

AN INTEGRATED DATA ENGINEERING AND MACHINE LEARNING FRAMEWORK FOR ENERGY PREDICTION AND FAULT DETECTION IN SMART BUILDINGS

Prof. Shilpa M S

Assistant Professor

Dept. of AIML, CBIT

Kolar

Shilpams05@gmail.com

Chethan G C

Dept. of AIML, CBIT

Kolar

chethangcchethan2004@gmail.com

Janardhan B G

Dept. of AIML, CBIT

Kolar

ganeshprabhavathi9945@gmail.com

Manu S

Dept. of AIML, CBIT

Kolar

Manusm151@gmail.com

Varun Kumar

Dept. of AIML, CBIT

Kolar

varunkumarvarunkumar204@gmail.com

Abstract: With the increasing demand for energy efficiency, this work proposes an advanced data-driven framework for energy prediction and fault detection in smart buildings. The system integrates multi-source data, including energy consumption, weather, and occupancy, to create a unified analytical pipeline. Advanced preprocessing techniques such as Z-score, IQR, STL, Rolling Z-score, and DBSCAN are applied for outlier detection, while linear and STL-based imputation ensure data continuity and reliability. For prediction, models including XGBoost, Random Forest, LSTM, Prophet, and MSTL are evaluated, where XGBoost achieves the highest accuracy and LSTM effectively captures temporal patterns. Fault detection is performed using Bollinger Bands, which identify anomalies based on deviations from expected energy trends. The proposed framework provides a robust, scalable, and efficient solution for intelligent energy management and proactive fault detection in modern smart building systems.

Keywords: Smart Buildings; Energy Prediction; Fault Detection; Data Engineering; XGBoost; Anomaly Detection

I. INTRODUCTION

The rapid growth in global energy consumption has made efficient energy management in smart buildings a critical challenge. Modern buildings are equipped with IoT sensors and monitoring systems that generate high-frequency time-series data, including energy usage, weather conditions, and occupancy patterns. However, extracting meaningful insights from this heterogeneous and dynamic data remains complex. Energy consumption is influenced by multiple factors and exhibits non-linear, seasonal, and time-dependent behavior, making traditional rule-based or static monitoring approaches inadequate for accurate prediction and analysis.

A major challenge in this domain is the early detection of faults and anomalies in energy usage. These anomalies may arise due to equipment malfunction, sensor noise, or irregular operational behavior. Since energy data follows temporal patterns such as daily and seasonal cycles, abnormal variations are often difficult to distinguish from normal fluctuations. Conventional threshold-based methods fail to adapt to such dynamic patterns, leading to missed detections or false alarms. Failure to identify these issues in time can result in increased energy consumption, higher operational costs, and reduced system efficiency.

To address these challenges, this paper proposes an integrated data engineering and machine learning framework for energy prediction and fault detection. The system integrates multi-source data, including energy consumption, weather conditions, and occupancy information, into a unified analytical pipeline. A robust preprocessing stage is implemented using techniques such as Z-score, IQR, STL decomposition, Rolling Z-score, and DBSCAN. For predictive modeling, algorithms including XGBoost, Random Forest, LSTM, Prophet, and MSTL are employed to capture complex energy consumption patterns. Fault detection employs Bollinger Bands to identify deviations from expected energy trends, enabling effective anomaly detection without requiring labeled fault data.

II. LITERATURE SURVEY

Recent advancements in smart building energy management have leveraged data-driven technologies and machine learning algorithms to improve energy efficiency and fault detection. Various studies have explored different aspects such as data integration, prediction models, and anomaly detection techniques.

Billanes et al. [1] reviewed smart building technologies such as Artificial Intelligence (AI), Internet of Things (IoT), cloud computing, and big data analytics for energy optimization, though no specific machine learning algorithms or experimental validation were provided. Horvath et al. [2] proposed a cloud-based DevOps architecture for energy fault detection enabling real-time monitoring, but it focuses on infrastructure and lacks detailed ML models and predictive analysis. Muniandi et al. [3] discussed AI-based systems using predictive analytics and rule-based optimization for HVAC and smart lighting, with issues of privacy, interoperability, and scalability left unaddressed.

Wu et al. [4] compared ML algorithms such as XGBoost, SVM, and ANN for energy prediction, finding XGBoost achieved high-accuracy regression; however, missing key features such as weather data reduces prediction accuracy. Liu et al. [5] reviewed ML models including LSTM, Random Forest, and XGBoost for time-series forecasting, but the study remains theoretical with no real implementation or validation. Xiu et al. [6] proposed an unsupervised framework using concept drift detection

algorithms for streaming energy data, though with limited evaluation on diverse real-world datasets.

Zhelun et al. [7] analyzed supervised and unsupervised methods for fault detection, noting that the lack of labeled fault data restricts supervised approaches. Seem et al. [8] applied pattern recognition for energy usage baseline modeling, limited in scalability for complex multi-building environments. Nascimento et al. [9] applied Random Forest for prediction and Z-score methods for anomaly detection, relying on synthetic data rarely available in practice. Alimohammadi et al. [10] evaluated KNN, DBSCAN, Z-score, and IQR for outlier detection, mainly on synthetic datasets lacking real-world validation.

Despite extensive research, existing works lack a unified framework combining data engineering, ML-based prediction, and fault detection in a single deployable system. Most studies address prediction or anomaly detection independently, without considering real-time deployment.

III. PROBLEM STATEMENT

Smart buildings generate large volumes of sensor and system data, but much of it remains fragmented and underutilized. Existing systems lack an integrated data engineering framework to efficiently process, store, and analyze this data in real time. Accurate energy consumption prediction is challenging due to dynamic environmental conditions and varying occupancy patterns. Additionally, fault detection in building systems is often delayed, leading to energy inefficiencies and increased maintenance costs. Therefore, there is a need for a unified AI and machine learning-based framework that enables efficient data integration, precise energy prediction, and early fault detection in smart buildings.

IV. OBJECTIVES

The key objectives of this work are: (i) to design an integrated data engineering pipeline for smart building data; (ii) to predict hourly energy consumption using machine learning models; (iii) to detect energy anomalies and equipment faults proactively; and (iv) to deploy a cloud-based dashboard for real-time monitoring and alerts.

V. EXISTING SYSTEM

Existing systems for energy prediction and fault detection in smart buildings primarily rely on traditional monitoring and basic data analysis techniques. These systems typically collect energy consumption data from sensors and building management systems, but the data is often stored in isolated formats without proper integration with weather conditions and occupancy patterns. Most existing approaches use simple statistical methods or standalone machine learning models

that do not consider the time-series nature of energy data, making them unable to handle seasonal variations, trends, and dynamic changes in building operations.

Fault detection in these systems is usually reactive, meaning faults are identified only after significant deviations or failures occur, leading to increased energy costs and operational inefficiencies. Many systems also suffer from poor data quality issues such as missing values, noise, and outliers. Moreover, these systems lack scalability and automation, making it difficult to handle large volumes of data from multiple buildings in real time.

VI. PROPOSED SYSTEM

The proposed system is an integrated data engineering and machine learning framework designed to improve energy prediction and fault detection in smart buildings. It combines multiple data sources—energy consumption, weather data, and occupancy data—into a unified cloud-based system for efficient processing and analysis. The system uses automated data pipelines for continuous data collection, cleaning, and integration, ensuring high-quality data for modeling.

Advanced machine learning algorithms, especially XGBoost, are used to accurately predict future energy consumption, while anomaly detection techniques like Bollinger Bands are applied to identify faults or abnormal energy usage patterns. The framework supports both general and category-based models, making it adaptable to different building types. It is scalable, automated, and includes a human-in-the-loop approach for validating detected faults, ensuring reliability and real-time decision-making.

VII. METHODOLOGY

A. Smart Building Data Input & Collection

The system collects data from IoT sensors, smart meters, weather systems, and Building Management Systems (BMS). This data includes energy consumption, environmental conditions, and occupancy information in both real-time and historical form. All collected data is transmitted and stored in a centralized cloud-based platform to ensure scalability and accessibility.

B. Data Preprocessing

The collected data is preprocessed to improve quality and reliability. Noise and inconsistencies are removed, missing values are handled using interpolation techniques, and outliers are detected and treated. Data formatting and normalization are performed to bring uniformity across datasets using Z-score, IQR, STL decomposition, Rolling Z-score, and DBSCAN.

C. Feature Extraction

Important features are extracted from the processed data including energy-related attributes, temporal patterns such as hour, day, and seasonal variations, weather conditions, and occupancy indicators. Feature engineering identifies meaningful patterns and relationships within the data while reducing redundancy.

D. Machine Learning Model Development

Machine learning models including Random Forest, XGBoost, and LSTM are trained using the prepared dataset to learn relationships between input features and output values. Performance evaluation is carried out using metrics such as RMSE and prediction accuracy. XGBoost achieves superior forecasting accuracy while LSTM effectively captures temporal dependencies.

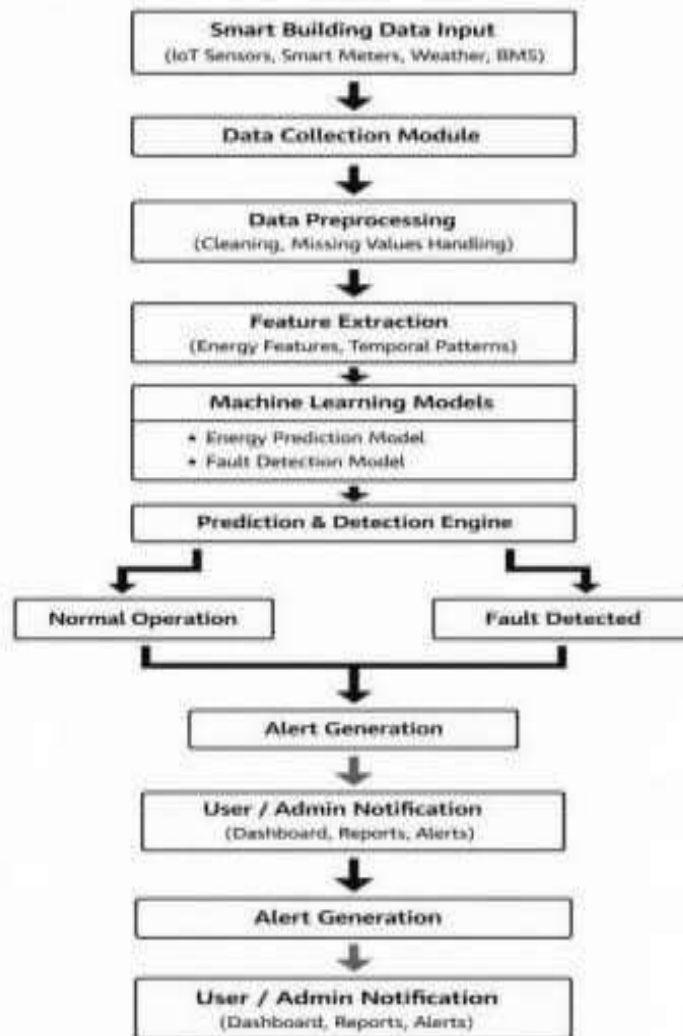
E. Prediction, Detection & System Output

The trained model generates predictions for energy consumption. Predicted values are compared with actual observed values to identify deviations. If deviation remains within acceptable limits, the system is considered to be operating normally. A significant deviation triggers fault or anomaly classification using Bollinger Bands.

F. Alert Generation & User Notification

When a fault is detected, the system generates alerts to notify users or administrators about abnormal conditions. Notifications are delivered through dashboards, reports, or alert messages, enabling quick awareness and corrective action. System events are logged for future analysis and model improvement.

FIGURE I: FLOW DIAGRAM



VIII. CONCLUSION

This paper presents an integrated data engineering and machine learning framework for energy prediction and fault detection in smart buildings. By combining multi-source data integration, advanced preprocessing, machine learning-based prediction with XGBoost and LSTM, and Bollinger Band-based anomaly detection, the proposed system addresses key limitations of existing approaches. The framework is scalable, cloud-enabled, and supports both real-time and batch processing, providing a reliable solution for proactive energy optimization and fault detection in modern smart building environments. Future work will focus on extending the framework to multi-building environments and incorporating federated learning for privacy-preserving data analysis.

REFERENCES

- [1] M. Billanes et al., "Data-Driven Technologies for Energy Optimization in Smart Buildings: A Scoping Review," *IEEE Access*, vol. 11, pp. 102345-102360, 2024.
- [2] A. Horvath et al., "Cloud-Based DevOps Framework for Energy Fault Detection in Smart Buildings," *IEEE Trans. Sustain. Comput.*, vol. 9, no. 1, pp. 55-67, 2024.
- [3] K. Muniandi et al., "AI-Driven Energy Management Systems for Smart Buildings: A Review," *Renew. Sustain. Energy Rev.*, vol. 182, p. 113456, 2024.
- [4] Y. Wu et al., "Comparative Analysis of Machine Learning Models for Building Energy Consumption Prediction," *Energy Build.*, vol. 290, p. 113210, 2024.
- [5] H. Liu et al., "Building Energy Consumption Prediction Using Machine Learning: A Comprehensive Review," *IEEE Access*, vol. 12, pp. 44567-44589, 2024.
- [6] Z. Xiu et al., "Concept Drift-Aware Anomaly Detection Framework for Energy Data Streams," *IEEE Trans. Big Data*, vol. 10, no. 2, pp. 789-802, 2024.
- [7] L. Zhelun et al., "Data-Driven Fault Detection and Diagnostics in Building Energy Systems: A Review," *Appl. Energy*, vol. 335, p. 120678, 2024.
- [8] J. Seem et al., "Pattern Recognition-Based Fault Detection in Building Energy Systems," *ASHRAE Trans.*, vol. 130, no. 1, pp. 210-220, 2024.
- [9] D. Nascimento et al., "Random Forest-Based Energy Prediction and Anomaly Detection Using Statistical Residual Analysis," *Energy Rep.*, vol. 10, pp. 456-468, 2024.
- [10] M. Alimohammadi et al., "Comparative Evaluation of Outlier Detection Techniques for Energy Data," *Expert Syst. Appl.*, vol. 235, p. 121234, 2024.