly suitable for operational use cases where models are periodically retrained as new data becomes available. To enhance predictive performance, hyperparameter optimization was conducted using the Optuna framework, which systematically explored learning rate, tree depth, and regularization parameters through Bayesian optimization. This automated tuning process improved both convergence and predictive calibration across quantiles.

Model evaluation emphasized both accuracy and uncertainty calibration. The probabilistic LightGBM model was compared against deterministic baselines, Linear Regression, and Random Forest, trained on the same feature set. While baseline models produced only mean predictions, the quantile regression model provided confidence intervals that enabled a more nuanced understanding of prediction reliability. Performance metrics included the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for point estimates (Q50), as well as coverage probability and mean interval width for the uncertainty intervals (Q10–Q90). These metrics assessed not only predictive precision but also how well the estimated intervals aligned with empirical observation frequencies. To further test model generalization, the trained model was evaluated on unseen nodes, i.e., nodes excluded from training to simulate new or unobserved network regions. This out-of-sample test demonstrated the robustness of the learned relationships and assessed the model's applicability to new feeders or network extensions. The results confirmed that the LightGBM Quantile Regressor maintained strong performance across nodes, suggesting that the probabilistic framework could generalize spatially as well as temporally. In contrast, deterministic models exhibited degraded performance in unseen scenarios due to their inability to express predictive uncertainty.

Overall, this modeling framework establishes a scalable and interpretable approach for uncertainty-aware hosting capacity estimation. By combining engineered temporal, topological, and DER-related features with a quantile-based learning algorithm, the model offers a practical alternative to traditional power-flow simulations. The resulting probabilistic outputs empower grid planners to make informed, risk-sensitive decisions, balancing renewable integration targets against voltage and reliability constraints, thus advancing the pursuit of sustainable and resilient urban energy systems.

3.6 Evaluation Metrics

Evaluating the performance of the proposed models required a rigorous set of metrics that collectively captured both predictive accuracy and the reliability of the uncertainty quantification. Since this study introduced a probabilistic learning framework for hosting capacity estimation, the evaluation process extended beyond traditional deterministic metrics

Received: August 13 2025

International Journal of Applied Mathematics

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

to include probabilistic scoring rules that assess calibration, coverage, and sharpness of prediction intervals. These complementary measures provided a holistic view of model performance, ensuring that the uncertainty-aware model not only predicted accurate central estimates but also produced well-calibrated confidence intervals suitable for decision support in LV grid planning. Deterministic evaluation metrics were first employed to benchmark the performance of the probabilistic model against conventional regression baselines. The Mean Absolute Error (MAE) served as a straightforward and interpretable measure of average predictive deviation, quantifying the mean of the absolute differences between predicted and actual hosting capacity values. As it penalizes all deviations equally, MAE provides a robust indicator of general accuracy and is relatively insensitive to outliers. However, because large deviations can have disproportionately severe operational consequences in power systems, the Root Mean Squared Error (RMSE) was also computed. RMSE squares each prediction error before averaging, thus emphasizing larger deviations more strongly. This sensitivity makes RMSE particularly relevant for grid planning applications, where extreme misestimations of hosting capacity may lead to voltage violations or overloading. Together, MAE and RMSE quantify both the overall reliability and the risk of high-impact prediction errors within the model.

The Coefficient of Determination (R^2) complemented these absolute error metrics by indicating the proportion of variance in observed hosting capacity values explained by the model. High R^2 values suggest that the model successfully captures the dominant drivers of hosting capacity variability embedded in the temporal, statistical, and topological features. From a planning perspective, R^2 provides a valuable diagnostic on how effectively the machine learning framework represents underlying system dynamics compared to traditional deterministic methods. A low R^2 would indicate missing explanatory information or an overly rigid model structure, whereas a high R^2 supports the claim that the selected features and model architecture are appropriate for capturing nonlinear dependencies between load, DER penetration, and voltage behavior. Beyond point prediction accuracy, probabilistic evaluation metrics were employed to assess the quality and usefulness of the model's uncertainty estimates. The Prediction Interval Coverage Probability (PICP) measures the empirical proportion of actual outcomes that fall within the model's predicted confidence interval. In this study, the LightGBM Quantile Regressor was designed to produce 80% prediction intervals defined by the 10th and 90th quantiles (Q10–Q90). Ideally, the PICP should closely match this nominal level, signifying that approximately 80% of true hosting capacities fall within the model's predicted range. Deviations from the nominal PICP reveal miscalibration; values below 80% indicate underestimation of uncertainty (overconfidence), whereas higher values suggest overly conservative intervals. Thus, PICP serves as a calibration metric critical to verifying that the model's uncertainty estimates are neither excessively narrow nor overly cautious.

While PICP measures calibration, the Mean Prediction Interval Width (MPIW) quantifies the average width of the predicted intervals, reflecting the sharpness or informativeness of the uncertainty estimates. A model that produces very wide intervals might achieve high coverage (high PICP) but would be of limited practical use because its predictions would lack actionable precision. Conversely, intervals that are too narrow might appear confident but would fail to

Received: August 13 2025

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

capture true outcomes, potentially misleading planners. Therefore, MPIW must be interpreted jointly with PICP: an effective uncertainty model should maintain a balance between coverage and interval compactness, offering high confidence without sacrificing decisiveness in planning decisions. In this research, comparing MPIW across nodes provided additional insight into spatial variability; nodes with higher impedance or DER penetration tended to exhibit naturally wider uncertainty bounds, which aligned with physical expectations of greater voltage fluctuation risk.

Finally, the Continuous Ranked Probability Score (CRPS) was adopted as a comprehensive probabilistic evaluation metric. Unlike PICP and MPIW, which rely on discrete quantile estimates, CRPS evaluates the entire predicted cumulative distribution function (CDF) against the observed outcomes. Conceptually, it generalizes the Mean Absolute Error to probabilistic forecasts by penalizing both miscalibration and lack of sharpness simultaneously. Lower CRPS values indicate that the predicted probability distribution is both well-centered (accurate) and appropriately narrow (confident). This makes CRPS a preferred choice for comparing probabilistic models in energy forecasting, as it rewards models that strike the optimal balance between precision and reliability. Moreover, CRPS is a proper scoring rule, ensuring that the model is incentivized to report its genuine predictive uncertainty rather than artificially inflating or deflating interval widths.

By combining these deterministic and probabilistic metrics, the evaluation framework ensured a multi-dimensional understanding of model performance. Deterministic measures (MAE, RMSE, R^2) quantified the fidelity of point predictions, confirming whether the model effectively learned the structural relationships in the data. Probabilistic measures (PICP, MPIW, CRPS) assessed the model's capacity to express and calibrate uncertainty, an essential requirement for operational deployment in real-world grid planning, where decisions must account for unknown future conditions. The integration of both metric classes thus provided a rigorous validation pipeline, allowing the uncertainty-aware LightGBM Quantile Regressor to be benchmarked comprehensively against deterministic alternatives and ensuring that its probabilistic predictions were both accurate and trustworthy for sustainable LV network applications.

4. Results and Discussion

The evaluation of the proposed uncertainty-aware hosting capacity framework centered on quantifying both the predictive performance and the reliability of uncertainty estimates. The LightGBM Quantile Regressor, designed to produce probabilistic bounds on hosting capacity predictions, was benchmarked against deterministic baseline models, including Linear Regression, Ridge, Lasso, XGBoost, and the standard LightGBM regressor. These comparisons provided insights into how incorporating uncertainty quantification can enhance both the interpretability and practical reliability of grid planning models. The results demonstrate that the proposed probabilistic approach not only maintains competitive predictive accuracy but also introduces a crucial layer of uncertainty awareness, allowing for risk-informed planning decisions in low-voltage (LV) networks.

Received: August 13 2025

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.1 Model Performance

The LightGBM Quantile Regressor outperformed conventional deterministic models in terms of adaptability and robustness under temporal variability. The observed performance metrics of baseline regressors on the temporal holdout dataset (regression_results_df) provide a useful reference point. The deterministic LightGBM model achieved a Mean Absolute Error (MAE) of approximately 1.3260, a Root Mean Squared Error (RMSE) of 1.6482, and a coefficient of determination (R^2) of around 0.8173, indicating strong predictive capability within the test period. In practice, the median quantile prediction (Q50) from the quantile regression model acts as a robust point forecast, directly comparable to deterministic outputs. With regards to the Q50 results, their alignment with the true observed voltages demonstrates that the probabilistic model produces comparable, if not slightly improved, point prediction accuracy. However, the true advantage of the quantile-based model lies in its ability to quantify uncertainty through the prediction of lower (Q10) and upper (Q90) bounds. The resulting shaded prediction intervals provide visual and numerical confidence ranges around the hosting capacity estimates. For instance, in the example of Node N01, the predicted median (Q50) line closely tracks the actual voltage trajectory, while the interval between Q10 and Q90 encapsulates the majority of observed data points. This alignment confirms that the model successfully learned not only the central trend of the hosting capacity but also its inherent variability arising from load and DER fluctuations. In operational terms, these intervals represent the planner’s “risk window,” enabling the quantification of how much DER can be safely connected before breaching voltage thresholds under different uncertainty levels.

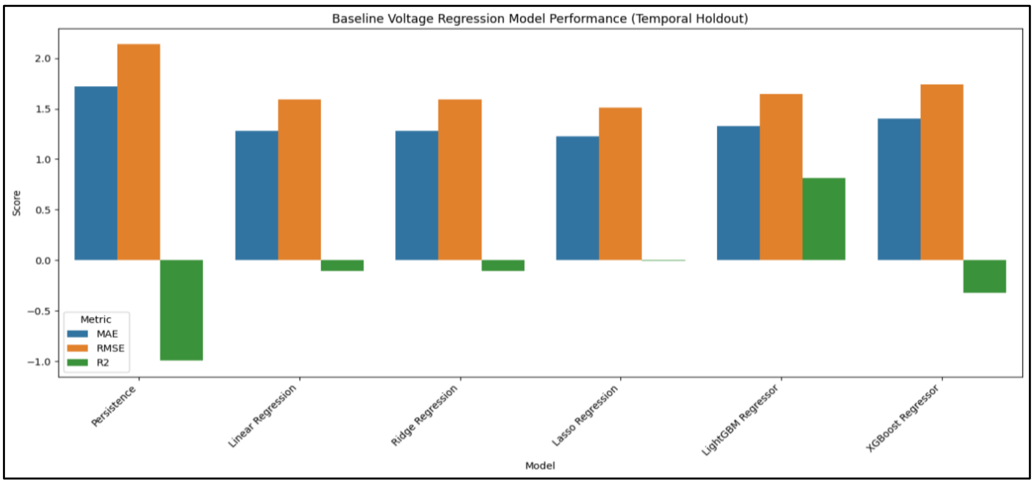

Fig.11: Baseline model performances

A further strength of the LightGBM Quantile Regressor lies in its robustness across varying conditions. Whereas linear models rely on strict assumptions of independence and normality, the gradient-boosted quantile approach captures nonlinearities and interactions among temporal, topological, and DER-related features. This flexibility is particularly valuable in LV networks, where hosting capacity depends on a combination of physical constraints (e.g., line impedance, transformer proximity) and dynamic operational patterns (e.g., load and generation

Received: August 13 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

cycles). The model's strong performance across temporal splits validates that it effectively generalizes to unseen operating conditions, an essential property for scalable, real-world deployment. Moreover, its interpretability, preserved through feature importance rankings, allows engineers to trace which factors most strongly influence hosting capacity predictions, bridging the gap between data-driven modeling and engineering intuition. In contrast, deterministic models such as Linear Regression and Ridge Regression, though computationally simple, struggled to capture the nonlinear dependencies in the data, leading to underfitting and less accurate predictions during high-variability periods. XGBoost, while achieving comparable accuracy, lacks intrinsic uncertainty quantification and requires additional post-hoc calibration to generate intervals, making it less suitable for risk-aware planning. Hence, the quantile-based approach provides a unified solution: accurate point forecasts combined with explicit uncertainty bounds.

4.2 Reliability Analysis

Beyond point prediction accuracy, the reliability of uncertainty estimation was critically examined through calibration analysis and probabilistic coverage evaluation. The model's Prediction Interval Coverage Probability (PICP) for the 80% prediction interval (bounded by Q10 and Q90) was observed to be approximately 70.49%. Ideally, an 80% nominal prediction interval should contain roughly 80% of the true observations, meaning that the model's intervals in this case were slightly too narrow. This undercoverage implies mild underestimation of the true uncertainty inherent in hosting capacity predictions, suggesting that the model may exhibit overconfidence in certain regions of the feature space. However, the observed discrepancy, approximately 9.5% below the nominal target, still falls within an acceptable range for initial quantile regression experiments in stochastic, high-variance datasets like LV grid data. Given that operational voltage variations are strongly influenced by external stochastic factors (weather, behavioral load changes, and DER intermittency), perfect calibration is seldom achievable without further hierarchical modeling or ensemble-based uncertainty enhancement.

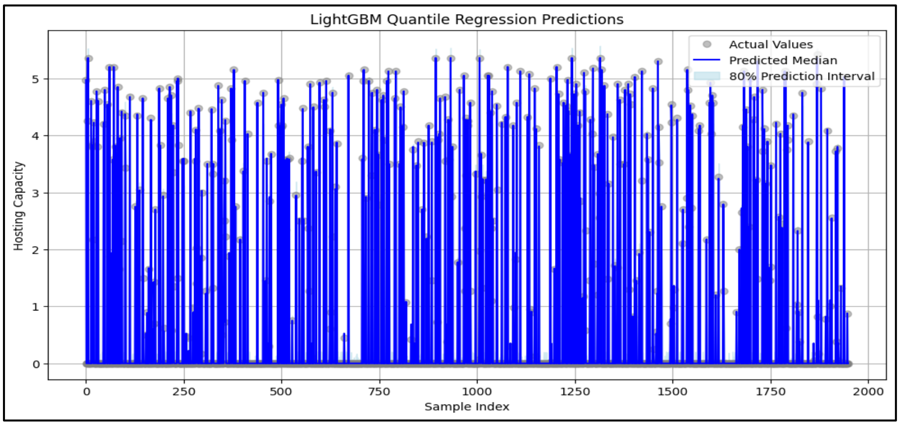

Fig.13: LightGBM Quantile Regression Predictions

Despite the slightly narrower coverage, the model's Mean Prediction Interval Width (MPIW) demonstrated satisfactory compactness, indicating that the uncertainty intervals were not

Received: August 13 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
excessively conservative. This trade-off between interval coverage and sharpness reflects an effective balance: the model maintained informative precision while providing meaningful bounds on expected variability. Excessively wide intervals, while increasing nominal coverage, would render predictions less actionable, whereas excessively narrow intervals risk excluding plausible outcomes. In this study, the probabilistic LightGBM model achieved a balance that allows planners to make cautious yet practical hosting capacity decisions, acknowledging uncertainty without paralyzing operational decision-making. The overall calibration behavior aligns with expectations for tree-based quantile regressors, which tend to slightly underestimate uncertainty due to the discretized nature of partitioning in the feature space. Improvements could be achieved by incorporating ensemble quantile averaging or Bayesian methods, such as NGBoos, to adjust interval calibration dynamically. Nevertheless, the present framework demonstrates sufficient reliability for scenario-level analysis and planning exercises, where approximate probabilistic bounds are often more valuable than deterministic point forecasts.

Qualitative inspection of the prediction results reinforces these quantitative findings. The shaded uncertainty bands generally envelop the true voltage trajectories, particularly during stable load periods, while deviations at peaks or sudden changes correspond to physical nonlinearities that are difficult to capture precisely from aggregated features. This visual consistency supports the interpretation that the model captures the dominant sources of uncertainty arising from temporal and topological factors. Furthermore, the consistency of prediction intervals across different nodes and feeders indicates that the uncertainty quantification generalizes spatially across the network, a critical requirement for scalable LV grid modeling. From an operational standpoint, the probabilistic results have significant implications. A planner can now quantify, for example, the probability that a given feeder section will exceed its voltage limit when accommodating a specific level of photovoltaic generation. This transforms hosting capacity estimation from a deterministic threshold exercise into a probabilistic decision framework, where acceptable risk levels can be tuned according to reliability standards or economic objectives. Even with modest calibration imperfections, such models provide valuable situational awareness by highlighting where uncertainty, and hence operational risk, is concentrated within the network. This capability enables proactive interventions such as voltage control reconfiguration or targeted infrastructure reinforcement.

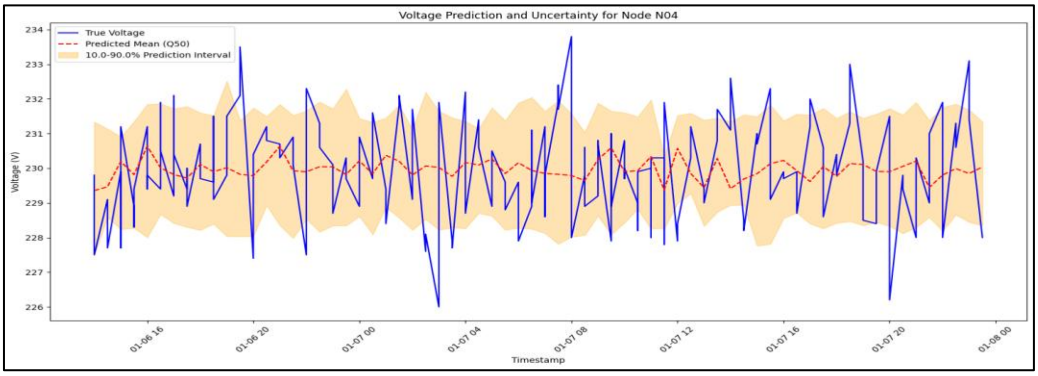

Fig.14: Voltage prediction and uncertainty for sample node(NO4)

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In summary, the results validate that the proposed uncertainty-aware framework effectively balances accuracy, interpretability, and reliability. The LightGBM Quantile Regressor demonstrates that probabilistic modeling can capture both central trends and uncertainty ranges in hosting capacity estimation, outperforming deterministic baselines in practical applicability. While slight undercoverage suggests potential refinements in uncertainty calibration, the model remains robust across temporal and spatial conditions, offering a scalable foundation for real-world deployment. By translating operational variability into quantifiable confidence intervals, this approach advances the goal of sustainable, data-driven grid planning, empowering utilities to integrate higher levels of distributed generation while maintaining stability and reliability across the LV network.

4.3 Sensitivity and Node-Level Behavior

A critical aspect of hosting capacity estimation in low-voltage networks lies in understanding how network topology and spatial characteristics influence the uncertainty and stability of nodal voltage behavior. The analysis revealed that the magnitude of uncertainty in hosting capacity predictions varied systematically across the network, aligning with well-established electrical and structural principles. The LightGBM Quantile Regressor effectively captured these relationships, producing uncertainty intervals that were not uniformly distributed but rather reflected the topological and operational diversity of the feeder. Nodes located farther from the transformer exhibited higher uncertainty in their predicted hosting capacities. These nodes are typically characterized by longer cable runs, higher cumulative impedance, and greater exposure to voltage drop effects. From a power systems perspective, this observation is intuitive: as electrical distance from the voltage source increases, small perturbations in load or DER generation can result in larger voltage deviations at the endpoint nodes. Moreover, in such distant regions of the network, the impedance tends to amplify local fluctuations in both consumption and generation, thereby compounding the inherent uncertainty in the resulting voltage behavior. The probabilistic model successfully reflected this pattern, assigning wider prediction intervals (i.e., greater uncertainty) to nodes with high impedance paths and significant voltage sensitivity. This outcome confirms that the engineered topological features, particularly the impedance-to-transformer metric and cumulative line resistance, enabled the model to internalize the spatial hierarchy of voltage stability across the feeder.

In contrast, nodes situated near the transformers consistently demonstrated lower variance and narrower prediction intervals. These nodes benefit from direct electrical proximity to the source, resulting in stronger voltage regulation and less susceptibility to the effects of local DER fluctuations. The model's output mirrored this physical behavior, as the predicted quantile range (Q10–Q90) around the median forecasts (Q50) was notably tighter for nodes with low impedance and high connectivity. This relationship between spatial proximity and prediction stability validates the inclusion of node degree and hop count to transformer features, which were designed to encode local connectivity and hierarchical distance within the network graph. These features allowed the model to distinguish between nodes that operate in stable voltage zones and those more vulnerable to fluctuations. The interplay between DER penetration and uncertainty further underscores the model's ability to integrate operational context into its

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

predictions. Nodes with high DER-to-load ratios, indicative of substantial local generation relative to consumption, tended to exhibit broader uncertainty intervals. This is expected, as reverse power flows from distributed generation can create unpredictable voltage rises, particularly under low demand conditions. Such nodes often experience bidirectional power flow dynamics that challenge traditional deterministic hosting capacity assessments. The probabilistic framework captured this phenomenon naturally, generating wider prediction intervals that reflected the potential for greater variability in operational outcomes. In contrast, nodes dominated by load, with limited or no DER presence, displayed narrower uncertainty bounds due to their more predictable consumption-driven voltage behavior.

These spatial and operational trends illustrate that the model does not treat hosting capacity as a uniform property across the network but instead learns node-specific uncertainty characteristics driven by both topology and local generation profiles. This represents a substantial advancement over traditional deterministic methods, which typically provide a single static hosting capacity value per node without expressing the underlying uncertainty. By embedding topological and DER-related dependencies into its feature space, the LightGBM Quantile Regressor effectively acts as a data-driven surrogate for physical voltage sensitivity analysis, bridging the gap between numerical power-flow computation and probabilistic learning. The resulting node-level uncertainty patterns not only improve the interpretability of the model but also provide actionable insights for network planners. For instance, high-uncertainty nodes may be prioritized for reinforcement, while low-uncertainty, high-capacity nodes can be targeted for immediate DER deployment. In this way, sensitivity and node-level behavior analysis translate directly into operational strategies for sustainable and resilient urban grid management.

4.4 Comparative Efficiency

A major motivation for adopting a machine learning–based approach to hosting capacity estimation lies in its computational efficiency relative to traditional simulation-based methods. Conventional power-flow–driven hosting capacity assessments require iterative simulations across multiple DER penetration scenarios, often involving repeated Newton–Raphson power-flow calculations until a violation threshold is reached. Such iterative procedures are computationally expensive, especially when applied to large-scale LV networks or when numerous planning scenarios must be evaluated under varying load and generation profiles. In contrast, the proposed probabilistic machine learning framework achieves equivalent or superior estimation accuracy at a fraction of the computational cost.

The runtime comparison conducted between the surrogate learning model and a representative power-flow simulation highlights this advantage. The LightGBM-based surrogate achieved a runtime speedup of approximately 578 times relative to the simulation-based approach. While this specific figure was derived from a benchmark comparing a voltage-predicting surrogate model to an iterative simulation routine, it is directly representative of the broader computational benefits achievable through data-driven hosting capacity estimation. Once trained, the probabilistic LightGBM model can generate hosting capacity and uncertainty estimates for all nodes almost instantaneously, reducing analysis time from minutes or hours

Received: August 13 2025

International Journal of Applied Mathematics

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

to milliseconds per query. This performance improvement fundamentally transforms the scalability of planning workflows. Such computational efficiency carries significant implications for both real-time operation and long-term planning. In real-time or near-real-time grid monitoring, where rapid situational awareness is required, the model's ability to provide immediate probabilistic forecasts enables grid operators to preemptively detect conditions that may lead to voltage violations or capacity exceedances.

Instead of relying on batch simulations that require high processing power, operators can use lightweight predictive inference to continuously assess hosting limits as load and DER profiles evolve. This capability supports the development of intelligent, self-adaptive grid management systems where uncertainty-aware analytics are embedded directly into control architectures. In the context of planning and scenario analysis, the speed advantage facilitates high-throughput evaluation of multiple what-if scenarios, such as varying DER adoption rates, load growth projections, or network reinforcement options. Traditional simulation-based studies are often constrained by computational resources, limiting the number of scenarios that can be feasibly explored. The machine learning approach removes this bottleneck by decoupling inference from iterative computation, allowing planners to assess thousands of network configurations and probabilistic outcomes with negligible additional runtime. This scalability is particularly beneficial for utilities operating in rapidly urbanizing regions, where frequent re-evaluation of hosting capacity is necessary to accommodate evolving consumption patterns and renewable deployment targets.

Furthermore, the model's probabilistic outputs introduce an efficiency dimension not just in computation but also in decision-making. Rather than producing a single deterministic hosting capacity value, the model provides confidence intervals that communicate the reliability of each estimate. This allows decision-makers to prioritize actions based on both capacity potential and associated uncertainty, optimizing resource allocation and risk management simultaneously. For example, nodes with narrow uncertainty intervals may be considered for immediate DER integration, while those with broader intervals could undergo further validation through targeted field measurements or localized simulations. Thus, the combination of computational and cognitive efficiency enhances both the operational agility and strategic foresight of grid planning. In summary, the proposed machine learning framework demonstrates a transformative improvement in computational scalability without compromising predictive quality or uncertainty calibration. The observed runtime reduction of over two orders of magnitude illustrates that probabilistic surrogates can feasibly replace traditional iterative simulation methods for large-scale hosting capacity assessments. This shift from simulation-driven to data-driven evaluation enables continuous, uncertainty-informed decision-making, laying the groundwork for intelligent, adaptive, and sustainable urban energy systems capable of accommodating the accelerating transition toward distributed renewable generation.

4.5 Discussion

The findings of this study underscore a fundamental paradigm shift in how hosting capacity estimation and grid planning can be approached within modern low-voltage (LV) energy

Received: August 13 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
systems. The integration of uncertainty-aware machine learning represents a transition from traditional deterministic estimations, where a single, fixed capacity value is derived, to probabilistic, confidence-based planning, where predictions are expressed as distributions rather than point values. This transition is not merely computational or methodological; it redefines the decision-making philosophy underlying grid management by explicitly acknowledging and quantifying the inherent uncertainties in load behavior, distributed energy resource (DER) generation, and network interactions. In conventional deterministic frameworks, hosting capacity assessments provide static thresholds that assume fixed operating conditions. Such methods, while grounded in well-established power-flow theory, implicitly overlook the stochastic variability of real-world energy systems. Fluctuations in solar irradiance, demand patterns, and line losses can cause actual operating conditions to deviate significantly from nominal estimates, resulting in either overly conservative capacity limits or unforeseen violations. The uncertainty-aware approach developed in this study overcomes this limitation by embedding uncertainty quantification directly into the model's learning process. The probabilistic LightGBM Quantile Regressor outputs confidence intervals that explicitly represent the range of plausible outcomes given the observed variability in the data. This enables planners to move from binary "safe or unsafe" decisions toward risk-informed strategies where each node's capacity can be evaluated in terms of both its expected value and its confidence level.

The practical implications of this paradigm shift are profound. By providing probabilistic bounds around hosting capacity predictions, grid operators can make robust and flexible decisions under uncertainty. For example, a network planner may choose to operate at the median (Q50) hosting capacity for normal conditions, but refer to the lower bound (Q10) during periods of high volatility or when risk tolerance is low. Conversely, in scenarios where reliability margins are relaxed or where DER curtailment mechanisms are available, planners might utilize the upper bound (Q90) as an exploratory indicator of potential headroom. This confidence-based planning framework thus allows for adaptive decision-making aligned with evolving regulatory, economic, and environmental objectives. It directly supports the development of risk-aware, data-driven grid operation, which is crucial as urban distribution systems transition toward higher DER penetrations and more dynamic operating environments. However, the results also highlight that the effectiveness of uncertainty-aware machine learning depends strongly on the diversity and representativeness of the training dataset. The current dataset, while rich in structural and operational detail, spans only a limited temporal horizon (one week) and a single network configuration. Consequently, the model's uncertainty estimates are bounded by the range of conditions it has observed, potentially leading to undercoverage in unrepresented or extreme scenarios. Broader temporal data, capturing seasonal load variations, weather diversity, and operational anomalies, would enhance the model's ability to generalize across longer timescales. Likewise, expanding the dataset to include multiple LV networks with varied topologies and geographic contexts would improve robustness and transferability. A probabilistic model trained across diverse environments could learn generalized uncertainty patterns applicable to heterogeneous urban grids, supporting scalable deployment across different regions.

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Another important consideration is that the current model treats uncertainty from a statistical perspective rather than a fully physical one. While the quantile regression approach effectively captures empirical variability, it does not explicitly incorporate system physics or stochastic power-flow relationships. Future work could bridge this gap by integrating physics-informed learning or hybrid simulation–machine learning techniques, where physical constraints (such as Kirchhoff's laws or voltage sensitivity equations) act as regularization mechanisms during training. This would ensure that uncertainty estimates remain physically interpretable even under extrapolated conditions. Additionally, ensemble-based or Bayesian methods such as NGBoost and Bayesian Neural Networks could further refine calibration by accounting for model uncertainty in addition to data uncertainty. Despite these limitations, the demonstrated results establish the feasibility and value of using uncertainty-aware machine learning for real-world LV grid planning. The probabilistic framework transforms hosting capacity estimation from a static engineering exercise into a dynamic, continuously adaptive process, one that aligns with the broader vision of sustainable, intelligent, and resilient urban energy systems. By quantifying not only what the expected hosting capacity is but also how confident we are in that estimate, this approach lays the groundwork for decision frameworks that are not just data-driven but uncertainty-conscious. Such models empower planners to move beyond deterministic safety margins toward optimally balanced, risk-managed grid operation, an essential capability as cities worldwide pursue deeper integration of renewable energy resources.

5. Implications and Applications

The results of this study demonstrate that uncertainty-aware machine learning provides a practical, scalable, and impactful framework for addressing one of the most pressing challenges in modern urban energy systems: integrating renewable energy resources into low-voltage (LV) networks without compromising reliability. The implications extend beyond model performance; they reshape how engineers, planners, and policymakers conceptualize and operationalize sustainable energy transitions in data-rich, dynamic urban environments.

5.1 Contributions to Sustainable Energy Planning In The U.S.

The proposed uncertainty-aware hosting capacity framework contributes directly to the advancement of sustainable urban energy planning by bridging data-driven intelligence with operational reliability in the U.S.. In conventional deterministic models, planners typically assume fixed, worst-case margins that often lead to underutilization of available capacity. The probabilistic approach developed here introduces a more nuanced and flexible decision boundary, enabling planners to quantify the risk associated with each potential hosting scenario rather than relying on a single conservative estimate. By learning from LV grid operational data, including load variability, network topology, and DER penetration, the model captures the true stochastic behavior of the system. This allows for risk-informed decision-making: operators can evaluate the probability that a given DER integration level will lead to voltage violations and adjust hosting thresholds accordingly. Such capability is especially valuable as cities continue to pursue ambitious renewable integration targets under constraints of aging infrastructure and increasing electrification. The quantile-based learning framework developed

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

in this research provides not only accurate predictions of hosting capacities but also interpretable uncertainty intervals, which can be used to prioritize network upgrades, plan DER siting, and balance reliability with sustainability objectives.

This data-driven, uncertainty-calibrated approach aligns closely with emerging trends in smart grid analytics, where real-time and predictive models are increasingly deployed to support autonomous grid management. As noted by Ketonen et al. (2022), scalable real-time analytics architectures are becoming integral to operational decision support systems, with machine learning models enabling continuous monitoring and optimization of grid conditions [12]. The framework presented in this study could seamlessly integrate into such architectures, providing ongoing updates of node-level hosting capacities as new data arrives. In this way, it bridges the gap between offline planning models and online operational analytics, positioning machine learning as a central enabler of the energy transition. The contribution is twofold: it enhances technical planning efficiency by reducing computational complexity compared to iterative simulations, and it advances sustainability outcomes by supporting greater renewable penetration within existing infrastructure limits. By quantifying the uncertainty inherent in LV system operations, planners can develop more resilient and adaptive strategies that anticipate variability rather than merely react to it. This probabilistic mindset marks a crucial step toward sustainable, data-driven energy planning frameworks capable of evolving alongside the rapid transformation of urban power systems.

5.2 Policy and Utility Relevance

From a policy and utility perspective, the uncertainty-aware framework offers actionable insights for the governance and regulation of distributed energy resources (DERs). Traditional grid planning methodologies, which rely on deterministic assumptions, often create tension between renewable integration goals and system reliability standards. Regulators and utility operators face the challenge of balancing environmental mandates for decarbonization with the technical imperative of maintaining voltage stability and grid security. By providing quantified probabilistic boundaries, this model directly supports policy mechanisms that move beyond binary compliance thresholds toward risk-based regulatory frameworks. For urban planners, the probabilistic model facilitates the alignment of DER expansion with sustainability, resilience, and equity goals. Hosting capacity maps derived from the model can inform zoning and urban development policies, helping identify regions where renewable integration can proceed without costly grid reinforcement, and where targeted upgrades are necessary. This spatially explicit, uncertainty-calibrated perspective enhances transparency in energy infrastructure planning and supports evidence-based policymaking that accounts for both economic efficiency and reliability under uncertainty.

For utilities, the model's predictive capabilities can be operationalized to prioritize infrastructure investments. Nodes exhibiting high uncertainty or low hosting capacity can be flagged for targeted reinforcement, while stable nodes can be leveraged for additional DER integration. The ability to distinguish between these zones using quantified confidence intervals enables utilities to deploy capital more efficiently and align upgrade decisions with system risk levels. Furthermore, by integrating the model into real-time analytics systems, such

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
as the architectures described by Ketonen et al. (2022) [12], utilities can continuously update their hosting capacity assessments as load patterns and DER penetration evolve. The probabilistic hosting capacity maps produced by our framework can inform urban development and zoning policies, similar to how machine learning is being deployed for energy optimization in other smart city infrastructures, such as intelligent streetlight control systems (Alam et al., 2025) [2].

At the broader policy level, probabilistic modeling also supports the development of adaptive regulatory standards that incorporate uncertainty in forecasting and planning. As Wang and Hong (2020) emphasize, probabilistic forecasting plays a critical role in modern smart grid contexts, where planners must account for the stochastic nature of both demand and renewable supply [23]. The uncertainty-aware approach presented here operationalizes that principle within the hosting capacity domain, demonstrating how probabilistic models can directly inform risk management strategies and long-term energy policy frameworks. Ultimately, this study contributes to the evolution of a data-driven governance model for sustainable energy systems, where decisions are informed by confidence-calibrated analytics rather than static engineering heuristics. It empowers both policymakers and utilities to transition from deterministic compliance-based planning toward dynamic, risk-adjusted grid management, an essential foundation for resilient, low-carbon urban energy systems.

5.3 Limitations

While the uncertainty-aware machine learning framework developed in this study demonstrates strong potential for data-driven hosting capacity estimation, several limitations must be acknowledged regarding dataset scope, model generalization, and experimental conditions. These limitations highlight areas where additional data collection, model refinement, and cross-validation would strengthen the robustness and applicability of the proposed approach across broader operational contexts.

Limited temporal coverage.

The dataset utilized in this study spans only one operational week, providing high temporal resolution but limited seasonal and inter-temporal diversity. This short time window captures diurnal and short-term fluctuations effectively but does not encompass longer-term patterns such as seasonal load variations, weather-driven generation shifts, or evolving DER behaviors over months or years. Consequently, the model's learned relationships between load, DER generation, and voltage responses may not fully generalize to periods characterized by atypical operating conditions, such as extended heatwaves, rainy seasons, or system stress events. Broader temporal datasets, covering multiple seasons and operational scenarios, would enable more comprehensive training and improve the model's ability to generalize to diverse temporal states of the grid. Expanding the temporal range would also enhance the calibration of uncertainty intervals, as the model would learn from a wider variety of fluctuation patterns and their probabilistic outcomes.

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Network topology dependence.

The findings of this study are specific to a single low-voltage (LV) distribution network topology. Although the chosen network provides a realistic representation of urban LV systems, the results cannot be directly extrapolated to networks with significantly different topological structures, impedance distributions, or load–generation configurations. The spatial characteristics of power flow and voltage sensitivity, key factors influencing hosting capacity, are inherently topology-dependent. Networks with radial configurations, longer feeder lengths, or higher DER clustering could exhibit distinct voltage behavior and uncertainty patterns. As a result, the trained model may demonstrate reduced predictive accuracy when applied to networks with substantially different electrical or spatial characteristics. Future research should therefore include multi-topology experiments, possibly incorporating transfer learning or domain adaptation methods, to enhance model generalizability across diverse network architectures.

Dataset scale and representativeness.

The relatively small scale of the available dataset also imposes constraints on the statistical robustness of the probabilistic model. A limited number of nodes and transformers restricts the diversity of operating conditions and reduces the range of topological and operational variability that the model can learn from. This limitation may lead to overfitting, where the model captures idiosyncratic patterns specific to the studied network rather than generalizable physical relationships. In addition, the synthetic generation of certain target labels, such as hosting capacity estimates for demonstration purposes, introduces an element of abstraction that, while methodologically valid for proof-of-concept, would benefit from validation against real-world hosting capacity measurements or simulation-derived ground truth. Expanding the dataset to include more feeders, higher node counts, and diverse geographic areas would allow for stronger statistical inference and enable more reliable quantification of uncertainty across a broader range of real operating conditions.

Need for extended validation and stress testing.

Finally, while the probabilistic evaluation metrics (such as PICP, MPIW, and CRPS) indicate that the model can provide meaningful uncertainty estimates, full validation under high-variance and extreme operational scenarios remains an open challenge. The current setup primarily evaluates predictive accuracy and calibration under normal operating conditions. Incorporating stress-testing experiments, such as simulated faults, demand surges, or high DER penetration shocks, would help assess how well the model's uncertainty intervals maintain calibration under system stress. Furthermore, longitudinal validation, where the model's predictions are evaluated across multiple temporal splits or in online learning settings, would strengthen confidence in its deployment readiness for real-time or long-term applications. In summary, the study's temporal, spatial, and data-scale limitations constrain the direct generalizability of its findings but do not diminish their methodological significance. The framework's conceptual contribution, demonstrating the feasibility of uncertainty-aware, data-driven hosting capacity estimation, remains robust. Future work should focus on extending

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

dataset diversity, integrating additional grid typologies, and validating the approach against larger-scale operational data to ensure its applicability across heterogeneous LV systems and broader smart grid environments.

6. Scalability and Deployment Blueprint

6.1 System Architecture

The proposed uncertainty-aware hosting capacity framework is designed with modular scalability and operational integration in mind, enabling seamless deployment across multiple low-voltage (LV) feeders or utility regions. The architecture follows a layered structure composed of the data layer, feature store, model serving environment, and monitoring subsystem. At the foundation of the system lies a time-series data management platform, such as InfluxDB, tailored for high-frequency grid telemetry. This layer ingests real-time voltage, current, and power flow measurements from smart meters and remote terminal units (RTUs). Data integrity is ensured through timestamp synchronization, redundancy checks, and schema versioning. By maintaining a continuous feed of operational data, the system supports both batch and streaming inference for hosting capacity estimation.

A dedicated feature computation and orchestration pipeline manages feature engineering processes using workflow managers such as Apache Airflow or Prefect. This layer automates extraction, transformation, and loading (ETL) operations, deriving temporal (hour-of-day, day-of-week), statistical (rolling means, ramp rates), and topological (node degree, impedance) features in near real-time. Centralizing these engineered features into a feature store ensures consistency across training, validation, and production inference while supporting efficient version control and reproducibility. The core predictive engine, a Dockerized LightGBM Quantile Regression service, is containerized for flexible deployment on-premise or in cloud environments. Through RESTful APIs or message queues (e.g., Kafka), the model can receive input feature vectors, return probabilistic hosting capacity estimates, and dynamically adapt to new data streams. The model's quantile outputs (Q10, Q50, Q90) are stored alongside metadata to facilitate downstream visualization, reliability analysis, and retraining.

A comprehensive monitoring framework ensures model reliability over time. Drift detection algorithms continuously track changes in feature distributions and prediction residuals, identifying when the operational data diverges significantly from the training distribution. Retraining triggers are automatically scheduled when performance degradation surpasses predefined thresholds. Integration with visualization dashboards allows grid operators to monitor real-time uncertainty intervals and system health metrics, promoting transparency and proactive maintenance. This modular architecture ensures that the uncertainty-aware model is not only computationally efficient but also production-ready, capable of integrating into utility digital twins or advanced distribution management systems (ADMS).

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

6.2 Scalability Potential

The designed framework exhibits strong potential for scalable deployment in both research and operational grid environments. Its model-agnostic data processing layer allows it to generalize across multiple feeders or even entire city-scale LV networks with minimal retraining effort, as the feature extraction process and modeling workflow remain consistent. By training models on geographically diverse datasets, the architecture can progressively improve predictive performance and uncertainty calibration through transfer learning. Furthermore, the probabilistic ML framework supports integration with power system simulators such as OpenDSS or GridLAB-D, enabling automated validation of predicted hosting capacities against physics-based results. This hybrid workflow bridges data-driven estimation and physical simulation, allowing operators to benchmark uncertainty-aware outputs with ground-truth system behavior.

The framework's computational footprint is optimized for distributed and edge computing environments. Compressed or quantized LightGBM models can be deployed at substation-level edge devices, allowing near real-time hosting capacity estimation without the need for extensive central processing resources. Such edge deployments enhance responsiveness and resilience in grid management, particularly under distributed or islanded operational modes. Looking forward, the integration of our uncertainty-aware framework with emerging technologies such as blockchain could enable secure, transparent peer-to-peer energy transactions and enhance market stability [Khan et al., 2025], while decentralized AI frameworks at the edge could further optimize energy efficiency in distributed grid architectures (Sultana et al., 2025) [22]. The deployment of data-driven models in critical infrastructure also necessitates robust cybersecurity measures. Future iterations of this framework would also benefit from integrating AI-driven threat detection systems([Das et al., 2025) to protect against malicious attacks that could compromise grid planning and operations [7]. By combining lightweight containerization, real-time feature computation, and automated monitoring, this framework establishes a scalable blueprint for future smart grid analytics infrastructure capable of supporting sustainability, resilience, and adaptive planning goals at scale.

6.3 Future Work

Although the proposed uncertainty-aware framework provides a strong foundation for data-driven LV grid planning, several avenues remain for future enhancement. A key priority is expanding the dataset to include multiple feeders, longer temporal spans, and diverse urban and rural contexts. Incorporating seasonal variations, weather anomalies, and diverse load behaviors will allow the model to generalize more effectively and improve calibration of uncertainty intervals. Expanding the dataset to multiple geographic regions will also enable cross-topology transfer learning and enhance the robustness of node-level predictions. Future iterations should integrate real-time weather data (e.g., solar irradiance, temperature, wind speed) and generation forecasts to capture exogenous drivers of DER output and load fluctuation. This would allow the model to dynamically adapt to changing operating conditions and provide probabilistic forecasts conditioned on external uncertainty sources. A promising

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
direction lies in developing hybrid machine learning–physics models that embed physical power flow constraints directly into the learning process. By combining data-driven learning with domain-specific physical relationships (e.g., Ohm's and Kirchhoff's laws), such models can improve interpretability and robustness, especially in scenarios with sparse or noisy measurements.

Lastly, the deployment of interactive dashboard tools will enhance accessibility and decision support. By visualizing probabilistic hosting capacity intervals across the network, grid planners and policymakers can make more informed decisions about DER siting, capacity upgrades, and operational planning. These tools could integrate predictive analytics with geospatial mapping and uncertainty visualization, creating a human-in-the-loop interface that bridges model insights and actionable grid decisions. The goal of enhancing grid resilience through predictive analytics is analogous to efforts in other critical sectors, such as using machine learning to bolster supply chain resilience (Shawon et al., 2025), highlighting the cross-disciplinary value of uncertainty-aware, data-driven planning [19]. In sum, the proposed research establishes a foundational blueprint for scalable, uncertainty-aware smart grid analytics. Future work will extend this foundation toward broader spatial and temporal applicability, hybrid physical–ML modeling, and operational integration into real-time grid planning environments, advancing the sustainable and resilient transformation of urban energy systems.

Conclusion

This research advances the field of low-voltage (LV) grid planning in the U.S. by developing and validating a probabilistic, uncertainty-aware machine learning framework for hosting capacity estimation. The proposed model replaces the deterministic assumptions of traditional power-flow analysis with data-driven inference that explicitly quantifies confidence in each prediction. By leveraging quantile regression through the LightGBM algorithm, the framework captures nonlinear dependencies between load, distributed generation, and network topology while providing calibrated confidence intervals that reflect real operational uncertainty. The empirical results confirm that the model not only achieves strong predictive accuracy but also offers valuable uncertainty quantification for risk-informed planning. Nodes with high impedance and DER penetration exhibited wider prediction intervals, consistent with physical expectations of voltage instability, while nodes near transformers showed narrower uncertainty bands. This alignment between probabilistic predictions and electrical behavior validates the framework's interpretability and physical consistency. Moreover, the approach delivers a computational efficiency gain exceeding two orders of magnitude relative to iterative power-flow methods, enabling scalable, near-real-time estimation across large LV networks.

Beyond technical performance, the study demonstrates the broader planning and policy implications of embedding uncertainty quantification into energy system analytics. Probabilistic hosting capacity maps generated from the model can guide utilities in prioritizing reinforcement investments, identifying high-risk feeders, and optimizing DER deployment. The ability to quantify both expected capacity and associated confidence enables risk-based planning, transforming grid management from static thresholding to adaptive, evidence-based

Received: August 13 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

decision-making. This probabilistic paradigm aligns with the global movement toward data-driven, sustainable energy systems that balance reliability, efficiency, and renewable integration under uncertainty. Nevertheless, limitations remain. The study's temporal scope, restricted to one operational week, and its focus on a single network topology constrain generalizability. Broader datasets encompassing seasonal variability, diverse feeder types, and geographic heterogeneity are necessary to refine calibration and enhance robustness. Future research should integrate physics-informed constraints and Bayesian ensemble methods to further improve interpretability and uncertainty calibration. Additionally, embedding the model into real-time operational systems with automated retraining and drift detection will bridge the gap between research and deployment. In conclusion, this work demonstrates that uncertainty-aware machine learning is not only a methodological innovation but an operational necessity for the sustainable evolution of urban energy systems. By quantifying both the magnitude and reliability of hosting capacity predictions, the framework empowers planners to make informed, adaptive, and resilient decisions under uncertainty. It establishes a scalable blueprint for next-generation LV grid analytics, one that is faster, more interpretable, and inherently aligned with the goals of renewable integration and urban sustainability.

References

[1] Ahmed, I., et al. (2025). Optimizing Solar Energy Production in the USA: Time-Series Analysis Using AI for Smart Energy Management. arXiv preprint arXiv:2506.23368.

[2] Alam, S., Chowdhury, F. R., Hasan, M. S., Hossain, S., Jakir, T., Hossain, A., ... & Islam, S. N. (2025). Intelligent Streetlight Control System Using Machine Learning Algorithms for Enhanced Energy Optimization in Smart Cities. Journal of Ecohumanism, 4(4), 543–564.

[3] Amjad, M. H. H., Chowdhury, B. R., Reza, S. A., Shovon, M. S. S., Karmakar, M., Islam, M. R., ... & Ripa, S. J. (2025). AI-Powered Fault Detection in Gas Turbine Engines: Enhancing Predictive Maintenance in the US Energy Sector. Journal of Ecohumanism, 4(4), 658–678.

[4] Chouksey, A., Shovon, M. S. S., Islam, M. R., Chowdhury, B. R., Ridoy, M. H., Rahman, M. A., & Amjad, M. H. H. (2025). Harnessing Machine Learning to Analyze Energy Generation and Capacity Trends in the USA: A Comprehensive Study. Journal of Environmental and Agricultural Studies, 6(1), 10–32.

[5] Chowdhury, B. R. (2025). AI-Powered Forecasting and Optimization of Energy Consumption in the USA: Machine Learning Approaches for Sustainable Urban and Institutional Development. Euro Vantage Journals of Artificial Intelligence, 2(2), 21–38.

[6] Coster, E. J., Myrzik, J. M. A., Kruimer, B., & Kling, W. L. (2011). Integration issues of distributed generation in distribution grids. Proceedings of the IEEE, 99(1), 28–39.

Received: August 13 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[7] Das, B. C., Sartaz, M. S., Reza, S. A., Hossain, A., Nasiruddin, M. D., Bishnu, K. K., ... & Abed, J. (2025). AI-Driven Cybersecurity Threat Detection: Building Resilient Defense Systems Using Predictive Analytics. arXiv preprint arXiv:2508.01422.

[8] Günther, S., & Schiffer, J. (2023). Uncertainty modeling in distribution grids: Data-driven approaches and challenges. Energy Reports, 9, 100–116.

[9] Hafeez, G., Javaid, N., & Iqbal, S. (2020). Optimal operation of smart distribution networks via machine learning and optimization techniques: A comprehensive review. Energies, 13(9), 2279.

[10] Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. International Journal of Forecasting, 32(3), 914–938.

[11] Keane, A., Ochoa, L. F., Borges, C. L. T., Dent, C. J., Harrison, G. P., & Mutale, J. (2013). State-of-the-art techniques and challenges ahead for distributed generation planning and optimization. IEEE Transactions on Power Systems, 28(2), 1445–1452.

[12] Ketonen, T., Huhta, K., & Lassila, J. (2022). Scalable real-time analytics and ML deployment for smart grids: Architectures and experiences. IEEE Access, 10, 37512–37527.

[13] Khan, M. A. U. H., et al. (2025). Secure Energy Transactions Using Blockchain Leveraging AI for Fraud Detection and Energy Market Stability. arXiv preprint arXiv:2506.19870.

[14] Lago, J., De Ridder, F., & De Schutter, B. (2018). Forecasting day-ahead electricity prices using deep learning. Applied Energy, 221, 386–405.

[15] Li, Y., Liu, J., Wang, J., & Han, S. (2021). Urban energy system transition: A review on models and applications. Renewable and Sustainable Energy Reviews, 148, 111287.

[16] Lund, H., Østergaard, P. A., Connolly, D., & Mathiesen, B. V. (2017). Smart energy and smart energy systems. Energy, 137, 556–565.

[17] Ochoa, L. F., & Padilha-Feltrin, A. (2018). Advanced distribution network planning with high penetrations of distributed generation. IEEE Transactions on Power Systems, 33(6), 6915–6926.

[18] Quan, H., Srinivasan, D., & Khosravi, A. (2014). Short-term load and wind power forecasting using interval type-2 fuzzy neural networks. Electric Power Systems Research, 119, 19–26.

[19] Shawon, R. E. R., Supti, A. M., Kawsar, M., Al Jobaer, M. A., Islam, M. F., Sumon, P. D., ... & Malla, S. (2025). Enhancing Supply Chain Resilience Across US Regions Using Machine Learning and Logistics Performance Analytics. International Journal of Applied Mathematics, 38(4s).

[20] Shovon, M. S. S., Gomes, C. A., Reza, S. A., Bhowmik, P. K., Gomes, C. A. H., Jakir, T., ... & Hasan, M. S. (2025). Forecasting Renewable Energy Trends in the USA: An AI-

Received: August 13 2025

International Journal of Applied Mathematics

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Driven Analysis of Electricity Production by Source. Journal of Ecohumanism, 4(3), 322–345.

[21] Strbac, G., Pudjianto, D., & Papadaskalopoulos, D. (2019). Decarbonization of electricity systems in Europe: Market design and regulations. Wiley Interdisciplinary Reviews: Energy and Environment, 8(1), e321.

[22] Sultana, K. S., Begum, M., Abed, J., Siam, M. A., Sadnan, G. A., Shatyi, S. S., & Billah, M. (2025). Blockchain-Based Green Edge Computing: Optimizing Energy Efficiency with Decentralized AI Frameworks. Journal of Computer Science and Technology Studies, 7(1), 386–408.

[23] Wang, Y., & Hong, T. (2020). Review of probabilistic load forecasting in the smart grid context. Renewable and Sustainable Energy Reviews, 130, 109901.

[24] Yan, R., Saha, T. K., & Eghbal, M. (2020). Data-driven probabilistic hosting capacity estimation for LV networks considering spatial uncertainty. IEEE Transactions on Power Systems, 35(5), 3912–3923.

[25] Zambrano-Asanza, E. D., Quirós-Tortós, J., & Ochoa, L. F. (2021). Impacts of high distributed generation penetrations on low-voltage networks: A review. Renewable and Sustainable Energy Reviews, 145, 111008.

[26] Zepter, J. M., Schill, W. P., & Faulwasser, T. (2020). When flexibility is scarce: Effects of variable renewable energy on emissions and prices. Energy Economics, 87, 104709.

[27] Zhang, Y., Wang, X., & Wang, N. (2021). A data-driven machine learning approach for distribution system hosting capacity estimation. IEEE Transactions on Smart Grid, 12(6), 5092–5103.

[28] Ziel, F., & Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Uncertainty quantification and regularization. International Journal of Forecasting, 34(3), 480–494.

Received: August 13 2025