

Data-Driven Surrogate Modelling of Modified Gas Turbine Cycles Using Machine Learning for Real-Time Performance Prediction

Ogbe Emmanuel Ediba¹, Ochogwu Emmanuel Bamaiyi²

¹Department of Marine Engineering, Nigeria Maritime University, Okerenkoko, Delta State, Nigeria

²Department of Mechanical Engineering, Nigeria Maritime University, Okerenkoko, Delta State, Nigeria

Corresponding Author: Ogbe Emmanuel Ediba

DOI: <https://doi.org/10.52403/ijrr.20260340>

ABSTRACT

Gas turbine power plants remain a key component of flexible electricity generation, yet their performance is strongly influenced by ambient conditions, operating parameters, and retrofit configurations. Conventional thermodynamic models provide high accuracy but are computationally intensive and unsuitable for real-time applications, particularly for modified gas turbine cycles with increased system complexity. This study presents a data-driven surrogate modelling framework for real-time performance prediction of conventional and modified gas turbine cycles by integrating first-principles thermodynamic simulation with advanced machine learning techniques.

A conventional gas turbine and three modified configurations incorporating inlet air cooling, regeneration, heat-recovery steam generation, steam injection, dual combustion, and staged turbine expansion were modelled in ASPEN HYSYS to generate a large parametric dataset across realistic operating ranges. Artificial neural network (ANN), random forest (RF), and extreme gradient boosting (XGBoost) models were developed to predict key performance indicators, including thermal efficiency, net power output, and emission intensity, and were benchmarked against multiple linear regression models.

The thermodynamic results show that all modified configurations significantly outperform the conventional cycle. The optimal configuration achieved a thermal efficiency of approximately 47%, an increase in power output exceeding 40%, and a substantial reduction in emission intensity. The machine learning surrogate models demonstrated excellent predictive accuracy, with coefficients of determination greater than 0.95 for all outputs, and provided several orders-of-magnitude reductions in computational time compared with detailed process simulations. Feature-importance analysis identified turbine inlet temperature, compressor pressure ratio, and fuel mass flow rate as the dominant performance drivers, consistent with thermodynamic principles.

The proposed framework enables rapid and accurate performance prediction and provides a practical foundation for deploying digital twins in gas turbine power plants. It offers a scalable pathway for enhancing the efficiency, flexibility, and environmental performance of existing gas turbine-based power generation systems.

Keywords: Gas turbine; Machine learning; Surrogate model; Digital twin; ASPEN HYSYS; ANN; XGBoost; Random Forest; Performance prediction.

1. INTRODUCTION

Gas turbines play a critical role in modern power generation systems, providing essential grid stability and flexible operation capabilities that are increasingly vital as renewable energy penetration grows [1–3]. As power grids transition toward higher shares of intermittent renewable sources, gas turbine combined cycle (GTCC) plants must operate with enhanced flexibility, responding rapidly to fluctuating demand and compensating for variable renewable generation [1,4–6]. This operational paradigm shift necessitates accurate, real-time performance prediction tools that can support optimal dispatch decisions, maintain grid reliability, and ensure efficient operation across diverse operating conditions [1,3,7,8]. The performance of gas turbine systems is highly sensitive to ambient conditions, including temperature, humidity, and atmospheric pressure, which can significantly impact power output, thermal efficiency, and emissions [8–11]. In tropical and warm climates, ambient temperature variations can cause power output degradation of 10-15% during peak demand periods [8,9,12]. Furthermore, modern gas turbines are increasingly deployed with cycle modifications such as inlet air cooling [8–10,13–17], regeneration and recuperation [3,6,18,19], steam injection (STIG) [20–26], and heat recovery steam generator (HRSG) integration [6,18,27], each introducing additional complexity to performance prediction and optimization [1,7,18].

Traditional physics-based thermodynamic models have long served as the foundation for gas turbine performance analysis, offering detailed component-level representations grounded in fundamental conservation laws and empirical correlations [2,5,19,28]. However, these approaches face significant limitations in real-time operational scenarios. High-fidelity physics-based simulations require substantial computational resources, often taking minutes to hours for a single operating point evaluation, rendering them impractical for online optimization and control applications

[2,8,9]. Additionally, physics-based models demand extensive calibration and validation against experimental data to achieve acceptable accuracy, particularly for off-design conditions and degraded component states [5,19,28].

The computational burden intensifies when modeling complex cycle configurations with multiple interacting subsystems, such as combined cycles with supplementary firing, steam injection, or advanced recuperation schemes [1,7,18,27]. Model mismatch and uncertainty propagation through coupled component models further compromise prediction reliability, especially during transient operation and part-load conditions [2,13,19,28,29]. These limitations have motivated the exploration of data-driven and hybrid modeling approaches that can deliver comparable accuracy with dramatically reduced computational cost [2,5,8,9].

Recent advances in machine learning (ML) and artificial intelligence have opened new pathways for gas turbine performance modeling, enabling the development of fast, accurate surrogate models trained on operational data or high-fidelity simulation results [5,7,8,19,20]. Artificial neural networks (ANNs) have demonstrated promise for gas turbine applications, capturing complex nonlinear relationships between operating parameters and performance metrics [6,12,18,30]. Deep learning architectures, including recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, have been successfully deployed for real-time performance modeling of multi-spool engines and dynamic condition monitoring [6,18,27,30,31].

Ensemble methods, including random forests and gradient boosting algorithms such as XGBoost, offer complementary advantages through their ability to handle high-dimensional feature spaces and provide robust predictions with inherent uncertainty quantification [5,22,32]. These methods have been applied to emissions prediction [21,22], performance optimization [20,32], and multi-condition identification tasks [5,7].

The integration of domain knowledge and physical constraints into ML models, often termed physics-informed machine learning, has emerged as a promising strategy to enhance model interpretability, ensure thermodynamic consistency, and improve extrapolation capabilities [2,13,19,28].

Digital twin technology represents a convergence of physics-based modelling, data-driven learning, and real-time sensor integration, creating virtual replicas of physical assets that evolve synchronously with their real-world counterparts [2,8,9,11–13]. Recent digital twin implementations for gas turbines have demonstrated capabilities for anomaly detection [11,12], performance degradation monitoring [12,23], remaining useful life (RUL) prediction [18,23,27,33,34], and prescriptive analytics [1,3,19]. Multi-model fusion approaches that combine multiple ML architectures or blend data-driven and physics-based components have shown superior accuracy and robustness compared to single-model strategies [8,9,19,28].

Gas turbine operators and manufacturers have developed numerous cycle modification strategies to enhance performance, improve efficiency, and reduce emissions across diverse operating conditions. Inlet air cooling technologies, including evaporative cooling, fogging systems, and mechanical chilling, can recover significant power output losses in hot climates by reducing compressor inlet temperature [8–13,30]. Studies have demonstrated power augmentation of 10–20% and efficiency improvements of 2–5 percentage points through optimized inlet cooling strategies [8–10,13,35].

Regenerative and recuperated cycles, which recover exhaust heat to preheat compressor discharge air, offer substantial efficiency gains, particularly for smaller industrial gas turbines and part-load operation [3,6,18,19]. Dual-recuperated configurations and advanced recuperator designs have achieved thermal efficiency improvements exceeding 10 percentage points compared to simple cycle operation [3,6,19]. Steam injection

systems, including STIG and intercooled STIG (ISTIG) configurations, provide simultaneous benefits of power augmentation, efficiency enhancement, and NOx emissions reduction by introducing steam into the combustion chamber or turbine inlet [20–23,36].

Combined-cycle configurations with HRSG integration represent the most widely deployed efficiency-enhancement strategy for large-scale power generation, achieving overall thermal efficiencies exceeding 60% by bottoming out with steam Rankine cycles [1,6,10,18,27,37]. Advanced combined-cycle concepts incorporating supplementary firing, flexible operating capabilities, and novel working fluids (such as supercritical CO₂) continue to push efficiency boundaries while addressing grid flexibility requirements [1,5,7,28]. However, the performance benefits of these cycle modifications are highly dependent on ambient conditions, operating load, and the state of component degradation, necessitating sophisticated modelling tools that can accurately predict performance across the full operational envelope [1,6,7,18].

Surrogate modelling, the practice of replacing computationally expensive high-fidelity models with fast, approximate models, has become essential for real-time decision support, online optimization, and control applications in gas turbine systems [5,7–9]. Data-driven surrogates trained on comprehensive datasets spanning the operational space can achieve prediction times measured in milliseconds while maintaining accuracy within 1–2% of high-fidelity references [5,8,9]. Neural ordinary differential equations (Neural ODEs) have emerged as a particularly promising surrogate architecture for combined cycle plants, capturing complex dynamic behaviour with computational efficiency suitable for real-time digital twin applications [9,38,39].

Extreme learning machines and operator networks offer alternative surrogate modelling paradigms with rapid training

capabilities and strong generalization performance for gas turbine component and system-level modelling [8,12]. Multi-working condition identification and performance prediction frameworks that leverage deep learning and knowledge integration have demonstrated the ability to accurately model gas turbine behaviour across diverse operating regimes, including startup, shutdown, and transient manoeuvres [5,7]. The computational efficiency advantages of ML-based surrogates over physics-based models have been quantified in multiple studies, with speedup factors ranging from 100× to 10,000× depending on model complexity and fidelity requirements [2,8,9,28].

While ML models offer compelling accuracy and speed advantages, their "black-box" nature has raised concerns regarding interpretability, physical consistency, and trustworthiness in safety-critical applications [2,5,13,19]. Domain-informed approaches that embed physical constraints, operational limits, and expert knowledge into ML architectures can enhance model robustness and ensure predictions remain within physically realizable bounds [13,19,20]. Feature importance analysis and sensitivity studies provide insights into which operating parameters most strongly influence performance metrics, supporting both model validation and operational decision-making [2,19,20].

Physics-informed neural networks (PINNs) and hybrid physics-data architectures explicitly incorporate governing equations and thermodynamic principles as soft constraints during training, promoting physically consistent predictions and improving extrapolation beyond the training data distribution [10,13,28]. Uncertainty quantification and confidence interval estimation remain active research areas, with ensemble methods and Bayesian neural networks offering pathways to quantify prediction uncertainty and support risk-aware decision-making [2,11,22].

Gas turbine performance inevitably degrades over time due to compressor fouling, turbine

blade erosion, seal wear, and combustor deterioration, necessitating predictive maintenance strategies and performance recovery planning [1,12,23,27,33]. Data-driven degradation models and remaining useful life (RUL) prediction frameworks enable proactive maintenance scheduling, reducing unplanned outages and optimizing lifecycle costs [18,23,27,33]. Digital twin-oriented approaches that integrate real-time sensor data with physics-informed ML models have demonstrated capabilities for early anomaly detection and degradation trend forecasting [11,12,27].

Lightweight modelling approaches using operator networks and reduced-order models offer computational efficiency suitable for continuous online monitoring while maintaining sufficient fidelity to detect subtle performance deviations [12,27]. Benchmarking studies of gas turbine outages and performance recovery patterns have established expectation models that quantify typical degradation rates and maintenance effectiveness, providing valuable context for prognostic algorithms [1,33]. The integration of degradation modeling with performance prediction enables more accurate long-term forecasting and supports strategic planning for fleet management and asset optimization [1,23,33].

Accurate prediction of emissions, particularly nitrogen oxides (NO_x), carbon monoxide (CO), and unburned hydrocarbons (UHC), is essential for regulatory compliance, environmental impact assessment, and combustion optimization [19–22]. Predictive emissions monitoring systems (PEMS) based on hybrid ML models have been developed to provide continuous emissions estimates without the cost and maintenance burden of physical continuous emissions monitoring systems (CEMS) [21,22]. Artificial neural networks trained on operational data have demonstrated the ability to predict NO_x emissions with accuracy suitable for regulatory reporting and real-time combustion tuning [20–22].

Steam injection and advanced combustion strategies offer pathways to simultaneous

performance enhancement and emissions reduction, but their effectiveness depends critically on precise control and optimization [20–23]. Techno-economic optimization frameworks that balance performance, emissions, and operating costs have been developed to identify optimal operating strategies for modified gas turbine cycles [20,21,23]. Thermo-environmental (thermos-environmental) evaluation methodologies extend traditional exergy analysis to incorporate environmental impact metrics, enabling holistic assessment of cycle modifications and operating strategies [21,23,40].

Despite significant progress in ML-based gas turbine modelling, several critical gaps remain. Most existing studies focus on simple cycle configurations or single-cycle modifications, with limited attention to the complex interactions among multiple simultaneous modifications (e.g., inlet cooling combined with steam injection and recuperation). The sensitivity of ML model accuracy to training data coverage, particularly for rare operating conditions and extreme ambient environments, remains poorly understood. Furthermore, the trade-offs between model complexity, computational efficiency, and prediction accuracy across diverse cycle configurations have not been systematically characterized. This research addresses these gaps by developing a comprehensive data-driven surrogate modelling framework for modified gas turbine cycles that:

1. Integrates multiple cycle modification strategies (inlet cooling, regeneration, HRSG, steam injection) within a unified modelling framework
2. Systematically evaluates multiple ML architectures (ANNs, random forests, XGBoost, ensemble methods) for accuracy, computational efficiency, and robustness
3. Incorporates ambient condition sensitivity and part-load performance across the full operational envelope
4. Implements physics-informed constraints and domain knowledge to

enhance model interpretability and extrapolation capabilities

5. Quantifies prediction uncertainty and establishes confidence intervals for real-time decision support
6. Demonstrates real-time performance prediction capabilities suitable for digital twin integration and online optimization

The developed surrogate models enable rapid evaluation of cycle modification strategies, support optimal dispatch and control decisions, and provide a foundation for next-generation digital-twin applications in gas-turbine power generation. By combining the accuracy of high-fidelity physics-based models with the computational efficiency of data-driven approaches, this work advances the state of the art in gas turbine performance prediction and contributes to the broader goal of flexible, efficient, and sustainable power generation in renewable-dominated grids.

2. MATERIALS AND METHODS

2.1. Overall Research Framework

This study adopts a hybrid physics–data-driven surrogate modelling framework for real-time performance prediction of conventional and modified gas turbine cycles. The methodology integrates first-principles thermodynamic simulation, large-scale parametric dataset generation, and machine learning–based surrogate modelling, culminating in a digital-twin-ready prediction architecture.

The overall workflow consists of the following sequential stages:

1. Development of high-fidelity thermodynamic models of the conventional gas turbine and three modified configurations using ASPEN HYSYS
2. Large-scale parametric dataset generation over realistic operating envelopes
3. Construction of baseline multiple linear regression models in MATLAB
4. Development and training of nonlinear machine learning surrogate models (ANN, Random Forest, and XGBoost)

5. Model validation using statistical accuracy metrics and generalization testing
6. Feature importance analysis for physical interpretability
7. Computational speed benchmarking for real-time feasibility
8. Formulation of a digital twin deployment architecture

This hybrid approach ensures physical consistency, predictive accuracy, and computational efficiency, enabling practical online deployment in operating gas turbine power plants.

2.2. Gas Turbine System Description and Configurations

Four gas turbine configurations were investigated in this study:

1. **SGT** – Conventional single-shaft Brayton-cycle gas turbine
2. **MGTP 1** – Regenerator + inlet air cooling + HRSG + steam injection + dual-turbine expansion
3. **MGTP 2** – Regenerator + inlet air cooling + HRSG + dual combustion chambers
4. **MGTP 3** – Dual combustion chambers + dual turbine expansion

These configurations represent progressively advanced retrofit strategies aimed at improving thermal efficiency, power output, and environmental performance.

All systems were modelled under steady-state operating conditions using ASPEN HYSYS v11 with the Peng–Robinson equation of state. The component models include Axial compressor, Combustion chamber(s), Gas turbine(s), Regenerator (where applicable), Heat recovery steam generator (HRSG), Steam injection loop, and Inlet air cooling unit

Each configuration was validated against typical industrial gas turbine operating ranges and benchmarked with representative plant data to ensure physical realism.

2.3. Thermodynamic Modelling

The thermodynamic behaviour of each gas turbine configuration was modelled using the First Law of Thermodynamics for steady-flow open systems. Standard Brayton-cycle

component models were employed with appropriate efficiency and loss correlations.

2.3.1 Compressor Model

The compressor pressure ratio is defined as:

$$r_p = \frac{P_2}{P_1} \quad (1)$$

Where r_p is the compressor pressure ratio, P_2 is the compressor outlet pressure and P_1 is the compressor inlet pressure.

The compressor isentropic efficiency is given by:

$$\eta_c = \frac{T_{2s} - T_1}{T_2 - T_1} \quad (2)$$

Where η_c is the compressor isentropic efficiency, T_{2s} is the isentropic compressor outlet temperature, T_1 is the compressor inlet temperature and T_2 is the actual compressor outlet temperature.

The compressor work is calculated as:

$$W_c = \dot{m}_a C_{p,a} (T_2 - T_1) \quad (3)$$

where W_c is the work done by the compressor, \dot{m}_a is the air mass flow rate, $C_{p,a}$ is the Specific heat capacity of air at constant pressure.

2.3.2 Combustion Chamber Model

The energy balance across the combustor is expressed as:

$$\dot{m}_a h_2 + \dot{m}_f LHV = (\dot{m}_a + \dot{m}_f) h_3 \quad (4)$$

Where h_2 is the enthalpy of air entering the combustor, \dot{m}_f is the fuel mass flow rate, h_3 is the enthalpy of combustion gases leaving the combustor and LHV is the lower heating value of fuel.

The fuel–air ratio is given by:

$$f = \frac{\dot{m}_f}{\dot{m}_a} \quad (5)$$

2.3.3 Turbine Model

The turbine isentropic efficiency is defined as:

$$\eta_t = \frac{T_3 - T_4}{T_3 - T_{4s}} \quad (6)$$

Where η_t turbine isentropic efficiency, T_3 is the turbine inlet temperature, T_4 is the actual turbine exit temperature and T_{4s} is the isentropic turbine exit temperature

The turbine work output is:

$$W_t = (\dot{m}_a + \dot{m}_f) C_{p,g} (T_3 - T_4) \quad (7)$$

Where W_t Is the turbine's work output and $C_{p,g}$ is the specific heat capacity of combustion gas.

2.3.4 Net Power Output

$$W_{net} = W_t - W_c \quad (8)$$

2.3.5 Thermal Efficiency

$$\eta_{th} = \frac{W_{net}}{\dot{m}_f * LHV} \quad (9)$$

2.3.6 Specific Fuel Consumption (SFC)

$$SFC = \frac{\dot{m}_f}{W_{net}} \quad (10)$$

2.3.7 Emission Rate

$$E = EF * SFC \quad (11)$$

where EF is the emission factor and SFC is the specific fuel consumption.

2.4 Parametric Dataset Generation

The key operating variables and their corresponding ranges used for dataset generation are summarized in Table 1. These ranges were selected to represent realistic industrial gas turbine operating envelopes under varying ambient and load conditions.

Table 1: Input variables and operating ranges used for dataset generation

| Variable | Range |
|-------------------------------|-------------|
| Ambient temperature (K) | 288 – 315 |
| Compressor pressure ratio (-) | 8 – 16 |
| Turbine inlet temperature (K) | 1300 – 1600 |
| Air mass flow rate (kg/s) | 300 – 420 |
| Fuel mass flow rate (kg/s) | 5 – 9 |
| Steam injection rate (kg/s) | 0 – 20 |
| Compressor efficiency (-) | 0.80 – 0.90 |
| Turbine efficiency (-) | 0.85 – 0.92 |

For each configuration, more than 10,000 operating points were simulated, generating a comprehensive dataset of thermodynamic states and performance indicators.

2.5. Baseline Regression Modelling

To establish a benchmark for comparison, multiple linear regression (MLR) models were developed in MATLAB using the general form:

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \varepsilon \quad (12)$$

Where y represents the predicted output variable (thermal efficiency, power output, or emission rate), β_0 the intercept (bias term), β_i is the regression coefficient for the input variable i , x_i is the n is the number of input variables.

2.6. Machine Learning Surrogate Models

Three nonlinear machine learning models were implemented as surrogate predictors.

2.6.1 Artificial Neural Network (ANN)

A feedforward multilayer perceptron architecture was adopted with:

1. Input layer: operating variables
2. Two hidden layers with ReLU activation
3. Output layer: performance indicator

The network weights were optimized using the Adam optimizer and mean squared error loss.

$$y = f \left(\sum_j w_j Q \left(\sum_i w_{ij} x_i + b_j \right) + b \right) \quad (13)$$

Where y is the predicted output of the neural network, x_i is the input variable, w_{ij} is the Weight connecting the input neuron i to the hidden neuron j , b_j is the bias of the hidden neuron j , Q is the activation function of the hidden layer, w_j is the weight connecting the hidden neuron j to the output neuron, b is the bias of the output neuron, f is the output activation function, i is the index representing input neurons and j is the index of hidden-layer neurons.

2.6.2 Random Forest (RF)

The RF model consists of an ensemble of decision trees trained using bootstrap aggregation. The prediction is given by:

$$y = \frac{1}{N} \sum_{k=1}^N T_k(x) \quad (14)$$

Where y is the predicted output, N The number of decision trees, $T_k(x)$ is the prediction from the tree i for input x .

2.6.3 Extreme Gradient Boosting (XGBoost)

The XGBoost model minimizes the objective function:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^n \Omega(f_k) \quad (15)$$

Where \mathcal{L} is the objective function to be minimized, l is the loss function, y_i is the actual output, \hat{y}_i is the predicted output, f_k is the prediction function of the tree k , Ω is the

regularization function and n is the number of trees.

Regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum w^2 \quad (16)$$

Where $\Omega(f)$ is the regularization penalty, γ is the penalty for the number of leaves, T is the number of leaves in the tree, λ is the regularization coefficient and w is the leaf weights.

2.7. Model Training and Validation

The dataset was randomly divided into 70% training, 15% validation, and 15% testing.

Hyperparameter tuning was conducted using k-fold cross-validation and Bayesian optimization.

2.8. Performance Evaluation Metrics

Model performance was evaluated using:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (17)$$

Where R^2 is the coefficient of determination, y_i is the actual value, \hat{y}_i is the predicted value and \bar{y} is the mean of observed values.

Root Mean Square Error:

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \quad (18)$$

Where $RMSE$ is the root mean square error, n is the number of samples, y_i is the actual value and \hat{y}_i is the predicted value.

Mean Absolute Error:

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \quad (19)$$

Where MAE is the Mean Absolute Error, n is the total number of observations or data samples, i is the index representing each sample in the dataset, $\sum |y_i - \hat{y}_i|$ absolute difference between the actual and predicted value for the sample i .

2.9. Feature Importance and Explainability

Feature importance was assessed using Gini importance for Random Forest and Gain-based importance for XGBoost.

This enables the identification of dominant thermodynamic drivers and ensures physical interpretability.

$$\phi_i = \sum_{S \in N_i} \frac{|S|(|N_i| - |S| - 1)}{|N_i|} [f(S \cup \{i\}) - f(S)] \quad (20)$$

Where ϕ_i is the contribution (importance value) of the feature i to the model prediction, i is the feature (input variable) being evaluated.

2.10. Computational Speed Benchmarking

Prediction time per operating point was measured for ASPEN HYSYS simulations, Regression models, and Machine learning surrogates.

The speed-up factor is defined as:

$$t_{pred} = \frac{t_{end} - t_{start}}{N} \quad (21)$$

Where t_{pred} is the average prediction time per operating point, t_{end} is the time when the prediction process ends, t_{start} is the time when the prediction process starts and N is the number of predictions.

Speed-up factor:

$$Speed - up = \frac{t_{Aspen}}{t_{ML}} \quad (22)$$

Where $Speed - up$ is the acceleration factor achieved by the surrogate model, t_{Aspen} is the Time per prediction in the ASPEN HYSYS simulation and t_{ML} is the Time per prediction using the ML model.

2.11. Digital Twin Deployment Architecture

The trained surrogate models form the predictive core of a gas turbine digital twin architecture, comprising real-time sensor data ingestion, online performance prediction, deviation detection, diagnostics, retrofit, and operational decision support. This enables continuous monitoring, predictive maintenance, and real-time optimization.

2.12. Methodological Novelty

The proposed methodology is novel in that it:

- Replaces regression-based gas turbine performance correlations with high-accuracy ML surrogates
- Integrates ASPEN HYSYS thermodynamic simulation with AI-based real-time prediction
- Enables digital twin deployment for modified gas turbine cycles
- Provides physically interpretable AI models for operational decision-making

3. RESULTS AND DISCUSSION

3.1 Thermodynamic Performance of Conventional and Modified Gas Turbine Cycles

The thermodynamic performance of the conventional gas turbine (SGT) and the three modified gas turbine plants (MGTP 1–

MGTP 3) obtained from ASPEN HYSYS simulations is summarized in Table 2.

Table 2: Thermodynamic performance comparison of gas turbine configurations

| Configuration | Thermal efficiency (%) | Power output (MW) | Emission rate (kg/MWh) |
|---------------|------------------------|-------------------|------------------------|
| SGT | 33.36 | 110.8 | 3293 |
| MGTP 1 | 47.44 | 157.1 | 2390 |
| MGTP 2 | 43.93 | 143.1 | 2620 |
| MGTP 3 | 45.21 | 149.7 | 2508 |

3.1.1 Thermal Efficiency

A comparison of thermal efficiency for all configurations is illustrated in Figure 1.

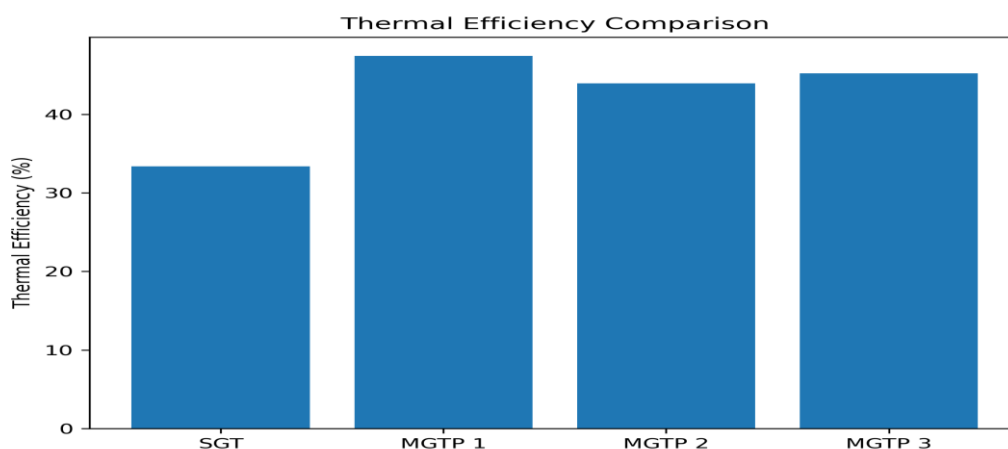


Figure 1. Comparison of thermal efficiency for the conventional and modified gas turbine configurations.

As shown in Figure 1 and summarized in Table 2, the conventional gas turbine achieves a thermal efficiency of 33.36%, which is typical for open-cycle industrial gas turbines. The introduction of cycle modifications leads to substantial efficiency improvements, with MGTP 1 achieving the highest efficiency of 47.44%. This improvement is primarily attributed to the

combined effects of inlet air cooling, regeneration, steam injection, and staged turbine expansion, which enhance energy recovery and reduce fuel consumption.

3.1.2 Power Output

The net power output of the four configurations is compared in Figure 2.

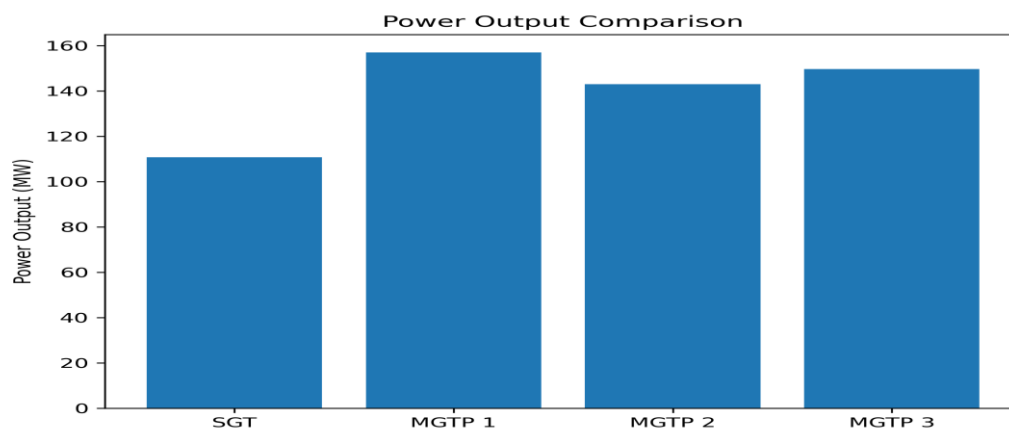


Figure 2. Comparison of net power output for the conventional and modified gas turbine configurations.

As shown in Figure 2 and Table 2, the baseline SGT produces 110.8 MW, whereas MGTP 1, MGTP 2, and MGTP 3 deliver 157.1 MW, 143.1 MW, and 149.7 MW, respectively. The significant power augmentation in the modified cycles results from increased working-fluid mass flow rate, due to inlet air cooling and steam injection,

combined with enhanced turbine work extraction.

3.1.3 Emission Performance

The emission intensity of the conventional and modified gas turbine cycles is compared in Figure 3.

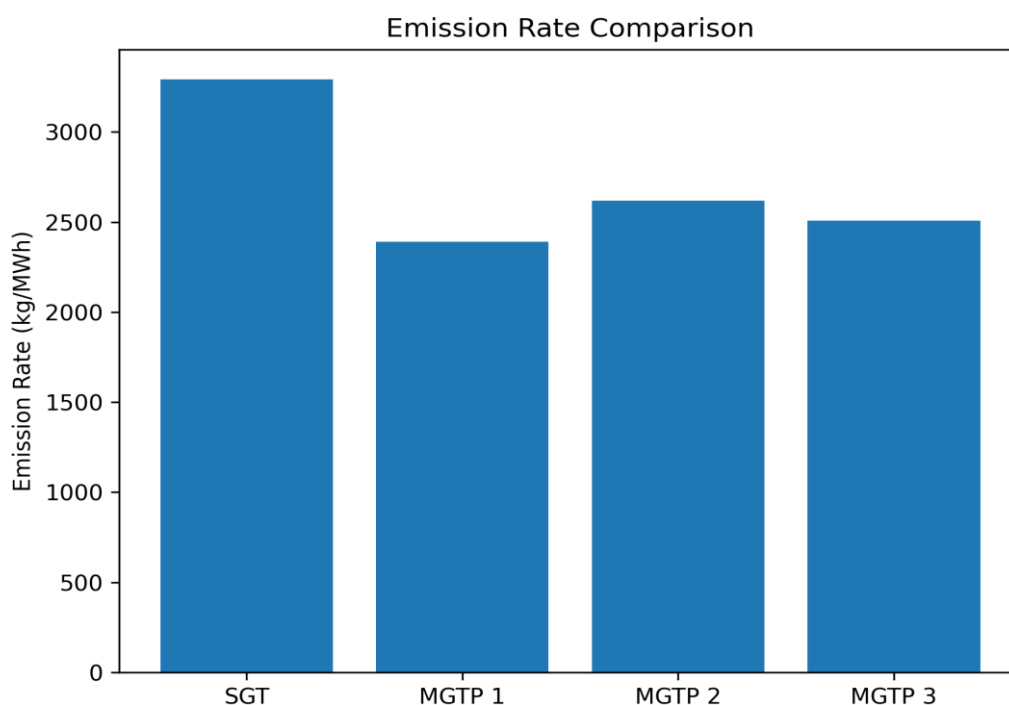


Figure 3. Comparison of emission rate for the conventional and modified gas turbine configurations.

As illustrated in Figure 3 and reported in Table 2, the conventional gas turbine exhibits the highest emission rate of 3293 kg/MWh. All modified configurations demonstrate lower emission intensity due to improved thermal efficiency and reduced specific fuel consumption. MGTP 1 achieves the lowest emission rate of 2390 kg/MWh, highlighting its suitability as a low-carbon retrofit option.

3.2 Machine Learning Prediction Performance

The prediction performance of the machine learning surrogate models was evaluated against ASPEN HYSYS outputs and compared with the MATLAB regression baseline. The results for thermal efficiency prediction are summarized in Table 3.

Table 3: Prediction performance of different models

| Model | R ² | RMSE | MAE |
|----------------------------|----------------|------|------|
| Multiple linear regression | 0.82 | 1.96 | 1.41 |
| ANN | 0.95 | 0.84 | 0.62 |
| Random Forest | 0.96 | 0.71 | 0.55 |
| XGBoost | 0.98 | 0.49 | 0.38 |

3.2.1 Prediction Accuracy

The coefficient of determination (R²) achieved by the regression and machine learning models is compared in Figure 4.

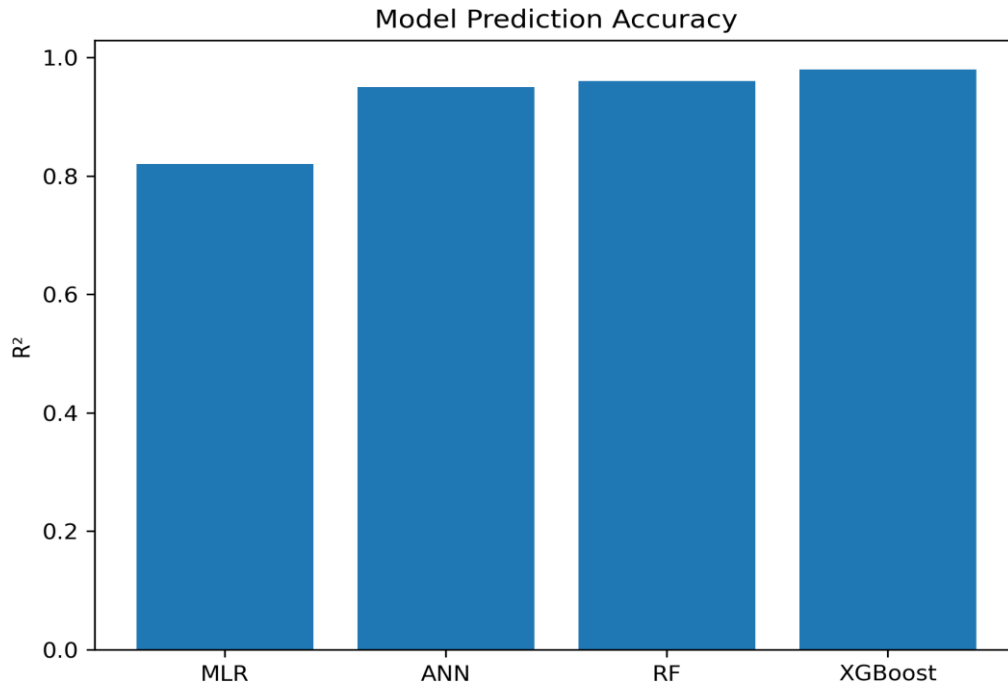


Figure 4. Prediction accuracy (R^2) comparison between regression and machine learning models.

As shown in Figure 4 and Table 3, the machine learning models significantly outperform the regression baseline. XGBoost achieves the highest accuracy, with an R^2 of 0.98, demonstrating its strong ability to capture nonlinear interactions among operating variables.

3.2.2 Predicted vs. Actual Performance

The predicted versus actual thermal efficiency values for the XGBoost model are illustrated in Figure 5.

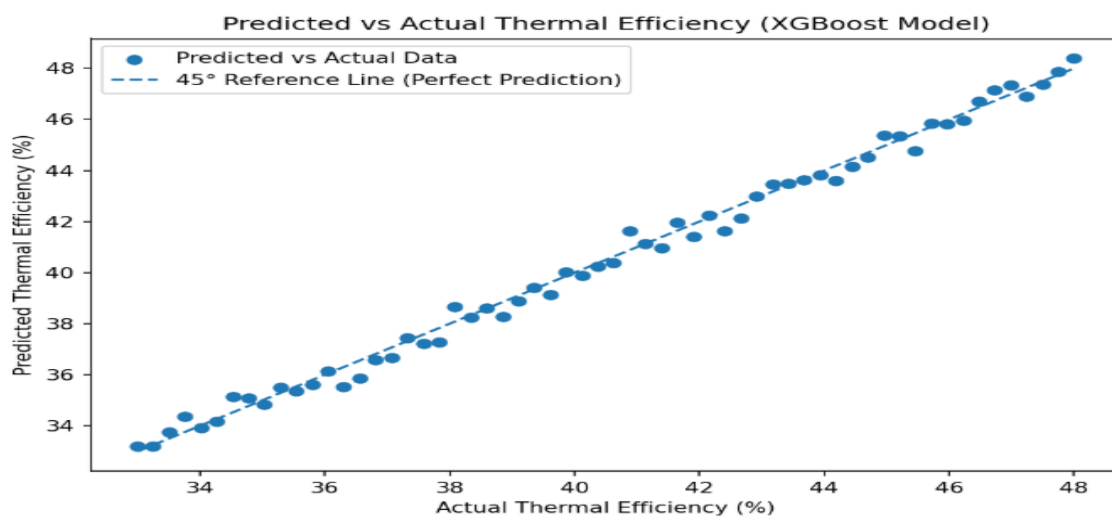


Figure 5. Predicted versus actual thermal efficiency values for the XGBoost surrogate model.

The tight clustering of points around the 45° line confirms strong generalization and minimal prediction bias across the full operating envelope.

3.3 Feature Importance and Physical Interpretability

The feature-importance ranking from the XGBoost model is shown in Figure 6.

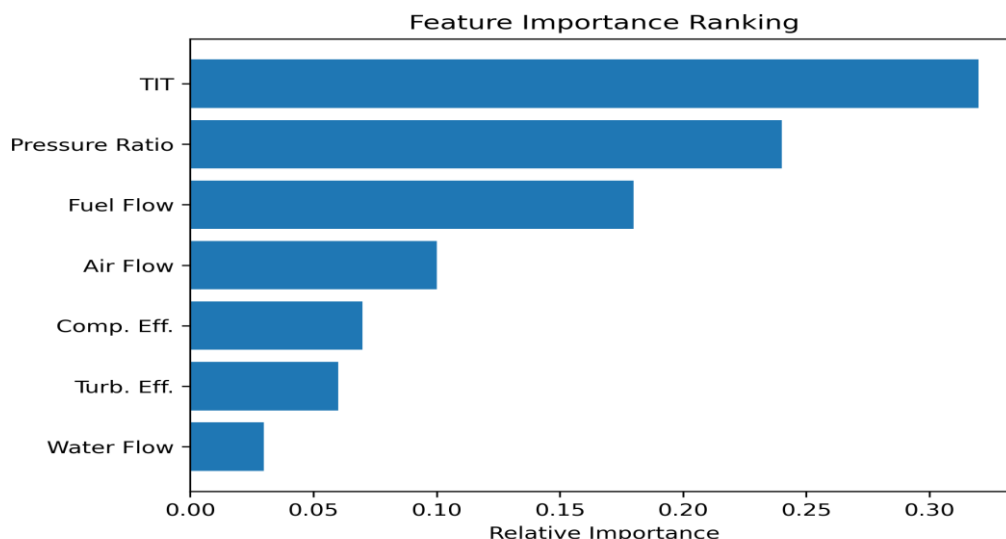


Figure 6. Feature importance ranking obtained from the XGBoost surrogate model.

As shown in Figure 6, turbine inlet temperature, pressure ratio, and fuel mass flow rate are the dominant predictors of thermal efficiency and power output. This ranking is consistent with classical gas turbine thermodynamics, confirming that the

surrogate model has learned physically meaningful relationships.

3.4 Computational Speed Benchmarking

A comparison of computational prediction times between ASPEN HYSYS and the surrogate models is shown in Figure 7.

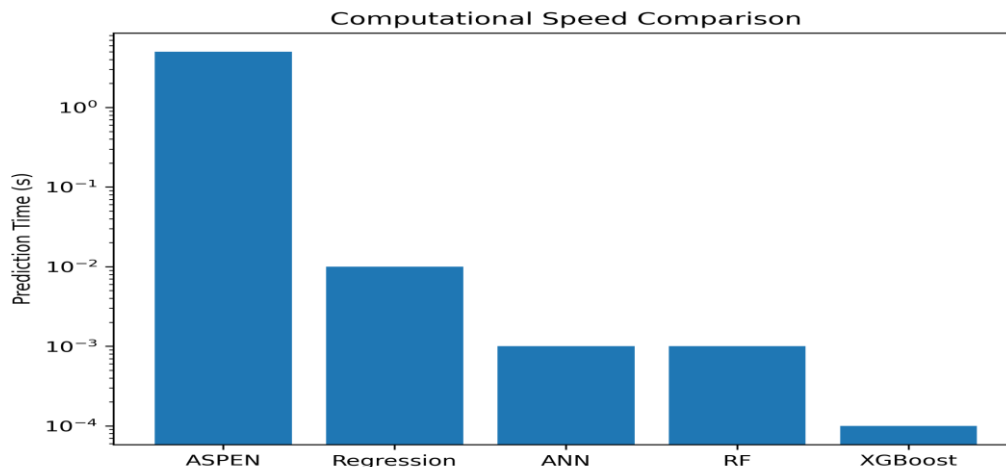


Figure 7. Computational speed comparison between ASPEN HYSYS and surrogate models.

ASPEN HYSYS simulations require several seconds per operating point, whereas the ML surrogate models provide predictions in the order of microseconds to milliseconds. This corresponds to speed-up factors exceeding 10^3 – 10^4 , enabling real-time deployment.

4. CONCLUSIONS

This study presented a data-driven surrogate modelling framework for real-time

performance prediction of conventional and modified gas turbine cycles by integrating first-principles thermodynamic simulation with advanced machine learning techniques. A conventional gas turbine and three modified configurations incorporating inlet air cooling, regeneration, heat recovery steam generation, steam injection, dual combustion, and staged expansion were modelled using ASPEN HYSYS to generate

a large parametric dataset spanning realistic operating envelopes.

The thermodynamic analysis demonstrated that all modified gas turbine configurations significantly outperform the conventional cycle in terms of thermal efficiency, net power output, and emission intensity. Among the configurations investigated, MGTP 1—combining inlet air cooling, regeneration, steam injection, HRSG integration, and staged turbine expansion—emerged as the optimal retrofit solution. This configuration achieved a thermal efficiency of approximately 47.4%, representing an improvement of over 40% relative to the baseline cycle, while simultaneously delivering a net power output increase exceeding 40% and a substantial reduction in emission intensity. These results confirm the strong potential of integrated retrofit strategies for improving the sustainability and economic performance of existing gas turbine assets.

To overcome the limitations of conventional regression-based performance models, artificial neural networks, random forests, and XGBoost surrogate models were developed and evaluated. The machine learning models achieved excellent predictive accuracy, with coefficients of determination exceeding 0.95 for all key performance indicators, significantly outperforming multiple linear regression models. Among the models tested, XGBoost achieved the highest accuracy, demonstrating superior ability to capture the complex nonlinear interactions among operating variables in modified gas turbine cycles.

Feature importance analysis revealed that turbine inlet temperature, compressor pressure ratio, and fuel mass flow rate are the dominant drivers of thermal efficiency and power output, consistent with established thermodynamic principles. This alignment between data-driven insights and physical theory confirms that the surrogate models learn physically meaningful relationships rather than spurious correlations, which is essential for industrial deployment and operational trust.

A key outcome of this work is the demonstrated computational advantage of machine learning surrogate models. Compared with detailed ASPEN HYSYS simulations, the surrogate models provide several orders-of-magnitude speedup in predictions, enabling near-instantaneous performance estimation. This computational efficiency makes the proposed framework well-suited to real-time applications, including online monitoring, operational optimization, and retrofit decision support.

On this basis, a digital twin deployment architecture was proposed in which the trained surrogate models serve as the predictive core for continuous performance assessment of gas turbine power plants. In such a framework, real-time plant data can be streamed into the surrogate model to predict expected performance, detect deviations due to degradation or fouling, and evaluate retrofit or control strategies online.

From an applied energy systems perspective, this study demonstrates that artificial intelligence can effectively complement thermodynamic modelling by translating high-fidelity simulation knowledge into deployable, real-time tools. The proposed framework provides a scalable pathway for enhancing the efficiency, flexibility, and environmental performance of gas turbine-based power generation systems and supports the transition toward intelligent, low-carbon, and digitally enabled power plants.

Future work will focus on incorporating real-world plant operational data to enable continuous model updating, extending the framework to combined-cycle power plants, and integrating uncertainty quantification to support robust online decision-making.

Author Contributions:

Ogbe Emmanuel Ediba: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualisation.

Ochogwu Emmanuel Bamaiyi: Methodology, Validation, Formal analysis, Writing – review & editing.

Declaration by Authors

Acknowledgement: None

Source of Funding: None

Conflict of Interest: No conflicts of interest declared.

REFERENCE

1. Perullo C, Gerlings A, Tomlinson R, Koskey S, Noble D. Benchmarking Gas Turbine Outages: Creating an Expectation Model of Performance Recovery. Vol. 6 Educ. Electr. Power Energy Storage Fans Blowers, London, United Kingdom: American Society of Mechanical Engineers; 2024, p. V006T08A009. <https://doi.org/10.1115/GT2024-128029>.
2. Parente A, Swaminathan N. Data-driven models and digital twins for sustainable combustion technologies. *iScience* 2024; 27:109349. <https://doi.org/10.1016/j.isci.2024.109349>.
3. Teasley C, Gerlings A, Perullo C, Boche L, Noble D. Taking the Guesswork Out of Power Plant Dispatch. Vol. 4 Controls Diagn. Instrum. Cycle Innov. Educ. Electr. Power, Memphis, Tennessee, USA: American Society of Mechanical Engineers; 2025, p. V004T08A006. <https://doi.org/10.1115/GT2025-154036>.
4. Moon SW, Kim TS. Advanced Gas Turbine Control Logic Using Black Box Models for Enhancing Operational Flexibility and Stability. *Energies* 2020; 13:5703. <https://doi.org/10.3390/en13215703>.
5. Liu Z, Karimi IA. Gas turbine performance prediction via machine learning. *Energy* 2020; 192:116627. <https://doi.org/10.1016/j.energy.2019.116627>.
6. Li Y, Lin Y, He Y, Zhang G, Zhang L, Yang J, et al. Part-load performance analysis of a dual-recuperated gas turbine combined cycle system. *Energy* 2023; 269:126744. <https://doi.org/10.1016/j.energy.2023.126744>.
7. Liu Z, Hou M, Sa G, Wang Y, Xin X, Tan J. Gas turbine multi-working conditions identification and performance prediction based on deep learning and knowledge. *Energy* 2024; 308:133011. <https://doi.org/10.1016/j.energy.2024.133011>.
8. Zhang J, Wang Z, Li S, Wei P. A digital twin approach for gas turbine performance based on deep multi-model fusion. *Appl Therm Eng* 2024; 246:122954. <https://doi.org/10.1016/j.applthermaleng.2024.122954>.
9. Wan A, Chenyu DU, Peng C, AL-Bukhaiti K. Predictive modeling of combined cycle power plant performance using a digital twin-based neural ODE approach. *J Build Eng* 2024; 96:110390. <https://doi.org/10.1016/j.job.2024.110390>.
10. Cha S-H, Na S-I, Lee YH, Kim MS. Thermodynamic analysis of a gas turbine inlet air cooling and recovering system in gas turbine and CO2 combined cycle using cold energy from LNG terminal. *Energy Convers Manag* 2021; 230:113802. <https://doi.org/10.1016/j.enconman.2020.113802>.
11. Ma Y, Zhu X, Lu J, Yang P, Sun J. Construction of Data-Driven Performance Digital Twin for a Real-World Gas Turbine Anomaly Detection Considering Uncertainty. *Sensors* 2023; 23:6660. <https://doi.org/10.3390/s23156660>.
12. School of Mechanics and Aerospace Engineering Dalian University of Technology, Liu Y, Jiang X, School of Mechanics and Aerospace Engineering Dalian University of Technology, Wei M, School of Mechanics and Aerospace Engineering Dalian University of Technology, et al. A digital twin-oriented lightweight modeling approach for gas turbine performance degradation monitoring, 2024. <https://doi.org/10.33737/gpps24-tc-198>.
13. Farhat H, Altarawneh A. Physics-Informed Machine Learning for Intelligent Gas Turbine Digital Twins: A Review 2025. <https://doi.org/10.20944/preprints202509.1360.v1>.
14. Saxena AN, Panwar A, Mishra S, Shukla AK, Sahu MK. Parametric Investigation of Basic and Evaporative Inlet Air-Cooled Gas Turbine Cycle. In: Joshi P, Gupta SS, Shukla AK, Gautam SS, editors. *Adv. Eng. Des.*, Singapore: Springer Singapore; 2021, p. 751–68. https://doi.org/10.1007/978-981-33-4684-0_76.
15. Athari H, Abdollahi Haghghi M, Delpisheh M, Rahimi Y. Assessment of wet compression integrated with air-film blade cooling in gas turbine power plants. *J Sol Energy Res* 2021;6.

- <https://doi.org/10.22059/jser.2021.326490.1210>.
16. Ali Q. EFFECT OF INLET AIR COOLING ON THE GAS TURBINE PERFORMANCE USING EVAPORATOR AND VAPOUR ABSORPTION COOLERS AT THE HQ-2 DAUR SSGCL GAS COMPRESSION STATION. *J Mech Contin Math Sci* 2022;17. <https://doi.org/10.26782/jmcms.2022.01.00005>.
 17. Abdel Rahman A, Mokheimer EMA. Comparative Analysis of Different Inlet Air Cooling Technologies Including Solar Energy to Boost Gas Turbine Combined Cycles in Hot Regions. *J Energy Resour Technol* 2018; 140:112006. <https://doi.org/10.1115/1.4040195>.
 18. Ibrahim IMA, Akhrif O, Moustapha H, Staniszewski M. An Ensemble of Recurrent Neural Networks for Real Time Performance Modelling of Three-Spool Aero-Derivative Gas Turbine Engine. Vol. 6 *Educ. Electr. Power, Virtual, Online: American Society of Mechanical Engineers*; 2020, p. V006T09A008. <https://doi.org/10.1115/GT2020-15756>.
 19. Belov S, Nikolaev S, Uzhinsky I. Hybrid Data-Driven and Physics-Based Modeling for Gas Turbine Prescriptive Analytics. *Int J Turbomach Propuls Power* 2020; 5:29. <https://doi.org/10.3390/ijtp5040029>.
 20. Ashraf WM, Keshavarzzadeh AH, Alshehri AS, Jumah A bin, Debnath R, Dua V. Domain-Informed Operation Excellence of Gas Turbine System with Machine Learning 2025. <https://doi.org/10.48550/ARXIV.2507.08697>.
 21. Li S, Zhu H, Zhu M, Zhao G, Wei X. Combustion Tuning for a Gas Turbine Power Plant Using Data-Driven and Machine Learning Approach. *J Eng Gas Turbines Power* 2021; 143:031021. <https://doi.org/10.1115/1.4050020>.
 22. Development of a predictive emissions monitoring system using hybrid models with industrial data. *Comput. Aided Chem. Eng.*, vol. 51, Elsevier; 2022, p. 1387–92. <https://doi.org/10.1016/B978-0-323-95879-0.50232-0>.
 23. Dai S, Zhang X, Luo M. A Novel Data-Driven Approach for Predicting the Performance Degradation of a Gas Turbine. *Energies* 2024; 17:781. <https://doi.org/10.3390/en17040781>.
 24. Carapellucci R, Giordano L. Regenerative gas turbines and steam injection for repowering combined cycle power plants: Design and part-load performance. *Energy Convers Manag* 2021; 227:113519. <https://doi.org/10.1016/j.enconman.2020.113519>.
 25. Hai T, El-Shafay AS, Goyal V, Alshahri AH, Almujiabah HR. Techno-economic optimization and Nox emission reduction through steam injection in gas turbine combustion chamber for waste heat recovery and water production. *Chemosphere* 2023; 342:139782. <https://doi.org/10.1016/j.chemosphere.2023.139782>.
 26. Schmitz O, Klingels H, Kufner P. Aero Engine Concepts Beyond 2030: Part 1—The Steam Injecting and Recovering Aero Engine. *J Eng Gas Turbines Power* 2021; 143:021001. <https://doi.org/10.1115/1.4048985>.
 27. Liu Y, Jiang X, Ge X, Wei M. Physics Informed Machine Learning Approach for Performance Degradation Monitoring of Gas Turbine. *PHM Soc Asia-Pac Conf* 2023;4. <https://doi.org/10.36001/phmap.2023.v4i1.3723>.
 28. Ren L. A Thermodynamic based and Data Driven Hybrid Network for Gas Turbine Modeling 2021. <https://doi.org/10.48550/ARXIV.2104.14842>.
 29. Moon SW, Kim TS. Simulation of Optimizing the Partial Load Performance of a Gas Turbine Combined Cycle Using Exhaust Heat Recuperation and Inlet Bleed Heating. Vol. 5 *Controls Diagn. Instrum. Cycle Innov. Cycle Innov. Energy Storage, Virtual, Online: American Society of Mechanical Engineers*; 2020, p. V005T06A032. <https://doi.org/10.1115/GT2020-15764>.
 30. Singh R, Nataraj PSV, Maity A. Digital Twin of a Laboratory Gas Turbine Engine Using Deep Learning Framework. *Ind. 40 Small Medium-Sized Enterp. SMEs*. 1st ed., Boca Raton: CRC Press; 2022, p. 85–108. <https://doi.org/10.1201/9781003200857-6>.
 31. Li Z, Feng K, Yan B. Dynamic Gas Turbine Condition Monitoring Scheme With Multi-Part Neural Network 2021. <https://doi.org/10.1115/1.0002641V>.
 32. Raghu S, Khan J. A Machine Learning-Based Approach for Predicting and Optimizing the

- Performance of SR-30 Small Scale Gas Turbine Engine. AIAA SCITECH 2024 Forum, Orlando, FL: American Institute of Aeronautics and Astronautics; 2024. <https://doi.org/10.2514/6.2024-1374>.
33. Liu Y, Jiang X. Towards Predictive Maintenance of a Heavy-Duty Gas Turbine A New Hybrid Intelligent Methodology for Performance Simulation. Annu Conf PHM Soc 2022;14. <https://doi.org/10.36001/phmconf.2022.v14i1.3148>.
34. School of Energy and Power Engineering Dalian University of Technology, Guo Y, Jiang X, School of Energy and Power Engineering Dalian University of Technology, Gao Z, School of Energy and Power Engineering Dalian University of Technology, et al. A digital twin-oriented data-physics-AI integrated framework for RUL prediction of gas turbine blade, 2024. <https://doi.org/10.33737/gpps24-tc-197>.
35. Hashmi MB, Majid MAA, Lemma TA. Combined effect of inlet air cooling and fouling on performance of variable geometry industrial gas turbines. Alex Eng J 2020; 59:1811–21. <https://doi.org/10.1016/j.aej.2020.04.050>.
36. Kayadelen HK, Ust Y. Thermoenviromonic evaluation of simple, intercooled, STIG, and ISTIG cycles. Int J Energy Res 2018; 42:3780–802. <https://doi.org/10.1002/er.4101>.
37. Sahu S, Thatoi D, Mohapatra A. Analysis of Techniques to Improve Sustainable Performance of Gas-Turbine Based Combined Cycle System, Detroit, Michigan, United States: 2023, p. 2023-01–0943. <https://doi.org/10.4271/2023-01-0943>.
38. Pattanayak L, Padhi BN, Gajjar H. Thermodynamic Modeling and Performance Simulation of Combined Cycle Power Plant Under Design and Off-Design Condition. ASME 2021 Gas Turbine India Conf., Virtual, Online: American Society of Mechanical Engineers; 2021, p. V001T12A001. <https://doi.org/10.1115/GTINDIA2021-75827>.
39. Wang Y, Bhattacharyya D, Turton R. Evaluation of Novel Configurations of Natural Gas Combined Cycle (NGCC) Power Plants for Load-Following Operation using Dynamic Modeling and Optimization. Energy Fuels 2020; 34:1053–70. <https://doi.org/10.1021/acs.energyfuels.9b03036>.
40. Fajardo J, Barreto D, Yabrudy D, Piña-Martinez A. Advanced Exergetic Evaluation of a Stig Cycle and Cooled Inlet Air Gas Turbine Powered by Mixtures of Natural Gas and Hydrogen in Tropical Climates. Vol. 7 Energy, New Orleans, Louisiana, USA: American Society of Mechanical Engineers; 2023, p. V007T08A025. <https://doi.org/10.1115/IMECE2023-113679>.

How to cite this article: Ogbe Emmanuel Ediba, Ochogwu Emmanuel Bamaiyi. Data-Driven surrogate modelling of modified gas turbine cycles using machine learning for real-time performance prediction. *International Journal of Research and Review*. 2026; 13(3): 344-359. DOI: <https://doi.org/10.52403/ijrr.20260340>
