IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 5 June 2026, accepted 23 July 2026, date of publication 27 July 2026, date of current version 31 July 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3717298

RESEARCH ARTICLE

Sensor-Limited Critical-Bus Voltage Forecasting for Virtual Metering in DER-Rich Distribution Networks

GISEOP NOH[1] AND HAN PYO LEE[2], (Senior Member, IEEE)
[1]Department of Software and Communications Engineering, Hongik University, Sejong 30016, Republic of Korea
[2]Department of Electronic and Electrical Convergence Engineering, Hongik University, Sejong 30016, Republic of Korea

Corresponding author: Han Pyo Lee (hplee@hongik.ac.kr)

This work was supported by 2026 Hongik University Research Fund and Korea Electric Power Corporation under Grant R25XO03-3.

ABSTRACT Distribution networks with high photovoltaic generation, electric-vehicle charging, and customer-side secondary circuits require better voltage awareness, but dense voltage sensing at every critical location is rarely practical. This paper formulates sensor-limited critical-bus voltage forecasting as a virtual metering task: historical voltage measurements from a small set of observed feeder locations are used to forecast future voltage trajectories at unmetered critical buses. The study uses a DER-rich modified IEEE 123-bus distribution-system dataset with secondary-circuit effects and constructs a reproducible benchmark pipeline with fixed chronological train/validation/test splits, metadata-based signal resolution, direct multi-output forecasting, and target-wise and horizon-wise evaluation. Four representative temporal baselines are evaluated: persistence, multilayer perceptron (MLP), one-dimensional convolutional neural network (CNN1D), and long short-term memory (LSTM). Sparse observation cases use only one to six voltage sensors, while a dense non-target observation case is included as a near full-observation reference rather than as a practical deployment scenario. The results show that accurate 24-hour multi-step voltage forecasting is feasible under low observability. In the primary sparse cases, CNN1D and LSTM achieve test mean absolute error near 0.0048 p.u. across three unmetered target buses, with stable multi-seed behavior. The dense non-target reference provides no large accuracy gain over the strongest sparse learned models, indicating that the evaluated sparse measurements already contain substantial predictive information for the selected critical buses. Target-wise and horizon-wise analyses further show that forecasting difficulty varies across buses and lead times. The paper therefore provides a defensible benchmark and feasibility evidence for sensor-limited virtual metering in DER-rich distribution networks, while explicitly limiting claims beyond the evaluated dataset, sensor sets, and baseline models.

INDEX TERMS Critical-bus voltage forecasting, distribution system monitoring, limited observability, secondary circuits, sensor-limited virtual metering, time-series forecasting, voltage inference.

I. INTRODUCTION

Distribution networks are becoming increasingly active because photovoltaic (PV) generation, electric vehicle (EV) charging, heat pumps, and other distributed energy resources (DERs) are being connected close to customers. These resources can improve decarbonization and flexibility, but they also change feeder operating conditions by introducing reverse power flow, sharper net-load ramps, phase imbalance, equipment loading stress, and voltage deviations [1]. Voltage monitoring is therefore a basic requirement for operating DER-rich distribution systems. However, distribution grids contain many buses, laterals, phase connections, and customer-side secondary circuits, so installing high-resolution voltage sensors at all potentially critical locations is rarely economical or practical.

The associate editor coordinating the review of this manuscript and approving it for publication was Mouloud Denai.

The low-observability nature of distribution systems has motivated extensive work on distribution system state estimation (DSSE), pseudo-measurements, sensor placement, and data-driven monitoring. Recent reviews emphasize that DSSE remains challenging under high DER penetration because distribution measurements are sparse, heterogeneous, and often less redundant than transmission-system measurements [2]. Related smart-meter-based methods also show that useful voltage and current estimates can be obtained from a limited number of measurement locations, but they typically require explicit estimation procedures, topology assumptions, or additional power-flow-related measurements [3]. In parallel, sensor placement studies have examined how a limited number of voltage sensors can support detection of voltage violations caused by stochastic DER and load variations [4]. These studies support the broader premise that limited measurements can still provide operationally valuable voltage information, but they do not directly answer how well future critical-bus voltages can be forecast from a small set of available voltage observations.

This paper studies that question from a virtual metering perspective. Instead of assuming that every critical bus is directly measured, we formulate sensor-limited critical-bus voltage forecasting as a supervised time-series problem: past voltage measurements from a small observation set are used to forecast future voltage trajectories at selected critical buses. The term virtual metering is used here in a pragmatic sense: the model provides estimated voltage time series at buses where physical sensors are unavailable or assumed to be absent. The focus is forecasting rather than static state estimation, because future voltage trajectories are useful for planning monitoring actions, ranking critical locations, and assessing whether limited observability is sufficient before more complex control or mitigation functions are considered.

The experimental setting is a modified IEEE 123-bus distribution system with PV, EV charging, and secondary circuits. The modified system extends the standard IEEE 123-bus feeder by incorporating DER-rich operating conditions, including PV and EV profiles, and low-voltage secondary-circuit representations downstream of distribution transformers. The IEEE radial distribution test feeders provide a widely used basis for evaluating distribution-system algorithms [5], and the modified setting used in this study extends that basis toward a more realistic DER-rich and customer-side voltage-inference problem. The available dataset contains time-series voltage, active-power, and reactive-power signals generated from this system. In the first-stage IEEE Access paper, we intentionally avoid claims that require unavailable scenario-generation information, such as PV/EV penetration robustness, OpenDSS-based counterfactual simulations, or voltage-violation mitigation. Instead, the paper asks a narrower and directly testable question: under fixed train/validation/test splits and practical sensor constraints, how accurately can baseline temporal models forecast multi-step critical-bus voltages?

This scope is deliberate. Many modern forecasting and learning-based monitoring studies emphasize new model architectures, but a credible limited-sensor voltage inference study also needs strong baselines, transparent preprocessing, fixed data splits, and target-wise and horizon-wise evaluation. Sequence-aware models such as long short-term memory (LSTM) networks are designed to capture temporal dependencies in sequential data [6], but their value in this setting should be evaluated against simpler alternatives, including persistence, multilayer perceptrons, and convolutional temporal baselines. This distinction from conventional distribution system state estimation is important because classic and review DSSE studies primarily address present-state monitoring under sparse measurements, whereas this paper evaluates future voltage-trajectory forecasting at unmetered critical buses [7], [8]. The goal of this paper is therefore not to introduce a new deep learning architecture. It is to provide a reproducible benchmark and feasibility study for sensor-limited critical-bus voltage forecasting in a DER-rich distribution-network dataset with secondary-circuit effects.

The empirical results further show that the dense non-target observation reference does not produce a large accuracy gain over the strongest sparse learned models. This finding supports the practical premise of the paper: for the selected critical buses and dataset, the dominant predictive information can be captured from a very small number of voltage sensors.

The main contributions of this paper are as follows:

1. **A sensor-limited voltage forecasting formulation.** We formulate critical-bus voltage forecasting as a virtual metering problem in a DER-rich modified IEEE 123-bus distribution system with secondary circuits.
2. **A reproducible benchmark pipeline.** We develop a reproducible pipeline for limited-sensor voltage forecasting, covering dataset processing, fixed train/validation/test splits, baseline training, inference, and target-wise and horizon-wise evaluation. The core implementation and a compact sample dataset are made publicly available on GitHub for reproducibility, subject to full-data sharing constraints.
3. **An empirical feasibility analysis under low observability.** We quantify the performance of representative temporal baselines, including persistence, MLP, CNN1D, and LSTM models, under practical sparse-sensor constraints using only 1–6 voltage sensor locations; use a dense non-target observation case as a near full-observation reference; and show that accurate multi-step critical-bus voltage forecasting is feasible even with a small number of voltage sensors.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the modified IEEE 123-bus system and dataset. Section IV formulates the sensor-limited forecasting problem. Section V presents the experimental pipeline and baseline models. Section VI reports the empirical results. Section VII discusses implications and limitations. Section VIII describes threats to validity. Section IX concludes the paper.

II. RELATED WORKS

A. VOLTAGE ESTIMATION AND FORECASTING UNDER LOW OBSERVABILITY

Distribution-network voltage monitoring has traditionally been studied through distribution system state estimation (DSSE), voltage estimation, and measurement placement. Early DSSE work formulated real-time monitoring for distribution systems under sparse and imperfect measurements, recognizing that distribution feeders differ from transmission grids because they are radial or weakly meshed, less instrumented, and often unbalanced [7]. Later reviews further emphasize that DSSE must handle pseudo-measurements, limited real-time sensors, heterogeneous smart-meter data, and topology-dependent uncertainty [8]. Recent surveys of emerging DSSE concepts also identify low observability, DER variability, and limited measurement redundancy as persistent barriers to reliable distribution-grid monitoring [2]. Recent learning-based studies for unobservable low-voltage grids further show that neural estimators are being developed specifically to compensate for limited measurement infrastructure in asymmetric three-phase LV systems [9].

Smart-meter and virtual-metering approaches address related observability gaps by using available measurements to infer electrical quantities at locations without dedicated sensors. For example, smart-meter-based voltage and current estimation methods show that sparse measurement infrastructure can still provide useful information about unmeasured parts of active distribution networks [3]. Sensor placement studies similarly show that a limited number of voltage sensors can be selected to improve the detection of voltage violations in distribution systems [4]. These studies support the premise that complete measurement coverage is not always required for useful voltage awareness.

However, state estimation and forecasting solve different operational problems. DSSE estimates the present system state from measurements, pseudo-measurements, and network models, whereas this paper forecasts future critical-bus voltage trajectories from historical voltage observations at a small set of sensors. The distinction matters because a forecasting benchmark must evaluate temporal generalization, multi-step horizon behavior, target-wise errors, and sensor-set dependence under fixed data splits. This paper therefore builds on the low-observability monitoring literature but focuses on a data-driven virtual metering task in which limited past voltage measurements are used to predict future voltages at unmeasured critical buses.

B. IMPACT OF DER PENETRATION AND SECONDARY CIRCUITS ON VOLTAGE PROFILES

PV generation, EV charging, heat pumps, and other low-carbon technologies change distribution-voltage behavior by altering both the magnitude and direction of power flow. High PV output can produce local voltage rise and reverse power flow, while clustered EV charging can increase evening peak demand and cause voltage drops. A recent overview of PV, EV, and heat-pump impacts reports that these technologies can intensify voltage deviations, phase imbalance, transformer loading, and feeder congestion in distribution grids [1]. Low-voltage distribution-grid reviews reach similar conclusions, emphasizing that high PV and EV penetration can create voltage-limit violations, asymmetric loading, congestion, and phase-balancing challenges close to end users [10].

These effects are particularly important in low-voltage secondary circuits. Secondary circuits connect distribution transformers to customer-side loads and DERs through service conductors, and their voltage profiles can differ from primary feeder voltages because of service-line impedance, single-phase connections, customer diversity, and unbalanced loading. Prior work on plug-in battery EV charging in secondary distribution systems shows that residential EV charging can increase voltage unbalance, undervoltage risk, and neutral current in secondary networks [11]. This indicates that feeder-level voltage observations alone may not fully represent customer-side voltage behavior when DERs are connected near end users.

The present study uses a DER-rich modified IEEE 123-bus dataset that includes secondary-circuit effects, so the forecasting task is closer to practical customer-side virtual metering than a feeder-only voltage prediction problem. At the same time, the scope is deliberately limited: this paper evaluates forecasting accuracy under the available DER-rich simulation condition and does not claim robustness to changing PV or EV penetration levels. Such scenario robustness analysis requires explicit PV/EV-to-bus mappings and network model files to regenerate counterfactual power-flow simulations, which is beyond the scope of the available static dataset. Instead, the focus of this study is strictly placed on verifying the temporal learning capabilities of baseline models under a fixed, high-penetration operational scenario.

C. DATA-DRIVEN TEMPORAL MODELING AND POSITIONING OF THIS STUDY

Data-driven time-series forecasting has been widely studied using both simple baselines and deep learning models. General surveys of deep learning for time-series forecasting identify feedforward networks, convolutional models, recurrent networks, encoder–decoder structures, and attention-based models as common families for learning temporal dependencies from historical data [12]. In energy systems, comparative studies of electric-load forecasting show that feedforward, recurrent, sequence-to-sequence, and temporal convolutional models are all relevant baseline classes, and that careful empirical comparison is needed because model performance depends on the dataset, horizon, input representation, and evaluation protocol [13]. These observations are directly relevant to voltage forecasting, where strong temporal autocorrelation can make persistence competitive while nonlinear and sequence-aware models may still improve multi-step prediction.

The baseline selection in this paper follows this pragmatic view. Persistence is included as a deterministic lower-complexity baseline because voltage time series are often strongly autocorrelated; in this study, it is implemented as a last-observation persistence variant that repeats the most recent input-voltage vector over the forecast horizon, with a linear input-to-target projection when the input and target dimensions differ. The MLP baseline tests whether a non-sequential neural model can learn useful mappings from flattened historical windows. The CNN1D baseline captures local temporal patterns through convolution over the lookback window, while the LSTM baseline represents recurrent sequence-aware modeling and is motivated by its ability to learn long-term temporal dependencies [6].

The goal is not to prove that any one architecture is universally superior, but to establish whether standard temporal baselines can support reliable critical-bus voltage forecasting under limited observability. Although a wider range of advanced sequence models, including Transformer-based architectures, could be incorporated into the benchmark, this paper does not aim to exhaustively compare state-of-the-art forecasting architectures. Instead, its primary objective is to establish a rigorous baseline framework and to verify the fundamental predictability of critical-bus voltage trajectories under sparse sensing constraints.

This positioning differentiates the present study from model-novelty-driven work. Many data-driven forecasting papers emphasize new architectures or architecture-specific improvements. In contrast, this paper focuses on problem formulation, benchmark construction, and empirical feasibility in a DER-rich modified IEEE 123-bus setting with secondary-circuit effects. The resulting contribution is a reproducible sensor-limited virtual metering benchmark: fixed observation cases, fixed target buses, fixed train/validation/test splits, standard temporal baselines, and target-wise and horizon-wise evaluation for multi-step critical-bus voltage forecasting.

This scope is also consistent with recent distribution-grid monitoring studies that emphasize practical observability limits. Smart-meter-based DSSE and two-layer MV/LV estimation studies show that partial measurement coverage can improve distribution-grid visibility, but they still depend on measurement availability, model assumptions, and communication constraints [14], [15]. Low-voltage voltage-estimation studies using alternative sensing sources or neural estimators further demonstrate that data-driven methods can infer unmeasured voltages in DER-rich settings, while requiring careful validation on realistic data [16], [17]. Recent work on smart-meter timeliness, voltage-sensitivity estimation, and secondary-system sensitivity structure also reinforces that low-voltage and secondary-circuit monitoring is constrained by measurement access, topology uncertainty, and customer-side circuit behavior [18], [19], [20].

III. SYSTEM AND DATASET DESCRIPTION

A. MODIFIED IEEE 123-BUS DISTRIBUTION SYSTEM

The test system used in this study is a modified IEEE 123-bus distribution system developed for this collaborative research. The original IEEE 123-bus feeder is widely used as a benchmark for distribution-system analysis because it contains long radial laterals, unbalanced phase sections, voltage regulators, and heterogeneous load locations. The modified system extends this benchmark setting by incorporating DER-rich operating conditions and low-voltage secondary-circuit representations, making it suitable for studying sensor-limited voltage inference rather than feeder-level power-flow behavior.

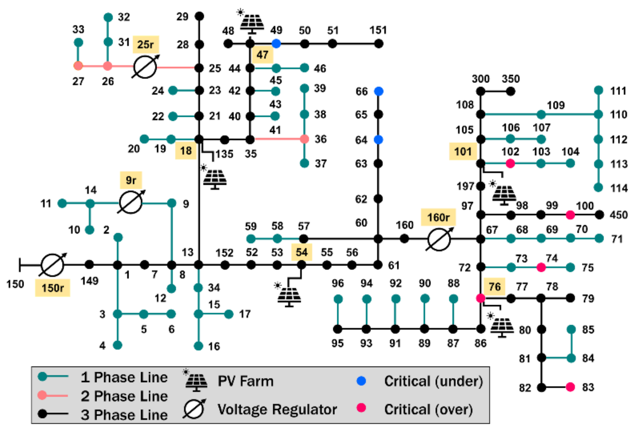

FIGURE 1. Modified IEEE 123-bus distribution system used in this study, including PV units, voltage regulators, and critical-bus locations.

Figure 1 shows the system overview used for the forecasting study. The diagram identifies primary feeder buses, phase sections, voltage regulators, PV-connected locations, and critical buses. The voltage regulators provide practical measurement locations because they are operationally meaningful and more likely to be monitored than arbitrary customer-side nodes. PV-connected buses are also relevant observation candidates because local generation can substantially affect nearby voltage profiles. In contrast, the critical buses are treated as target locations whose future voltage trajectories must be inferred from a limited set of sensor measurements. Figure 1 primarily depicts the primary feeder-level layout; the secondary circuits are modeled downstream of selected primary buses and are represented in the dataset through x. . . and sx. . . node identifiers rather than being individually expanded in the system overview.

The available time-series dataset was generated from this modified system and contains voltage, active-power, and reactive-power signals at 15-minute resolution. In the first-stage benchmark reported in this paper, the forecasting problem is defined on phase-A voltage signals, with selected regulator and PV-connected buses used as sensor inputs and selected critical buses used as prediction targets. Although the dataset contains voltage, active-power, and reactive-power time series, the benchmark models in this paper use only past voltage measurements as inputs; active- and reactive-power signals are retained as dataset context and for possible future extensions. This setting represents the virtual metering concept: the model relies solely on past voltage measurements from a few physical sensors to forecast future voltage trajectories at unmetered critical locations.

B. SECONDARY CIRCUIT REPRESENTATION

A key distinction of the modified system is that the dataset includes not only primary feeder buses but also low-voltage secondary-side nodes. In distribution systems, secondary circuits are the customer-side networks connected downstream of distribution transformers. They contain service conductors, customer load connections, and locations where customer-side DERs such as rooftop PV or EV charging loads can affect local voltages. As a result, voltage behavior in secondary circuits can deviate from the voltage profile observed at the upstream primary feeder bus.

This distinction is reflected in the dataset identifiers. Primary feeder nodes are represented by bus-oriented names such as 49.1.2.3, where the suffix indicates the available phases. In contrast, identifiers beginning with x or sx are interpreted as transformer secondary-side or service/customer-side nodes. For example, nodes such as x49a.1.0, sx49 a1.1.2, and sx49a2.1.2 are associated with the secondary circuit connected around primary bus 49 on phase A.

These identifiers are not arbitrary measurement labels; they indicate that the simulated voltage time series includes low-voltage locations downstream of the primary feeder. Figure 2 provides a representative secondary-circuit structure to illustrate how transformer secondary-side and customer-side voltage locations arise downstream of a primary feeder bus. The x-node and sx-node labels in the figure should be interpreted as guide labels for the naming convention, not as an exhaustive mapping from every diagram element to every dataset column.

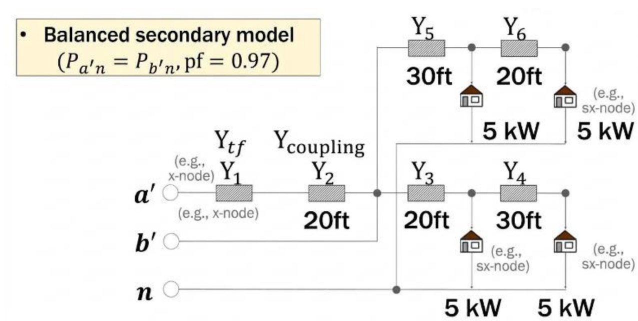

FIGURE 2. Representative balanced low-voltage secondary circuit model used to illustrate transformer secondary-side and customer-side voltage locations. The x-node and sx-node labels are naming-convention guides; dataset node identifiers such as x... and sx... indicate simulated secondary-side voltage locations rather than a one-to-one mapping to every element in the diagram.

Including secondary circuits is important for the virtual metering task considered in this paper. Customer-side voltage can be affected by service-line impedance, local load diversity, single-phase connections, EV charging, and distributed generation near end users. Therefore, a forecasting model evaluated only on primary feeder buses may underrepresent the voltage variability experienced by customers.

The modified dataset used here provides a more realistic setting because it contains voltage trajectories from both primary feeder nodes and secondary-side nodes, even though the first-stage benchmark focuses on selected phase-A sensor and target locations. Although Figure 2 shows a balanced secondary model for clarity, it is used only as a representative structural example; the dataset-level voltage behavior reflects the broader modified feeder simulation, including phase-specific and customer-side voltage variations.

The present paper does not use detailed secondary-circuit topology, transformer parameters, or service-line impedances as model inputs. Instead, the secondary-circuit effects are implicitly included in the simulated P/Q/V time-series data generated from the modified system.

This choice keeps the first-stage IEEE Access study focused on sensor-limited temporal forecasting and reproducible baselines. More detailed topology-aware modeling of transformer-secondary relationships, customer connections, and line parameters is left for future work when complete network-model metadata are available.

C. DATASET SIGNALS AND TEMPORAL RESOLUTION

The dataset contains time-series simulation outputs for voltage magnitude, active power, and reactive power at the nodes of the modified distribution system. The raw signals are organized by electrical phase and measurement type, so phase-specific voltage matrices such as V_A can be processed independently from the corresponding active- and reactive-power matrices. This structure allows the same system to support multiple future inference settings, including voltage-only forecasting, power-assisted forecasting, and phase-specific analysis.

In this paper, the benchmark is restricted to phase-A voltage forecasting. Although active-power and reactive-power signals are available in the dataset, they are not used as model inputs in the first-stage experiments. The input tensor contains only historical voltage measurements from the selected sensor locations, and the output tensor contains future voltage trajectories at the selected critical buses.

This design is consistent with the intended virtual metering setting, in which a small number of physical voltage sensors are available while the target critical buses remain unmetered during inference. Figure 3 summarizes this voltage-only data-processing procedure, including phase-specific signal selection, sliding-window sample construction, and chronological train/validation/test splitting.

As shown in Figure 3, the raw simulation outputs first provide phase- and signal-specific matrices, from which the phase-A voltage matrix is selected for the present benchmark. The selected sensor columns form the input tensor X, whereas the critical-bus voltage columns form the

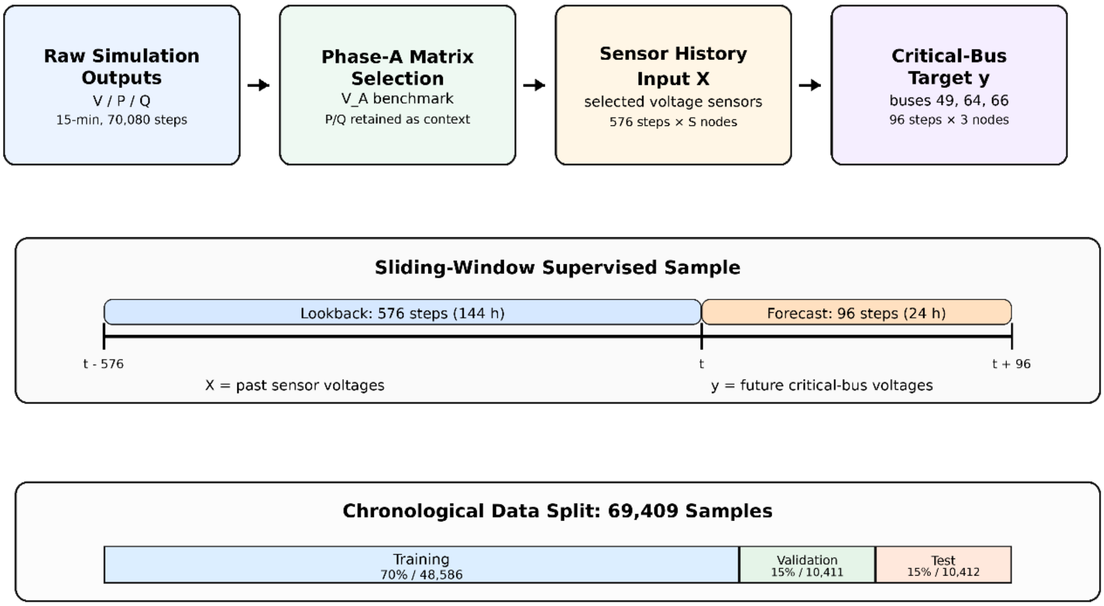

FIGURE 3. **Dataset processing pipeline for the phase-A voltage-only benchmark. The raw simulation data include voltage, active-power, and reactive-power signals, but the first-stage models use only selected sensor-voltage histories as inputs; active- and reactive-power signals are retained as dataset context and possible future extensions. Here, S denotes the case-dependent number of sensor locations.**

target tensor y. Sliding windows are then generated without random shuffling, and the resulting samples are divided chronologically to preserve the temporal ordering of the forecasting problem.

The time resolution is 15 minutes, and the available sequence spans 70,080 time steps, corresponding to 730 days of data. Sliding-window samples are generated using a lookback length of 576 steps and a forecast horizon of 96 steps. At 15-minute resolution, these correspond to a 144-hour historical window (6 days) and a 24-hour prediction horizon (1 day), respectively. The resulting processed dataset contains 69,409 supervised samples, which are split chronologically into training, validation, and test subsets with a 70%/15%/15% ratio.

The 144-hour (6-day) lookback window is intentionally reported in steps, hours, and days to avoid ambiguity. Earlier planning notes described a 24-hour lookback, but the current processed benchmark configuration uses lookback_steps = 576. Therefore, all results reported from the present processed tensors should be interpreted as six-day-history to one-day-ahead voltage forecasting unless the preprocessing configuration is changed in a later experiment.

D. SENSOR AND TARGET DEFINITIONS

The forecasting task separates observed sensor nodes from unobserved critical target buses. The sensor set represents the limited set of voltage measurements assumed to be available to the operator during inference, whereas the target set represents critical buses whose future voltage trajectories must be inferred virtually. This separation is important because the target-bus voltages are not included in the input tensor during inference; they are used only as supervised learning targets during training and as reference values during evaluation.

For the phase-A benchmark, the critical target buses are fixed as buses 49, 64, and 66. In the processed voltage matrix, these targets correspond to the phase-A voltage columns resolved from 49.1.2.3, 64.1.2.3, and 66.1.2.3, respectively. These buses are treated as unmetered critical locations in the virtual metering task. Therefore, the model is evaluated on its ability to forecast the future voltage trajectories of these three buses from measurements collected elsewhere in the network.

Two sparse observation settings are currently prepared from the processed dataset. The regulator-only case uses three voltage sensor locations, 150r, 9r, and 25r, which correspond to the resolved columns 150r, 9r.1, and 25r.1. These locations are voltage-regulator-related nodes rather than arbitrary measurement points. Because voltage regulators are operational devices that shape feeder voltage profiles, their associated measurements are treated as engineering-informed sparse sensors that can summarize important voltage-control behavior in the feeder. The expanded sparse case adds three additional feeder locations, 18, 47, and 54, resolved as 18.1.2.3, 47.1.2.3, and 54.1.2.3, resulting in six input sensor locations. In both cases, the input feature is voltage magnitude only, and the same three critical target buses are used. This design allows the experiments to quantify how additional sparse measurements affect critical-bus voltage forecasting accuracy while keeping the target definition fixed.

The resulting tensor dimensions follow directly from the selected sensor and target sets. For a sensor set of size S, each input sample has shape $576 \times S$, corresponding to a 144-

hour (6-day) voltage history at 15-minute resolution. Each target sample has shape 96×3, corresponding to a 24-hour (1-day) forecast horizon for the three critical buses. Thus, the regulator-only case uses input tensors with $S = 3$, whereas the expanded sparse case uses $S = 6$. Table 1 summarizes the main dataset and experiment configuration used throughout this study.

TABLE 1. Dataset and experiment configuration.

Item	Value
Test system	Modified IEEE 123-bus distribution system
DER components	PV, EV charging loads
Secondary circuits	Included
Time resolution	15 minutes
Input signal	Voltage magnitude
Source phase	Phase A, unless otherwise stated
Lookback window	576 steps (144 h, 6 days)
Forecast horizon	96 steps (24 h, 1 day)
Target buses	49, 64, 66
Train/validation/test split	70% / 15% / 15% chronological split

E. DATA LIMITATIONS

The dataset used in this paper consists of pre-generated time-series simulation outputs from the modified IEEE 123-bus distribution system. The available files provide voltage, active-power, and reactive-power trajectories, which are sufficient for defining and evaluating the sensor-limited voltage forecasting task considered here. However, the complete simulation-generation package, including detailed OpenDSS model files, PV and EV placement metadata, scenario-generation scripts, and all secondary-circuit parameter files, is not assumed to be available in the present study.

This limitation affects the scope of the claims made in this paper. The experiments evaluate whether future critical-bus voltages can be forecast from sparse voltage measurements under the provided DER-rich simulated operating conditions. They do not attempt to quantify robustness across arbitrary PV penetration levels, EV adoption scenarios, secondary-circuit design variants, or alternative protection and control settings. Such scenario-sensitivity studies require direct access to the full simulation model and scenario-generation process and are therefore left for future work.

The same limitation also motivates the voltage-only benchmark design. Although active- and reactive-power signals are available in the dataset, the first-stage experiments intentionally use only historical voltage measurements from selected sensor locations. This choice avoids relying on unavailable deployment assumptions about real-time power measurements at every node and keeps the benchmark aligned with the virtual metering objective. Consequently, the reported results should be interpreted as empirical evidence for sensor-limited temporal predictability within the given simulated dataset, not as a universal performance guarantee for all distribution feeders or all DER deployment scenarios.

IV. PROBLEM FORMULATION

A. SENSOR-LIMITED VIRTUAL METERING TASK

The objective of the proposed task is to forecast future voltages at unmetered critical buses by using only historical voltage measurements from a limited set of observed sensor locations. Let $\mathcal{S}$ denote the set of observed sensor nodes and $\mathcal{C}$ denote the set of critical target buses. The two sets are disjoint, so that $\mathcal{S} \cap \mathcal{C} = \emptyset$. This disjoint definition reflects the virtual metering setting: target-bus voltages are available during training and evaluation, but they are not used as input measurements during inference.

For each time index t, the input is a lookback window of phase-A voltage measurements from the sensor set, as defined in (1):

$$\boldsymbol{X}_t = [\boldsymbol{v}^{\mathcal{S}}_{t-L+1}, \boldsymbol{v}^{\mathcal{S}}_{t-L+2}, \ldots, \boldsymbol{v}^{\mathcal{S}}_{t}]^{\top} \in \mathbb{R}^{L\times\mathcal{S}}, \tag{1}$$

where L is the lookback length and $S =| \mathcal{S} |$ is the number of available sensor locations.

Here, $\mathbf{v}^{\mathcal{S}}_t \in \mathbb{R}^S$ denotes the voltage-magnitude vector at time step t for the sensor locations. The corresponding prediction target is the future voltage trajectory at the critical buses, as defined in (2):

$$\boldsymbol{Y}_t = [\mathbf{v}^{\mathcal{C}}_{t+1}, \mathbf{v}^{\mathcal{C}}_{t+2}, \ldots, \mathbf{v}^{\mathcal{C}}_{t+H}]^{\top} \in R^{H\times C}, \tag{2}$$

where H is the forecast horizon and $C =| \mathcal{C} |$ is the number of critical buses.

Similarly, $\mathbf{v}^{\mathcal{C}}_t \in \mathbb{R}^C$ denotes the voltage-magnitude vector at time step t for the target locations. In the present benchmark, $L = 576$, $H = 96$, and $C = 3$, corresponding to a 144-hour (6-day) input history, a 24-hour (1-day) forecast horizon, and the three target buses 49, 64, and 66. The value of S depends on the observation case: for example, $S = 3$ for the regulator-only case and $S = 6$ for the expanded sparse case.

The learning problem is to estimate a forecasting function f_θ that maps the sensor-voltage history in (1) to the future critical-bus voltage trajectory in (2), as defined in (3):

$$\widehat{\boldsymbol{Y}}_t = f_\theta\left(\boldsymbol{X}_t\right), \tag{3}$$

where θ denotes the trainable model parameters.

The model is trained by minimizing a supervised forecasting loss between $\widehat{\boldsymbol{Y}}_t$ and $\boldsymbol{Y}_t$ over the training windows; the specific loss and evaluation metrics are defined in the following subsections. This formulation is model-agnostic: f_θ can be instantiated as a persistence baseline, MLP, CNN1D, LSTM, or another temporal forecasting model, while the input-output definition remains unchanged.

B. OBSERVATION CASES

The observation case determines which voltage signals are available to the forecasting model. For each case k, let $\mathcal{S}_k$ denote the corresponding sensor set and let $S_k =| \mathcal{S}_k |$ denote the number of observed sensor locations. The critical target set is fixed across all cases as $\mathcal{C} = \{49, 64, 66\}$, and the no-leakage condition in (4) is enforced for every case:

$$\mathcal{S}_k \cap \mathcal{C} = \emptyset. \tag{4}$$

Accordingly, under the no-leakage condition in (4), changing the observation case changes only the input dimension S_k of $\mathbf{X}_t \in \mathbb{R}^{L \times S_k}$, while the output dimension remains $\mathbf{Y}_t \in \mathbb{R}^{H \times 3}$. This preserves the forecasting interface introduced in (1)–(3). This design isolates the effect of sensor availability on critical-bus voltage forecasting accuracy. In other words, all cases predict the same target buses over the same 24-hour horizon, but they differ in how much real-time voltage information is assumed to be available from the feeder.

The primary practical cases are the regulator-only case and the expanded sparse case. The regulator-only case, denoted as reg_3, uses the three voltage-regulator-related nodes 150r, 9r, and 25r. This case represents a highly sparse monitoring condition in which only a few operationally meaningful feeder measurements are available. The choice of regulator-related nodes is intentional: these locations are associated with feeder voltage-control equipment and are therefore more informative than randomly selected nodes for tracking feeder-wide voltage behavior. The expanded sparse case, denoted as reg_pv_6, adds three additional feeder locations, 18, 47, and 54, to the regulator-related sensors. These additional locations are treated as PV-connected or DER-relevant observation points in the benchmark, allowing the experiments to test whether a modest increase in sensor coverage improves multi-step voltage forecasting performance.

Two ablation cases are also included to separate the effect of sensor type and sensor count. The reg_1 case uses only 150r and represents an extreme low-observability setting. The pv_3 case uses only 18, 47, and 54, thereby testing whether DER-relevant observation points alone provide useful predictive information for the same critical target buses. These ablations are not intended to represent complete utility deployment plans; they are included to quantify how different sparse measurement choices affect predictability.

Finally, the dense_oracle case is included as an explicitly defined dense non-target observation reference. In this case, all available non-target phase-A voltage nodes are used as inputs, while the critical target buses remain excluded from the input set. The processed phase-A voltage matrix contains 690 voltage measurement columns. After excluding the three target buses, 687 non-target input nodes remain for the dense reference case. The term oracle is used only in an experimental sense: the case is not a realistic sensor deployment and is not proposed as an operational monitoring architecture. Instead, it provides a dense non-target observation reference for interpreting whether sparse practical sensing loses substantial accuracy relative to a nearly full-observation setting. Table 2 summarizes the observation cases used in this study.

TABLE 2. Sensor observation cases.

Case	Sensor nodes	Num. sensors	Description
reg_3	150r, 9r, 25r	3	Regulator-only practical sparse baseline
reg_pv_6	150r, 9r, 25r, 18, 47, 54	6	Expanded sparse case with regulator and DER-relevant feeder observations
reg_1	150r	1	Extreme low-observability ablation
pv_3	18, 47, 54	3	DER-relevant observation-point ablation
dense_oracle	All non-target phase-A voltage nodes	687	Dense non-target observation reference, not a practical deployment case

C. EVALUATION METRICS

The models are trained as supervised multi-step regressors. For a training set containing M sliding-window samples, the default optimization objective is the mean squared error (MSE) in (5), computed over all forecast horizons and target buses:

$$\mathcal{L}_{MSE} = \frac{1}{MHC} \sum_{m=1}^{M} \sum_{h=1}^{H} \sum_{c=1}^{C} \left(\widehat{Y}_{m,h,c} \mid Y_{m,h,c}\right)^2, \tag{5}$$

where $\widehat{Y}_{m,h,c}$ and $Y_{m,h,c}$ denote the predicted and reference voltage magnitudes for sample m, forecast step h, and target bus c, respectively.

The loss in (5) is used only for model training and validation selection. Final performance is reported using complementary error metrics to capture both average and tail-error behavior.

The primary aggregate metrics are mean absolute error (MAE), root mean square error (RMSE), and percentile absolute error. Let $N = MHC$ denote the total number of evaluated scalar voltage predictions after flattening all samples, horizons, and target buses. The aggregate MAE and RMSE are defined in (6) and (7), respectively:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} | \widehat{y}_i - y_i |. \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\widehat{y}_i \mid y_i)^2}. \tag{7}$$

Tail errors are summarized using the empirical percentile of the absolute error distribution in (8):

$$P_q = Q_q\left(\left\{|\widehat{y}_i - y_i|_{i=1}^{N}\right\}\right), \tag{8}$$

where $Q_q(\cdot)$ denotes the empirical q -th quantile.

In this paper, P95 is used as the main tail-error metric, and P99 may be reported when additional tail-risk detail is needed. Maximum absolute error is also tracked as a diagnostic quantity, but it is interpreted cautiously because it can be dominated by isolated outliers.

To understand where errors occur, the same metrics can be computed by forecast horizon and by target bus. The horizon-wise MAE for forecast step h is defined in (9):

$$\mathrm{MAE}_h = \frac{1}{MC}\sum_{m=1}^{M}\sum_{c=1}^{C} |\,\widehat{Y}_{m,h,c} - Y_{m,h,c}\,|. \tag{9}$$

and the target-wise MAE for target bus c is defined in (10):

$$\mathrm{MAE}_c = \frac{1}{MH}\sum_{m=1}^{M}\sum_{h=1}^{H} \left|\widehat{Y}_{m,h,c} - Y_{m,h,c}\right|. \tag{10}$$

The decomposed metrics in (9) and (10) are important because voltage forecasting difficulty may vary across both lead time and critical-bus location. Horizon-wise metrics show whether errors accumulate over the 24-hour prediction window, whereas target-wise metrics reveal whether particular critical buses are more difficult to infer from sparse observations.

When voltage-limit analysis is reported, violations are defined using nominal per-unit thresholds. A ground-truth voltage is marked as an undervoltage event if $Y_{m,h,c} < 0.95$ and as an overvoltage event if $Y_{m,h,c} > 1.05$. The violation mask is therefore defined in (11):

$$\mathcal{V} = \left\{(m, h, c) : Y_{m,h,c} < 0.95 \text{ or } Y_{m,h,c} > 1.05\right\}. \tag{11}$$

Violation-conditioned errors are computed only over the set in (11), as shown in (12):

$$\mathrm{MAE}_{\mathcal{V}} = \frac{1}{|\,\mathcal{V}|}\sum_{(m,h,c)\in\mathcal{V}} |\,\widehat{Y}_{m,h,c} - Y_{m,h,c}\,|. \tag{12}$$

If $|\ \mathcal{V}\ |=\ 0$ for a given split or target subset, the violation-conditioned metric in (12) is not interpreted as a meaningful performance indicator. In that case, the paper reports the absence of valid violation samples rather than drawing conclusions from threshold-conditioned errors.

V. EXPERIMENTAL PIPELINE AND BASELINE MODELS

A. DATA PROCESSING PIPELINE

The experimental pipeline converts raw simulation outputs into fixed supervised learning tensors through three reproducible stages: interim matrix construction, metadata-based column resolution, and sliding-window tensor generation. This separation is intentional. The raw files preserve the original simulation outputs, the interim matrices provide efficient numerical access, and the processed tensors define the exact input-output samples used by the baseline models.

In the first stage, raw no-header CSV files are converted into NumPy matrices. Each source file is loaded as a two-dimensional floating-point array and saved as an interim .npy matrix. For the present benchmark, the main source matrix is the phase-A voltage matrix V_A.npy, which has shape 70080 × 690. The first dimension corresponds to 70,080 time steps at 15-minute resolution, and the second dimension corresponds to available phase-A voltage columns. Additional active- and reactive-power matrices are retained in the interim dataset for documentation and future extensions, but they are not used as model inputs in the first-stage voltage-only benchmark.

In the second stage, requested sensor and target bus identifiers are resolved to concrete matrix columns using the signal metadata table. This step is necessary because raw column names may include phase suffixes or feeder-specific identifiers. For example, requested target buses 49, 64, and 66 are resolved to 49.1.2.3, 64.1.2.3, and 66.1.2.3, respectively. Similarly, regulator-related sensor requests such as 9r and 25r are resolved to their available phase-A voltage columns. The resolved column names and zero-based column indices are written to each processed case configuration file, so that every experiment can be traced back to the exact matrix columns used to build its tensors.

In the third stage, supervised samples are generated by sliding a fixed lookback-horizon window over the selected matrix columns. For a case with sensor set $\mathcal{S}_k$ and target set $\mathcal{C}$, sample m is constructed using (13) and (14):

$$\mathbf{X}^{(m)} = \mathbf{V}_a\left[m : m + L - 1, \mathcal{S}_k\right]. \tag{13}$$
$$\mathbf{Y}^{(m)} = \mathbf{V}_a\left[m + L : m + L + H - 1, \mathcal{C}\right]. \tag{14}$$

where $\mathbf{V}_a$ denotes the phase-A voltage matrix, $L = 576$, and $H = 96$. The interval notation in (13) and (14) follows the implemented preprocessing convention: the input window includes time indices from m to $m + L - 1$, and the target window includes indices from $m + L$ to $m + L + H - 1$. With $T = 70080$ total time steps, the number of supervised samples is computed using (15):

$$\begin{aligned} M &= T - L - H + 1 \\ &= 70080 - 576 - 96 + 1 \\ &= 69409 \end{aligned} \tag{15}$$

The resulting M samples from (15) are split chronologically rather than randomly. The first 70% of windows are used for training, the next 15% for validation, and the final 15% for testing. This yields 48,586 training samples, 10,411 validation samples, and 10,412 test samples. Chronological splitting is used to preserve the temporal forecasting setting and to avoid information leakage from future windows into model selection or training.

For each observation case, the pipeline writes processed input and target tensors for the training, validation, and test splits. The corresponding case-level configuration metadata records the resolution, lookback length, forecast horizon, requested sensor nodes, resolved sensor columns, requested target nodes, resolved target columns, split ratios, sample counts, and tensor shapes. This metadata is part of the reproducibility design: the same processed tensors can be regenerated from the interim matrix and metadata table without relying on implicit notebook state or manual column selection.

B. BASELINE MODELS

The benchmark uses four representative temporal forecasting baselines: persistence, MLP, CNN1D, and LSTM. These models are intentionally selected to cover a range of modeling assumptions without turning the paper into a new-architecture study. The neural baselines are chosen to represent common feed-forward, convolutional, and recurrent sequence-modeling families used in energy time-series forecasting [6], [12]. Persistence measures how much predictive information is already contained in the most recent voltage observation. MLP provides a non-sequential neural baseline after flattening the lookback window. CNN1D captures local temporal patterns through convolution along the time axis. LSTM provides a recurrent sequence model that can summarize the full historical window before producing a multi-step forecast.

All trainable baselines share the same input-output interface. For a sample $\mathbf{X}_t \in \mathbb{R}^{L\times S}$, each model outputs $\widehat{\boldsymbol{Y}}_t \in \mathbb{R}^{H\times C}$. The output layer is sized as $H \times C$, so the models directly produce the full 24-hour target trajectory rather than recursively feeding one-step predictions back into the input. This direct multi-output design avoids error accumulation caused by autoregressive rollout and keeps the comparison focused on the representation learned from the sensor-voltage history.

The persistence baseline repeats the most recent observed input vector over the forecast horizon after mapping it to the target dimension. Let $\mathbf{x}_t \in \mathbb{R}^S$ denote the last sensor-voltage vector in the lookback window. The persistence prediction is defined in (16):

$$\widehat{\mathbf{Y}}_{t,h} = \mathbf{P}\mathbf{x}_t, h = 1, 2, \ldots , H, \tag{16}$$

where $\mathbf{P} \in \mathbb{R}^{C\times S}$ is an identity mapping when $S = C$ and a linear input-to-target projection when the sensor and target dimensions differ.

This baseline is included because voltage time series often exhibit strong short-term autocorrelation, and any neural model should be compared against this simple temporal reference.

The MLP baseline flattens the full lookback window into a vector and applies two fully connected hidden layers with ReLU activations and dropout. Its mapping can be written as (17):

$$\widehat{\mathbf{Y}}_t = \text{reshape}\left(g_{\text{MLP}}(\text{vec}(\mathbf{X}_t)), H, C\right), \tag{17}$$

where $g_{MLP}(\cdot)$ denotes the feed-forward network.

Here, vec$(\cdot)$ linearizes the input matrix into a column vector, and reshape$(\cdot, H, C)$ maps the flattened prediction back to the horizon-target grid. In the default configuration, the hidden dimension is 256 and dropout is 0.1. This model tests whether a simple nonlinear mapping from the entire historical window is sufficient for critical-bus voltage forecasting.

The CNN1D baseline treats the sensor channels as input channels and applies one-dimensional temporal convolutions over the lookback axis. The input is transposed from $L \times S$ to $S \times L$, passed through two convolutional layers with ReLU activations, and then summarized using adaptive average pooling before the final linear output head. The default configuration uses 64 channels, a kernel size of 5, and dropout of 0.1. This model emphasizes local temporal motifs while remaining less complex than recurrent or attention-based architectures.

The LSTM baseline processes the lookback sequence in chronological order and uses the final recurrent state to produce the full multi-step output. The default LSTM uses two layers, hidden size 128, dropout 0.1, and a unidirectional recurrent encoder. Its prediction can be summarized as (18):

$$\widehat{\mathbf{Y}}_t = \text{reshape}\left(\mathbf{W}_o\mathbf{h}_L + \mathbf{b}_o, H, C\right), \tag{18}$$

where $\mathbf{h}_L$ is the final hidden representation after processing the L input steps, and reshape$(\cdot, H, C)$ converts the flattened output layer into the multi-step target grid.

The LSTM is included as a sequence-aware baseline rather than as a claim of architectural novelty. Table 3 summarizes the baseline model configurations used in the benchmark.

TABLE 3. Baseline model configuration.

Model	Role	Key configuration
Persistence	Last-observation temporal reference	Repeats the most recent sensor-voltage vector over the horizon after identity or linear input-to-target projection
MLP	Non-sequential neural baseline	Flattened $L \times S$ input; two hidden layers; hidden dimension 256; ReLU; dropout 0.1
CNN1D [12]	Local temporal pattern baseline	Two temporal Conv1D layers; 64 channels; kernel size 5; ReLU; adaptive average pooling; dropout 0.1
LSTM [6]	Sequence-aware recurrent baseline	Two-layer unidirectional LSTM; hidden size 128; dropout 0.1; final-state multi-output head

C. TRAINING AND INFERENCE SETUP

All baseline models are trained and evaluated using the same processed train, validation, and test tensors. The training split is shuffled at the mini-batch level, whereas the validation and test splits are evaluated without shuffling. This preserves the chronological split defined during preprocessing while still allowing stochastic mini-batch optimization on the training windows. Unless otherwise stated, the batch size is 128, the maximum number of epochs is 20, and the random seed is fixed to 42 for Python, NumPy, and PyTorch.

Trainable models are optimized using Adam with learning rate 10^{-3} and weight decay 0. The optimization criterion is MSE, while model selection is based on validation MAE. At every epoch, the pipeline records training and validation loss, MAE, and RMSE. The checkpoint with the lowest validation MAE is retained for final evaluation, and the test set is used only after model selection has been completed.

The persistence baseline follows the same input-output interface as the neural baselines. When the input and target dimensions are equal, the persistence mapping uses an identity projection and contains no trainable parameters. When the dimensions differ, the implementation uses a linear input-to-target projection before repeating the projected last observation over the forecast horizon. In such cases, the projection parameters are optimized on the training split using the same MSE objective, ensuring that the persistence baseline remains compatible with sparse sensor sets whose dimension differs from the target-bus dimension.

Device selection is automatic: the training script uses a CUDA device when available and otherwise falls back to CPU. The selected device, model configuration, best epoch, best validation MAE, validation metrics, test metrics, and elapsed runtime are stored as run-level metadata. Epoch-level training records are also retained for reproducibility and diagnostic analysis. In addition, the evaluation routine computes aggregate error metrics, tail-error metrics, voltage-threshold counts, and violation-conditioned metrics using the thresholds 0.95 p.u. and 1.05 p.u.

The implementation is organized to support a reproducible benchmark release. The preprocessing scripts generate fixed processed tensors and case-level configuration metadata, the model definitions implement the four baselines using a common tensor interface, and the metric computation is centralized in the evaluation module. This organization allows reviewers or follow-up users to rerun the benchmark from released processed tensors or a compact sample dataset without manually reconstructing data splits, sensor columns, or target columns.

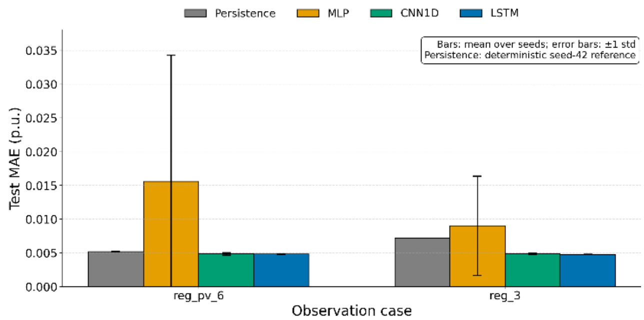

FIGURE 4. Multi-seed test MAE comparison across the two primary sparse observation cases and four baseline models. Bars indicate mean test MAE, and error bars indicate one standard deviation across three seeds for trainable models; persistence is shown as a deterministic reference.

TABLE 4. Multi-seed test performance across random seeds.

Case	Model	MAE	RMSE	P95 absolute error
reg_pv_6	Persistence	0.005201	0.006665	0.013451
reg_pv_6	MLP	0.015592 ± 0.018681	0.016702 ± 0.018241	0.025637 ± 0.021372
reg_pv_6	CNN1D	0.004852 ± 0.000171	0.006584 ± 0.000484	0.014722 ± 0.001220
reg_pv_6	LSTM	0.004807 ± 0.000013	0.006265 ± 0.000250	0.013680 ± 0.001042
reg_3	Persistence	0.007216	0.008416	0.014572
reg_3	MLP	0.009023 ± 0.007347	0.010782 \ ± 0.007716	0.020289 ± 0.010722
reg_3	CNN1D	0.004841 ± 0.000096	0.006202 ± 0.000116	0.013264 ± 0.000455
reg_3	LSTM	0.004770 ± 0.000009	0.006304 ± 0.000053	0.014046 ± 0.000208

VI. RESULTS

A. OVERALL MODEL COMPARISON

Table 4 reports the main test-set performance for the two primary sparse observation cases. The reg_pv_6 case corresponds to the expanded sparse setting with six sensor locations, whereas the reg_3 case corresponds to the regulator-only setting with three sensor locations. The values are reported in per-unit voltage magnitude. For the trainable neural baselines, the table reports mean and standard deviation over three random seeds. The persistence baseline is reported as a single deterministic reference using the fixed benchmark split.

Across both observation cases, the LSTM and CNN1D baselines achieve the strongest overall performance. In the expanded sparse case, LSTM obtains an MAE of 0.004807 ± 0.000013 p.u., while CNN1D remains close with 0.004852 ± 0.000171 p.u. The persistence baseline is also competitive in this case, with MAE 0.005201 p.u., indicating that the voltage trajectories contain strong short-term temporal auto-correlation. However, the MLP baseline shows substantially larger variance, suggesting that simply flattening the six-day lookback window is less robust for this forecasting task.

In the regulator-only case, LSTM again achieves the lowest MAE of 0.004770 ± 0.000009 p.u., while CNN1D produces a similar MAE of 0.004841 ± 0.000096 p.u. The persistence baseline degrades to 0.007216 p.u. MAE, showing that the learned temporal models extract more useful information

TABLE 5. Sensor observability comparison.

Sensor case	Number of sensors	Persistence MAE	CNN1D MAE	LSTM MAE	Interpretation
reg_1	1	0.004965 ± 0.000019	0.004786 ± 0.000023	0.004786 ± 0.000018	Extreme low-observability case; learned models remain accurate with one regulator-related measurement
reg_3	3	0.007216	0.004841 ± 0.000096	0.004770 ± 0.000009	Regulator-only sparse case; low error with three operational measurements
pv_3	3	0.006069 ± 0.000000	0.004780 ± 0.000030	0.004767 ± 0.000006	DER-relevant observation-only case; comparable learned-model accuracy
reg_pv_6	6	0.005201	0.004852 ± 0.000171	0.004807 ± 0.000013	Expanded sparse case; improves persistence but gives marginal learned-model gain
dense_oracle	687	0.005174 ± 0.000094	0.004792 ± 0.000050	0.004765 ± 0.000003	Dense non-target observation reference; no large gain over sparse learned models

TABLE 6. Horizon-wise MAE by 6-hour forecast block.

Case	Model	0–6 h	6–12 h	12–18 h	18–24 h
reg_pv_6	Persistence	0.004203	0.005554	0.005936	0.005110
reg_pv_6	CNN1D	0.004976	0.005090	0.005096	0.005035
reg_pv_6	LSTM	0.004844	0.004785	0.004783	0.004830
reg_3	Persistence	0.006825	0.007298	0.007483	0.007258
reg_3	CNN1D	0.004919	0.004921	0.004940	0.004980
reg_3	LSTM	0.004774	0.004768	0.004773	0.004795

from the limited three-sensor history than the last-observation reference. Overall, these results support the feasibility of forecasting future critical-bus voltages from a small number of voltage sensors, while also showing that sequence-aware or convolutional temporal structure is beneficial.

Figure 4 visualizes the multi-seed test MAE comparison for the two primary sparse observation cases. The bars show the mean over three random seeds for the trainable neural baselines, and the error bars show one standard deviation; persistence is shown as a deterministic reference. This presentation is used instead of relying on a single seed because the MLP baseline is substantially more sensitive to initialization than CNN1D and LSTM. The figure is consistent with the multi-seed numerical results in Table 4: CNN1D and LSTM remain stable and accurate, whereas MLP exhibits large variance. This motivates the sensor-observability analysis in the next subsection.

B. SENSOR OBSERVABILITY EFFECT

The sensor-observability effect is evaluated by comparing four sparse observation cases, reg_1, reg_3, pv_3, and reg_pv_6, with the dense non-target observation reference, dense_oracle. These cases use the same target buses, forecast horizon, chronological data split, and baseline model definitions, but differ in the number and type of observed voltage sensors. Therefore, the comparison isolates how sensor availability affects critical-bus voltage forecasting accuracy under otherwise identical experimental conditions.

Table 5 summarizes the MAE values for persistence, CNN1D, and LSTM. The single-regulator case, reg_1, uses only one voltage observation and still produces low MAE for the learned models: 0.004786 ± 0.000023 p.u. for CNN1D and 0.004786 ± 0.000018 p.u. for LSTM. The DER-relevant observation-only case, pv_3, also performs well, with LSTM achieving 0.004767 ± 0.000006 p.u. MAE. These results indicate that, for the selected target buses and dataset, highly sparse observations can contain sufficient temporal information for accurate critical-bus forecasting.

The comparison between reg_3 and reg_pv_6 shows that expanding the sensor set from three to six locations improves the persistence reference but provides only marginal gain for the learned temporal models. Persistence decreases from 0.007216 p.u. in reg_3 to 0.005201 p.u. in reg_pv_6, whereas CNN1D and LSTM remain near 0.0048 p.u. in both cases. This supports the main feasibility claim: accurate 24-hour critical-bus voltage forecasting is possible under low observability, and the strongest learned models do not require dense voltage measurements in the evaluated setting.

The dense non-target observation case, dense_oracle, uses 687 input voltage nodes while excluding the three target buses

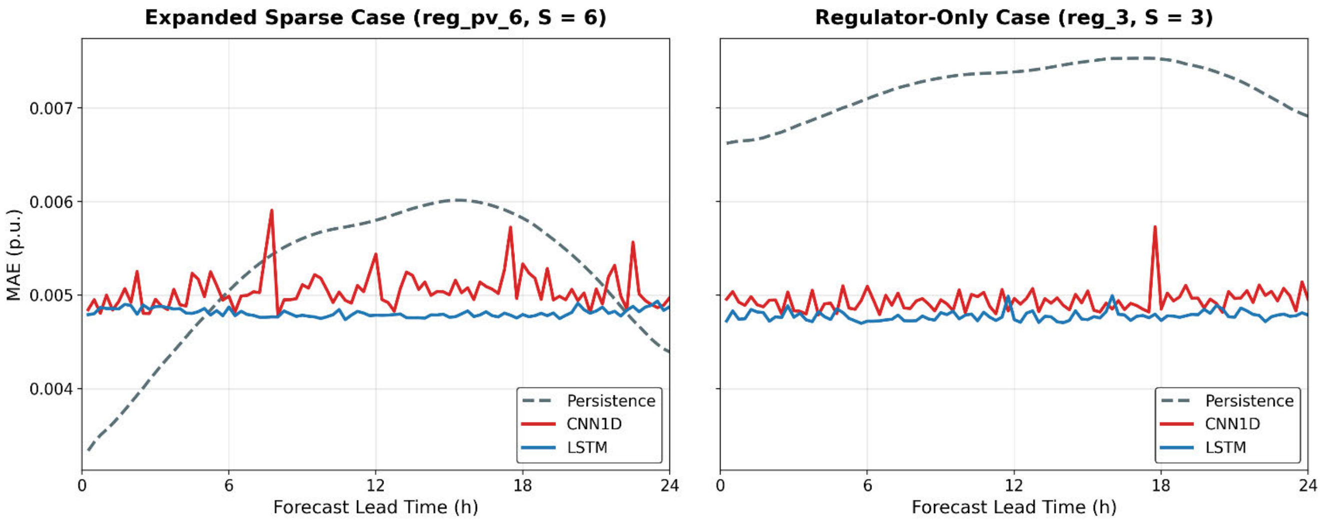

FIGURE 5. Horizon-wise MAE over the 24-hour multi-step forecast horizon for representative seed-42 runs. Errors are averaged over all test samples and the three critical target buses at each forecast step.

from the input. Its results are consistent with the sparse-case findings. Dense-oracle LSTM achieves 0.004765 ± 0.000003 p.u. MAE, which is nearly identical to the best sparse-case LSTM results. Dense-oracle CNN1D achieves 0.004792 ± 0.000050 p.u. MAE, also comparable to the sparse CNN1D cases. Thus, the dense reference does not provide a large accuracy gain over carefully selected sparse observations. This result strengthens the practical interpretation of the benchmark: the dominant predictive information for the selected critical buses appears to be captured by a very small number of voltage sensors.

The dense result should not be interpreted as a deployable monitoring architecture. It is an experimental reference that uses nearly all non-target phase-A voltage nodes and therefore represents a substantially higher observability condition than realistic sensor deployments. Its role is to contextualize the sparse-sensor results, not to define a practical sensing strategy.

C. HORIZON-WISE ERROR ANALYSIS

Horizon-wise error analysis examines whether forecasting accuracy degrades as the prediction lead time increases over the 24-hour horizon. Figure 5 shows the seed-42 horizon-wise MAE curves for the two primary sparse observation cases, reg_pv_6 and reg_3, using persistence, CNN1D, and LSTM. The curves are computed by averaging the absolute error over all test samples and all three target buses at each forecast step.

The learned temporal models show stable error profiles across the full horizon. In the reg_pv_6 case, LSTM remains between 0.004783 and 0.004844 p.u. across the four 6-hour blocks, while CNN1D remains close to 0.0050 p.u. In the reg_3 case, LSTM similarly stays near 0.00477–0.00480 p.u., and CNN1D stays near 0.00492–0.00498 p.u. This indicates that the direct multi-output forecasting formulation does not exhibit strong error accumulation over the 24-hour prediction window for the best-performing learned models.

The persistence baseline behaves differently. In the reg_3 case, persistence remains consistently higher than CNN1D and LSTM across all horizon blocks, with MAE increasing from 0.006825 p.u. in the first 6 hours to 0.007483 p.u. in the 12–18 h block. In the reg_pv_6 case, persistence is competitive in the earliest horizon block but becomes worse than LSTM in later blocks. This pattern is consistent with the overall results: last-observation information is useful, but learned temporal models provide more stable multi-step forecasts. Table 6 summarizes these horizon-wise MAE values by 6-hour forecast blocks.

D. TARGET-WISE CRITICAL-BUS ANALYSIS

Target-wise analysis evaluates whether the forecasting difficulty is uniform across the three unmetered critical buses. Table 7 reports the seed-42 MAE values separately for buses 49, 64, and 66, with errors averaged over all test samples and all 96 forecast steps. This decomposition is useful because the same observation set may provide different predictive information for different critical locations, depending on feeder position, proximity to observed nodes, and local voltage variability.

The learned models show the lowest error at bus 49 and higher errors at buses 64 and 66. For example, in the reg_pv_6 case, LSTM obtains MAE values of 0.003842 p.u. at bus 49, 0.005278 p.u. at bus 64, and 0.005310 p.u. at bus 66. A similar pattern appears in the reg_3 case, where LSTM achieves 0.003952 p.u. at bus 49 and approximately 0.0052 p.u. at buses 64 and 66. This suggests that buses 64 and 66 are more difficult target locations under the present sparse voltage-only observation settings.

The comparison with persistence also clarifies the role of learned temporal mapping. In the reg_3 case, persistence

TABLE 7. Target-wise MAE by critical bus.

Case	Model	Bus 49	Bus 64	Bus 66
reg_pv_6	Persistence	0.004419	0.005512	0.005669
reg_pv_6	CNN1D	0.004557	0.005276	0.005313
reg_pv_6	LSTM	0.003842	0.005278	0.005310
reg_3	Persistence	0.009136	0.006085	0.006423
reg_3	CNN1D	0.003958	0.005407	0.005455
reg_3	LSTM	0.003952	0.005180	0.005199

performs poorly at bus 49, with an MAE of 0.009136 p.u., whereas CNN1D and LSTM reduce the bus-49 error to approximately 0.0040 p.u. This indicates that simply repeating or projecting the most recent sensor measurements is insufficient for some target locations, even when the overall voltage time series is highly autocorrelated. In contrast, the learned temporal models produce more balanced target-wise errors, which supports their use as virtual metering baselines for unobserved critical buses.

E. TRAINING STABILITY AND RUNTIM

Training stability is evaluated using the three-seed results reported in Table 4. The MLP baseline shows the largest seed-to-seed variation, with test MAE standard deviations of 0.018681 p.u. in the reg_pv_6 case and 0.007347 p.u. in the reg_3 case. This instability is consistent with the role of MLP as a weak non-sequential baseline: after flattening the long lookback window, the model has limited temporal inductive bias and is more sensitive to initialization and optimization behavior.

In contrast, CNN1D and LSTM exhibit stable performance across seeds. For reg_pv_6, CNN1D obtains 0.004852 ± 0.000171 p.u. and LSTM obtains 0.004807 ± 0.000013 p.u. For reg_3, CNN1D obtains 0.004841 ± 0.000096 p.u. and LSTM obtains 0.004770 ± 0.000009 p.u. These small standard deviations indicate that the main findings are not driven by a single favorable random seed. They also support the interpretation that temporal structure, rather than model capacity alone, is important for the sensor-limited voltage forecasting task.

Figure 6 summarizes the runtime–accuracy tradeoff for the representative seed-42 runs. LSTM provides the lowest MAE among the trainable baselines, but it requires substantially longer training time: approximately 77.5 s for reg_pv_6 and 74.4 s for reg_3 on the recorded runs. CNN1D is less expensive, requiring approximately 24.1 s and 20.4 s for the same two cases, while producing MAE values close to LSTM. Therefore, CNN1D can be viewed as a practical lower-cost baseline, whereas LSTM provides the strongest recurrent reference. The final model choice should depend on whether the application prioritizes marginal accuracy improvement or computational efficiency.

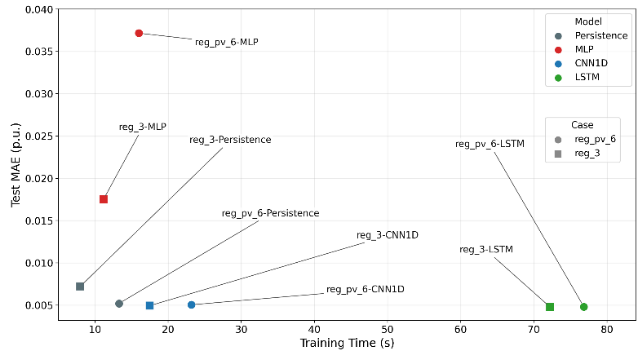

FIGURE 6. Runtime and accuracy tradeoff across baseline models for representative seed-42 runs.

F. SUMMARY OF EMPIRICAL FINDINGS

The experimental results lead to five main findings. First, accurate multi-step critical-bus voltage forecasting is feasible under sparse voltage sensing. Across the primary sparse cases, CNN1D and LSTM achieve test MAE values near 0.0048 p.u., even though the target buses are excluded from the input measurements. This supports the central virtual metering premise of the paper: future voltages at unmetered critical buses can be inferred from a limited set of observed feeder voltage histories.

Second, temporal inductive bias is important. Persistence remains a strong reference because distribution voltage time series exhibit temporal autocorrelation, but it is not uniformly reliable across sensor sets or target buses. MLP is less stable and less accurate because flattening a long historical window does not provide a suitable temporal structure for this task. In contrast, CNN1D and LSTM provide stable multi-seed performance and consistent horizon-wise behavior, indicating that explicitly modeling temporal patterns improves robustness.

Third, increasing observability does not automatically produce a large gain for the tested learned models. The dense non-target observation reference provides a useful comparison point, but its learned-model MAE is close to that of the sparse cases. This does not imply that sensor placement is unimportant; rather, it shows that the selected sparse voltage measurements already capture much of the predictive information needed for the evaluated target buses and forecast horizon. Fourth, target-wise analysis shows that forecasting difficulty is not uniform across critical buses, with buses 64 and 66 being more difficult than bus 49 in the present setup. Finally, the runtime comparison suggests a practical tradeoff: LSTM provides the strongest recurrent reference, while CNN1D offers similar accuracy with lower computational cost.

VII. DISCUSSION

A. IMPLICATIONS FOR VIRTUAL METERING

The results support the feasibility of virtual metering for critical-bus voltage forecasting under low-observability conditions. In the proposed task, the target buses are explicitly excluded from the input measurements, and their future voltage trajectories are inferred from voltage histories measured elsewhere in the feeder. The low MAE values obtained by CNN1D and LSTM indicate that sparse voltage measurements can contain sufficient temporal information to reconstruct the future behavior of selected unmetered critical locations over a 24-hour horizon.

This implication is practically relevant for distribution-system monitoring because dense voltage sensor deployment is often expensive and operationally difficult. The benchmark does not assume that every customer-side or critical bus is directly metered during inference. Instead, it evaluates whether a limited number of physically measured feeder signals can support forecasting at unmetered locations. Under this framing, virtual metering can be viewed as a complementary monitoring layer: it does not replace physical measurements where they are required for protection or billing, but it can help extend situational awareness to locations where instrumentation is unavailable.

The dense non-target observation reference further clarifies the role of sparse sensing. Since dense non-target inputs do not produce a large accuracy gain over the tested sparse learned models, the results suggest that carefully selected sparse measurements may be adequate for the evaluated target buses and dataset. This should not be interpreted as a general sensor-placement optimum. Rather, it motivates a practical research direction: benchmarked virtual metering can be used to quantify how much predictive value is obtained from a given sensor set before additional field instrumentation is deployed.

B. ROLE OF TEMPORAL INDUCTIVE BIAS

The model comparison indicates that temporal inductive bias is a key factor in this benchmark. The input window contains 144 hours (6 days) of voltage history, and the output is a direct 24-hour (1-day) multi-step forecast. A model must therefore extract useful temporal structure from a long sequence rather than simply memorize a static mapping between input columns and target columns. CNN1D and LSTM are better suited to this setting because they impose sequence-aware assumptions: CNN1D emphasizes local temporal motifs through shared convolutional filters, while LSTM summarizes the historical trajectory through recurrent state updates.

The weaker MLP results are consistent with this interpretation. After flattening the lookback matrix, the MLP treats the historical samples as a high-dimensional vector and does not explicitly encode time ordering, local temporal continuity, or recurrent dependence. This makes the model more sensitive to initialization and optimization, especially when the input dimension grows with the number of sensors and the long lookback length. The observed multi-seed variability therefore should not be interpreted merely as poor tuning; it reflects a mismatch between the flattened feed-forward structure and the sequential nature of the forecasting task.

Persistence remains competitive in several cases because voltage magnitude time series are temporally autocorrelated and often change smoothly over short intervals. This makes last-observation-based prediction a meaningful lower baseline rather than a trivial reference. However, persistence cannot learn target-specific temporal transformations from observed sensors to unobserved critical buses. Its target-wise and horizon-wise behavior shows that simple repetition or projection is insufficient when the relation between measured nodes and target buses changes over the forecast horizon. The learned temporal baselines improve on this limitation by using the historical sequence to infer more stable multi-step trajectories.

C. SENSOR PLACEMENT AND OBSERVABILITY

The observation-case results show that sensor composition matters, but they should not be interpreted as an optimized sensor placement study. The evaluated sensor sets were selected to represent practical and interpretable monitoring conditions: regulator-related measurements, DER-relevant feeder measurements, their combined sparse case, and a dense non-target reference. This design allows the benchmark to compare levels and types of observability while keeping the target buses, forecast horizon, preprocessing, and model configurations fixed.

The regulator-only and DER-relevant cases provide different forms of information. Regulator-related measurements are operationally meaningful because voltage regulators directly affect feeder voltage profiles and are plausible locations for utility-side monitoring. DER-relevant observations can capture local variability associated with PV or other distributed resources. The combined reg_pv_6 case tests whether these two information sources are complementary. The results indicate that adding more sparse sensors can improve some baselines, especially persistence, but the learned temporal models already achieve low error with smaller sensor sets. Therefore, the benefit of additional measurements is model-dependent and target-dependent rather than strictly proportional to sensor count.

The dense non-target observation case provides context for this interpretation. Because its learned-model performance is close to that of sparse cases, the experiment suggests that the selected sparse measurements contain substantial predictive information for the three evaluated target buses. This small sparse–dense gap should not be interpreted simply as evidence that the dataset is trivial. Distribution voltages are physically coupled through feeder topology, regulator actions, shared loading patterns, and DER-driven temporal behavior. Measurements at operationally informative locations, especially voltage-regulator-related nodes, can therefore contain strong predictive information about downstream critical-bus voltage trajectories. In this sense, the result is consistent with the virtual metering objective: a small number of informative physical sensors may be sufficient to infer selected unmetered voltages in the evaluated feeder.

However, this does not establish that the same sensor sets are optimal for other target buses, phases, feeder topologies, or DER scenarios. A full sensor placement study would require a separate optimization objective, candidate-location constraints, cost models, and validation across broader operating conditions. The present work instead provides a reproducible benchmark for evaluating how a given observation set affects virtual metering performance.

D. LIMITATIONS OF VIOLATION-CENTRIC EVALUATION

Although voltage monitoring is often motivated by under-voltage and over-voltage concerns, the present dataset does not provide enough actual violations of the 0.95/1.05 p.u. thresholds to support strong violation-detection claims. The experiments therefore focus on continuous forecasting accuracy, using MAE, RMSE, P95 absolute error, horizon-wise error, and target-wise error as the primary evidence. These metrics are appropriate for evaluating whether the future voltage trajectory can be predicted accurately, but they do not by themselves prove operational violation-detection performance.

This distinction is important because violation-centric evaluation requires sufficient positive samples near or beyond operational limits. If the test set contains few or no threshold-crossing events, classification-style metrics such as precision, recall, and F1 score become unstable or uninformative. A model can achieve low voltage-trajectory error while still not being validated for rare-event detection, especially when the rare events are absent from the evaluation data. For this reason, the reported violation metrics should be interpreted only as auxiliary diagnostics rather than as the main empirical claim of the paper.

The scope of the present conclusions is therefore limited to forecasting accuracy in the available modified IEEE 123-bus dataset, evaluated phase-A voltage signals, selected target buses, tested observation cases, and baseline models. Claims about robust voltage-violation detection, mitigation, alarm generation, or control support would require additional scenarios with sufficient under-voltage and over-voltage events, preferably generated under systematically varied DER penetration, loading, topology, and secondary-circuit conditions. Such violation-focused evaluation is an important direction for future work, but it is outside the evidence provided by the current benchmark.

VIII. THREATS TO VALIDITY

A. DATA AND EXPERIMENTAL VALIDITY

Several steps were taken to reduce experimental validity threats. The preprocessing pipeline resolves requested sensor and target identifiers to explicit voltage-matrix columns, generates fixed lookback-horizon tensors, and applies chronological train/validation/test splits. This reduces ambiguity in column selection and prevents random-window shuffling from leaking future information into training or model selection. The same processed tensors and split definitions are used across all baseline models, so model comparisons are based on identical input-output samples.

Training variance is another potential threat because neural baselines can be sensitive to initialization and optimization. To address this, the main trainable baselines are evaluated over multiple random seeds, and the reported results include mean and standard deviation values. The multi-seed results show that CNN1D and LSTM are stable, whereas MLP has larger variance. This supports the main empirical conclusions while also identifying where a baseline is less reliable. Nevertheless, the number of seeds remains limited, and small differences between CNN1D and LSTM should not be overinterpreted as statistically definitive.

A data-generation limitation remains. The available dataset consists of pre-generated voltage, active-power, and reactive-power simulation outputs for the modified IEEE 123-bus system. The current study does not include the full OpenDSS model files, PV/EV-to-bus mapping, or complete scenario-regeneration capability. Therefore, the experiments validate the proposed forecasting benchmark on the available fixed dataset rather than on independently regenerated operating scenarios. To support code-level reproducibility, the intended release focuses on the benchmark implementation, including preprocessing, model definitions, training, inference, metric computation, plotting scripts, fixed split metadata, and either processed tensors or a compact sample dataset when full raw-data release is constrained by size or sharing restrictions.

B. GENERALIZATION VALIDITY

The conclusions are limited to the evaluated modified IEEE 123-bus distribution system and should not be treated as universal evidence for all distribution networks. The benchmark focuses on phase-A voltage magnitude signals, three selected critical target buses, and the observation cases defined in this paper. Other phases, target locations, feeder topologies, regulator configurations, secondary-circuit layouts, and DER deployment patterns may exhibit different voltage correlations and different levels of predictability.

The present results also do not establish robustness across systematically varied PV penetration, EV charging penetration, loading conditions, or topology changes. Although the dataset includes DER-rich operating behavior and secondary-circuit effects, the experiments are conducted on the available fixed simulation outputs rather than on a broad scenario ensemble generated from controllable OpenDSS cases. Therefore, claims about topology-aware generalization, transfer to other feeders, robustness to unseen DER scenarios, or performance under sensor dropout remain outside the current evidence. This study should be viewed as a first-stage benchmark for testing the fundamental predictability of critical-bus voltage trajectories under sparse sensing, rather than as final evidence of deployment-ready generalization across large-scale or real utility feeders.

These limitations do not invalidate the benchmark contribution, but they define its scope. The study demonstrates that sensor-limited critical-bus voltage forecasting is feasible in the evaluated setting and provides a reproducible framework for comparing observation cases and temporal baselines. Extending the framework to additional feeders, phases, critical-bus definitions, PV/EV penetration levels, and OpenDSS-regenerated scenarios is necessary before making broader deployment-level claims.

C. METRIC AND CONCLUSION VALIDITY

The primary metrics, including MAE, RMSE, and P95 absolute error, quantify continuous voltage-trajectory accuracy but do not fully capture all operational requirements of distribution-system monitoring. For example, an operator may care about threshold crossings, voltage-margin ranking, alarm timing, uncertainty, or the cost of false alarms. The reported metrics are therefore appropriate for evaluating forecasting accuracy, but they should not be interpreted as complete evidence of operational readiness for protection, control, or alarm applications.

Voltage-violation metrics are particularly limited in the current dataset because there are not enough actual under-voltage or over-voltage samples relative to the 0.95/1.05 p.u. thresholds. As discussed in Section VII-D, classification-style violation metrics become unreliable when positive samples are rare or absent. The paper therefore avoids making strong claims about violation detection and instead treats violation-related quantities as auxiliary diagnostics. Robust violation-centric evaluation would require additional scenarios deliberately containing sufficient voltage-limit events.

Finally, the conclusions should not overstate small differences between the strongest learned baselines. The multi-seed results show that CNN1D and LSTM are both stable and effective, with LSTM often giving the lowest MAE and CNN1D offering a lower-cost alternative. However, their absolute differences are small, and the number of random seeds is limited. The defensible conclusion is that temporal baselines with appropriate sequence structure are effective for the proposed benchmark, not that one architecture is universally superior across all distribution-network forecasting settings.

IX. CONCLUSION

This paper formulated sensor-limited critical-bus voltage forecasting as a virtual metering task for a DER-rich modified IEEE 123-bus distribution system with secondary-circuit effects. The task explicitly separates observed sensor locations from unmetered critical target buses, preventing target-bus voltage leakage during inference. A reproducible benchmark pipeline was developed to process phase-A voltage signals, construct fixed lookback-horizon tensors, apply chronological data splits, train representative temporal baselines, and evaluate target-wise and horizon-wise forecasting performance. Among the evaluated baselines, LSTM achieved the best overall performance, with a test MAE of approximately 0.0048 p.u. while using only 1–6 voltage sensor locations. CNN1D produced comparable forecasting accuracy, indicating that reliable multi-step voltage prediction is feasible even under highly constrained sensing conditions.

The experimental results show that accurate 24-hour multi-step forecasting is feasible under sparse voltage sensing in the evaluated setting. CNN1D and LSTM achieve stable low-error performance across the primary sparse observation cases, while persistence provides a meaningful but limited temporal reference and MLP shows weaker stability due to its lack of temporal inductive bias. The dense non-target observation reference does not provide a large learned-model accuracy gain over the sparse cases, suggesting that the selected sparse measurements contain substantial predictive information for the evaluated critical buses. At the same time, target-wise and horizon-wise analyses show that forecasting difficulty varies across buses and forecast lead times, reinforcing the need for decomposed evaluation rather than relying only on aggregate error metrics.

The study should be interpreted as a benchmark and feasibility analysis rather than a universal sensor-placement or model-superiority claim. The conclusions are limited to the available modified IEEE 123-bus dataset, evaluated phase-A signals, selected target buses, tested observation cases, and baseline architectures. Future work will extend the framework to topology-aware models, sensor-dropout robustness, multi-phase forecasting, broader PV/EV penetration scenarios, additional feeders and target locations, and OpenDSS-based scenario regeneration for violation-rich evaluation.

APPENDIX

DATA AVAILABILITY

The core benchmark code and a compact sample dataset are made publicly available on GitHub at: https://github.com/kafa46/sensor-limited-voltage-forecasting.

The repository includes the preprocessing scripts, baseline model definitions, training and inference scripts, metric computation code, plotting scripts, fixed split metadata, configuration files, and sample tensors needed to reproduce the reported workflow at a compact scale.

Because the full simulation dataset is large, the public release includes only a limited-time sample sufficient for readers to run the pipeline, verify the data format, test the model scripts, and reproduce the workflow at a smaller scale. The full experimental dataset is not planned for direct public upload to GitHub because of storage size and data-sharing constraints.

The full dataset used in this study may be made available from the corresponding author upon reasonable request, subject to review and approval by the authors' institution. Access to the full dataset will therefore depend on institutional data-sharing approval and any applicable usage restrictions.

REFERENCES

[1] N. Damianakis, G. R. C. Mouli, and P. Bauer, "Grid impact of photovoltaics, electric vehicles and heat pumps on distribution grids—An overview," *Appl. Energy*, vol. 380, Feb. 2025, Art. no. 125000, doi: 10.1016/j.apenergy.2024.125000.

[2] A. P. Yadav, J. Nutaro, B. Park, J. Dong, B. Liu, S. B. Yoginath, H. Yin, J. Dong, Y. Dong, Y. Liu, T. Kuruganti, and Y. Xue, "Review of emerging concepts in distribution system state estimation: Opportunities and challenges," *IEEE Access*, vol. 11, pp. 70503–70515, 2023, doi: 10.1109/ACCESS.2023.3292793.

[3] M. M. Othman, M. H. Ahmed, and M. M. A. Salama, "A novel smart meter technique for voltage and current estimation in active distribution networks," *Int. J. Electr. Power Energy Syst.*, vol. 104, pp. 301–310, Jan. 2019, doi: 10.1016/j.ijepes.2018.07.015.

[4] P. Buason, S. Misra, S. Talkington, and D. K. Molzahn, "A data-driven sensor placement approach for detecting voltage violations in distribution systems," *Electr. Power Syst. Res.*, vol. 232, Jul. 2024, Art. no. 110387, doi: 10.1016/j.epsr.2024.110387.

[5] W. H. Kersting, "Radial distribution test feeders," in *Proc. IEEE Power Eng. Soc. Winter Meeting. Conf.*, vol. 2, Columbus, OH, USA, Oct. 2001, pp. 908–912, doi: 10.1109/PESW.2001.916993.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[7] M. E. Baran and A. W. Kelley, "State estimation for real-time monitoring of distribution systems," *IEEE Trans. Power Syst.*, vol. 9, no. 3, pp. 1601–1609, Aug. 1994, doi: 10.1109/59.336098.

[8] A. Primadianto and C.-N. Lu, "A review on distribution system state estimation," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3875–3883, Sep. 2017, doi: 10.1109/TPWRS.2016.2632156.

[9] A. Bragantini, A. Sumper, V. Gadelha, and S. Galceran-Arellano, "Neural network-based three-phase state estimation for unobservable low voltage grids," *Int. J. Electr. Power Energy Syst.*, vol. 167, Jun. 2025, Art. no. 110568, doi: 10.1016/j.ijepes.2025.110568.

[10] A. Kotsonias, L. Hadjidemetriou, M. Asprou, and C. G. Panayiotou, "Operational challenges and solution approaches for low voltage distribution grids—A review," *Electr. Power Syst. Res.*, vol. 239, Feb. 2025, Art. no. 111258, doi: 10.1016/j.epsr.2024.111258.

[11] M. K. Gray and W. G. Morsi, "Probabilistic quantification of voltage unbalance and neutral current in secondary distribution systems due to plug-in battery electric vehicles charging," *Electr. Power Syst. Res.*, vol. 133, pp. 249–256, Apr. 2016, doi: 10.1016/j.epsr.2015.12.022.

[12] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep learning for time series forecasting: A survey," *Big Data*, vol. 9, no. 1, pp. 3–21, Feb. 2021, doi: 10.1089/big.2020.0159.

[13] A. Gasparin, S. Lukovic, and C. Alippi, "Deep learning for time series forecasting: The electric load case," *CAAI Trans. Intell. Technol.*, vol. 7, no. 1, pp. 1–25, Mar. 2022, doi: 10.1049/cit2.12060.

[14] A. Al-Wakeel, J. Wu, and N. Jenkins, "State estimation of medium voltage distribution networks using smart meter measurements," *Appl. Energy*, vol. 184, pp. 207–218, Dec. 2016, doi: 10.1016/j.apenergy.2016.10.010.

[15] M. A. Khan and B. Hayes, "Smart meter based two-layer distribution system state estimation in unbalanced MV/LV networks," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 688–699, Jan. 2022, doi: 10.1109/TII.2021.3079267.

[16] M. Markovic, A. Sajadi, A. Florita, R. Cruickshank III, and B.-M. Hodge, "Voltage estimation in low-voltage distribution grids with distributed energy resources," *IEEE Trans. Sustain. Energy*, vol. 12, no. 3, pp. 1640–1650, Jul. 2021, doi: 10.1109/TSTE.2021.3060546.

[17] M. Pertl, P. J. Douglass, K. Heussen, and K. Kok, "Validation of a robust neural real-time voltage estimator for active distribution grids on field data," *Electric Power Syst. Res.*, vol. 154, pp. 182–192, Jan. 2018, doi: 10.1016/j.epsr.2017.08.016.

[18] M. S. Kemal, R. S. Martin-Loeches, R. L. Olsen, F. Iov, and H.-P. C. Schwefel, "On the trade-off between timeliness and accuracy for low voltage distribution system grid monitoring utilizing smart meter data," *Int. J. Electr. Power Energy Syst.*, vol. 121, Oct. 2020, Art. no. 106090, doi: 10.1016/j.ijepes.2020.106090.

[19] L. Fang, A. B. Pengwah, L. L. H. Andrew, R. Razzaghi, and M. A. Muñoz, "Three-phase voltage sensitivity estimation and its application to topology identification in low-voltage distribution networks," *Int. J. Electr. Power Energy Syst.*, vol. 158, Jul. 2024, Art. no. 109949, doi: 10.1016/j.ijepes.2024.109949.

[20] S. Talkington, S. Grijalva, M. J. Reno, J. A. Azzolini, and J. Peppanen, "Localized structure in secondary distribution system voltage sensitivity matrices," *Electr. Power Syst. Res.*, vol. 226, Jan. 2024, Art. no. 109788, doi: 10.1016/j.epsr.2023.109788.

GISEOP NOH received the B.S. degree in industrial engineering from Korea Air Force Academy, South Korea, the M.S. degree in computer science from the University of Colorado Denver, USA, in 2009, and the Ph.D. degree in computer science from Seoul National University, South Korea, in 2014.

From 1994 to 2005, he was an Officer with Republic of Korea Air Force, where he was responsible for the operation and maintenance of advanced weapon computer systems. He later worked at the Defense Acquisition Program Administration on aviation electronics and computer projects. From 2017 to 2018, he was a Faculty Member with the Department of Computer Science, Korea Air Force Academy. From 2018 to 2024, he was a Professor with the Department of Artificial Intelligence Software, Cheongju University, South Korea. Since 2025, he has been a Professor with the Department of Software Convergence, Hongik University, South Korea, where he established and directs the DeepShark Laboratory. His research interests include machine learning, computer vision, retrieval-augmented generation (RAG) systems, graph neural networks, and AI applications in semiconductor inspection and smart manufacturing. He is a Member of Korean Institute of Information Scientists and Engineers (KIISE), and Korea Information Processing Society (KIPS).

HAN PYO LEE (Senior Member, IEEE) received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, Republic of Korea, in 2008, the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2020, and the Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, NC, USA, in 2024, where he was a Member of the Future Renewable Electric Energy Delivery and Management (FREEDM) Systems Center. He is currently an Assistant Professor with the Department of Electronic and Electrical Convergence Engineering, Hongik University, Republic of Korea. His research interests include power system operation, renewable energy integration, and data-driven energy analytics.