Short-Term Load Forecasting for 11kV Distribution Feeder Using Machine Learning Ensemble Methods: A Case Study of Khaboorah Substation in Oman

Abdul Saleem Shaik
University of Technology and Applied Sciences (UTAS)
Sohar, Oman
abdulsaleem@utas.edu.om

***Abstract*—Short-term load forecasting (STLF) is essential for operational planning in electrical distribution systems, including generation scheduling, demand-side management, and grid stability enhancement. This paper presents a comprehensive machine learning approach for accurate load prediction at an 11kV distribution feeder in Khaboorah, Oman. A dataset spanning March 2025 to August 2025 was collected from SCADA measurements, incorporating hourly load readings from eight feeder lines, temporal features, and weather parameters. The proposed methodology employs sophisticated feature engineering including lag variables (1-168 hours), rolling statistics, and cyclical time encodings. Four machine learning models—XGBoost, LightGBM, Random Forest, and Ridge Regression—were systematically evaluated. Results demonstrate that among the tested models, XGBoost achieved superior performance with a Root Mean Square Error (RMSE) of 6.827 A, R^2 score of 0.8094, and Mean Absolute Percentage Error (MAPE) of 3.25%. Feature importance analysis revealed that rolling mean statistics and 24-hour lag features are the dominant predictors. Additionally, the study identified a significant 53% reduction in load during the Ramadan period, highlighting the importance of cultural factors in regional load forecasting applications.**

***Index Terms*—Short-term load forecasting, machine learning, XGBoost, distribution system, Oman, ensemble methods**

I. Introduction

Accurate short-term load forecasting (STLF) plays a pivotal role in modern power system operations, enabling utilities to optimize generation scheduling, manage demand response programs, and maintain grid stability [4]. With the increasing penetration of renewable energy sources and the growing complexity of distribution networks, the need for precise load predictions has become more critical than ever [5].

The Gulf Cooperation Council (GCC) region presents unique challenges for load forecasting due to extreme weather conditions, cultural practices affecting energy consumption patterns, and rapid economic development [6]. Oman, in particular, experiences significant seasonal variations in electricity demand, with peak loads occurring during summer months when air conditioning requirements are at their highest [15].

Traditional forecasting methods, including regression-based approaches and time series models, have been widely employed in the power industry [7]. However, these methods often struggle to capture the complex non-linear relationships between load and influencing factors. Machine learning (ML) techniques have emerged as powerful alternatives, demonstrating superior performance in handling high-dimensional data and capturing intricate patterns [10].

This paper presents a comprehensive study on short-term load forecasting for an 11kV distribution feeder at Khaboorah substationin Oman. The main contributions of this work are:

- Development of a robust feature engineering framework incorporating lag variables, rolling statistics, and cyclical temporal encodings for distribution-level load forecasting.
- Systematic evaluation of four machine learning models (XGBoost, LightGBM, Random Forest, and Ridge Regression) using real SCADA data from Oman's distribution network.
- Analysis of cultural factors, specifically the impact of Ramadan on electricity consumption patterns in the GCC region.
- Practical insights for implementing ML-based forecasting solutions in Middle Eastern distribution systems.

II. Literature Review

A. Machine Learning in Load Forecasting

The application of machine learning techniques to load forecasting has gained significant momentum in recent years. Chen et al. [2] demonstrated the effectiveness of Support Vector Regression (SVR) for calculating demand response baselines in commercial buildings, achieving accurate predictions with limited training data. The XGBoost algorithm, introduced by Chen and Guestrin [1], has become a benchmark for tabular data problems due to its scalability and regularization capabilities.

LightGBM, proposed by Ke et al. [3], offers computational advantages over traditional gradient boosting methods through its leaf-wise tree growth strategy. Kong et al. [8] explored deep learning approaches, specifically LSTM networks, for residential load forecasting, demonstrating the potential of recurrent architectures for capturing temporal dependencies.

Recent comprehensive reviews by Haben et al. [5], [9] have highlighted the growing importance of data-driven methods in low voltage load forecasting, emphasizing the need for feature engineering and model selection strategies tailored to specific application domains.

B. Load Forecasting in the GCC Region

Load forecasting in the GCC region presents unique challenges due to extreme temperatures and cultural factors. Swaroop [6] applied artificial neural networks for short-term load forecasting in the Al Batinah region of Oman, demonstrating the applicability of ML methods to local distribution networks. Al-Hamadi and Soliman [7] proposed a Kalman filtering approach with moving window weather models for short-term forecasting in similar climatic conditions.

The influence of Islamic holidays, particularly Ramadan, on electricity consumption patterns has been noted in regional studies but remains underexplored in the literature. This cultural factor significantly impacts daily load profiles and requires special consideration in forecasting models deployed in Muslim-majority countries.

III. Methodology

A. Data Collection and Description

The dataset used in this study was collected from the SCADA system at Khaboorah 33/11kV substation in Al Batinah North Governorate, Oman. The data spans from March 2025 to August 2025, comprising 4,193 hourly observations. The substation serves a mixed residential, commercial, and agricultural load through eight 11kV feeder lines:

- FDR-1 Khaboorah: Serving the main town center
- FDR-2 Ghilah: Agricultural and residential areas
- FDR-3 ONWJ (Old): Industrial water pumping
- FDR-4 Nuaman: Residential community
- FDR-5 ONWJ (New): Updated water infrastructure
- FDR-6 Qasab: Mixed residential and commercial
- FDR-7 KOM: Commercial zone
- FDR-8 PACA: Government and institutional loads

Weather data including temperature, humidity, and solar radiation was integrated from meteorological stations to enhance forecasting accuracy.

B. Feature Engineering

A comprehensive feature engineering approach was implemented to capture temporal patterns and dependencies in the load data:

1) Lag Features: Historical load values at various time intervals were incorporated as features:

$$X_{lag\_h} = y_{t-h}, \quad h \in \{1, 2, 3, 6, 12, 24, 48, 168\} \tag{1}$$

where y_{t-h} represents the load value h hours before the forecast time.

2) Rolling Statistics: Rolling window statistics were computed to capture recent trends:

$$\mu_{roll,w} = \frac{1}{w} \sum_{i=0}^{w-1} y_{t-i} \tag{2}$$

$$\sigma_{roll,w} = \sqrt{\frac{1}{w} \sum_{i=0}^{w-1} (y_{t-i} - \mu_{roll,w})^2} \tag{3}$$

with window sizes $w \in \{6, 12, 24, 48, 168\}$ hours.

3) Cyclical Temporal Features: Time-based features were encoded using sine and cosine transformations to preserve cyclical continuity:

$$hour_{sin} = \sin\left(\frac{2\pi \cdot hour}{24}\right) \tag{4}$$

$$hour_{cos} = \cos\left(\frac{2\pi \cdot hour}{24}\right) \tag{5}$$

Similar encodings were applied for day of week and month features.

C. Machine Learning Models

Four machine learning algorithms were evaluated in this study:

1) XGBoost: XGBoost implements gradient boosting with regularization terms to prevent overfitting [1]:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{6}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ is the regularization term.

2) LightGBM: LightGBM employs histogram-based splitting and leaf-wise tree growth for improved efficiency [3].

3) Random Forest: Random Forest constructs an ensemble of decision trees using bootstrap aggregation and random feature selection.

4) Ridge Regression: Ridge Regression serves as a baseline linear model with L2 regularization:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|^2 \right\} \tag{7}$$

D. Evaluation Metrics

Model performance was assessed using the following metrics:

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{8}$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{10}$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{11}$$

IV. Results

A. Data Exploration

Fig. 1 presents the exploratory analysis of the collected load data. The time series exhibits clear daily and weekly periodicity, with peak loads occurring during afternoon hours when cooling demands are highest. The distribution of total load shows a right-skewed pattern typical of electrical demand data.

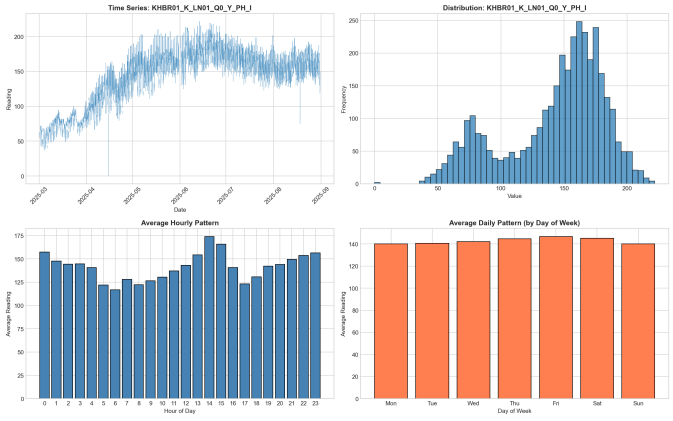

Fig. 1. Exploratory data analysis showing load patterns across the study period including time series trends and distribution characteristics.

B. Weather Features Analysis

Fig. 2 illustrates the weather parameters collected during the study period. Temperature shows strong correlation with load demand, with values ranging from 25°C to 45°C during the summer months. Humidity and solar radiation patterns also influence consumption behavior.

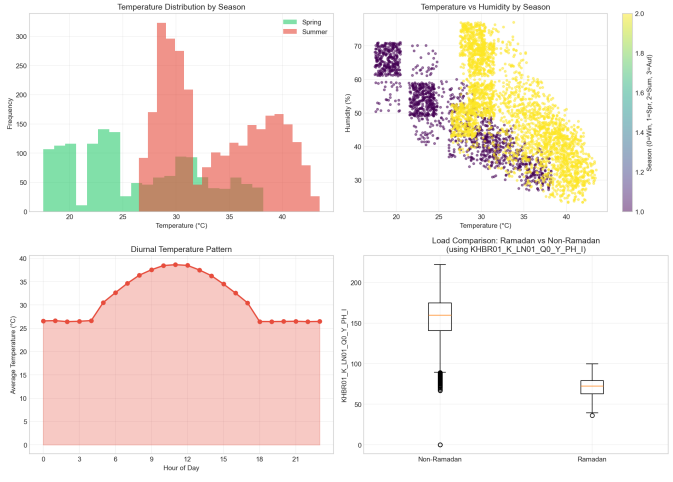

Fig. 2. Weather features visualization showing temperature, humidity, and solar radiation patterns during the study period.

C. Model Performance Comparison

Table I presents the comprehensive performance comparison of the four evaluated models. XGBoost achieved the best overall performance across all metrics on the test dataset.

TABLE I
Model Performance Comparison

Model	Test MAE	Test RMSE	Test R^2	MAPE
XGBoost	4.923	6.827	0.8094	3.25%
LightGBM	5.061	7.015	0.7988	3.35%
Random Forest	6.149	8.565	0.7000	4.04%
Ridge Regression	7.153	8.978	0.6704	4.59%

TABLE II
Training Performance Metrics

Model	Train MAE	Train RMSE	Train R^2
XGBoost	1.600	2.156	0.9965
LightGBM	3.048	4.232	0.9867
Random Forest	3.910	6.087	0.9725
Ridge Regression	10.598	14.476	0.8443

Fig. 3 provides a visual comparison of model predictions against actual load values. XGBoost demonstrates the closest alignment with measured data, particularly during peak demand periods.

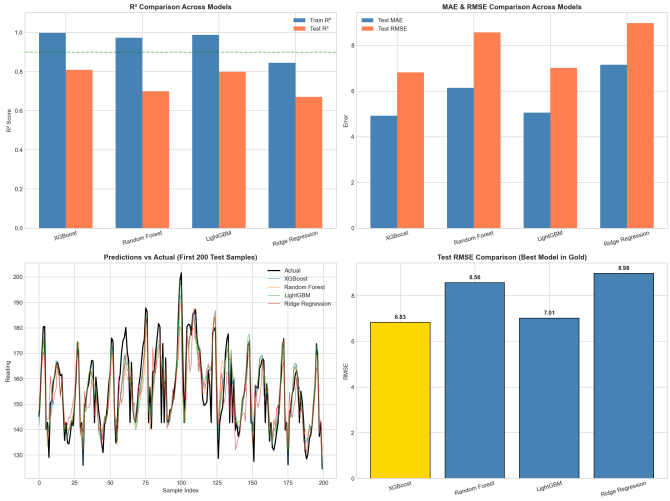

Fig. 3. Model comparison visualization showing predicted vs. actual load values for all evaluated models.

Fig. 4 presents detailed evaluation metrics including residual analysis and prediction accuracy distributions for each model.

D. Feature Importance Analysis

Fig. 5 displays the feature importance rankings from the tree-based models. Rolling mean statistics, particularly the 24-hour and 168-hour windows, emerged as the most influential predictors. The 24-hour lag feature also ranked highly, confirming the strong daily periodicity in load patterns.

E. Ramadan Impact Analysis

A significant finding of this study is the pronounced impact of Ramadan on electricity consumption patterns. During the Ramadan period (approximately March 10 to April 9, 2025), a 53% reduction in average load was observed compared to

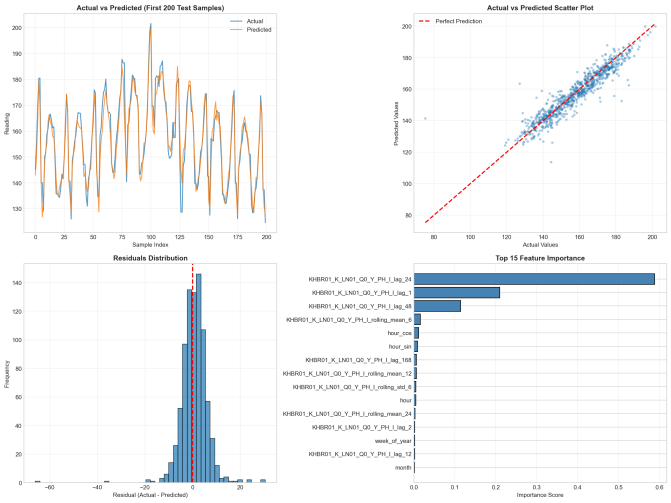

Fig. 4. Detailed model evaluation showing residual distributions and prediction accuracy analysis.

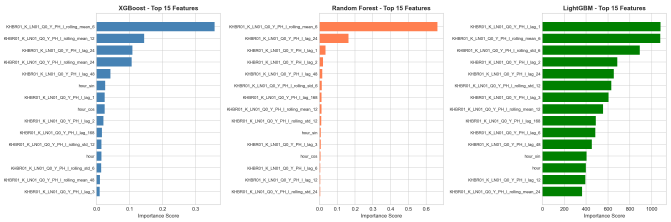

Fig. 5. Feature importance comparison across XGBoost, LightGBM, and Random Forest models showing the relative contribution of each feature to prediction accuracy.

non-Ramadan periods. This substantial decrease is attributed to:

- Altered daily activity schedules during fasting hours
- Reduced commercial and industrial operations
- Changes in cooking and household routines
- Modified work schedules in government and private sectors

This finding underscores the importance of incorporating cultural and religious factors into load forecasting models deployed in Muslim-majority regions.

V. Discussion

A. Model Performance Analysis

The superior performance of XGBoost can be attributed to several factors. First, the gradient boosting framework effectively captures non-linear relationships between features and load values. Second, the built-in regularization prevents overfitting despite the high-dimensional feature space. Third, XGBoost's handling of missing values and its robustness to outliers contribute to its reliability.

The gap between training and test performance for XGBoost (R^2 of 0.9965 vs. 0.8094) suggests some degree of overfitting, though the test performance remains acceptable for practical applications. LightGBM showed more consistent training-test performance, indicating better generalization at a slight cost to accuracy.

Ridge Regression's limited performance confirms that linear models are insufficient for capturing the complex temporal dependencies in distribution-level load data. This supports the adoption of ensemble methods for similar forecasting applications.

B. Practical Applications

The developed forecasting system has several practical applications for distribution system operators in Oman:

- **Operational Planning:** Day-ahead load forecasts enable optimal transformer and feeder loading decisions.
- **Maintenance Scheduling:** Predicted low-load periods can be identified for maintenance activities.
- **Demand Response:** Accurate forecasts support peak shaving and demand response program implementation.
- **Integration Planning:** Load predictions assist in planning distributed generation and storage integration.

C. Limitations and Future Work

This study has some limitations that suggest directions for future research. The six-month dataset, while sufficient for demonstrating methodology, may not capture full annual seasonal patterns. Future work should incorporate longer historical data spanning multiple years to improve model robustness.

Additionally, deep learning approaches such as LSTM [8] and attention mechanisms [13] could be explored for improved sequence modeling. Real-time forecasting with adaptive model updating represents another promising research direction.

VI. Conclusion

This paper presented a comprehensive machine learning approach for short-term load forecasting at an 11kV distribution feeder in Khaboorah, Oman. Using 4,193 hourly observations from SCADA measurements, four machine learning models were systematically evaluated with sophisticated feature engineering techniques.

Key findings include:

1) XGBoost achieved the best forecasting performance with RMSE of 6.827 A, R^2 of 0.8094, and MAPE of 3.25%.
2) Rolling statistics and lag features are the most important predictors for distribution-level load forecasting.
3) Cultural factors, specifically Ramadan, cause significant (53%) load reductions that must be accounted for in regional forecasting models.
4) Ensemble methods significantly outperform linear models for capturing complex load patterns.

The methodology developed in this study provides a practical framework for implementing ML-based load forecasting in distribution systems across the GCC region. Future work will focus on extending the approach to longer time horizons and incorporating deep learning techniques for improved accuracy.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to the University of Technology and Applied Sciences (UTAS), Sohar Campus, for providing the research facilities and support for this study. Special thanks are extended to Oman Electricity Transmission Company (OETC) for providing access to the SCADA data from Khaboorah substation. The author also acknowledges the valuable feedback from colleagues in the Engineering Department.

REFERENCES

[1] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[2] Y. Chen *et al.*, "Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings," *Applied Energy*, vol. 195, pp. 659–670, 2017.

[3] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.

[5] R. Haben, C. Singleton, and P. Grindrod, "Review of low voltage load forecasting: Methods, applications, and recommendations," *Applied Energy*, vol. 304, p. 117798, 2021.

[6] R. Swaroop, "Short-term load forecasting using artificial neural network for Al Batinah region in Oman," *J. Eng. Sci. Tech.*, vol. 7, no. 4, pp. 498–504, 2012.

[7] H. M. Al-Hamadi and S. A. Soliman, "Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model," *Electric Power Systems Research*, vol. 68, no. 1, pp. 47–59, 2004.

[8] W. Kong *et al.*, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[9] S. Haben *et al.*, "Review and integration of data-driven load forecasting using machine learning and deep learning," *Energy Reports*, vol. 7, pp. 5234–5252, 2021.

[10] M. Q. Raza and A. Khosravi, "A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings," *Renewable and Sustainable Energy Reviews*, vol. 50, pp. 1352–1372, 2015.

[11] A. Ahmad *et al.*, "An accurate and fast converging short-term load forecasting model for industrial applications in a smart grid," *IEEE Trans. Ind. Informatics*, vol. 13, no. 5, pp. 2587–2596, 2017.

[12] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecasting—A novel pooling deep RNN," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5271–5280, 2018.

[13] S. Wang *et al.*, "Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting," *Int. J. Electrical Power Energy Syst.*, vol. 109, pp. 470–479, 2019.

[14] E. Zivot and J. Wang, "Rolling analysis of time series," in *Modeling Financial Time Series with S-PLUS*. New York: Springer, 2006, pp. 313–360.

[15] Authority for Electricity Regulation Oman, "Annual Report 2024," Muscat, Oman, 2024.

[16] X. Zhang, J. Wang, and K. Zhang, "Short-term electric load forecasting based on singular spectrum analysis and support vector machine optimized by cuckoo search algorithm," *Electric Power Systems Research*, vol. 146, pp. 270–285, 2017.

[17] L. Hernandez *et al.*, "Artificial neural networks for short-term load forecasting in microgrids environment," *Energy*, vol. 75, pp. 252–264, 2014.

[18] OSTI, "A cross-dimensional analysis of data-driven short-term load forecasting," U.S. Dept. Energy, Tech. Rep. 2997924, 2025.