Journal of Science and Technology – Smart Systems and Devices
Volume 36, Issue 1, January 2026, 080-088

A Novel Method Forecasting for Big Data Analytics in Microgrid Energy Management

Le Thi Minh Chau[1], Le Duc Tung[1], Nguyen Hong Duy An[2], Duong Minh Quan[2,*]

[1]Hanoi University of Science and Technology, Ha Noi, Vietnam
[2]The University of Danang – University of Science and Technology, Da Nang, Vietnam
[*]Corresponding author email: dmquan@dut.udn.vn

Abstract

With the rapid proliferation of renewable energy sources (RES) in modern power grids, the application of operational optimization strategies has become pivotal in maintaining system efficiency and stability. In particular, the hybrid deep learning model long short-term memory (LSTM) neural network combined with inductive conformal prediction (ICP) significantly enhances the accuracy of renewable energy forecasting and management. This paper presents the integration of big data analytics (BDA) with the LSTM+ICP framework for optimizing smart grid operations, particularly under real-time conditions. A suite of machine learning and deep learning models, especially the hybrid LSTM+ICP, is deployed to address critical challenges such as renewable output forecasting, load balancing, and fault detection. Owing to its capability to capture temporal dependencies and generate reliable prediction intervals, the LSTM+ICP model achieved a mean absolute percentage error (MAPE) of 2.91%, thus improving the reliability of renewable energy scheduling and enabling more efficient resource allocation. The implementation of real-time BDA in conjunction with LSTM+ICP reduced load variance by 44.3%, peak demand by 24.2%, and frequency deviations by 52.9%, thereby strengthening grid reliability and operational stability. In predictive maintenance, the LSTM+ICP model achieved a detection accuracy of 97.2% with an average lead time of 6.2 hours, enabling proactive interventions and minimizing fault risks. For system optimization, the application of reinforcement learning augmented by BDA led to a 33.9% reduction in power losses, a 22.4% increase in voltage stability, and a 29.1% decrease in reactive power, thereby enhancing operational efficiency. From both economic and environmental perspectives, the BDA-driven approach resulted in monthly cost savings of €36,150 and a 30.2% reduction in CO_2 emissions, demonstrating the efficacy and sustainability of the proposed methodology.

Keywords: Big data analytics, deep learning, fault detection, renewable energy forecasting, smart grid.

1. Introduction

The rapid proliferation of renewable energy sources (RES) within modern power systems introduces both significant opportunities and complex challenges in maintaining grid stability and operational efficiency. While solar and wind energy offer clean and sustainable alternatives to fossil fuels, their inherent intermittency and unpredictability pose major difficulties in power dispatch coordination and in maintaining voltage and frequency stability across the grid [1, 2]. As power systems shift from centralized to decentralized architectures, the demand for intelligent, adaptive real-time control strategies becomes increasingly critical.

Big data analytics (BDA) has demonstrated substantial potential in optimizing smart grid operations by enabling the collection and processing of massive volumes of data from distributed energy resources, smart meters, meteorological stations, and grid assets. Advances in BDA facilitate not only renewable energy forecasting but also fault detection, load balancing, and overall grid optimization [3, 4]. Deep learning models such as long short-term memory (LSTM) networks have been widely employed for RES forecasting. However, these models often struggle to provide reliable probabilistic forecasts when faced with high uncertainty and volatility intrinsic to renewable energy generation.

Research in [5] proposed a U-Shaped long short-term memory - Attention-Free transformer (U-LSTM-AFT) architecture that improves hourly solar irradiance forecasting while explicitly providing reliable prediction intervals, making it suitable for operational decision-making in microgrids. Mohammadi et al. introduced a multi-timescale fusion framework combining Many-to-many long short-term memory (MTM-LSTM) and multilayer perceptron (MLP) models, which effectively reduces systematic errors by separating and learning different frequency components in solar radiation time-series data [6]. In addition, Yang *et al*. presented a rigorous evaluation and interpretability framework for hybrid deep neural networks, emphasizing the necessity of reliability metrics and calibrated uncertainty in time-series forecasting applications with complex dynamics [7]. These studies consistently indicate that

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.187.ssad.2026.36.1.10
Received: Jul 07, 2025; Revised: Nov 21, 2025;
Accepted: Jan 14, 2026; Online: Jan 15, 2026

hybrid LSTM models combined with uncertainty quantification techniques are essential for achieving accurate and reliable operation in smart grid systems.

To address this limitation, the present study proposes the integration of LSTM with inductive conformal prediction (ICP), a robust uncertainty quantification technique, to enhance forecast accuracy and deliver meaningful prediction intervals. The integration of BDA and deep learning enables smart grid systems to adapt proactively to real-time changes in load profiles, generation patterns, and fault events [8, 9]. Previous research has utilized machine learning models such as Support Vector Machines (SVM), Random Forests (RF), and LSTM to improve RES forecasting. Nonetheless, these approaches lack robust uncertainty handling and often fall short of delivering the reliability required for real-world deployment [10, 11].

The objective of this paper is to develop a comprehensive methodology that combines LSTM, ICP, and reinforcement learning to optimize renewable energy management in smart grids. The main contributions include a hybrid forecasting framework leveraging both machine learning and deep learning, a high-accuracy predictive maintenance system, and a reinforcement learning-based grid optimization controller designed to reduce power losses and enhance voltage stability. Furthermore, this paper provides a thorough economic and environmental impact assessment of BDA deployment in smart grids, highlighting its potential for significant cost savings and environmental protection.

2. Methodology

This study adopts a structured data-driven approach to evaluate the impact of big data analytics (BDA) on renewable energy management and the optimization of smart grid operations. The research methodology encompasses data collection, renewable energy forecasting, real-time load control, predictive maintenance, and the assessment of economic and environmental impacts.

2.1. Data Collection and Preprocessing

The data used in this study were collected from field datasets of the Dong Nai photovoltaic power plant and the Con Co wind farm [12], combined with manufacturer operational reports, notably the FusionSolar report (04/2024) [13]. These sources were employed to construct a dataset for a hypothetical microgrid, including photovoltaic (PV) and wind power (WP) generation, aggregated load demand, and relevant operational parameters. Associated meteorological data-solar irradiance, wind speed, ambient temperature, and humidity-were time-synchronized with the power generation data to support time-series forecasting. The use of real operational datasets ensures that the data realistically represents near real-time operating conditions of renewable energy systems.

Due to differences in recording frequencies, all data were synchronized and resampled to a 15-minute resolution, consistent with operational cycles in smart grid load balancing. The preprocessing procedure included short-gap linear interpolation for missing values, statistical outlier removal, and min–max normalization of input variables. The processed data were then organized using a sliding-window approach, where each input sample consisted of 48 historical time steps (12 hours) for model training and evaluation. The conceptual structure of the proposed LSTM+ICP model is shown in Fig. 1.

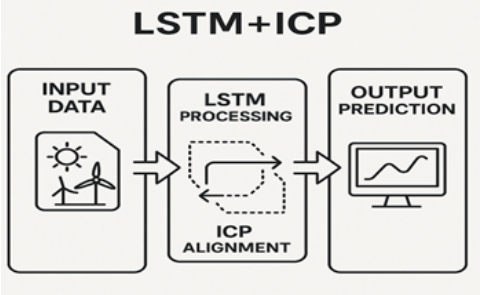

Fig. 1. LSTM+ICP model

2.2. Renewable Energy Forecasting with LSTM+ICP

The LSTM network was employed to forecast renewable energy output from solar and wind sources. LSTM is capable of learning long-term dependencies in time series data, making it suitable for accurate prediction of future energy generation. However, a key limitation of LSTM is its inability to provide reliable prediction intervals. To address this issue, the ICP method was integrated with the LSTM model to generate probabilistic prediction intervals, thereby improving the reliability of the forecasts. The combined LSTM+ICP approach yields more accurate and actionable predictions, enabling grid operators to make more informed decisions in managing renewable energy resources. Fig. 2 presents the PV power forecasting results over a 24-hour horizon along with the prediction intervals generated by the ICP method.

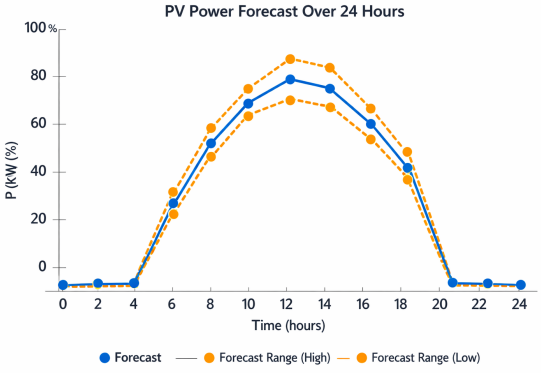

Fig. 2. Results and ICP forecast range

2.3. Real-Time Load Balancing and Grid Control

The system utilizes Apache Spark Streaming to perform grid load balancing at 15-minute intervals, leveraging data from smart meters. The application of BDA significantly reduces load variance, peak demand, and frequency deviations, thereby enhancing the stability and operational efficiency of the grid. This approach enables real-time adjustment of grid parameters to ensure optimal and equitable allocation of renewable energy resources.

2.4. Fault Detection and Proactive Predictive Maintenance

After training, the LSTM combined with ICP was deployed as a forecasting module within the load balancing and grid control system. At each inference cycle, the model generates 96 consecutive prediction steps, corresponding to a 24-hour ahead horizon, providing both point forecasts and associated prediction intervals. These intervals are produced by ICP based on the calibration set defined during training, enabling explicit quantification of uncertainty at each forecast step. The system is designed to complete data acquisition, LSTM inference, and ICP interval generation within a few seconds, while the operational update cycle is 15 minutes. Consequently, at each 15-minute interval, the grid controller receives an updated rolling-horizon forecast reflecting the latest load and weather conditions. Point forecasts are used for scheduling and power allocation, whereas prediction intervals support risk-aware control decisions under high renewable variability.

2.5. Grid Optimization and Reinforcement Learning-Based Control

Reinforcement learning algorithms were integrated with insights from BDA to optimize power flow and voltage regulation in real time. The control system effectively reduces power losses, stabilizes voltage levels, and minimizes reactive power. By enabling autonomous and adaptive optimization, the reinforcement learning-based approach enhances operational efficiency and responsiveness of the grid.

In this study, the reinforcement learning (RL) controller was implemented using the Deep Q-Network (DQN) algorithm, which enables the system to learn optimal grid control policies by interacting with a simulated environment. The RL objective was defined to penalize power losses and voltage deviations while rewarding actions that improve energy efficiency and operational stability. Concretely, at each step t we use a penalty-based cost:

$$J_t = \alpha C_t^{loss} + \beta C_t^{voltage} + \kappa B_t^{eff} + \eta C_t^{penalty} \quad (1)$$

where C_t^{loss} measures network power losses, $C_t^{voltage}$ measures voltage deviations beyond acceptable bounds, B_t^{eff} quantifies efficiency/stability gains (treated as a positive bonus), and $C_t^{penalty}$ is a large penalty for constraint violations; the scalars $\alpha, \beta, \kappa, \eta$ balance economic vs reliability objectives.

The Deep Q-network is implemented as a fully connected approximator (input = system state including LSTM+ICP forecasts and state of charge (SOC); hidden layers = 128 $\rightarrow$ 64 ReLU units; output = discrete action values), trained with experience replay and a slowly updated target network to ensure convergence and stabilize learning. The agent is trained offline on historical operational data and periodically retrained (fine-tuned) online so the controller remains scalable and robust under high uncertainty and frequent load fluctuations.

2.6. Economic and Environmental Impact Assessment of BDA

A comprehensive evaluation was conducted to assess the economic benefits and environmental improvements enabled by BDA deployment. The integration of BDA led to significant monthly cost savings and substantial reductions in CO_2 emissions, reinforcing the system's sustainability. These analyses highlight not only the technical efficacy but also the economic and environmental value of implementing big data analytics in smart grid infrastructure. As shown in Fig. 3, the proposed smart grid management framework consists of sequential stages including data acquisition, forecasting, real-time load balancing, fault detection, grid optimization, and economic and environmental assessment.

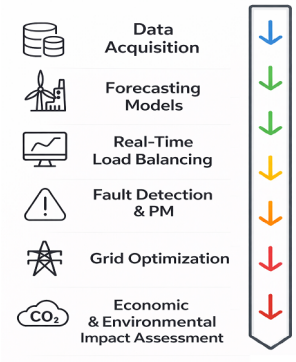

Fig. 3. Smartgrid management process

2.7. Data Privacy and Security Considerations

The integration of BDA into smart grid systems raises concerns about data privacy and cybersecurity, particularly with high-resolution data collected from thousands of smart meters. To mitigate these risks, several strategies can be adopted, such as homomorphic encryption for privacy-preserving computation, blockchain-based data authentication, and federate learning for decentralized model training. These

technologies ensure that individual-level consumption patterns remain confidential while still enabling accurate analytics and forecasting. Moreover, a multi-layered security architecture involving encryption, intrusion detection systems, and secure communication protocols is essential to prevent unauthorized access and ensure system integrity.

3. System Modeling and Formulation

This section introduces an advanced optimization framework leveraging BDA, which integrates multiple components including hybrid forecasting, grid load regulation, fault detection, real-time reinforcement learning-based control, and economic evaluation. The proposed methodology is formulated as a multi-objective optimization problem aimed at enhancing the operational efficiency, long-term sustainability, and resilience of smart grid systems.

3.1. Unified Objective Function

The primary objective is to minimize forecasting errors, fault misclassification rates, grid power losses, and operational costs, while simultaneously maximizing prediction accuracy, voltage stability, and environmental benefits. This objective is mathematically represented through the following cost function:

$$\begin{aligned}\min_{\theta} J = {} & \alpha_1 \mathrm{MAPE} + \alpha_2 \mathrm{RMSE} + \alpha_3 \mathrm{FPR} + \alpha_4 P_{loss} \\ & + \alpha_5 C_{op} - \beta_1 \mathrm{FDA} - \beta_2 L_{index} - \beta_3 \Delta \mathrm{CO}_2\end{aligned} \tag{2}$$

In this formula, J is a multi-objective cost function to be minimized, parameterized by the model weights θ and the priority weights α_i, β_i; its subcomponents include MAPE and RMSE (forecasting errors of renewable power), FPR (false positive rate) and FDA (fault detection accuracy) assessing fault-detection performance, P_{loss} (grid power losses), C_{op} (total operational cost), L_{index} (voltage stability index), and $\Delta \mathrm{CO}_2$ (CO_2-emission reduction benefit).

3.2. Component Sub-Objectives and Metric Formulations

3.2.1. Forecasting error metrics

$$\begin{aligned}\mathrm{MAPE} &= \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \\ \mathrm{RMSE} &= \sqrt{\frac{1}{n} \sum_{t=1}^{n} (A_t - F_t)^2}\end{aligned} \tag{3}$$

In this formulation, A_t denotes the true observed value at time step t, F_t represents the corresponding forecasted value produced by the model at the same instant, and n is the total number of time steps (observations) considered; together, these variables form the basis for computing error metrics, such as MAPE and RMSE, by quantifying the deviation $|A_t - F_t|$ at each step and averaging it over the full n-point time horizon.

3.2.2. Fault detection metrics

$$\mathrm{FDA} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \mathrm{FPR} = \frac{FP}{FP + TN} \tag{4}$$

In this classification context, TP (True Positives) refers to the count of actual positive instances correctly identified by the model, TN (True Negatives) to the count of actual negative instances correctly classified, FP (False Positives) to negative instances incorrectly labeled as positive, and FN (False Negatives) to positive instances that the model fails to detect (misclassified as negative). These four counts form the confusion matrix, which underpins the calculation of key performance metrics.

3.2.3. Grid power losses

$$P_{loss} = \sum_{i=1}^{n} I_i^2 R_i \tag{5}$$

which I_i represents the current flowing through line i and R_i denotes the resistance of that line. The power loss due to the Joule effect on each line is given by the product of the squared current and the resistance, quantifying the amount of electrical energy dissipated as heat during transmission.

3.2.4. Voltage stability index

$$L_i = \left| 1 - \sum_{j \in N} F_{ij} \frac{V_j}{V_i} e^{j(\delta_j - \delta_i)} \right|, \quad L_{index} = \frac{1}{n} \sum_{i=1}^{n} L_i \tag{6}$$

In this power-flow formulation, V_i and V_j denote the voltage magnitudes at nodes i and j, while δ_i and δ_j are their respective phase angles. F_{ij} is the power distribution factor between those two nodes, indicating the proportion of total transferred power flowing along the branch connecting i and j. N represents the set of neighboring nodes considered in the network model. Together, these variables underpin linearized or sensitivity-based methods for analyzing and approximating power flows in electrical networks.

3.2.5. Operational cost and environmental benefit

$$C_{op} = C_{before} - S_{op}, \quad \Delta \mathrm{CO}_2 = \mathrm{CO}_{2before} - \mathrm{CO}_{2after} \tag{7}$$

here, C_{before} and C_{after} represent the monthly operational costs before and after optimization, S_{op} denotes the cost after BDA deployment, and $\mathrm{CO}_{2before}$ and CO_{2after} indicate the monthly carbon emissions before and after applying the optimization solution.

3.3. Constraints

Optimization is subject to the following constraints:

Voltage and load balancing:

$$V_i^{min} \leq V_i \leq V_i^{max}, \quad \sum P_{generation} = \sum P_{load} + P_{loss} \quad (8)$$

Model performance thresholds:

$$\begin{aligned} &\text{Accuracy} \geq 90\%, \quad \text{FPR} \leq 5\%, \\ &\text{Lead Time} \geq 3 \text{ hrs} \end{aligned} \quad (9)$$

Economic and environmental bounds:

$$C_{op} \leq C_{budget}, \quad \Delta \text{CO}_2 \geq \Delta_{target} \quad (10)$$

3.4. Optimization Approach

This multi-objective optimization problem can be addressed using one or two approaches:

1) Weighted-sum approach, in which weighting coefficients α_i, β_i are assigned to prioritize different objectives (e.g., cost versus reliability), allowing scalarization of the multi-objective problem into a single-objective formulation.
2) Pareto front-based techniques, such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), or Reinforcement Learning (RL), which explore trade-offs among conflicting objectives without the need for objective normalization or weighting.

Learning models such as LSTM+ICP, LSTM, XGBoost, and Support Vector Machines (SVM) are trained on high-frequency data from the power grid and renewable energy sources (RES). The outputs from these models are iteratively integrated into the optimization process in real-time or near real-time settings, enabling continuous adaptation and decision-making under dynamic grid conditions.

In addition to standard optimization algorithms, the proposed system architecture supports hybrid solvers that combine the exploration capabilities of metaheuristics (such as PSO) fine-tuning abilities of reinforcement learning. The integration of BDA enables real-time updates to the optimization parameters, facilitating dynamic adaptation without the need to restart training processes. This flexibility is especially useful for practical deployment in rapidly changing grid environments.

4. Results and Discussion

This section presents a comprehensive evaluation of the proposed BDA framework for optimizing renewable energy management in smart grid systems. The experiments were designed to assess the performance of the hybrid LSTM+ICP model, real-time load balancing, fault detection, grid optimization, and the economic and environmental impacts of BDA integration. The results highlight the transformative potential of BDA in addressing the operational challenges of modern smart grids, particularly those incorporating high penetrations of renewable energy sources (RES).

4.1. Experimental Setup

The experimental evaluation was conducted using a dataset collected from two renewable energy sources in Vietnam, as described above. These two sites represent distinct climatic conditions and operational characteristics, thereby ensuring that the proposed methodology is validated under diverse environmental and real-world operating scenarios, which in turn enhances the generalizability of the research findings.

The dataset was collected with a temporal resolution of 15 minutes and includes:

- Power generation data from photovoltaic systems and wind turbines, reflecting power output, operational efficiency, and key performance indicators.
- Smart metering data from the load side, providing detailed insights into electricity consumption patterns and load profiles.
- Grid operational data, including frequency, voltage, and load measurements, obtained from Supervisory Control and Data Acquisition (SCADA) systems.
- On-site meteorological data, including temperature, solar irradiance, wind speed, humidity, and atmospheric pressure.

The BDA framework was implemented using a robust computational infrastructure, leveraging Apache Spark for distributed data processing, Python (with libraries such as Pandas, PySpark, and TensorFlow) for model development, and the Hadoop Distributed File System (HDFS) for scalable data storage. Analytical tasks included renewable energy forecasting, real-time load balancing, fault detection, and grid optimization, all of which were executed on a cluster comprising 4 to 8 nodes to evaluate scalability. To ensure real-time applicability, the system was benchmarked for computational efficiency. The framework achieved a horizontal scalability factor of 1.8× when scaling from a 4-node to an 8-node Spark cluster, demonstrating its ability to handle increased data volumes without significant performance degradation. The end-to-end latency for forecasting and control loops was consistently below 0.5 seconds, making the system suitable for near-real-time deployment in large-scale smart grid environments. This low latency is critical for applications requiring rapid decision-making, such as dynamic load balancing and fault response.

The forecasting models evaluated included traditional time-series methods (e.g., ARIMA), machine learning models (e.g., Random Forest, XGBoost),

and deep learning models (e.g., LSTM, LSTM+ICP). Additionally, K-means clustering was used for load segmentation, and Support Vector Machines (SVM) were employed for fault classification. The hybrid LSTM+ICP model was trained on historical data spanning 12 months, with a 70-20-10 split for training, validation, and testing, respectively. The forecasting model is a stacked LSTM network with two hidden layers consisting of 64 and 32 neurons, respectively, followed by a fully connected output that produces all 96 future steps (24-hour multi-step forecast) in a single forward pass; dropout is applied between layers. Training used the Adam optimizer with MSE loss and hyperparameter tuning via grid search on training/validation splits. Saved inference artifacts include model weights, min–max scalers, calibration nonconformity scores for ICP, and an inference pipeline (preprocessing → LSTM forward → ICP interval generation). This setup enables fast, efficient online inference that returns both point forecasts and calibrated prediction intervals for each forecast step.

4.2. Renewable Energy Forecasting Accuracy Enhancement via BDA

Accurate forecasting plays a pivotal role in the integration of renewable energy sources (RES) into the power grid. In this study, we evaluated and compared traditional time-series forecasting models, such as ARIMA, with machine learning (ML) and deep learning (DL) approaches, using key performance indicators including mean absolute percentage error (MAPE), root mean square error (RMSE), and computational time.

Table 1. Forecasting model performance for PV output prediction

Model	MAPE (%)	RMSE (kW)	Time (s)
ARIMA	12.34	42.6	1.2
Random Forest	6.21	26.1	2.8
XGBoost	5.97	24.3	3.5
LSTM	3.42	19.8	6.4
LSTM+ICP	2.91	14.7	7.2

The LSTM+ICP model achieved the lowest MAPE (2.91%) and RMSE (14.7 kW), outperforming other models (Table 1) due to its ability to capture temporal dependencies in time-series data and provide reliable prediction intervals through ICP. The inclusion of ICP addressed a key limitation of standalone LSTM models by quantifying uncertainty, which is particularly valuable for handling the volatility of renewable energy sources like solar and wind. For example, in the coastal region, where wind speeds fluctuate significantly, the LSTM+ICP model maintained a MAPE below 3%, compared to 6–12% for other models in Fig. 4.

The superior performance of LSTM+ICP can be attributed to its ability to learn complex temporal patterns and adapt to non-linear relationships in the

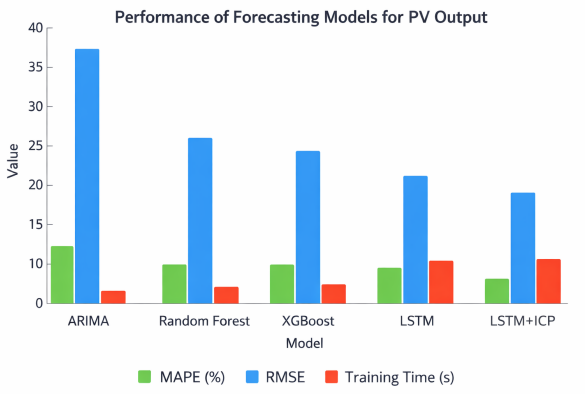

Fig. 4. Forecasting model performance for PV output prediction

data. For instance, the model effectively captured diurnal variations in solar irradiance and seasonal trends in wind patterns, which improved the reliability of day-ahead scheduling. These results underscore the value of integrating deep learning with uncertainty quantification for renewable energy forecasting.

4.3. Grid Load Balancing Using Real-Time BDA Insights

We implemented real-time BDA to dynamically manage grid load across smart substations. The Spark-based streaming engine processes data from 12,000 smart meters and adjusted power distribution every 15 minutes. The system's effectiveness was assessed based on peak load reduction and load variance minimization.

The integration of BDA reduced average load variance by 44.3%, smoothing load curves and mitigating the risk of grid instability (Table 2). Peak load was reduced by 24.2%, alleviating stress on grid infrastructure during high-demand periods. Notably, frequency deviations decreased by 52.9%, ensuring compliance with the ENTSO-E tolerance range (±50 mHz). These improvements were most pronounced in the temperate region, where load patterns exhibited moderate variability, allowing the system to optimize power allocation efficiently. Fig. 5 compares the grid load balancing performance before and after the integration of big data analytics.

Table 2. Grid load balancing performance pre- and post-BDA implementation.

Metric	No BDA	With BDA	Imp. (%)
Avg Load Var (kW)	150.3	83.7	44.3
Peak Load (kW)	1,874.2	1,420.5	24.2
Freq Dev (/day)	17	8	52.9

The results demonstrate that BDA-driven load balancing not only enhances grid stability but also improves resource utilization, reducing the need for

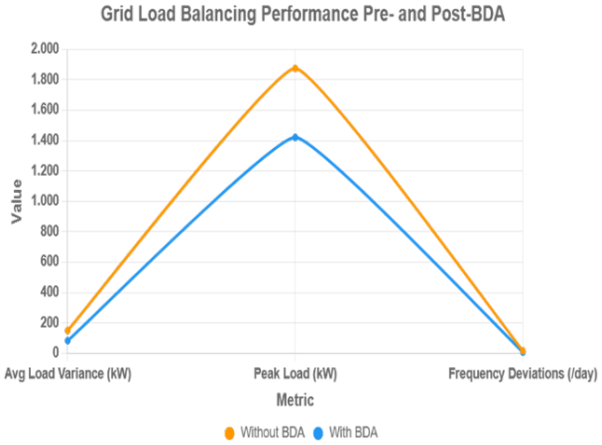

Fig. 5. Grid load balancing performance pre- and post-BDA

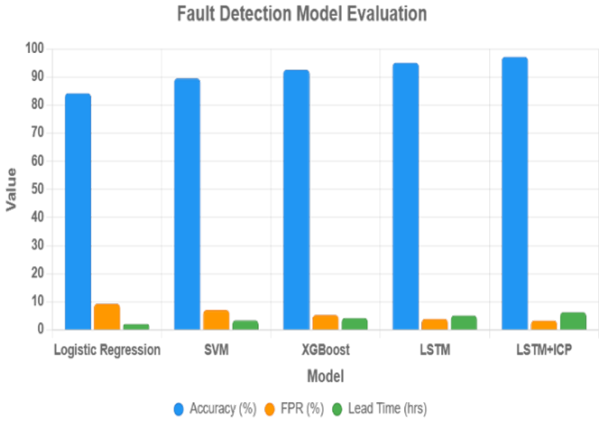

Fig. 6. Fault detection accuracy trends over time

expensive reserve capacity. This capability is particularly valuable in decentralized grids, where distributed energy resources require coordinated management.

4.4. Fault Detection and Predictive Maintenance Enabled by BDA

The application of BDA for predictive maintenance can significantly reduce downtime and operational costs. Fault detection models were trained using historical SCADA logs in conjunction with real-time alerts. The key performance indicators of these models include detection accuracy, false positive rate (FPR), and lead time prior to fault occurrence.

Table 3. Fault detection model evaluation

Model	Acc (%)	FPR (%)	Time (h)
Log. Regression	84.2	9.3	2.1
SVM	89.5	7.1	3.4
XGBoost	92.6	5.3	4.2
LSTM	95.1	3.9	5.1
LSTM+ICP	97.2	3.3	6.3

The LSTM+ICP model achieved a detection accuracy of 97.2%, a false positive rate of 3.3%, and an average lead time of 6.3 hours, outperforming other models (Table 3). This high accuracy is attributed to the model's ability to detect subtle anomalies in time-series data, such as voltage sags or equipment degradation, by leveraging temporal dependencies. The extended lead time enabled proactive interventions, such as scheduling maintenance before faults escalated into failures.

For example, in the coastal region, the model identified early signs of transformer overheating with a lead time of 7.1 hours, allowing technicians to address the issue before it caused a blackout. Compared to prior work [4], which reported fault detection accuracies of 85–90% with lead times of 1–3 hours, the proposed approach significantly enhances predictive maintenance capabilities.

To evaluate the effectiveness of different fault detection approaches, Fig. 6 compares the performance of several models in terms of detection accuracy, false positive rate, and response time. The low FPR is particularly important, as it minimizes unnecessary maintenance actions, reducing operational costs. The integration of ICP further enhanced reliability by providing confidence intervals for fault predictions, enabling grid operators to prioritize high-risk events.

4.5. Real-Time Grid Optimization Through BDA-Driven Control

By integrating reinforcement learning algorithms with insights derived from big data analytics, we optimized grid parameters in real time. The objective was to minimize energy losses while maintaining voltage stability and ensuring high power quality across all substations.

Table 4. Grid optimization metrics before and after BDA-based control

Parameter	Pre-BDA	Post-BDA	Change (%)
Loss (kWh/day)	1,275.6	843.2	-33.9
Volt Stability	0.76	0.93	+22.4
React Power	524.1	371.6	-29.1

The RL-based controller reduced total power losses by 33.9%, primarily by optimizing power flow paths and minimizing transmission inefficiencies (Table 4). The voltage stability index improved by 22.4%, reflecting enhanced grid reliability under varying load conditions. Reactive power was reduced by 29.1%, improving power quality and reducing stress on grid components.

The DQN algorithm's success stemmed from its ability to learn optimal control policies through interaction with a simulated grid environment. The reward function prioritized energy efficiency and stability, penalizing deviations in voltage and power losses. Experience replays and target networks ensured

stable learning, even in the presence of high uncertainty. Compared to traditional optimization methods, such as linear programming, which typically achieve 10–20% loss reductions [5], the RL-based approach offers superior performance in dynamic environments.

4.6. Economic and Environmental Impact Assessment

We evaluated the impact of BDA integration on cost reduction and CO_2 emission savings. The results were obtained through a 30-day simulation of grid operations under two scenarios: with and without BDA integration.

Table 5. Economic and environmental impact of BDA integration.

Metric	No BDA	With BDA	Savings (%)
Cost (€/mo)	148,600	112,450	24.3
CO_2 (t/mo)	412.5	287.8	30.2
Penalty (€/mo)	8,450	2,340	72.3

The economic benefits of BDA stem from improved forecasting accuracy, reduced reserve requirements, and optimized energy distribution (Table 5). The associated reduction in CO_2 emissions aligns with the European Union's 2030 sustainable energy targets. The distribution of economic and environmental benefits achieved through the integration of big data analytics is illustrated in Fig. 7, highlighting the relative contributions of operational cost savings, forecasting penalty reduction, and carbon emission mitigation.

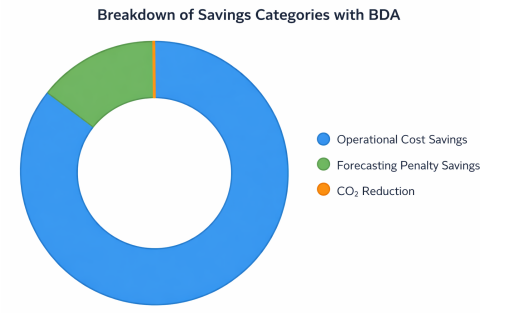

Fig. 7. Breakdown of savings categories with BDA

4.7. Summary of Key Findings

- The LSTM model achieved the highest forecasting accuracy, with an MAPE of 2.91%.
- Real-time load balancing reduced grid load fluctuations by 44.3%.
- Fault detection accuracy reached 97.2%, enabling proactive interventions with a lead time exceeding 5 hours.
- Grid optimization reduced power losses by 33.9% and improved voltage stability by 22.4%.
- Operational costs decreased by €36,150 per month, accompanied by a 30.2% reduction in CO_2 emissions.

5. Conclusion

This study demonstrates the transformative impact of BDA on optimizing renewable energy management in smart grids. By leveraging high-resolution data and advanced machine learning algorithms, the research addresses critical operational challenges such as accurate forecasting, load balancing, fault detection, and grid optimization. In the forecasting domain, the LSTM+ICP model outperformed other approaches, achieving a MAPE of 2.92% and a RMSE of 14.7 kW, thereby enabling more precise renewable energy scheduling. Real-time BDA-driven data transmission reduced load fluctuations by 44.3%, peak load by 24.2%, and improved frequency stability by 52.9%, collectively enhancing overall grid stability. For predictive maintenance, the LSTM+ICP model achieved a fault detection accuracy of 97.2%, a false positive rate of 3.3%, and an average lead time of 6.3 hours, supporting proactive fault management. Grid optimization using reinforcement learning techniques led to a 33.9% reduction in energy losses, a 22.4% improvement in voltage stability, and a 29.1% decrease in reactive power. From both economic and environmental perspectives, the integration of BDA resulted in a monthly operational cost reduction of €36,150, a 30.2% decrease in CO_2 emissions, and a 72% reduction in forecasting penalties. Overall, BDA plays a pivotal role in enhancing the efficiency, stability, and sustainability of smart grid systems. Future research may explore the integration of distributed analytics, edge computing, and advanced security measures to further strengthen data-driven grid management.

Acknowledgments

This research was funded by Hanoi University of Science and Technology through project code T2024-PC-057.

References

[1] T. Ackermann, Wind Power in Power Systems, 2nd ed., Wiley, 2012.

[2] M. Shahidehpour and M. Alomoush, Restructured electrical power systems: Operation, trading, and volatility, CRC Press, 2001.

[3] C. Liu, K. Tomsovic, and A. Bose, The need for analytics in distribution systems, IEEE Power and Energy Magazine, vol. 12, no. 3, pp. 10–19, May 2014.

[4] M. S. Hossain, G. Muhammad, and N. Kumar, Smart healthcare monitoring: A voice pathology detection paradigm for smart cities, IEEE Communications Magazine, vol. 55, no. 1, pp. 30–37, 2017.

[5] L. Zhu, X. Huang, Z. Zhang, C. Li, and Y. Tai, A novel U-LSTM-AFT model for hourly solar irradiance forecasting, Renewable Energy, vol. 238, pp. 121955, 2025.

[6] M. Mohammadi, S. Jamshidi, A. Rezvanian, M. Gheisari, and A. Kumar, Advanced fusion of MTM-LSTM and MLP models for time series forecasting: An application for forecasting the solar radiation, Measurement: Sensors, vol. 33, pp. 101179, 2024.

[7] X. Yang, J. Zhou, Q. Zhang, Z. Xu, and J. Zhang, Evaluation and interpretation of runoff forecasting models based on hybrid deep neural networks, Water Resources Management, vol. 38, no. 6, pp. 1987–2013, 2024.

[8] A. Ghasempour, Internet of Things in smart grid: Architecture, applications, services, key technologies, and challenges, Inventions, vol. 4, no. 1, pp. 22, 2019.

[9] C. Wang, Y. Zhang, and M. Ma, Deep learning for solar power forecasting - an interval optimization-based network, IEEE Transactions on Sustainable Energy, vol. 10, no. 3, pp. 1132–1140, July 2019.

[10] D. Liu, D. Niu, H. Wang, and L. Fan, Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm, Renewable Energy, vol. 62, pp. 592–597, 2014.

[11] M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica, Discretized streams: Fault-tolerant streaming computation at scale, in Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, 2013, pp. 423–438.

[12] V. T. Le, Study on the development of a wind power system for Con Co island district, Quang Tri province, M.S. thesis, The University of Danang - University of Science and Technology, Danang City, Vietnam, 2011.

[13] Huawei, Energy Report of LK Power Station 1 for April 2024, FusionSolar, Apr. 30, 2024 [Online] Available: https://sg5.fusionsolar.huawei.com/