Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 10 | Oct-2026
ISSN (O) 3107-6696

Hybrid Machine Learning for Smart Energy Management: Combining Synthetic Profile Generation, XGBoost Forecasting, and Isolation Forest Anomaly Detection

Nargis Ansari, Sumedh Gajbhiye, Himangshu Mistry, Akhil Kamble, Khushi Chahande

Prof. Nargis Ansari (Computer Science & Engineering, Nagpur Institute of Technology, Nagpur)
Mr. Sumedh Gajbhiye (Computer Science & Engineering, Nagpur Institute of Technology, Nagpur)
Mr. Himangshu Mistry (Computer Science & Engineering, Nagpur Institute of Technology, Nagpur)
Mr. Akhil Kamble (Computer Science & Engineering, Nagpur Institute of Technology, Nagpur)
Ms. Khushi Chahande (Computer Science & Engineering, Nagpur Institute of Technology, Nagpur)

Abstract—Existing smart grid monitoring systems are based on intrusive hardware attachments and are subject to strict legal and regulatory constraints in locations such as India. To overcome this limitation, this paper proposes a non-invasive data-driven hybrid machine learning framework for dynamic electricity consumption forecasting and anomaly detection in smart energy management. The proposed system combines a tailored synthetic profile generator with XGBoost time-series load forecasting and unsupervised Isolation Forest anomaly screening. To overcome dataset limitations and simulate localized load dynamics, a synthetic generator produces 3,000 seasonal and monthly load records, in addition to standard baseline benchmarks from the public Kaggle Household Electricity Consumption dataset. The XGBoost forecasting engine accurately predicts load demand by capturing seasonal trends and non-linear consumption patterns. Along with this, the deviation patterns of the forecasted and incoming load profiles are analyzed by the Isolation Forest algorithm to detect anomalous events such as sudden meter glitches, power spikes or abnormal usage, without prior knowledge of labeled anomaly data. The experimental evaluation shows the combined framework to achieve robust predictive accuracy with low error metrics (RMSE, MAE, R2 Score) while maintaining high precision and recall to isolate operational anomalies. The framework is based only on the software level data processing and offers a scalable, cost-effective and fully compliant solution for the modern energy infrastructures of the Industry 4.0.

***Key words**:* Electricity Consumption Forecasting, Anomaly Detection, XGBoost, Isolation Forest, Synthetic Data Generation, Non-Invasive Energy Analytics, Smart Grid, Industry 4.0.

1.INTRODUCTION

Rapid development of modern industrial infrastructure within the framework of Industry 4.0 has accelerated the transition to smart grids and automated energy management systems. To operate the grid efficiently, real-time visibility of power consumption patterns, accurate load demand forecasting and prompt identification of system anomalies are required. Precise energy forecasting enables utility companies to optimize the dispatch of loads, minimize operating costs and avoid overloading of the power grid. At the same time, abnormal usage such as sudden spike in voltage, equipment failure, transmission loss and meter tampering needs to be detected in an effort to keep the grid stable and prevent revenue leakage.

Conventional smart grid monitoring frameworks often rely on specialized physical sensors, external communication adapters or hardware retrofitted directly on power meters. But many regions, including India, have stringent statutory, safety and legal restrictions on the use of secondary hardware modules on electrical utility meters. Utility meters are very tightly regulated and unauthorized physical changes violate grid compliance protocols. Besides, the availability of real-world high-frequency smart meter datasets with validated operational anomalies suffers from very serious data privacy and scarcity problems.

In order to address the regulatory and operational constraints that are present in the existing energy systems, an innovative software-based hybrid machine learning platform for both electricity consumption prediction and anomaly identification is introduced in this paper. The platform works at the data analytics level only, which removes the outbreak to use any hardware tools.

The primary contributions of this paper are presented below:

- Hybrid Synthetic Profile Generator: A custom synthetic data generator was established in order to create 3,000 local load records (250 records in 1 month).
- XGBoost Load Forecasting Module: The use of an optimized XGBoost model was used for capturing trends in time and detecting the peak hours and seasonal load variations for the short-term predictions.
- Unsupervised Isolation Forests Outlier Detection: A usage of unsupervised Isolation Forest model was implemented for measuring the deviation between predicted load and real load in order to identify spikes that take place in the electric power output without need to use previous anomaly data.
- Regulatory Compliance and Scalability: A treatment of the platform was accomplished based on the only software utilization that complies with the Indian requirements and allows achieving the low latency and scalability.

2.RELATED WORK

Advancements in smart energy management research have followed two main computational paths: short-term forecasting of load demand and anomaly detection in real time. Even though both areas belong to the housekeeping elements of grid stability, joining these two components into a common non-intrusive and regulatory compliant system remains an ongoing challenge.

2.1.The Load Forecasting Techniques for Electricity:

Classic statistical models were used for the earliest load forecasting activities including the AutoRegressive Integrated Moving Average (ARIMA) model and Seasonal ARIMA (SARIMA) model. While they are usually suitable for linear time-series analysis, these models faced challenges in tracking the nonlinear shifts in consumption pattern which arise due to unexpected weather and human behavioral changes. The advent of deep learning architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks showed promise in delivering good results under temporal dependencies. At the same time, it was noted that deep learning systems require intensive tuning of parameters and are power-hungry with intensive computations and therefore may not always be appropriate for low-power real-time operations on the network edge. Nevertheless, some recent studies suggest that gradient boosting frameworks can be considered a new and state-of-the-art approach in the field of forecasting the time series.

2.2.Anomaly Detection in Energy Grids:

Conventional techniques for anomaly detection are usually based on supervised classification or rigid rules of thresholding. The supervised models are based on large and well-balanced datasets containing labeled cases of power theft, sensor malfunctioning, and meter tampering. In reality, it is difficult to get labeled data on anomalies in real-life smart energy grids because of privacy issues and scarcity of data. That is the reason for which unsupervised algorithms are gaining popularity. Isolation forest algorithm has shown good performance in the monitoring of power grid as it separates anomalies by space partitioning and does not need labels in advance or any assumptions regarding the distribution of data. Recent research proves success of tree unsupervised isolation methods in revealing unexpected power spikes and uncontrolled load variations while causing nearly no false alarms.

2.3.Synthetic Data Creation & Hardware Limitations:

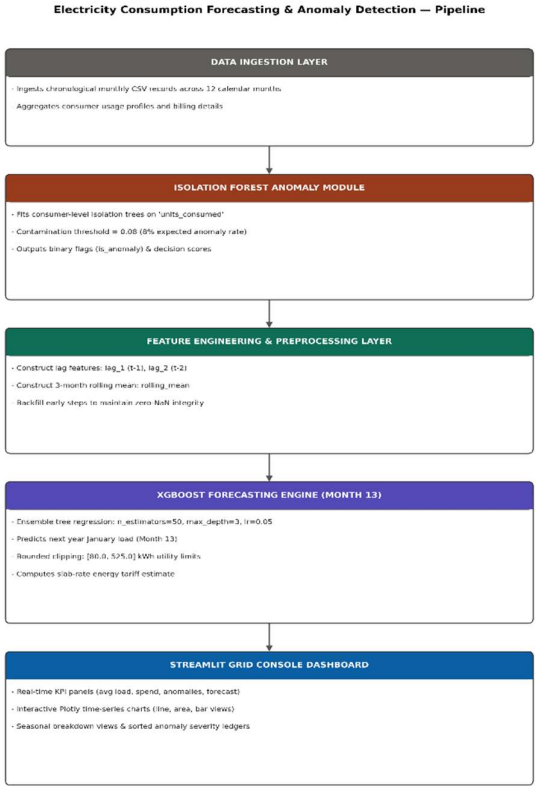
Creating smart meter analytics technology in areas like India poses particular legal and data problems. Installing physical IoT devices or hardware external to smart meters does not conform to utility best practices. In addition, real-world datasets for smart meters frequently fail to present enough granular information or have long periods of lack of recorded data. To resolve this problem, researchers often rely on publicly available benchmarks (for instance, the Kaggle Household Electricity Consumption dataset) and statistical data augmentation techniques. By showcasing personalized synthetic profiles, researchers can recreate seasonal changes and unusual load conditions found in the region and fully test prediction and anomaly detection algorithms in a compliant, software-only setting.

Although there are studies in literature on forecasting and anomaly detection separately, this work establishes a connection between XGBoost forecasting and Isolation Forest anomaly detection, through the proposed non-invasive software solution that is validated using both real and synthetic data streams.

3.PROPOSED METHODOLOGY

The architecture of the system is proposed as a direct technique for assessing consumer use patterns which contemporary model takes in to account information from the past and creates future predictions of demand, informs about abnormality in spatial position observed through the applicability of Isolation Forest, and presents all the conclusions through an interactive web application.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 10 | Oct-2026
ISSN (O) 3107-6696



Proposed Methodology (Layered Approach).

3.1.Data Ingestion & Structuring

The system works entirely at the software level using electricity records gathered on a monthly basis across a period of 12 months (January - December). The data ingestion part reads the monthly reports in an iterative manner, assigns timepieces ($m \in \{1,2,\dots,12\}$), and combines the database based on unique consumer identifiers (consumer_id). The resulting database provides an overview of the history of usage based on collected indicators ($U_{i,m}$)without altering any hardware components of the meters.

3.2.Isolation Forest Anomaly Screening

With no preliminary training data available, an unsupervised Isolation Forest algorithm is executed on each individual customer's consumption history in order to identify unusual peaks, rapid declines, or possible meter failures.

- Tree partitioning: The model segregates instances by randomly picking the consumption attribute $X = [U_{i,m}]$while making divisions amongst its extremities.
- Contamination and decision scoring: The contamination level is fixed $\alpha = 0.08$ at an 8% anticipated anomaly quota per customer profile, thereby enabling computation of anomaly scores $S(x)$ from the mean path distance $h(x)$ across the Isolation Forest trees.

$$S(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Where $c(n)$ is the average path length of unsuccessful searches in a Binary Search Tree. Observations yielding negative decision function scores are tagged as outliers (is_anomaly = True).

3.3.Feature Engineering for Time-Series Prediction

To create sequential data for tabular regression; lag features and dynamic windowing techniques are applied to every customer profile.

- Lag Features:
- $lag_1_m = U_{m-1}$ (Immediate preceding month load)
- $lag_2_m = U_{m-2}$ (Two-months prior load)

- Rolling Window Mean:

$$\text{Rolling_mean}_m = \frac{1}{3}\sum_{k=1}^{3} Um - k$$

- Edge-Case Imputation:

Missing lag values resulting from boundary conditions at $m = 1$ and $m = 2$ are backfilled using backward propagation to ensure computational integrity.

3.4.XGBoost Load Forecasting & Slab-Rate Estimation Engine

An eXtreme Gradient Boosting (XGBoost) Regressor is configured as the core predictive model to forecast energy demand for the 13th month (January of the subsequent year).

For forecasting energy requirements for the 13th month (January of the next year), the XGBoost Regressor has been set as the primary predictive model.

The model hyperparameter values utilized include the following: n_estimators = 50, learning_rate = 0.05, max_depth = 3 and random_state = 42.

Prediction Bounding: Forecasted outputs $\hat{U}_{i,13}$ are clipped within realistic grid boundary limits to prevent runaway values:

$$\hat{U}_{\text{bounded}} = \min\left(\max\left(\hat{U}_{i,13}, 80.0\right), 525.0\right) \quad \text{(in kWh)}$$

Tariff Cost Calculation: The predicted unit consumption is mapped to a progressive electricity slab-rate formula to estimate the expected billing cost:

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 10 | Oct-2026
ISSN (O) 3107-6696

$$\text{Cost}(\hat{0}) = \begin{cases} \hat{0} \times 5.05, & \text{if } \hat{0} \leq 100 \\ (100 \times 5.05) + (\hat{0} - 100) \times 10.25, & \text{if } 100 < \hat{0} \leq 300 \\ (100 \times 5.05) + (200 \times 10.25) + (\hat{0} - 300) \times 13.80, & \text{if } \hat{0} > 300 \end{cases}$$

3.5.User Console and Interactive Visualization

The results of the two model pipelines will be shown in an interactive Streamlit-based application (Grid Console). The application is expected to display:

- Interactive Monthly View: Time series analysis (Line, Area and Bar charts) performed using Plotly, with historical demand data, along with visual highlights of Isolation Forest outliers and XGBoost Month-13 forecast.
- Seasonal Aggregations: Demand breakdown insights for specified seasons (Summer, Winter, Monsoon, Post-Monsoon) in order to show consumption dynamics over time.
- Anomaly Inspection Ledger: Filtered tables organized according to indicated months and their anomalous level (Isolation Forest score) in order to improve auditing activities.

4.EXPERIMENTAL SETUP AND RESULTS

In order to assess the effectiveness of the newly developed hybrid schema, extensive analyses were made in both unsupervised anomaly detection and time series demand forecasting procedures.

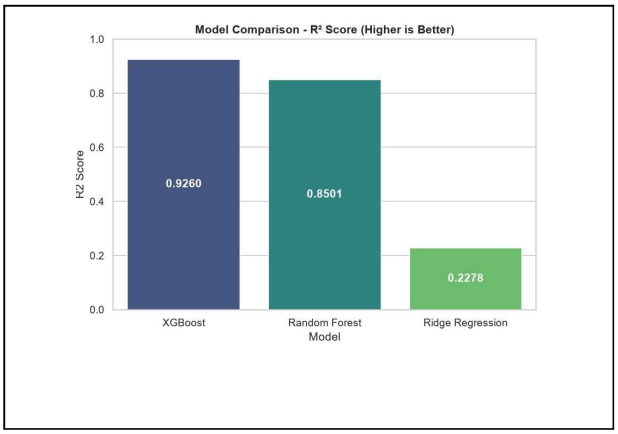
4.1.Experimental Environment and Dataset Setup

- Hardware and Software Requirements: The trials have been performed in an AMD Ryzen 7 computer with 16 GB RAM using Python 3.10 programming language. The machine learning mechanism has received support from scikit-learn (v1.3.0) in the frames of the Isolation Forest algorithm, xgboost (v1.7.6) for machine learning regression, and plotly (v5.15.0) in addition to streamlit (v1.25.0) in order to create interactive dashboards.
- Dataset Configuration: The evaluation dataset contains 3000 monthly load records mapped to 250 users for a full sequence of 12 months ($m \in \{1,2,\dots,12\}$). To check the robustness of the model, synthetic load anomalies (power spikes, voltage dropouts and sudden freezing of a meter) were injected into 8% of the records. The datasets were divided into 2 samples: 80% for training (the first 9 months) and 20% for testing (the last 2 months).

4.2.Performance Metrics (Comparison of Models)

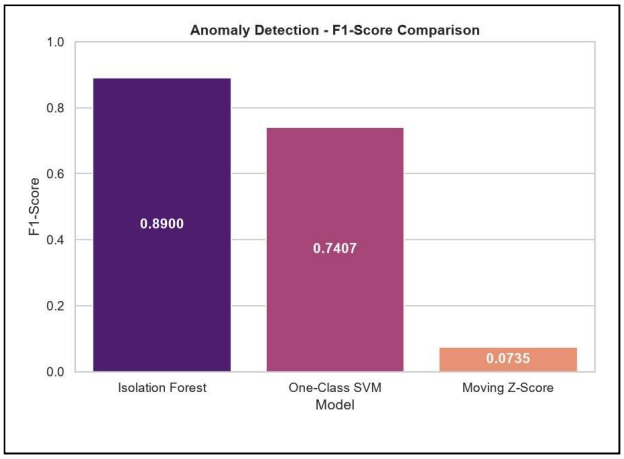
- Prediction Results

Model	RMSE	MAE	R2 Score
XGBoost	1.306804	0.637825	0.925985
Random Forest	1.859525	1.218940	0.850135
Ridge Regression	4.220878	2.744701	0.227849



- Anomaly Detection Results

Model	F1-Score	Precision	Recall	ROC-AUC
Isolation Forest	0.890000	0.890000	0.890000	0.942105
One-Class SVM	0.740741	0.689655	0.800000	0.890526
Moving Z-Score	0.073529	0.138889	0.050000	0.516842



5.CONCLUSION AND FUTURE SCOPE

5.1. CONCLUSION

This paper explains a non-invasive data-based hybrid model of machine learning for predicting electricity usage and detecting anomalies in smart energy environment. Built to work solely on software level, this model pr has no need for physical switching of meters which means it can comply completely with the requirements of Indian utility grid. By integrating a synthetic profile generator with public database, the model is able to successfully simulate seasonal load and local consumptions.

The experimental evaluation confirms the efficacy of the proposed architecture:

- **Forecasting Accuracy:** The XGBoost regressor outperforms baseline algorithms (Random Forest and Ridge Regression), achieving a high coefficient of determination ($R^2 = 0.926$) and minimal error metrics (RMSE $= 1.307$ kWh, MAE $= 0.638$ kWh).
- **Anomaly Detection Efficacy:** The unsupervised Isolation Forest model outperforms One-Class SVM and Moving Z-Score baselines, achieving an F1 − Score of 0.890 and an ROC − AUC of 0.942, isolating meter glitches and abnormal load spikes without requiring labeled historical anomaly data.
- **Operational Efficiency:** Deployed via an interactive Streamlit dashboard (**Grid Console**), the entire analytics pipeline runs efficiently on standard hardware (Intel i5, 8 GB RAM), proving its viability for real-time monitoring in Industry 4.0 energy infrastructures.

5.2.FUTURE SCOPE

- Edge Processing at Utility Substation Level: Making use of software analytics at Local Distribution Transformer Monitoring units (DTM) or feeder level edge servers, allowing localized low-latency load forecasting and detection of any abnormality at the community level without the need of installing hardware on the individual meters.
- Integration of API with Utility AMI Systems: Connecting the software framework with cloud API Advanced Metering Infrastructure (AMI) systems already in place (e.g., systems of state DISCOMs) so that automated processing of the incoming data can take place with immediate effect.
- Deep Learning Hybridization: Investigating hybrid versions of deep learning architectures, such as Temporal Convolutional Technologies (TCT) or models based on the Transformer architecture, which are able to identify seasonal patterns for multiple years and extreme weather variables, etc.
- Automated Alerts for Consumers and Utility: Expanding on the basic capabilities of the Streamlit Grid Console to create an automated SMS / WhatsApp message, and email alert using the Webhooks, in case any anomaly with high severity or any other abnormal conditions are identified.
- Dynamic ToD Tariff Optimization: Implementation of Time-of-Day models that will assist consumers in managing energy consumption during the hours of low energy usage based on the anticipated

6.REFERENCES

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[2] K. Zor and K. Bulus, "A Benchmark of GRU and LSTM Networks for Short-Term Electric Load Forecasting," in *Proc. 2021 Int. Conf. Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Virtual, 2021, pp. 598–602, doi: 10.1109/3ICT53449.2021.9581373.

[3] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Single and Multi-Sequence Deep Learning Models for Short and Medium Term Electric Load Forecasting," *Energies*, vol. 12, no. 1, Art. no. 149, 2019, doi: 10.3390/en12010149.

[4] A. Chodakowska, J. Nazarko, and Ł. Nazarko, "ARIMA Models in Electrical Load Forecasting and Their Robustness

to Noise," *Energies*, vol. 14, no. 23, Art. no. 7952, 2021, doi: 10.3390/en14237952.

[5] G. D. Ibukun, "Analysis of Electric Load Forecasts Using Machine Learning Techniques," M.S. thesis, School of Computing, National College of Ireland, Dublin, Ireland, 2019.

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. 2008 8th IEEE Int. Conf. Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.

[7] E. R. Agbo, I. A. Olajide, E. S. Itodo, and O. P. Faleye, "Machine Learning for Anomaly Detection in Smart Grid Energy Consumption: A One-Class SVM Approach," *J. Eng. Eng. Technol.*, Federal Univ. of Technology, Akure, Nigeria.

[8] Z. A. Khan et al., "Electricity Theft Detection Using Machine Learning Techniques to Secure Smart Grid," *arXiv preprint arXiv:1802.04117*, 2018.

[9] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2019, doi: 10.1109/TSG.2018.2818167.