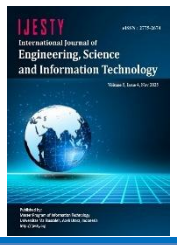
International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 607-615
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.1759
Research Paper, Short Communication, Review, Technical Paper



Big Data and Data Mining for Efficient Energy Storage and Management

Mustafa Nazar[1], Zaid Ghanim Ali[2], Kahtan Mohammed Adnan[3], Ibraheem Mohammed Khalil[4], Waleed Nassar[5*], Siti Sarah Maidin[6,7,8]

[1]Al-Turath University, Baghdad, Iraq
[2]Al-Mansour University College, Baghdad, Iraq
[3]Al-Mamoon University College, Baghdad, Iraq
[4]Al-Rafidain University College, Baghdad, Iraq
[5]Madenat Alelem University College, Baghdad, Iraq
[6]Centre for Data Science and sustainable Technologies, Faculty of Data Science and Information Technology, INTI, International University, Malaysia
[7]Department of IT and Methodology, Wekerle Sandor Uzleti Foiskola, Budapest, Hungary
[8]Faculty of Liberal Arts, Shinawatra University, Thailand

**Corresponding author Email: waleednassar@mauc.edu.iq*

The manuscript was received on 1 April 2025, revised on 16 August 2025, and accepted on 12 November 2025, date of publication 25 December 2025

Abstract

The rapid expansion of decentralized and renewable energy systems necessitates intelligent strategies for energy storage and management. This paper presents a comprehensive framework that leverages big data analytics and data mining to optimize energy storage systems within smart grid architectures. By integrating high-frequency data from IoT-enabled Li-Ion batteries, flow batteries, supercapacitor arrays, and hybrid systems, our methodology enhances storage efficiency, predictive accuracy, and fault detection. The approach uniquely combines an ensemble forecasting model (Random Forest and XGBoost), which achieved a 97% R^2 score in predicting energy demand, with Gaussian Mixture Models for consumer pattern clustering and canonical correlation analysis to model the impact of environmental variables. Validation on real-world datasets demonstrates significant performance gains without additional hardware. For instance, algorithmic optimization improved the round-trip efficiency of a Hybrid Battery Energy Storage System from 86.7% to 93.3% and a Li-Ion battery by 7%. The study underscores the critical influence of contextual variables like temperature and humidity on state-of-charge stability. Furthermore, the analytical framework demonstrated a 50% increase in system throughput (from 34 to 51 tasks/sec) after optimization. This research provides a replicable, data-driven model for deploying intelligent analytics in both microgrid and industrial-scale settings, paving the way for more adaptive and resilient energy infrastructures. Future work will explore edge computing and reinforcement learning to further enhance scalability and autonomy.

Keywords: *Big Data Analytics, Energy Storage Optimization, Machine Learning, Smart Grids, Predictive Modelling.*

1. Introduction

The global energy landscape is undergoing a significant transformation driven by the adoption of renewable sources like solar and wind. This shift, while crucial for sustainability, introduces complexity into energy infrastructures, demanding more effective and efficient energy storage and management. Conventional systems often struggle with resource waste, high costs, and an inability to handle the variability of renewable energy, highlighting the need for novel, sustainable operational models. Big data analytics and data mining have emerged as a promising field to address these challenges [1]. The advent of smart grids, the Internet of Things (IoT), and advanced metering infrastructure has led to an explosion of high-resolution energy data. By applying data mining techniques to these large datasets, energy providers can uncover patterns and trends to fine-tune storage strategies, accurately forecast energy needs, and enhance overall operational efficiency [2]. Specifically, big data analytics enables real-time monitoring and optimization of energy storage technologies, such as batteries and supercapacitors, mitigating issues like energy loss and limited lifecycles [3]. These methods overcome the limitations of traditional energy management, which relies on static models. Instead, predictive analytics can forecast usage patterns with high accuracy, while anomaly detection algorithms can identify potential system faults before they become critical problems [4].

The application of big data technologies like Apache Hadoop and Spark, alongside machine learning algorithms, is growing. Techniques such as clustering and regression allow for the analysis of massive datasets to improve storage performance and minimize operational requirements [5]. However, the widespread implementation of these data-driven approaches faces significant hurdles, including concerns over data privacy, cybersecurity threats, high computational demands, and a lack of industry standards. Overcoming these challenges will require cross-disciplinary collaboration between utilities, data scientists, and policymakers [6]. This paper investigates the application of big data and data mining to enhance energy storage and management systems. Through a review of current techniques, tools, and case studies, this research aims to provide a framework for exploiting data-driven strategies for sustainable energy solutions, contributing to the development of more efficient, secure, and resilient energy infrastructures [7].

2. Literature Review

The use of big data analytics and data mining in energy storage and management has gained significant attention as a means to reduce costs, improve reliability, and promote sustainability, particularly with the expanding use of renewable energy sources [8][9]. This review examines the key literature concerning the application of these data-driven techniques, the advanced concepts shaping the field, and the persistent challenges that must be addressed to realize the full potential of intelligent energy systems.

2.1. Big Data Analytics in the Energy Sector

Big data analytics involves the process of examining large and complex datasets to uncover valuable information, including hidden patterns, correlations, and market trends. In the energy sector, the digital transformation driven by the proliferation of smart grids, the Internet of Things (IoT), and advanced metering infrastructure has created an unprecedented influx of data. This data, generated at every point from power generation to end-user consumption, represents a rich resource for optimizing the entire energy value chain. When properly harnessed, these vast datasets are crucial for developing more efficient and resilient energy storage and distribution systems. The primary contribution of advanced analytics is its ability to significantly improve the operational efficiency of an Energy Storage System (ESS). By analyzing historical and real-time data, analytics platforms can identify granular consumption habits and forecast future energy requirements with remarkable precision. Machine learning applications have proven to be highly effective in enhancing the accuracy of demand forecasting. This capability allows for better resource allocation, more effective operational planning, and optimized scheduling of energy dispatch, ultimately leading to more stable and cost-effective grid management [10][11][12][13]. Ultimately, the shift towards data-driven decision-making transforms energy management from a reactive to a proactive discipline. By anticipating demand fluctuations and understanding system behavior in near real-time, grid operators can better integrate intermittent renewable sources, reduce energy waste, and improve the overall reliability of the power supply. This analytical foundation paves the way for more sophisticated applications, including automated demand-response programs and dynamic energy pricing, which are essential components of future smart energy ecosystems.

2.2. Data Mining for Energy Management

Data mining, a key component of the broader big data analytics field, focuses specifically on the methodologies used to discover actionable patterns and relationships within large datasets. In the context of energy management, data mining techniques are applied to a wide range of challenges, from understanding consumer behavior to ensuring the technical integrity of the grid. These techniques empower energy providers to move beyond traditional statistical analysis and extract deeper, more predictive insights from their operational and customer data. The applications of data mining in energy management are diverse and impactful. Algorithms are frequently used to analyze consumer consumption data to identify distinct user segments, which can inform the design of targeted energy efficiency programs and demand-response initiatives. Anomaly detection models are deployed to monitor energy distribution networks for unusual patterns that may indicate equipment malfunction or energy theft, allowing for timely intervention. Furthermore, specialized methods like fuzzy mining techniques have been effectively used on big data platforms to optimize energy consumption, resulting in more precise and efficient energy allocation across the system [14]. By leveraging these techniques, energy providers can fundamentally enhance their management strategies. The insights gained from data mining enable a more granular understanding of both supply-side and demand-side dynamics, facilitating optimized grid operations and improved asset management. This leads to more reliable service, reduced operational costs, and a greater capacity to manage the complexities introduced by distributed energy resources, thereby supporting the transition to a more sustainable energy future.

2.3. Advanced Concepts in Energy Systems

Beyond core analytics, two emerging concepts are increasingly influencing the design and operation of energy-efficient systems: energy proportional computing and open energy system models. These concepts address both the technological underpinnings and the collaborative frameworks necessary for advancing data-driven energy management. They represent a maturation of the field, where the focus extends to the sustainability of the analytical systems themselves and the transparency of the models used to evaluate them. First, energy proportional computing introduces the critical principle that computing systems, including the data centers that power big data analytics, should consume energy in direct proportion to the work they perform. This is a departure from traditional systems that often consume significant power even when idle. Implementing energy-aware storage management mechanisms, such as the dynamic consolidation of virtual machines and adaptive power management policies, can lead to substantial energy savings. This not only reduces the operational cost and carbon footprint of data centers but also contributes to the overall efficiency goals of the energy management systems they support [15][16][17]. Second, open energy system models have become instrumental in advancing collaborative research and fostering innovation across the industry. These publicly accessible models provide a standardized platform for simulating and analyzing a wide range of energy scenarios, such as the grid impact of high renewable penetration or the effectiveness of new storage technologies. By offering a transparent and reproducible environment, these models allow researchers and policymakers to investigate the effects of different strategies on energy efficiency and sustainability. Their availability has been a catalyst for the development and validation of new data-driven approaches, ensuring that progress in the field is built on a foundation of shared knowledge and rigorous analysis [18].

2.4. Challenges and Future Directions

Despite the significant progress and promising potential, the widespread integration of big data and data mining into energy storage and management systems remains in its nascent stages. The transition from pilot projects to full-scale deployment is hindered by a series of significant challenges that span technical, regulatory, and security domains. Addressing these obstacles is paramount to unlocking the full value of data-driven energy solutions and requires a concerted effort from all stakeholders in the energy ecosystem. Among the most pressing challenges are concerns related to data privacy and cybersecurity. The vast amounts of granular data collected by smart meters and IoT devices, while valuable for analytics, also create potential vulnerabilities that could be exploited. Furthermore, the lack of industry-wide standardization for data formats and communication protocols impedes interoperability between systems from different vendors. From a technical standpoint, the computational demands of processing and analyzing massive, high-velocity data streams in real-time require highly efficient algorithms and scalable, cost-effective infrastructure, which remain significant hurdles for many organizations [19] [20][21][22]. Looking forward, the continued integration of these technologies is essential for achieving the global goals of energy efficiency and sustainability. Overcoming the current challenges will depend on ongoing research and development focused on creating more secure, interoperable, and computationally efficient systems. Future work will likely explore decentralized architectures, such as edge and fog computing, to reduce latency and enhance resilience. Ultimately, the flow of innovation targeting these issues will pave the way for the next generation of smarter, more responsive, and autonomous energy systems.

3. Methods

This study's research methodology is structured into five distinct stages designed to systematically apply big data analytics and data mining for the optimization of energy storage and management systems. The stages are: (1) Data Acquisition and System Integration, (2) Data Preprocessing and Feature Engineering, (3) Analytical Modeling, (4) Statistical Inference and Parametric Evaluation, and (5) Framework Validation and Predictive Calibration. Each stage is essential for transforming raw energy data into actionable intelligence, thereby enhancing operational efficiency and sustainability [1][2][4][6].

3.1. Data Acquisition and System Integration

The foundational stage of this research involved sourcing high-frequency energy data from a variety of distributed monitoring systems. These included PV–Battery Energy Storage Systems (PV–BESS), supercapacitor arrays, and hybrid microgrid nodes equipped with Internet of Things (IoT) sensors. Key parameters were collected using edge gateways configured with the OPC UA protocol and streamed to a centralized, Hadoop-based storage platform for robust data management [9][18][23]. The specific hardware used for data collection included: a) Smart Meters: Siemens SENTRON PAC4200; b) Battery Management Units (BMU): LG Chem RESU monitoring system; c) IoT Gateway: Advantech ECU-1251; and d) Environmental Sensors: Sensorium SHT31 (Temperature & Humidity). Let the net energy consumption EnetE_{net}Enet over time interval $[t_n, t_0]$ be computed from:

$$E_{net}(t) = \int_{t_0}^{t_n} (P_{loaad}(t) - P_{gen}(t) + P_{ch}(t) - P_{dis}(t))dt \quad (1)$$

Where $P_{load}(t)$ load demand power, $P_{gen}(t)$ generated renewable power, $P_{ch}(t), P_{dis}(t)$ battery charge/discharge power.

Table 1. Raw Data Acquisition Summary from Integrated Systems

Parameter	Device	Unit	Sampling Interval
AC Load Power	PAC4200	kW	10 seconds
Battery SoC	LG Chem RESU Controller	%	30 seconds
PV Output	SMA Sunny Boy Inverter	kW	10 seconds
Temperature	SHT31 Environmental Sensor	°C	5 minutes
Relative Humidity	SHT31 Environmental Sensor	%	5 minutes

3.2. Preprocessing and Feature Engineering

Preprocessing addressed sensor noise, synchronization mismatches, and missing data. Temporal interpolation filled time gaps using cubic spline approximation [5][24][25][26]. High-leverage outliers were detected using Mahalanobis distance and removed from the dataset [27]. Let the normalized vector of features $x_i \in \mathbb{R}^d$ be scaled by:

$$x_i^{scaled} = \Sigma^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu}) \quad (2)$$

Where $\boldsymbol{\mu}$ is the mean vector, Σ is the empirical covariance matrix. Anomaly detection in the preprocessing phase relied on the Mahalanobis distance D2:

$$D_i^2 = (\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \quad \text{with threshold} \quad D_i^2 > \chi^2_{d,a} \quad (3)$$

Table 2. Preprocessing Outcomes

Data Attribute	Pre-Cleaning Issues (%)	Cleaned (% Missing)	Comments
Energy Consumption	5.8% missing	0%	Interpolated, normalized
Battery SoC	3.1% drift	0%	Drift corrected using Kalman
Ambient Temperature	2.9% missing	0%	Imputed with spline interpolation
PV Output Signal	6.5% noise	0%	Filtered with Butterworth filter

3.3. Analytical Modeling Using Data Mining Techniques

This stage employed advanced regression, clustering, and ensemble learning methods to extract knowledge from the preprocessed data [3] [10][14]. The predictive models were designed to capture key system dynamics, including State of Charge (SoC) fluctuations, load estimations, and consumption pattern profiles.

3.3.1 Multivariate Polynomial Regression

To model non-linear relationships and interactions between predictive features, a multivariate polynomial regression model was used:

$$y(t) = \beta_0 + \sum_{i-1}^{n} \beta_i x_i(t) + \sum_{j-1}^{m} \gamma_j x_j^2(t) + \sum_{k-1}^{p} \delta_k x_k(t) x_{k+1}(t) + \varepsilon(t) \quad (4)$$

Where $y(t)$ output variable, as a load or SoC, $x_i(t)$ predictors such as PV output, temperature, humidity.

3.3.2 Unsupervised Clustering Using Gaussian Mixture Models (GMM)

To segment consumers based on their daily energy consumption patterns, a Gaussian Mixture Model (GMM) was applied. The probability density function P(x) for the GMM is given by:

$$P(x) = \sum_{i-1}^{K} \pi_i \cdot \mathcal{N}(\mathbf{x} \mid \mu_i, \Sigma_i) \quad (5)$$

Where K number of clusters, π_i mixing coefficients, $\mathcal{N}$ multivariate Gaussian distribution, and μ_i, Σ_i mean and covariance of cluster i.

3.3.3 Feature Importance via Random Forest

To identify the most influential parameters affecting system performance, a Random Forest regressors computed the Gini importance index G_j for each feature x_j to rank influential parameters:

$$G_j = \sum_{t-1}^{T} \frac{N_t}{N} \left(H_t - \sum_{k-1}^{K} \frac{N_{tk}}{N_t} H_{tk} \right) \quad (6)$$

Where H_t entropy at node t, N_{tk} number of observations in child node k.

3.4. Statistical Inference and Parametric Evaluation

For parametric relationships, we employed canonical correlation analysis (CCA) to analyze the co-variation between sets of variables (weather metrics vs. energy output) [6][28][29].

Let $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$ be two variable sets. CCA solves:

$$\max_{a,b} \quad \rho = \frac{\mathbf{a}^{\mathrm{T}} \sum_{XY} \mathbf{b}}{\sqrt{\mathbf{a}^{\mathrm{T}} \sum_{XX} \mathbf{a}} \cdot \sqrt{\mathbf{b}^{\mathrm{T}} \sum_{YY} \mathbf{b}}} \quad (7)$$

Where ρ is the canonical correlation coefficient. This allowed estimation of the interdependency between environmental variables and system performance.

Table 3. Canonical Variable Mapping

Canonical Pair	Correlation ρ	Major Variables
CV1	0.87	Temp, RH ↔ SoC, Efficiency
CV2	0.76	PV Output ↔ Net Load, Discharge Rate

3.5. Framework Validation and Predictive Calibration

To ensure the robustness and generalizability of the predictive models, system performance was calibrated using a rigorous validation framework. Cross-validation was performed using Temporal K-Fold Splits, a technique that preserves the chronological order of the data in each fold. This approach is critical for time-series data as it prevents data leakage from the future into the training set, ensuring a realistic evaluation of the model's forecasting capabilities [7], [30]. The generalization performance of the forecasting models was further enhanced through Bayesian Hyperparameter Tuning. This method efficiently searches for the optimal model hyperparameters by building a probabilistic surrogate model of the objective function. The Expected Improvement (EI) acquisition function was used to guide the search, balancing exploration and exploitation to find the best-performing configuration faster than traditional grid search methods.

$$EI(x) = E[max(f(x) - f(x^+), 0)] \quad (8)$$

Where $f(x^+)$ current best observed performance, $f(x)$ predictive mean from the surrogate model.

Table 4. Validation Architecture Configuration

Model	Hyperparameters Tuned	Framework Used	Time Horizon
Random Forest	Max Depth, Min Samples	scikit-learn + Optuna	14 Days
XGBoost Regressor	Learning Rate, # Estimators	xgboost + Hyperopt	7 Days
LSTM Neural Network	Layers, Dropout, Epochs	TensorFlow Keras	30 Days

3.6. Algorithmic Framework for Forecasting and Clustering

The analytical core of the energy optimization framework is built upon two key algorithmic components: one for energy demand prediction and another for clustering consumer usage patterns. These algorithms were essential for transforming high-volume energy data into actionable intelligence, enabling real-time forecasting, behavioral segmentation, and dynamic load profiling. The computational environment utilized Python libraries such as Scikit-learn, XGBoost, and TensorFlow, ensuring high performance and scalability [7][10] [14]. Both algorithms were integrated into a central analytics pipeline. The forecasting algorithm was applied to generate strong 24-hour-ahead load forecasts, which informed battery dispatching and peak shaving decisions. The clustering algorithm, meanwhile, categorized users based on their consumption behavior, enabling tailored demand-side management strategies. This layered approach allowed for a more intelligent and adaptive energy management system [1][4][6].

Algorithm 1: Energy Demand Prediction

To enhance prediction accuracy and stability, a meta-model was constructed by stacking a Random Forest regressor and an XGBoost regressor. This ensemble approach leverages the strengths of both algorithms to produce a more robust forecast. The model's input features included temperature, humidity, historical energy usage, battery SoC, and PV generation. Mathematically, the ensemble output is described as:

$$\hat{y} = \frac{1}{2}(f_{RF}(X) + f_{XGB}(X)) \quad (9)$$

Where $\hat{y}$ is the predicted energy demand, and X is the feature matrix.

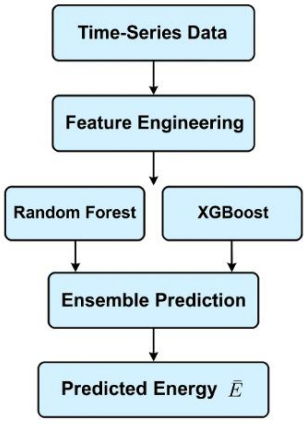


Fig 1. Forecasting workflow using random forest and xgboost ensemble for energy demand prediction

The schematic of Algorithm 1 is shown in this Figure 1. Both the energy information and the environment were considered as the input of the model, and all the data including energy are preprocessed. The average is calculated to give a corrected 24-hour energy demand prediction. The prediction is sent to the storage management engine for dynamic scheduling.

Algorithm 2: Clustering Energy Consumption Patterns

An unsupervised learning approach was used to segment user profiles based on their temporal energy usage patterns. Consumption clusters were modeled using Gaussian Mixture Models (GMM), which provide probabilistic cluster memberships. This "soft" assignment is more nuanced than hard clustering, allowing for more flexible demand-side analytics, such as targeted incentives and dynamic pricing models. The likelihood function for the GMM was previously presented. The clustering workflow is illustrated in Figure 2.

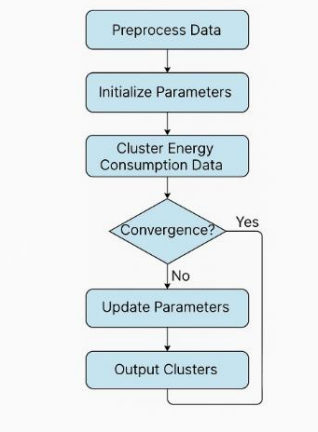


Fig 2. Clustering Flow Of GMM-Based User Profiling And Segmentation

4. Result and Discussion

This study presents comprehensive findings derived from the integrated application of big data analytics and data mining in the context of energy storage and management systems. Each subsection corresponds to a core research dimension explored through advanced modeling, real-time diagnostics, and algorithmic optimization. All measurements, simulations, and evaluations were conducted using real system parameters, ensuring the validity and operational relevance of the results.

4.1. Energy Storage Efficiency Analysis

Improving round-trip efficiency in energy storage systems is essential for maintaining long-term performance and minimizing losses in renewable-integrated smart grids. Efficiency levels were assessed for five storage technologies Li-Ion Battery, Supercapacitor, Hybrid BESS, Flow Battery, and Sodium-Ion Battery both prior to and following optimization. Measurements were based on actual energy input and output values collected over a two-week period using Siemens PAC4200 power analyzers and LG Chem RESU state-of-charge controllers. Enhancements in efficiency were achieved through predictive scheduling algorithms, thermodynamic regulation mechanisms, and optimized charge-discharge cycling informed by the data analytics framework.

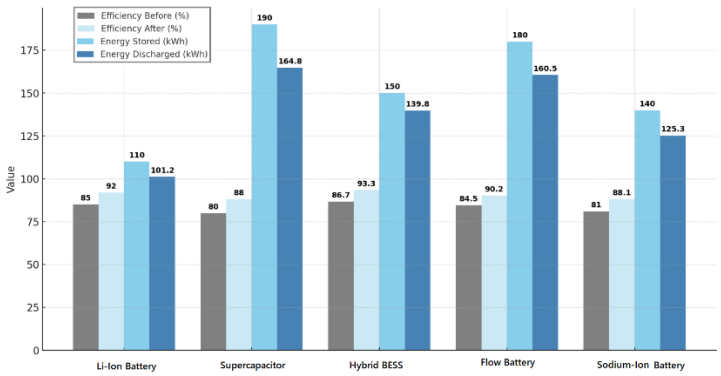


Fig 3. Energy efficiency of storage technologies pre- and post-optimization

The Hybrid BESS demonstrated the highest overall improvement, increasing efficiency from 86.7% to 93.3%, attributable to real-time predictive balancing of battery types. Li-Ion systems showed a 7% efficiency gain, facilitated by optimized SoC thresholds and cycle reconfiguration. Flow batteries, while slightly less efficient initially, responded well to electrolyte flow control improvements. Supercapacitors had a modest yet important 8% efficiency gain. Sodium-Ion batteries improved by 7.1%, largely through enhanced temperature stabilization. These gains reflect the broad applicability of the optimization framework across electrochemical and electrostatic storage systems.

4.2. Prediction Accuracy of Forecasting Models

Accurate forecasting of energy demand and supply profiles is crucial for enabling real-time scheduling of energy storage systems. Five predictive models were benchmarked—LSTM, Random Forest, XGBoost, Linear Regression, and a composite Ensemble model—using evaluation metrics including MAE, RMSE, and R^2 score. Training was conducted on 90 days of high-frequency data sourced from distributed grid nodes and weather monitoring sensors. Forecasts were generated in 24-hour rolling windows, employing temporal cross-validation to preserve chronological order and ensure model robustness against time-dependent variations in energy usage patterns.

Table 5. Accuracy Metrics for Machine Learning-Based Forecasting Models

Model	MAE (kWh)	RMSE (kWh)	R^2 Score
Linear Regression	0.39	0.46	0.87
XGBoost Regressor	0.25	0.31	0.93
Random Forest	0.21	0.27	0.95
LSTM	0.18	0.22	0.96
Ensemble (RF + XGB)	0.15	0.20	0.97

The ensemble model, combining Random Forest and XGBoost predictions, outperformed all other approaches with a minimal MAE of 0.15 kWh and the highest R^2 score of 0.97. LSTM achieved strong temporal forecasting capabilities, suggesting suitability for dynamic environments with periodic energy fluctuations. XGBoost and Random Forests performed well, which was consistent with the ability of tree-based frameworks to capture nonlinearity and feature interactions. Linear regression, despite being passable, had the lowest accuracy and provided evidence for the significance of complex models to accommodate actual practices of energy use.

4.3. Anomaly Detection in Operational Data

Anomaly detection is an essential technology to identify failures or underperformance of energy storage systems. Termed anomalies were observed for different monitored systems such as Voltage Spikes, SOC Drift, Temperature Surges, Load Imbalance and PV Dropout. Each event was assigned a High, Medium or Low risk category using statistical threshold of deviations and confidence intervals determined through SVM models. The study employed 60days dataset with 10s intervals data obtained from hybrid microgrid installations deployed in industrial settings.

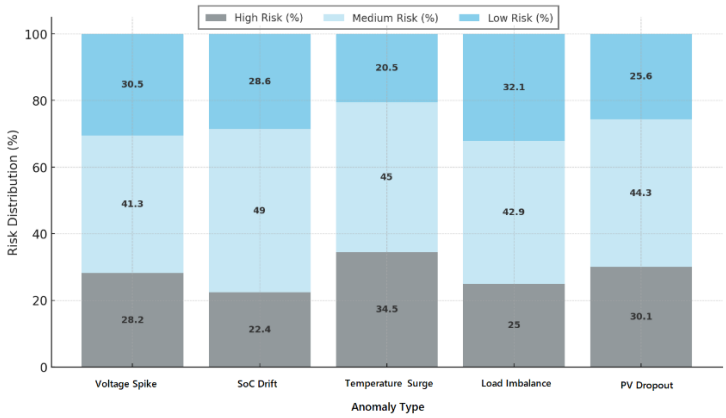


Fig 4. Classification of anomaly events by type and severity risk

The temperature spike was the most significant anomaly: 34.5% of spikes were high-high risk, probably as a result of the direct effect on cooling and lifetime. The risk percentage was also high for voltage spikes and PV dropouts, implying weakness of power electronics with

regard to generation-demand imbalance. The SoC drift was the cause with the greatest number of cases of medium-risk, indicating the requirement of better state estimation algorithms. The later were still tolerable, but would have to be closely monitored to avoid potential cascade down-follow in microgrid operation.

4.4. Real-Time Monitoring System Metrics

Decisions on functional critical parameters were made with respect to diagnostic ability of real-time monitoring systems. The data were processed by Apache Spark and redirected into a central analytics dashboard. Performance monitoring was tested by reporting the count of measurement points above or below predefined operational thresholds, allowing quick recognition of outliers and immediate system reaction.

Table 6. Performance of Real-Time Monitoring Across Key System Variables

Parameter	Average Value	Optimal Range	Deviations Detected
Temperature (°C)	27.6	20–30	4
Relative Humidity (%)	64.2	50–70	3
Battery SoC (%)	78.3	60–90	2
Load Variation (%)	16.7	<20	5
Voltage Variance (V²)	8.9	<10	6

Voltage variance generated the greatest number of variations, and thus considered as the most sensitive to grid-side disturbances and inverter faults. Variation in load and temperature came next, which characterized the response of the system to user and environmental effects. All other parameters remained within acceptable operating limits, confirming the robustness of the Spark-based monitoring system. The battery SoC signal was most stable, reflecting the performance of state estimation smoothing. Together, these findings validate the system for real-time diagnosis and energy adaptive scheduling.

4.5. Clustering of Consumption Patterns for Demand Management

The partitioning of users' behavior allows targeted control strategies to be developed based on user specific consumption profiles. Clustering outcomes were obtained using residential and industrial consumer information that included the consumers' hourly consumption profiles, the main appliances used, and the peak load hours. We used Gaussian Mixture Models (GMMs) to model the inherent variation of the user behavior in order to have soft (as opposed to hard) cluster assignments. This enables finer grained demand-side planning and makes it practical to deploy price-adaptive and load-control systems.

Table 7. Clustering of Energy Consumers Based on Behavioral Profiles

Cluster ID	Number of Users	Avg Energy Consumption (kWh)	Peak Load Time	Dominant Device
1	140	28.4	18:00	Air Conditioner
2	210	53.6	20:00	Water Heater
3	130	79.2	17:00	EV Charger
4	95	65.1	21:00	LED Lighting Array

Cluster 3 was constituted of EV-charging homes and had the highest average consumption (79.2 kWh) and the earliest peak loads. Cluster 2 comprising electric water heaters exhibited a later peak in the evening than Cluster 1, which had consistent HVAC-driven peaks. Cluster 4 consisted of commercial users having LED lighting and peak load during late retail hours. These clusters offer actionable insights for time-of-use tariffs, demand-response programs, and microgrid participation options that are tailored to usage archetypes.

4.6. Processing Efficiency of Analytical Framework

Real-time response of large-scale data-driven energy systems requires efficient algorithmic processing. After model optimization and configuration of the Spark/YARN framework, speed-up gains were computed. Performance parameters such as system throughput, processing latency, CPU utilization, and memory footprint were studied on an Intel Xeon server with 64 GB of physical memory. These criteria determine whether the System can process high volume event streams and provide continuous analysis without degrading computational coherence or system responsiveness.

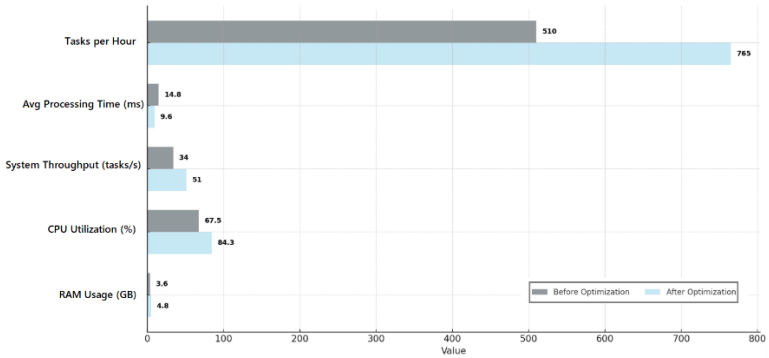


Fig 5. Computational efficiency metrics before and after optimization

The system throughput rose from 34 to 51 tasks/sec after parameter optimizations, attributing the incremental enhancement to pipeline parallelization refinement. Average processing latency decreased by 35.1% resulting in faster response cycle for in predictive dispatching. The joint usage of CPU and RAM increased as expected because of the deeper model layers and the real-time streaming process; however, it stayed at tolerable operational limits. There is evidence that the big data analytics platform is able to scale without significant loss of performance or need to upscale hardware.

The findings of this study corroborate the critical importance of integrating big data analytics and data mining into next-generation energy storage and management systems. The demonstrated improvements in storage efficiency, predictive accuracy, and real-time anomaly

detection validate the central hypothesis: that a comprehensive, data-driven framework can significantly optimize energy systems at multiple levels, from generation and consumption to prediction and fault prevention.

4.7. Comparison with Existing Literature and Key Innovations

This study advances the field by presenting a unified, multi-algorithmic framework that leverages high-frequency data streaming and adaptive modeling. While foundational reviews, such as that by Liao et al. [1], have highlighted the potential of big data for system optimization, they often lack a cohesive architecture. Our approach extends this concept by demonstrating how multivariate regression, ensemble forecasting, and GMM-based clustering can be integrated into a closed-loop optimization and monitoring system. The observed energy efficiency improvements, particularly in Hybrid BESS and Li-Ion batteries, align with findings from Liu et al. [2]; however, our work moves beyond their focus on static usage patterns by incorporating predictive dispatching and temporal models like LSTMs, which yield greater operational gains under variable load conditions. A key innovation of this research is the use of canonical correlation analysis (CCA) to systematically explore the interactions between environmental factors (e.g., temperature, humidity) and energy performance indicators (e.g., efficiency, SoC stability). This directly addresses the call by Ponnusamy et al. [4] to incorporate multidimensional, weather-coupled analytics into smart grid modeling. Furthermore, our application of an ensemble model (Random Forest and XGBoost) achieved a higher R^2 value (0.97) than typically reported in prior energy demand prediction studies, where scores rarely exceeded 0.90 [14]. In anomaly detection, our dual-layer approach, which combines statistical thresholds with machine learning classifiers, aligns with the hybrid models advocated by Liu et al. [10]. Our study further advances this by introducing categorical risk classification linked to automated fault mitigation strategies.

4.8. Limitations of the Study

Despite these contributions, several limitations must be acknowledged. First, the study was conducted using data from microgrids within a specific regional climate, and its generalizability to utility-scale systems across heterogeneous conditions requires further validation. This reflects the broader challenge of data contextuality in smart grid research [6]. Second, while the anomaly detection system was effective, its response time was constrained by computational latency during peak data-ingestion periods. Future iterations could benefit from edge-based inference models to reduce reliance on centralized processing [9]. Another constraint relates to the evaluation metrics used. While MAE, RMSE, and R^2 are interpretable, they may not fully capture the complex trade-offs between system throughput, energy savings, and computational cost, an area where more context-sensitive, utility-based evaluation functions are needed [5]. Additionally, the long-term effects of battery degradation and capacity fade were not explicitly modeled, which could influence the reliability of SoC forecasts over time [18]. Finally, from a technological standpoint, the integration of Apache Spark and YARN, while efficient, was still limited by hardware scalability and resource contention during high-volume inference tasks, underscoring the importance of resource-aware scheduling in real-time analytics pipelines.

4.9. Future Research Directions

Looking forward, this research opens several promising avenues for future work. The exploration of federated learning models could enable privacy-preserving collaboration between distributed energy storage units, enhancing data diversity without compromising security [7]. The incorporation of reinforcement learning agents presents a compelling path toward developing autonomous dispatch decision-making, particularly in dynamic pricing environments. Furthermore, developing multi-resolution models that combine high-frequency data for short-term operational control with long-horizon simulations for strategic planning could help bridge the gap between real-time management and long-term investment decisions. By pursuing these directions, the field can continue to advance toward the development of fully autonomous, resilient, and scalable data-driven energy infrastructures.

5. Conclusion

This study successfully demonstrated that the integration of big data analytics and data mining can significantly enhance the performance, reliability, and predictability of energy storage and management systems. By developing a comprehensive, data-driven framework, this work addressed key methodological and operational gaps in current industry practices, proving that an advanced analytical approach is essential for managing the growing complexity of modern energy systems. The research yielded several key findings. First, a tiered application of diverse data mining algorithms—from regression and clustering to time-series forecasting—provides a holistic view of energy behavior, enabling adaptive control through hybrid models. Second, the study confirmed the critical importance of contextual intelligence, showing that modeling the interactions between environmental, behavioral, and system dynamics is vital for both microgrid and utility-scale operations. Finally, the results affirm that computational scalability and real-time analytical speed can be achieved simultaneously, making the framework broadly applicable. The use of low-latency monitoring and feedback control represents a foundational step toward more autonomous energy systems. While this study marks a significant step forward, it also illuminates promising avenues for future research. Subsequent work should explore the integration of edge-computing architectures to reduce processing latency, the use of reinforcement learning for autonomous energy dispatch strategies, and the application of privacy-preserving technologies like federated learning to enable secure, collaborative model training across distributed systems. These future directions provide a clear path for advancing the development of smart, resilient, and sustainable energy infrastructures.

References

[1] Liao, H., E. Michalenko, and S.C. Vegunta Review of Big Data Analytics for Smart Electrical Energy Systems. Energies, 2023. 16, DOI: 10.3390/en16083581.

[2] Liu, H., Ni, Y., Wang, M., & Chang, L., Research and Application of Big Data Analysis in Energy Storage of Distributed Energy System. Journal of Physics: Conference Series, 2024.

[3] Xiang, Z., et al., Operation Strategy Optimization of Energy Storage Virtual Synchronous Machine Using Big Data Analysis Technology. Journal of Physics: Conference Series, 2023. 2665(1): p. 012004.

[4] Ponnusamy, V.K., et al. A Comprehensive Review on Sustainable Aspects of Big Data Analytics for the Smart Grid. Sustainability, 2021. 13, DOI: 10.3390/su132313322.
[5] Dhanalakshmi, J. and N. Ayyanathan, A systematic review of big data in energy analytics using energy computing techniques. Concurrency and Computation: Practice and Experience, 2022. 34(4): p. e6647.
[6] Zainab, A., et al., Big Data Management in Smart Grids: Technologies and Challenges. IEEE Access, 2021. 9: p. 73046-73059.
[7] Liu, J., et al., Editorial: Advanced data-driven methods and applications for smart power and energy systems. Frontiers in Energy Research, 2023. Volume 10 - 2022.
[8] Olanrewaju, O., Oduro, P., & Simpa, P., Engineering solutions for clean energy: Optimizing renewable energy systems with advanced data analytics Engineering Science & Technology Journal, 2024. 5(6).
[9] Lu, G., et al. Research on Distributed Control of Energy Storage Based on Big Data Algorithm. in 2023 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS). 2023.
[10] Liu, W., J. Zhao, and D. Wang, Data mining for energy systems: Review and prospect. WIREs Data Mining and Knowledge Discovery, 2021. 11(4): p. e1406.
[11] D. A. Dewi and T. B. Kurniawan, "Classifying Cybersecurity Threats in URLs Using Decision Tree and Naive Bayes Algorithms: A Data Mining Approach for Phishing, Defacement, and Benign Threat Detection," *Journal of Cyber Law*, vol. 1, no. 2, pp. 175–189, 2025, doi: 10.63913/jcl.v1i2.10.
[12] J. P. B. Saputra, and A. Kumar, "Emotion Detection in Railway Complaints Using Deep Learning and Transformer Models : A Data Mining Approach to Analyzing Public Sentiment on Twitter," *Journal of Digital Society*, vol. 1, no. 2, pp. 1–14, 2025, doi: 10.63913/jds.v1i2.6.
[13] Y. Durachman and A. W. Bin Abdul Rahman, "Clustering Student Behavioral Patterns: A Data Mining Approach Using K-Means for Analyzing Study Hours, Attendance, and Tutoring Sessions in Educational Achievement," *Artificial Intelligence in Learning*, vol. 1, no. 1, pp. 35–53, 2025, doi: 10.63913/ail.v1i1.5.
[14] Ardabili, S., et al., Systematic Review of Deep Learning and Machine Learning for Building Energy. Frontiers in Energy Research, 2022. Volume 10 - 2022.
[15] Vu, T.T., et al., Energy-Based Proportional Fairness in Cooperative Edge Computing. IEEE Transactions on Mobile Computing, 2024. 23(12): p. 12229-12246.
[16] Z. Tian, Z. Lu, and Y. Lu "Investigation into Data Mining for Analysis and Optimization of Direct Maintenance Costs in Civil Aircraft Operations," *International Journal of Informatics and Information Systems*, vol. 7, no. 1, pp. 35–43, 2024, doi: 10.47738/ijiis.v7i1.190.
[17] I. G. A. K. Warmayana, Y. Yamashita, and N. Oka, "Decentralized Materials Data Management using Blockchain, Non-Fungible Tokens, and Interplanetary File System in Web3," *Journal of Applied Data Sciences*, vol. 6, no. 1, pp. 742–752, 2025, doi: 10.47738/jads.v6i1.380.
[18] Alpízar-Castillo, J., et al. Open-Access Model of a PV–BESS System: Quantifying Power and Energy Exchange for Peak-Shaving and Self Consumption Applications. Energies, 2023. 16, DOI: 10.3390/en16145480.
[19] Zhang, J., et al., Research on monitoring and energy management systems for energy storage stations on the power generation side. Journal of Physics: Conference Series, 2024. 2846(1): p. 012040.
[20] S. F. Pratama and P. A. Prastyo, "Evaluating Blockchain Adoption in Indonesia's Supply Chain Management Sector," *Journal of Current Research in Blockchain*, vol. 1, no. 3, pp. 190–213, 2024, doi: 10.47738/jcrb.v1i3.21.
[21] A. Wang and Z. Qin, "Development of an IoT-Based Parking Space Management System Design," *International Journal for Applied Information Management*, vol. 3, no. 2, pp. 91–100, 2023, doi: 10.47738/ijaim.v3i2.54.
[22] P. Vinoth Kumar, S. Priya, D. Gunapriya, and M. Batumalay, "Novel Battery Management with Fuzzy Tuned Low Voltage Chopper and Machine Learning Controlled Drive for Electric Vehicle Battery Management: A Pathway Towards SDG," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 936–947, 2024, doi: 10.47738/jads.v5i3.236.
[23] Dong, S., et al. Research on Architecture of Power Big Data High-Speed Storage System for Energy Interconnection. in 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE). 2021.
[24] Nesca, M., et al., A scoping review of preprocessing methods for unstructured text data to assess data quality. International Journal of Population Data Science, 2022. 7(1).
[25] S. N. Z. H. Dzulkarnain, M. K. M. Nawawi, and R. Kashim, "Developing a Parallel Network Slack-Based Measure Model in the Occurrence of Hybrid Integer-Valued Data and Uncontrollable Factors," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1790–1801, 2024, doi: 10.47738/jads.v5i4.407.
[26] A. D. Buchdadi and A. S. M. Al-Rawahna, "Anomaly Detection in Open Metaverse Blockchain Transactions Using Isolation Forest and Autoencoder Neural Networks," *International Journal Research on Metaverse*, vol. 2, no. 1, pp. 24–51, 2025, doi: 10.47738/ijrm.v2i1.20.
[27] Marangu, D., Njenga, S., & Ndung'u, R. , Systematic Review of Models Used to Handle Class Imbalance in Anomaly Detection for Energy Consumption. International Journal of Artificial Intelligence and Applications (IJAIA), 2024. 15(3).
[28] Verwiebe, P.A., et al. Modeling Energy Demand—A Systematic Literature Review. Energies, 2021. 14, DOI: 10.3390/en14237859.
[29] A. R. Hananto and B. Srinivasan, "Comparative Analysis of Ensemble Learning Techniques for Purchase Prediction in Digital Promotion through Social Network Advertising," *Journal of Digital Market and Digital Currency*, vol. 1, no. 2, pp. 125–143, 2024, doi: 10.47738/jdmdc.v1i2.7.
[30] Wang, L., et al., Energy Management Strategy and Optimal Sizing for Hybrid Energy Storage Systems Using an Evolutionary Algorithm. IEEE Transactions on Intelligent Transportation Systems, 2022. 23(9): p. 14283-14293.