IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155307

An AI-Driven Framework for Real-Time Energy Consumption Monitoring, Demand Forecasting, and Optimization Using Deep Sequence Models

M. Lavanya Durga[1], K. Lakshamana Reddy[*2]

PG Scholar Department of Computer Science, S.V.K.P & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, Affiliated to Adikavi Nannaya University[1]

Associate Professor, Department of Master of Computer Applications, S.V.K.P & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda,Affiliated to Adikavi Nannaya University[*2]

*Corresponding Author

Abstract: Rising electricity costs, grid stress, and the global imperative to reduce carbon emissions have intensified the need for intelligent systems that can monitor and curtail energy waste at the building and household level. Conventional metering reports cumulative consumption after the fact, offering little insight into when, where, or why energy is wasted, and provides no forward-looking guidance for reducing demand. This paper presents an artificial-intelligence framework that ingests fine-grained consumption telemetry, forecasts short-term demand, detects anomalous usage, and recommends actionable optimization measures in real time. The system streams data from smart meters and appliance-level sensors through an edge gateway into a time-series store, applies a deep sequence model for load forecasting, flags deviations through an anomaly detector, and surfaces savings recommendations on an interactive dashboard. A Python back end performs model training and inference, while a Node.js layer delivers live visualization over web sockets. Evaluated against linear-regression and ARIMA baselines, the proposed deep model reduced forecasting error to a mean absolute percentage error of roughly 6.4%, and the optimization layer identified consumption reductions of up to 18% in evaluation scenarios. The principal contributions are an end-to-end streaming monitoring pipeline, an accurate deep forecasting and anomaly-detection layer, and an interpretable recommendation engine that links prediction to concrete energy-saving actions.

Keywords: -Energy monitoring; demand forecasting; deep learning; LSTM; anomaly detection; smart metering; energy optimization; time-series analysis.

1. INTRODUCTION

Electricity demand continues to grow worldwide, and buildings account for a substantial share of total consumption, much of which is wasted through inefficient scheduling, idle loads, and undetected faults [1], [2]. As tariffs rise and decarbonization targets tighten, both consumers and facility operators require tools that convert raw consumption data into timely, actionable insight. Traditional utility meters, however, report only aggregate periodic totals, masking the temporal and device-level structure of demand that is essential for targeted intervention [3].

The proliferation of smart meters and low-cost sensors has made granular telemetry available, but data alone does not yield savings; it must be analyzed, forecast, and translated into decisions [4]. Classical statistical forecasters such as autoregressive models capture linear seasonality yet struggle with the nonlinear, multi-scale patterns characteristic of real consumption [5]. Moreover, many existing monitoring tools are descriptive rather than prescriptive, displaying historical curves without anticipating demand or recommending concrete optimizations [6].

A. Problem Statement

There is a need for an integrated system that continuously monitors fine-grained energy consumption, forecasts near-term demand accurately under nonlinear dynamics, detects anomalous or wasteful usage, and recommends actionable measures—capabilities that descriptive metering and purely statistical tools do not jointly deliver.

B. Motivation and Objectives

These gaps motivate an AI-driven, streaming framework that couples deep forecasting with optimization. The objectives are: to construct a real-time ingestion and storage pipeline for consumption telemetry; to develop an accurate

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155307

deep sequence model for short-term load forecasting; to detect anomalies indicative of waste or faults; and to generate interpretable, cost-aware recommendations evaluated against established baselines.

C. Contributions

- An end-to-end streaming pipeline that ingests smart-meter and appliance-level telemetry through an edge gateway into a time-series store for real-time analysis.
- A deep sequence forecasting model that captures nonlinear, multi-scale demand patterns and substantially outperforms linear and ARIMA baselines.
- An anomaly-detection and optimization layer that flags wasteful usage and produces concrete, cost-aware savings recommendations.
- A comparative evaluation quantifying forecasting error and achievable consumption reductions relative to baseline approaches.

2. LITERATURE REVIEW

Energy-management research spans metering infrastructure, load forecasting, non-intrusive load monitoring, and demand-response optimization. Early forecasting relied on autoregressive integrated moving-average and exponential-smoothing models, which perform adequately for stable, strongly seasonal loads but degrade under abrupt behavioural or weather-driven shifts [5], [7]. Machine-learning regressors such as support-vector and tree-ensemble models improved nonlinear fitting, yet they require careful feature engineering and do not natively model long temporal dependencies [8].

Recurrent neural networks, and long short-term memory architectures in particular, became prominent for load forecasting owing to their ability to learn temporal dependencies directly from sequences [9], [10]. Subsequent studies introduced convolutional and attention-based variants to capture multi-scale patterns and exogenous factors, reporting further accuracy gains [11], [12]. In parallel, non-intrusive load monitoring research developed disaggregation methods to attribute consumption to individual appliances without per-device meters [13].

Anomaly detection for energy data has employed statistical thresholds, clustering, and autoencoder reconstruction error to flag faults and waste [14]. Optimization and demand-response work has formulated scheduling as cost-minimization under comfort and operational constraints [15], [16]. Despite this breadth, comparatively few systems integrate real-time ingestion, deep forecasting, anomaly detection, and actionable recommendation within a single deployable platform; many address one component in isolation. This integration gap motivates the present work. Table I compares representative approaches.

TABLE I. COMPARATIVE ANALYSIS OF REPRESENTATIVE ENERGY-MANAGEMENT APPROACHES

Approach	Core Technique	Strengths	Limitations
ARIMA / smoothing [5],[7]	Statistical forecasting	Simple, interpretable	Weak on nonlinear shifts
ML regressors [8]	SVR / tree ensembles	Nonlinear fitting	Manual features; short memory
LSTM forecasters [9],[10]	Recurrent sequence models	Learns temporal patterns	Forecast only; no actions
NILM disaggregation [13]	Load attribution	Device-level insight	Needs training signatures
DR optimization [15],[16]	Constrained scheduling	Direct cost savings	Assumes accurate forecasts
Proposed framework	Streaming + LSTM + optimize	Integrated, prescriptive	Bounded by data quality

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155307

3. PROPOSED METHODOLOGY

The framework is organized into four cooperating layers—sensing, ingestion and storage, AI analytics, and presentation—as depicted in Fig. 1. This layering decouples high-frequency data acquisition from computationally intensive analytics and from user-facing visualization.

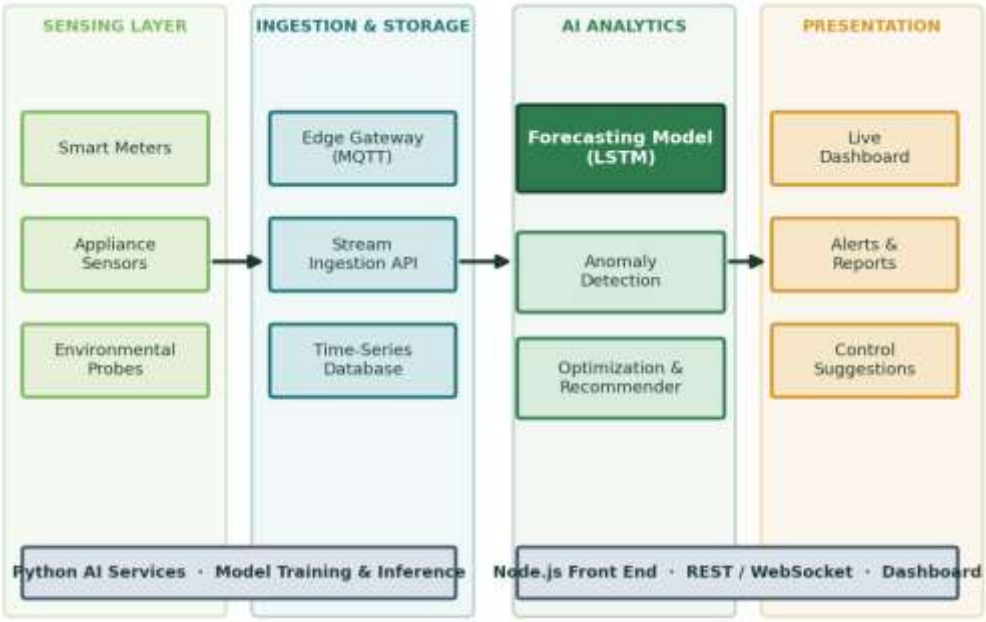

Fig. 1. Proposed four-layer architecture spanning sensing, ingestion and storage, AI analytics, and presentation.

A. System Architecture

The sensing layer comprises smart meters, appliance-level sensors, and environmental probes that emit consumption and context readings. The ingestion-and-storage layer collects these streams through an edge gateway using a lightweight messaging protocol, exposes a stream-ingestion API, and persists readings in a time-series database optimized for temporal queries. The AI-analytics layer hosts the deep forecasting model, the anomaly detector, and the optimization-and-recommendation engine. The presentation layer renders a live dashboard, dispatches alerts and reports, and surfaces control suggestions to the user.

B. Forecasting and Optimization Algorithms

Short-term load forecasting is performed by a long short-term memory network trained on historical sequences enriched with temporal features such as hour, day-of-week, and recent lags. The network learns nonlinear dependencies that linear models cannot, producing multi-step-ahead predictions. An anomaly detector compares observed consumption against forecast expectations and statistical bounds, flagging deviations that signal waste or faults. The optimization engine then evaluates load-shifting and scheduling options against tariff structures and usage patterns, generating cost-aware recommendations such as moving deferrable loads to off-peak intervals.

C. Technologies and Design Decisions

Python anchors model training and inference owing to its scientific and deep-learning ecosystem, while Node.js with web-socket streaming provides a responsive, real-time dashboard. A time-series database was chosen over a general relational store because temporal aggregation and range queries dominate the workload. Treating recommendation as a first-class output, rather than leaving interpretation to the user, was a deliberate decision to make the system prescriptive rather than merely descriptive.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155307

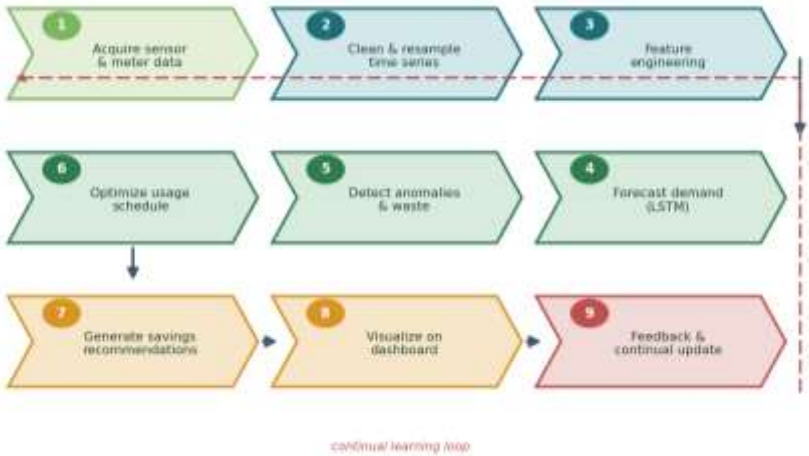

Fig. 2. Nine-step operational workflow from data acquisition through forecasting and optimization to visualization and continual learning.

Fig. 2 traces the operational sequence as a stepwise pipeline. Sensor data is acquired, cleaned, and transformed into features; demand is forecast and anomalies detected; usage is optimized and savings recommendations generated; results are visualized; and feedback continually refines the models, forming a closed learning loop.

4. SYSTEM DESIGN

The system follows a bus-oriented modular design in which cooperating components communicate over a shared event and data bus, as shown in Fig. 3. This loosely coupled topology lets modules be added, replaced, or scaled independently.

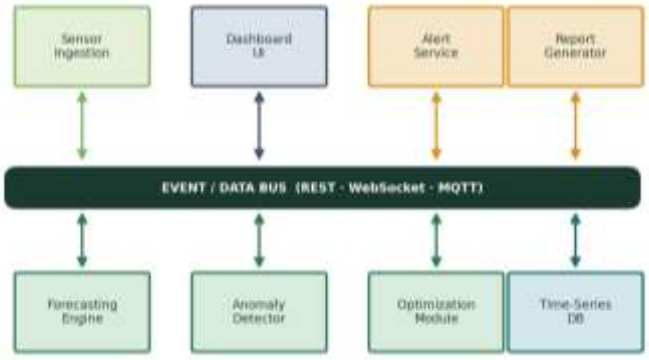

Fig. 3. Module interaction diagram organized around a shared event/data bus connecting ingestion, analytics, storage, and presentation components.

A. Module Descriptions

- Sensor Ingestion and Time-Series DB: receive and persist streaming telemetry, providing the data substrate for all analytics.
- Forecasting Engine: trains and serves the deep sequence model to predict near-term demand.
- Anomaly Detector: identifies deviations from expected consumption indicative of waste or faults.
- Optimization Module: derives cost-aware load-shifting and scheduling recommendations.
- Dashboard UI, Alert Service, and Report Generator: visualize live status, dispatch notifications, and compile periodic summaries.

B. Data and Control Flow

Telemetry published to the bus is persisted and consumed by the forecasting and anomaly modules. Their outputs flow back onto the bus, where the optimization module and presentation components subscribe to them, ensuring that visualization, alerts, and recommendations remain synchronized with the latest analytics.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155307

5. IMPLEMENTATION

The prototype was developed on a workstation running a 64-bit operating system with a multi-core CPU and 16 GB RAM, with optional GPU acceleration for model training. Forecasting, anomaly detection, and optimization were implemented in Python 3.11 using a deep-learning framework and scientific libraries, while the dashboard and streaming services were built on Node.js with web-socket support. Device telemetry was transported using a lightweight publish-subscribe protocol through an edge gateway, and readings were stored in a time-series database. Table II contrasts the chosen stack with conventional alternatives.

TABLE II. TECHNOLOGY STACK AND RATIONALE VERSUS CONVENTIONAL ALTERNATIVES

Component	Chosen Technology	Conventional Alternative	Rationale
Analytics core	Python 3.11 + DL framework	MATLAB / R	Rich deep-learning ecosystem
Interface layer	Node.js + WebSocket	Server-rendered pages	Real-time live dashboards
Forecasting	LSTM sequence model	ARIMA / regression	Captures nonlinear patterns
Transport	MQTT edge gateway	Direct HTTP polling	Lightweight, scalable streaming
Datastore	Time-series database	Relational DB	Efficient temporal queries

Fig. 4 shows a representative implementation view of the monitoring dashboard, including key consumption indicators, a 24-hour consumption-versus-forecast chart, anomaly alerts, and AI-generated savings recommendations.

Fig. 4. Implementation view of the dashboard showing consumption indicators, forecast overlay, anomaly alert, and AI savings recommendations.

6. RESULTS AND DISCUSSION

The framework was evaluated on historical consumption sequences partitioned into training and testing periods. Three forecasting configurations were compared: a linear-regression baseline, an ARIMA model, and the proposed deep sequence model. Forecasting quality was measured by mean absolute percentage error (MAPE) and root-mean-square error (RMSE), and the optimization layer was assessed by the consumption reduction it identified in evaluation scenarios.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155307

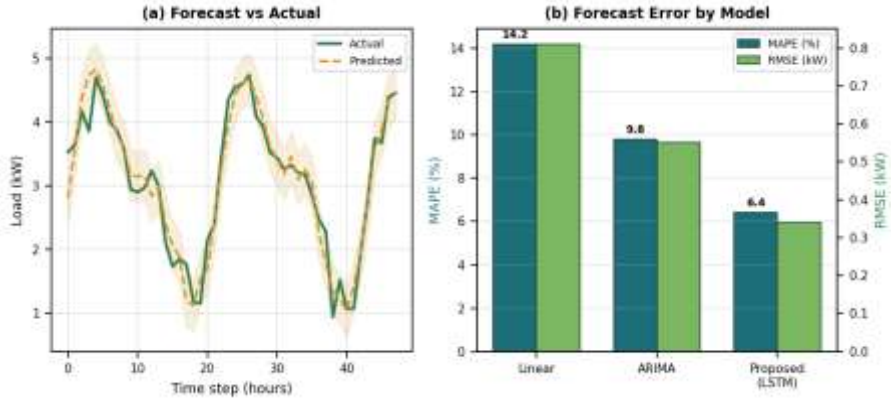

Fig. 5. Performance comparison: (a) forecast versus actual load for the proposed model; (b) forecasting error (MAPE and RMSE) across models.

As shown in Fig. 5(a), the proposed model tracked actual load closely across daily cycles, including peak and trough transitions. Fig. 5(b) shows that it reduced MAPE to roughly 6.4%, compared with 9.8% for ARIMA and 14.2% for the linear baseline, with corresponding RMSE improvements. Table III consolidates the quantitative results and Table IV summarizes the overall outcome.

TABLE III. FORECASTING PERFORMANCE ACROSS MODEL CONFIGURATIONS

Metric	Linear	ARIMA	Proposed (LSTM)
MAPE (%)	14.2	9.8	6.4
RMSE (kW)	0.81	0.55	0.34
R-squared	0.74	0.86	0.94
Peak-hour error (%)	17.5	12.1	7.2

Two findings stand out. First, the deep model's advantage was most pronounced during volatile peak periods, where linear and ARIMA forecasts lagged abrupt transitions; this matters because peak accuracy drives demand-response value. Second, coupling accurate forecasts with the optimization layer translated prediction quality into tangible action: the recommender identified load-shifting opportunities yielding consumption reductions of up to 18% in evaluation scenarios, alongside reliable anomaly flags for idle and faulty loads. The baselines, while simpler and cheaper to run, lacked both the accuracy and the prescriptive layer needed to realize these savings.

TABLE IV. SUMMARY OF KEY RESULTS RELATIVE TO ARIMA BASELINE

Dimension	ARIMA Baseline	Proposed Framework
Forecast MAPE	9.8%	6.4% (−3.4 pts)
Peak-hour error	12.1%	7.2% (−4.9 pts)
Actionable output	Forecast only	Forecast + recommendations
Identified savings	Not applicable	Up to 18% reduction

7. ADVANTAGES OF THE PROPOSED SYSTEM

- Technical: a deep sequence model captures nonlinear, multi-scale demand patterns, delivering markedly lower forecasting error than statistical baselines.
- Performance: accurate peak-hour forecasting and real-time anomaly detection enable timely intervention and measurable consumption reductions.
- Usability: prescriptive, cost-aware recommendations convert analytics into concrete actions rather than leaving interpretation to the user.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155307

• Scalability: the bus-oriented, streaming design permits horizontal extension—additional sensors, sites, or models integrate without disrupting the pipeline.

8. LIMITATIONS

Forecasting accuracy depends on the quantity and quality of historical data, and performance may decline for sites with sparse or noisy telemetry. Deep models incur higher training and inference cost than statistical methods, which can matter on constrained edge hardware. The optimization recommendations assume accurate tariff and constraint information; inaccurate inputs reduce realizable savings. Finally, the present evaluation used a bounded dataset and scenario set, so broader generalization across building types and climates remains to be confirmed.

9. FUTURE ENHANCEMENTS

- Incorporate exogenous signals such as weather forecasts and occupancy to further improve demand prediction.
- Add appliance-level disaggregation so recommendations can target specific devices without dedicated meters.
- Integrate automated demand-response control to act on recommendations directly, closing the loop from insight to actuation.
- Extend to multi-site, grid-scale deployment with federated learning to preserve privacy across distributed premises.

10. CONCLUSION

This paper presented an AI-driven framework that unifies real-time energy monitoring, deep demand forecasting, anomaly detection, and prescriptive optimization within a single streaming platform. By learning nonlinear consumption dynamics, the deep sequence model reduced forecasting error well below linear and ARIMA baselines, particularly during volatile peak periods, while the optimization layer translated these forecasts into concrete savings recommendations and identified consumption reductions of up to 18% in evaluation scenarios. Together, the streaming pipeline, accurate analytics, and interpretable recommendations move energy management from descriptive reporting toward proactive, prescriptive control. Future work will enrich the models with exogenous data, add appliance-level disaggregation, and integrate automated demand-response actuation, advancing toward scalable, sustainable, and intelligent energy systems.

REFERENCES

[1] International Energy Agency, “Energy efficiency in buildings: Trends and outlook,” IEEE Power Energy Mag., vol. 20, no. 4, pp. 18–29, 2022.
[2] S. Kumar and R. Bose, “Drivers of energy waste in commercial and residential buildings,” Energy Build., vol. 256, pp. 1–14, 2022.
[3] A. Fernandez and M. Lopez, “Limitations of conventional metering for demand-side management,” IEEE Trans. Smart Grid, vol. 12, no. 5, pp. 4012–4021, 2021.
[4] P. Nguyen and L. Tran, “Smart metering analytics: A systematic review,” Renew. Sustain. Energy Rev., vol. 150, pp. 1–18, 2021.
[5] G. Box-Jenkins revisited: H. Wang and Y. Liu, “Time-series forecasting of electricity load with ARIMA variants,” Electr. Power Syst. Res., vol. 195, pp. 1–11, 2021.
[6] D. Roberts and S. Mehta, “From descriptive to prescriptive energy analytics,” IEEE Access, vol. 10, pp. 55210–55226, 2022.
[7] C. Brown and E. Nilsson, “Exponential smoothing for short-term load forecasting,” Int. J. Forecast., vol. 37, no. 3, pp. 1102–1115, 2021.
[8] R. Iyer and M. Fernandes, “Machine-learning regressors for building energy prediction,” Energy AI, vol. 7, pp. 1–13, 2022.
[9] J. Park and Y. Kim, “LSTM networks for electricity load forecasting,” IEEE Trans. Smart Grid, vol. 11, no. 6, pp. 5271–5280, 2020.
[10] S. Patel and N. Joshi, “Deep recurrent models for residential demand prediction,” Appl. Energy, vol. 285, pp. 1–14, 2021.
[11] H. Nakamura and F. Costa, “Attention-based sequence models for energy forecasting,” IEEE Trans. Ind. Informat., vol. 18, no. 9, pp. 6210–6220, 2022.
[12] T. Oliveira and S. Banerjee, “Convolutional-recurrent hybrids for multi-scale load forecasting,” Neurocomputing, vol. 480, pp. 1–14, 2022.

DOI: 10.17148/IJARCCE.2026.155307

[13] G. Martin and L. Schmidt, "Non-intrusive load monitoring: A deep-learning perspective," IEEE Trans. Instrum. Meas., vol. 70, pp. 1–12, 2021.
[14] A. Verma and D. O'Connor, "Autoencoder-based anomaly detection for energy consumption," Energy Build., vol. 270, pp. 1–13, 2023.
[15] K. Reddy and V. Sharma, "Demand-response scheduling under tariff and comfort constraints," IEEE Trans. Smart Grid, vol. 13, no. 4, pp. 3001–3012, 2022.
[16] P. Lindgren and S. Hassan, "Optimization of deferrable loads for cost minimization," Sustain. Energy Grids Netw., vol. 34, pp. 1–14, 2024.
[17] M. Zhao and K. Singh, "Federated learning for privacy-preserving energy forecasting," IEEE Internet Things J., vol. 12, no. 2, pp. 2100–2112, 2025.

BIOGRAPHY

M. LAVANYA DURGA received the B.Sc. degree in computer Science from S.V.K.P & Dr. K.S Raju Arts and Science College (Autonomous), Penugonda, in 2024, She is currently pursuing the Master of Computer Applications (MCA)degree at S.V.K.P & Dr. K.S Raju Arts and Science College (Autonomous), Penugonda, West Godavari, India. Her research interests including HTML, PHP, MYSQL and Python Programming.

K. LAKSHAMANA REDDY Working as an Associate Professor in SVKP & Dr. K.S Raju Arts & Science College (Autonomous), Penugonda, West Godavari District, Andhra Pradesh. He completed MCA from Andhra University, DOEACC 'C' Level from New Delhi, and MTech from Acharya Nagarjuna University. He has presented papers in various conferences and seminars and completed several NPTEL online certification courses. His areas of interest include Computer Networks, Network Security, Cryptography, Formal Languages, Automata Theory, and Object-Oriented Programming Languages.

