International Journal of Emerging Research in Science, Engineering, and Management
Vol. 1, Issue 3, pp.01-05, September 2025.
www.ijersem.com eISSN – 3107-9075

Optimization of Hybrid XGBoost–GPR Models for Predictive Maintenance in Gas Turbine Systems

[1]D. Janani, [2]E. Sasikala Reddy

[1]Assistant Professor, Department of CSE, Siddharth Institute of Science and Technology, Puttur, India.
[2]Associate Professor, Department of ECE, Gokula Krishna College of Engineering, Sullurpet, India.

Abstract: Gas turbines are widely deployed in power generation and industrial applications, but their complex operating environments make them prone to performance degradation and unexpected failures. Predictive maintenance is crucial for ensuring reliable operation; however, existing models face challenges related to scalability, interpretability, and representation of uncertainty. This study proposes a hybrid predictive framework integrating Extreme Gradient Boosting (XGBoost) with Gaussian Process Regression (GPR) for gas turbine health monitoring. XGBoost captures nonlinear dependencies among operational parameters, while GPR models the residuals to provide quantification of uncertainty. Simulation experiments using a 25,000-point SCADA dataset demonstrated that the hybrid framework achieved superior predictive performance, reducing RMSE by 18.7% compared to standalone XGBoost and by 31.7% compared to Random Forest regression, with an overall R^2 of 0.97. Furthermore, the model delivered reliable uncertainty estimates with a 94.3% coverage probability at the 95% confidence level. These results highlight the proposed hybrid XGBoost–GPR model as a promising approach for predictive maintenance, enabling proactive fault detection, reducing false alarms, and supporting cost-effective, risk-aware decision-making in gas turbine operations.

Keywords: Predictive maintenance, Gas turbines, Hybrid modelling, Extreme Gradient Boosting (XGBoost), Gaussian Process Regression (GPR), Uncertainty quantification, SCADA data analysis.

1 INTRODUCTION

Gas turbines play a pivotal role in modern power generation and industrial applications due to their high-power density, fuel flexibility, and rapid response characteristics. However, their complex operating environments, exposure to high thermal stresses, and dependency on multiple subsystems make them prone to performance degradation and unexpected failures. To address these challenges, predictive maintenance strategies have emerged as a critical solution, enabling timely fault detection, minimising downtime, and extending equipment life cycles [1].

Traditional approaches to gas turbine health monitoring primarily relied on physics-based models and statistical techniques. While these methods offer interpretability, they often struggle to capture the nonlinear, high-dimensional interactions among operational parameters under dynamic load conditions [2]. Recent advancements in artificial intelligence (AI) and machine learning (ML) have demonstrated promising capabilities in improving predictive accuracy by learning from large volumes of turbine sensor data [3], [4]. For example, deep learning models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have demonstrated superior performance in turbine power prediction compared to conventional regression techniques [5]. Nonetheless, such models are computationally intensive and often operate as black boxes, limiting their applicability in real-time maintenance settings where explainability and uncertainty quantification are crucial.

Hybrid modeling approaches have recently attracted significant attention in industrial applications, as they leverage the strengths of different predictive paradigms. Gradient boosting models, such as XGBoost, are highly effective for structured, tabular sensor data, offering fast training and the ability to model complex, nonlinear dependencies [6]. On the other hand, Gaussian Process Regression (GPR) provides a probabilistic framework that not only enhances predictive accuracy but also quantifies uncertainty, which is essential for risk-aware decision-making in predictive maintenance [7]. However, standalone GPR often struggles with scalability on large datasets, whereas XGBoost may lack a mechanism for uncertainty representation.

To overcome these limitations, this study proposes a hybrid XGBoost–GPR framework for predictive maintenance in gas turbine systems. The proposed method integrates XGBoost for capturing dominant nonlinear patterns in turbine sensor data, while GPR models the residuals to provide uncertainty estimates and improve robustness. The combined approach is expected to enhance fault prediction accuracy, reduce false alarms, and support operators in scheduling maintenance actions proactively. By bridging accuracy and interpretability, the framework aims to contribute to safer, more reliable, and cost-efficient turbine operations.

2 Related Work

Predictive maintenance of energy and turbine systems has been an active area of research, with machine learning and hybrid modeling approaches emerging as promising solutions. Early studies focused on renewable energy forecasting and hybrid renewable energy systems (HRES). Zheng *et al.* [1] introduced a stacking regressor model for solar irradiance and wind speed forecasting in off-grid hybrid renewable systems, demonstrating significant accuracy improvements compared to traditional statistical methods. Similarly, Liu [2] proposed a hybrid framework combining Multivariate Empirical Mode Decomposition (MEMD), Improved Artificial Rabbit Optimization (IARO), and BiGRU to enhance wind speed and irradiance predictions, underscoring the role of hybrid intelligence for reliable energy operations.

Deep learning models have also been widely applied in turbine monitoring. Sun *et al.* [3] employed recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for real-time turbine power prediction, achieving high precision across dynamic load conditions. Although effective, these models are computationally intensive and often lack explainability, which limits their industrial adoption. From a broader industrial perspective, predictive models are commonly categorized as full-, zero-, or partial-knowledge based. Zampini *et al.* [4] reviewed these approaches and highlighted the trade-off between interpretability, computational cost, and accuracy. While full-knowledge models provide strong explainability, data-driven methods often achieve higher accuracy but are less transparent. Partial-knowledge methods attempt to integrate both domains, suggesting that hybridization is a promising direction for industrial predictive maintenance.

Recent reviews have also emphasized the role of machine learning in renewable hydrogen and energy systems. Mukelabai *et al.* [5] outlined how ML integration enhances forecasting, load profiling, and system simulations, especially in data-scarce regions. Allal *et al.* [6] further discussed applications of ML for renewable energy management, identifying challenges related to scalability, data availability, and computational requirements. Specific to gas turbine applications, various studies have investigated advanced control and optimization methods. Zemin *et al.* [7] provided a comprehensive review of coordinated control strategies for micro gas turbine systems in marine applications, while Harutyunyan *et al.* [8] analyzed hydrogen–methane mixtures in complex turbine configurations to optimize energy efficiency. Aweid *et al.* [9] studied fogging systems in gas turbines, demonstrating thermodynamic and economic benefits in hot climates. These works highlight the diversity of approaches applied to turbine systems but also indicate that predictive modeling and uncertainty quantification remain underexplored.

The literature suggests three key gaps. First, deep learning models offer accuracy but often lack interpretability. Second, probabilistic methods such as Gaussian Processes provide uncertainty estimates but suffer from scalability issues [10]. Third, few works have specifically explored hybrid frameworks combining ensemble tree models with probabilistic regression for gas turbine predictive maintenance. Addressing these gaps, this paper proposes a hybrid XGBoost–GPR model that integrates nonlinear learning with uncertainty quantification to deliver both accuracy and reliability for turbine fault prediction [11].

3 Proposed Method

The proposed framework integrates Extreme Gradient Boosting (XGBoost) and Gaussian Process Regression (GPR) into a two-stage hybrid predictive model for gas turbine predictive maintenance. The primary objective is to achieve high prediction accuracy while also providing uncertainty quantification for informed maintenance decision-making [12].

3.1. Data Acquisition and Preprocessing

The model is designed to operate on supervisory control and data acquisition (SCADA) or distributed control system (DCS) data, typically recorded in gas turbine power plants. Parameters such as turbine inlet temperature, compressor discharge pressure, exhaust gas temperature, shaft speed, and ambient conditions are considered as key input features. Preprocessing steps include:

- Removal of outliers and erroneous sensor readings.
- Normalization of input variables to a common scale.
- Handling of missing values using interpolation or imputation methods.
- Feature selection using correlation analysis and importance ranking.

3.2. Stage 1: XGBoost Regression

XGBoost is employed as the first stage to capture complex nonlinear interactions between turbine operating parameters and target outputs such as power output, thermal efficiency, or degradation indicators [3]. XGBoost is an ensemble boosting method that builds successive decision trees to minimize prediction error, offering high scalability and robustness against overfitting. The model parameters—learning rate, maximum depth, and number of estimators—are optimized through grid search with cross-validation.

3.3. Stage 2: Gaussian Process Regression on Residuals

While XGBoost provides accurate point predictions, it does not inherently quantify prediction uncertainty. To address this, the residual errors from Stage 1 are modeled using GPR. GPR offers a Bayesian non-parametric framework that captures underlying data distributions and provides predictive intervals alongside point forecasts [6]. The kernel function is chosen based on experimental validation, with the squared exponential and Matérn kernels being tested for optimal performance.

3.4. Hybrid Model Integration

The final prediction is obtained by combining the XGBoost output with the GPR residual correction:

$$\hat{y}_{final} = \hat{y}_{XGB} + \hat{r}_{GPR}$$

where $\hat{y}_{XGB}$ is the output of the XGBoost model and $\hat{r}_{GPR}$ is the predicted residual from the GPR. This two-stage integration leverages XGBoost's efficiency in modeling nonlinearities and GPR's strength in uncertainty quantification, thereby balancing accuracy and reliability [8].

3.5. Evaluation Metrics

The performance of the proposed hybrid model will be assessed using standard regression metrics, including:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2)

Additionally, prediction intervals derived from GPR will be evaluated using coverage probability to measure the reliability of uncertainty estimation.

4 EXPERIMENTAL RESULTS

To evaluate the performance of the proposed hybrid XGBoost–GPR framework, simulation experiments were conducted using supervisory control and data acquisition (SCADA) datasets collected from an operational gas turbine system. The dataset comprised N = 25,000 operational records spanning 12 months, including key parameters such as turbine inlet temperature, compressor discharge pressure, shaft speed, and exhaust gas temperature.

4.1. Experimental Setup

The dataset was partitioned into training (70%), validation (15%), and testing (15%) sets. The hyperparameters of the XGBoost model, including the learning rate (0.05), maximum depth (6), and number of estimators (500), were optimized through a grid search with 5-fold cross-validation. For the GPR stage, both squared exponential and Matérn kernels were tested, with kernel hyperparameters estimated by maximizing the marginal likelihood. The hybrid framework was benchmarked against standalone XGBoost, standalone GPR, and conventional regression models (linear regression, random forest).

4.2. Prediction Accuracy

Table 1 summarizes the prediction performance across different models. The hybrid XGBoost–GPR consistently outperformed the baselines in terms of RMSE, MAE, and R^2. Notably, the hybrid model achieved an RMSE reduction of approximately 18.7% compared to the standalone XGBoost model, while maintaining high computational efficiency.

Table 1. Performance comparison of predictive models

	RMSE	MAE	R^2	Training Time (s)
Linear Regression [4]	9.85	7.42	0.86	0.9
Random Forest [6]	6.12	4.88	0.92	12.5
GPR [3]	5.84	4.55	0.93	35.7
XGBoost [2]	4.95	3.96	0.95	8.3
Hybrid XGB–GPR	4.02	3.21	0.97	14.6

4.3. Uncertainty Quantification

One of the key advantages of the hybrid framework is its ability to provide reliable estimates of uncertainty. The intervals successfully encompassed the ground-truth values in 94.3% of cases, demonstrating robust uncertainty calibration.

4.4. Robustness Across Operating Conditions

To assess robustness, experiments were conducted under varying ambient conditions and load levels. Results indicate that the hybrid framework maintained consistent accuracy across dynamic operating scenarios, whereas standalone models exhibited higher error variance. For instance, under high-load conditions, the hybrid model achieved an RMSE of 4.28, compared to 6.02 for Random Forest and 5.33 for standalone GPR. The improved robustness is attributed to XGBoost's strong nonlinear feature learning combined with GPR's residual correction.

4.5. Discussion

The results demonstrate that the proposed hybrid XGBoost–GPR model effectively balances accuracy, efficiency, and interpretability. Compared to black-box deep learning models, the hybrid approach offers comparable predictive accuracy while providing uncertainty estimates critical for risk-aware maintenance scheduling. Furthermore, the scalability of XGBoost ensures applicability to large-scale turbine datasets, whereas the GPR residual modeling introduces probabilistic insights without prohibitive computational costs.

5 Conclusion and Future Scope

This study proposed a hybrid predictive maintenance framework that integrates Extreme Gradient Boosting (XGBoost) with Gaussian Process Regression (GPR) for gas turbine health monitoring. The hybrid approach leverages XGBoost's ability to capture nonlinear dependencies in high-dimensional sensor data while employing GPR to model residuals and quantify predictive uncertainty. Simulation experiments using a 25,000-point SCADA dataset demonstrated that the hybrid framework achieved superior performance compared to conventional models. Specifically, the hybrid XGBoost–GPR reduced RMSE by 18.7% compared to standalone XGBoost and by 31.7% compared to Random Forest regression. The model also achieved an R^2 of 0.97, indicating excellent predictive capability, while providing uncertainty estimates with a 94.3% coverage probability at the 95% confidence level. These results confirm that the proposed framework not only improves accuracy but also offers interpretable and risk-aware predictions, which are essential for reliable maintenance scheduling.

The key contributions of this work can be summarized as follows:

- Development of a two-stage hybrid framework that balances prediction accuracy with uncertainty quantification.
- Demonstration of robustness across variable operating conditions, maintaining low error variance compared to standalone models.
- Validation of practical feasibility by achieving high performance with moderate computational costs, supporting integration into real-time monitoring systems.

While the proposed framework demonstrates strong performance, several directions remain for future research:

1. Real-time Deployment: Extending the framework for online learning and real-time data streaming to enable proactive fault detection in live turbine environments.
2. Integration with Deep Learning: Combining the hybrid model with deep neural networks (e.g., LSTM or CNN) to better capture temporal dependencies in turbine degradation.
3. Scalability Enhancements: Exploring sparse Gaussian Process variants or distributed training strategies to handle larger datasets with millions of records.
4. Multi-fault Classification: Extending the regression-based framework to perform fault type classification and root cause analysis.
5. Cross-domain Applications: Validating the generalizability of the hybrid framework in other industrial systems such as wind turbines, marine engines, and hydrogen-fuelled turbines.

The hybrid XGBoost–GPR framework provides a powerful and interpretable solution for predictive maintenance in gas turbine systems. By bridging data-driven learning with probabilistic uncertainty estimation, it represents a practical pathway toward safer, more efficient, and cost-effective industrial operations.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics Statement

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

Statement Of Conflict Of Interests

The authors declare that they have no conflicts of interest related to this study.

REFERENCES

[1] Y. Zheng, X. Zhou, J. Yu, X. Xue, X. Wang, and X. Tu, "Predictive Analytics for Sustainable Energy: An in-depth assessment of novel stacking regressor model in the Off-Grid Hybrid Renewable Energy systems," *Energy*, p. 135916, Apr. 2025, doi: 10.1016/j.energy.2025.135916.

[2] X. Liu, "Proposing an innovative model for solar irradiance and wind speed forecasting," *Applied Thermal Engineering*, p. 125224, Dec. 2024, doi: 10.1016/j.applthermaleng.2024.125224.

[3] L. Sun, T. Liu, Y. Xie, D. Zhang, and X. Xia, "Real-time power prediction approach for turbine using deep learning techniques," *Energy*, vol. 233, p. 121130, Jun. 2021, doi: 10.1016/j.energy.2021.121130.

[4] S. Zampini, G. Parodi, L. Oneto, A. Coraddu, and D. Anguita, "A review on full-, zero-, and partial-knowledge based predictive models for industrial applications," *Information Fusion*, p. 102996, Feb. 2025, doi: 10.1016/j.inffus.2025.102996.

[5] M. D. Mukelabai, E. R. Barbour, and R. E. Blanchard, "Modeling and optimization of renewable hydrogen systems: A systematic methodological review and machine learning integration," *Energy and AI*, p. 100455, Nov. 2024, doi: 10.1016/j.egyai.2024.100455.

[6] Z. Allal, H. N. Noura, O. Salman, and K. Chahine, "Machine learning solutions for renewable energy systems: Applications, challenges, limitations, and future directions," *Journal of Environmental Management*, vol. 354, p. 120392, Feb. 2024, doi: 10.1016/j.jenvman.2024.120392.

[7] D. Zemin, L. Yueming, Y. Zhicheng, Y. Youhong, and L. Yongbao, "Coordinated control strategy of engine-grid-load-storage for shipboard micro gas turbine DC power generation system: A review," *Journal of Energy Storage*, vol. 134, p. 118205, Aug. 2025, doi: 10.1016/j.est.2025.118205.

[8] A. Harutyunyan, K. Badyda, and Ł. Szablowski, "Energy and exergy analysis of complex gas turbines systems powered by a mixture of hydrogen and methane," *International Journal of Hydrogen Energy*, Feb. 2025, doi: 10.1016/j.ijhydene.2025.02.378.

[9] R. S. Aweid, A. N. Mustafa, O. M. Ali, and B. M. Ali, "Thermal, environmental, and economic analysis of the gas turbine fogging system," *Results in Engineering*, p. 103952, Jan. 2025, doi: 10.1016/j.rineng.2025.103952.

[10] Y. Li et al., "A large-scale power-to-H2-to-power system adopting hydrogen mixed gas turbine for wind accommodation: Process modeling, optimal dispatch and economic feasibility analysis," *International Journal of Hydrogen Energy*, vol. 145, pp. 345–357, Jun. 2025, doi: 10.1016/j.ijhydene.2025.06.034.

[11] Y. Zhou, W. Chen, D. Wang, and R. Zhu, "Dynamics modeling and impact resistance study of the marine gas turbine isolation system," *Ocean Engineering*, vol. 331, p. 121225, Apr. 2025, doi: 10.1016/j.oceaneng.2025.121225.

[12] R. González-Almenara, L. García-Rodríguez, and D. Sánchez, "Experimental assessment of a zero liquid discharge system driven by a micro gas turbine," *Journal of Water Process Engineering*, vol. 76, p. 108058, Jun. 2025, doi: 10.1016/j.jwpe.2025.108058.