JEEE-U
Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

A Data-Driven Digital Twin for Gas Turbine Power Output Estimation and Performance Deviation Monitoring

Digital Twin Berbasis Data untuk Estimasi Keluaran Daya Turbin Gas dan Pemantauan Penyimpangan Kinerja

Asaad Saber Hamad Ameen [1]*, Ibrahim Ismail Hamarash[2]
[1,2] *Department of Electrical Engineering, Salahaddin University, Iraq*

[1]*engasaad86@gmail.com*
[2]*ibrahim.hamarash@su.edu.krd*

Abstract _ *Digital twin which creates virtual representation of physical systems using real-time data are increasingly employed for performance monitoring, operational optimization, and decision support in the energy sector in power generation , the electrical power output in gas turbines varies significantly with environmental and operational conditions. Accurate estimation of the expected power output (MW) is indispensable for benchmarking and degradation detection. This study proposes a data-driven digital twin framework for predicting power output and monitoring performance deviation for an operational GE Frame 9E gas turbine power plant. Real-time data is acquired from a Mark VIe control system and archived in the AVEVA PI historian (PI Process Book).Four machine learning algorithms, Support Vector Regression (SVR), Random Forest regression, Extreme Gradient Boosting, and a feed-forward Artificial Neural Network (ANN) were implemented and evaluated to find the most suitable model for the digital twin, The XGBoost-based digital twin achieved the best overall accuracy (RMSE = 1.4056 MW and R^2 = 0.9125) and generated the most stable deviation behavior, supporting performance benchmarking in industrial operation. By calculating the deviation (DMW) between actual and expected power production, the digital twin makes it possible to continuously and discreetly monitor changes im performance. This paper offers a framework for deploying digital twins for condition monitoring in gas turbine operations by combining real-time data with machine learning models.*

Keywords: *Digital Twin (DT), Gas Turbine (GT), Machine Learning, Power Output Estimation, Industry 4.0, anomaly detection*

Abstrak_Digital twin, yang menciptakan representasi virtual dari sistem fisik menggunakan data waktu nyata, semakin banyak diterapkan untuk pemantauan kinerja, optimasi operasional, dan dukungan pengambilan keputusan di sektor energi. Dalam pembangkitan tenaga listrik, keluaran daya listrik turbin gas sangat bervariasi akibat kondisi lingkungan dan operasional. Oleh karena itu, estimasi yang akurat terhadap keluaran daya yang diharapkan sangat penting untuk pembandingan kinerja dan deteksi dini degradasi. Penelitian ini mengusulkan kerangka kerja digital twin berbasis data untuk memprediksi keluaran daya dan memantau penyimpangan kinerja pada pembangkit listrik turbin gas GE Frame 9E yang sedang beroperasi. Data waktu nyata diperoleh dari sistem kontrol Mark VIe dan diarsipkan dalam AVEVA PI historian melalui PI ProcessBook. Empat algoritma pembelajaran mesin, yaitu Support Vector Regression (SVR), Random Forest Regression, Extreme Gradient Boosting (XGBoost), dan feed-forward Artificial Neural Network (ANN), diterapkan dan dievaluasi untuk mengidentifikasi model yang paling sesuai bagi digital twin. Digital twin berbasis XGBoost mencapai kinerja keseluruhan terbaik, dengan RMSE sebesar 1,4056 MW dan R^2 sebesar 0,9125, serta menghasilkan perilaku penyimpangan yang paling stabil untuk pembandingan kinerja industri. Dengan menghitung penyimpangan antara produksi daya aktual dan yang diharapkan, digital twin yang diusulkan memungkinkan pemantauan perubahan kinerja secara berkelanjutan dan tidak langsung. Makalah ini menyajikan kerangka kerja praktis untuk penerapan digital twin berbasis pembelajaran mesin dalam pemantauan kondisi dan deteksi anomali pada operasi turbin gas.

Kata Kunci: Digital Twin; Turbin Gas; Pembelajaran Mesin; Estimasi Keluaran Daya; Industri 4.0; Deteksi Anomali

I. INTRODUCTION

Digital Twin (DT) was introduced by Dr Michael Grieves in 2003 as a virtual counterpart of a physical asset when he is enabling data flow between real and virtual spaces. it can be defined as a dynamic virtual representation of a physical asset that continuously mirrors its behavior using real-time or historical data[1]. Recently, digital twin have seen rapid adoption across various sectors, including renewable energy systems[2].

Gas turbine power plants play a significant role in modern power systems due to their high flexibility, fast response, and capability to work under a wide range of load and weather conditions. However, the performance of gas turbines, particularly net power output, varies significantly with ambient temperature, pressure, humidity, and operating parameters. These variations make it challenging for operators to distinguish between expected performance changes and actual performance degradation caused by factors such as compressor fouling, inlet losses, or component aging[3, 4].

Traditionally, the performance monitoring on gas turbine power plants has depended on physics-based models, correction curves, or reference performance maps provided by original equipment manufacturers (OEMs). While these approaches are valuable, they often require detailed design information, simplifying assumptions, and frequent recalibration. Moreover, they may not fully capture the complex nonlinear interactions present in real industrial operation, especially when plants operate far from design conditions or experience gradual degradation[3, 5].

Within the context of high-performance industry 4.0 technologies such as Cyber-Physical Production Systems CPPS system, the internet of things (IOT), and big data analysis, data-driven technologies in the power generation sector, digital twin have gained significant attention in the power generation industry. in gas turbines, a performance digital twin focuses on estimating the expected power output under given operating and ambient conditions, rather than predicting future values or directly controlling the plant[6, 7] [8, 9].

Significant research has studied data-driven digital twin-based machine learning methods for smart manufacturing systems and predicting power in combined cycle power plants. Power output and gas emissions have been modelled with statistical methods: linear regression, multilayer perception, K-nearest neighbors, artificial neural networks (ANN), and ARIMA. Analysis also showed that nonlinear models like ANN and ensemble methods such as Random Forests usually perform better than simple regression techniques, due to their ability to capture complete interactions between features [10-12]. Some neural network-based models, such as recursive neural networks (RNNs) and transformer-based deep neural networks, have also been used for enhancing predictive performance. Specifically, both models have been shown to be highly effective at capturing temporal patterns and at modeling the nonlinear nature of operations data across large [13, 14]. It was shown that ensemble and hybrid models like Support Vector Regression (SVR), Gaussian Process Regression (GPR), Decision Trees (DT), Gradient-boosted

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

Generalized Additive Models (GAM) and Backpropagation Neural Networks (BPNN) outperforming standard regression methods on high-dimensional or non-linear datasets [15-17]. The use of ML with CCPP power is doubtful because the features that contribute to output power levels are nonlinear. Researchers recommended that features such as Ambient Temperature, Barometric Pressure, Relative Humidity, and Vacuum are useful in the prediction of a combined cycle power output. Several studies have modeled these features and quantified their prevalence with varying degrees of success rate depending on the accuracy and quality of the dataset [18].

The digital twin modeling analysis for power plant performance optimizations is proposed for a power plant smart management in China[19]. The framework for the data-driven approach that drives advancements in machine learning and process data-mining techniques, as well as continuous model improvement and validation[20].

The literature demonstrates that machine-learning-based digital twins have several benefits. They may be implemented without requiring invasive changes to current control systems, understand intricate nonlinear relationships directly from plant data, and adapt to site-specific features. When integrated with plant historians, such digital twins can operate in parallel with the physical plant, providing continuous performance benchmarking and deviation analysis[21-23]. However, integrating these models into a persistent, online digital twin using real-time data historian systems remains a practical challenge and a key research gap.

In addition, rather than using continuous real industry operating data, the majority of earlier studies depend on benchmark or laboratory-scale datasets. Additionally, only a small amount of research has looked into the integration of online performance digital twins with industrial historian platforms for ongoing monitoring of performance deviations under practical operating conditions. As a result, a workable and scalable digital twin system that can function with real-time industrial data while preserving high prediction accuracy and interpretability is still required.

This paper addresses this gap by presenting the development ad validation of a data-driven performance digital twin for an industrial GE frame 9E gas turbine using real operational data collected from a GE Mark VIe algorithms and archived in the PI historian (PI Process Book).

The main contributions of this work are:

- The development of a non-intrusive data-driven performance digital twin is developed for a real industrial GE Frame 9E gas turbine, using online PI historian data.
- A comparative evaluation of four machine-learning regression techniques—SVR, Random Forest, XGBoost, and ANN—are compared as alternative digital twin core models, , using time- aware testing to maintain realistic operation circumstances.
- To detect performance changes and degradation patterns, a continuous performance deviation monitoring system ($\Delta MW = MW\_actual - MW\_expected$) is applied.
- The application of important future analysis to boost the digital twins' physical sincerity and interpretability in industrial use.

The rest of this paper is organized in the following way. The system description and digital twin architecture is section 2. Section 3 outlines the suggested digital twin architecture and data collection model. Section 4 elaborates on data-driven modeling approach. Section 5 presents the results and performance deviation analysis. Finally, Section 6 concludes the paper and outlines future research directions.

II. SYSTEM DESCRIPTION AND DIGITAL TWIN ARCHITECTURE

1. POWER PLANT DESCRIPTION AND DATA INFRASTRUCTURE

The case study considered in this study is a real gas turbine power plant located in the Kurdistan region, Iraq. The plant is equipped with a modern control and monitoring infrastructure based on the GE Speedtronic Mark VIe control system. The Mark VIe controller controls the entire operation of the turbine itself and the auxiliary systems. It controls the received Natural Gas fuel system air inlet and exhaust system, axial compressor operation, turbine thermal management, and continuously measures all parameters from the turbine, such as temperature, pressure, flow, level, speed, vibration, and switches. Furthermore, Mark Vie controller supervises and monitors the grid and generator synchronization process, the overspeed (over frequency) process, and the startup and shutdown of the gas turbine unit. all historian data can be monitored and archived through the communication protocol PDH (plant Data Highway) in the AVEVA PI historian (PI Process Book). The data from the PI system (see figure 1.) will be used to build a performance model that represents the correlation between the inputs and the power outputs from the synchronous generator.

[Figure 1. about here]

The PI historian provides long-term storage, data integrity, and accessibility for both offline analysis and online monitoring. For the development and validation of the proposed digital twin, online real-time and historical plant data is exported from the PI historian in CSV format. This approach enables seamless integration with data-driven modeling tools while preserving traceability to the original plant measurements.

In the projected framework, the PI system works as the main component of the digital twin, empowering continuous synchronization between the physical plant and its virtual counterpart without interfering with plant control logic.

2. DIGITAL TWIN

A digital twin is a live adaptive computer model of a physical asset that uses real operational data to predict, simulate, supervise and optimize the asset behavior. Unlike the other methods for prediction and simulation a digital twin is continuously updated with the system parameter variations. Allowing it to mirror the current state of its physical counterpart.

Based on the modeling approach, the digital twin is classified into three main models[24, 25]:

2.1. Physics-based digital twin: Rely on the mathematical model to visualize the actual physical behaviors of the machines by using the mathematical engineering equations.

2.2 Data-Driven Digital Twin: learn the behavior of a real system entirely from data using AI and machine learning algorithms.

2.3 Hybrid Digital twin: combines physical asset and data-driven digital twin to use the strength of both models.

3. PROPOSED DIGITAL TWIN ARCHITECTURE AND WORKFLOW

This study focuses on data-driven performance digital twins. The proposed architecture is illustrated in Figure 2. Operates on a three-layer paradigm:

3.1 physical layer: the gas turbine power plant generates real-time data during changing ambient and load conditions.

3.2 Data layer: which contains the Mark VIe automation control system and PI historian data acquisition, communication, and storage for data.

3.3 Virtual layer: is a data-driven digital twin model that estimates the expected net electrical power output based on online operation conditions.

[Figure 2. about here]

The workflow is as follows; The live operational data are obtained from the real physical power plant across the automation control and supervised systems (PI historian) . This data processed through the AI/machine learning algorithms pre-processing procedure for data cleaning, validation, and feature selection for data superiority and reliability. The processed data are used to train machine-learning models that learn the nonlinear relationships between operating conditions and power output, producing an AI-based virtual representation of the plant. When employed, the trained digital twin continuously gets operational data and provides real-time estimates of the power output under normal operation and environmental conditions. The measured plant outputs are compared with these assessed values to measure the deviations, which allow early identification of performance degradation, abnormal behavior or sensor anomalies. The projected workflow is described in Figure 3.

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

[Figure 3. about here]

The intended architecture is used for both offline and online implementation. In off-line scenario, the historical data from the PI system is applied for ML model training, validation, and benchmarking. In the online mode, the digital twin is combined with the PI system to operate in real time, incessantly updating performance estimates and performance deviation signs. This procedure is continuously compared with actual measured power to compute the performance deviation using Equation 1.

$$\Delta MW = MW_{\text{actual}} - MW_{\text{expected}} \quad \dots\dots\dots\dots (1)$$

A deviation value near zero indicates normal operation. Negative performance may signal performance loss (e.g. compressor fouling, inlet pressure losses, or component aging) Positive deviations could indicate favorable environmental or operational circumstances. This deviation signal offers continuous, non- intrusive performance monitoring and degration detection.

4. DIGITAL TWIN ONLINE OPERATION AND SERVICES

The gas turbine unit and the suggested data-driven performance digital twin are intended to run together. Real-time measurements are obtained from the GE Mark VIe control system during online operation and stored in the AVEVA PI historian system via the Plant Data Highway (PDH). The digital twin collects the most recent operational and ambient input data at each sampling instant (4 minutes) and calculates the anticipated net power output under typical operating conditions. A deviation indicator is then calculated by continually comparing the plant's measured net power output with the anticipated expected power output. offers a direct health-related signal for identifying anomalous behavior, performance changes, and possible degradation trends.

In the offline mode, the machine-learning regression models are trained, validated, and benchmarked using historical PI data. The chosen digital twin model can be used in online mode for ongoing performance assessment and deviation tracking. This supports condition monitoring and instantaneous decision-making in industrial operations by allowing operators to differentiate between output changes caused by the environment and the actual performance losses. Table 1 summarizes the services provided in the proposed framework.

Table 1. about here

III. DATA-DRIVEN DIGITAL TWIN MODELING METHODOLOGY

The goal of our data-driven digital twin is to estimate the gas turbine's expected net electrical power output under specific operating and ambient conditions, then utilize this estimate as a benchmark for performance monitoring. Figure 4. Shows the screenshot of the proposed digital twin.

[Figure 4. about here]

For predictive modeling and performance monitoring of industrial gas turbines, this work uses a quantitative data-driven research methodology based on supervised machine learning algorithms. The study creates an intelligent framework for evaluating projected gas turbine power output under various operating conditions by combining nonlinear regression modeling, digital twin principles, and historical operational data analysis.

The digital twin learns a nonlinear mapping function, f1 from a set of x1 input variables to the expected net power output.

This relationship is expressed as:

$$MW_{\text{expected}} = f(\mathbf{x})$$

where x is a vector of environmental and operational input variables, and MW "expected" is the expected net power output as determined by the digital twin. A performance indicator is the discrepancy between the measured and anticipated power output (ΔMW). It formulae are the root of continuous performance monitoring and degradation exposure. The function is learned by machine learning models trained on historical data.

1. Data Collection and Preprocessing

The time- stamped measurements obtained from the Mark VIe control system were collected and archived by the PI historian system, forming the dataset we refer to here as the DT-Gas Dataset. key preprocessing steps ensured data quality are:

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

- Removing non-representative operation times, including travel circumstances, shutdown, and startup. Managing missing numbers and eliminating glaring anomalies.
- All input variables are synchronized and aligned in time.
- filtering steady-state operations to lessen the impact of sudden changes in load.

The operational data were collected directly from the GE Mark VIe control system installed at the industrial gas turbine unit. real-time industrial tags were used to continually archive real-time process metrics using the PI Historian (PI Process Book) platform. Environmental, operational, compressor, and turbine-related metrics captured under actual plant operating conditions during a prolonged operational time are included in the gathered information.

2. Input Feature Selection and Target Variable

The availability of data and the fundamental principles of gas turbine thermodynamics guided the selection of input features. The inputs for the digital twin consist of Table 2.:

- **Ambient parameters:** ambient temperature, ambient pressure, and humidity
- **Operating parameters:** fuel flow and grid frequency
- **Compressor and turbine parameters:** compressor pressure ratio, compressor discharge pressure, and exhaust gas temperature

The target variable for the digital twin is the measured net electrical power output, expressed in megawatts (MW). All error metrics and performance deviation values derived from this target variable are therefore reported in EP (MW).

[Table 2. about here]

In this study, machine learning algorithms act as predictive analytical instruments within the digital twin framework. Their role is to learn the nonlinear relationship between operational input variables and the expected power output, enabling performance estimation and deviation analysis under varying operating conditions.

3. Machine Learning Models

Four machine-learning regression models were implemented and compared.

3.1. Support Vector Regression

SVR is a type of regression that is the most proper algorithm for nonlinear continuous supervised problems with a large scale of data. The (SVR) in sci-kit-learn is a part of the ‘sklearn’. The Support Vector Regression technique is implemented by the SVR module, which is also capable of working with numerous kernel functions, comprising Radial Basis Function (RBF), Polynomial, and linear kernels. Its objective is to fit the data within a specified margin ξ_i of error, as detailed in its standard formulation.[26-28]

3.2 Random Forest Regression

One popular and effective ensemble learning approach for handling classification and regression problems is Random Forest (RF). A flexible ensemble learning technique, Random Forest, is frequently applied to regression and classification issues. To calculate the average predictions at the regression output, training entails building a huge number of decision trees. Every tree in the forest is constructed using a sample of the training set and characteristics to lower the risk of overfitting and increase generality. Regarding a model: When you add randomization to the trees, the ensemble model becomes more durable and accurate than a single decision tree, adding diversity to the model.

Random Forest often performs well for large (in relation to the number of features) dimensional spaces and does a respectable job of capturing non-linear interactions/correlations between the features because a major portion of the data is categorical and contains many characteristics. Additionally, it produces feature importance metrics that help determine which factors have the biggest impact on the model's predictions.

The RF algorithm is detailed in [29, 30].

3.3 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) was employed as the primary digital twin model due to its strong performance in handling nonlinear relationships and noisy industrial datasets. It is a popular gradient boosting method that has shown remarkable performance in numerous ML domains. The time and computational resources required to optimize the model can be significantly decreased by using meta-learning to suggest hyperparameters for XGBoost. The process of determining the optimal hyperparameters for XGBoost can be automated by utilizing meta-learning approaches, resulting in more effective and efficient model optimization. Because of its ease of use, scalability, and capacity to effectively handle both regression and ranking issues utilizing the Gradient Boosting framework, the XGBoost algorithm has gained favor recently[31-33].

3.4 Artificial Neural Network

A neural network is a machine learning technique made up of neurons with highly weighted connections. it consists of three layers: input, hidden, and output. The input layer is where the network receives data from the growth of each information unit. The input from the info layer is linked to a nonlinear change in the hidden layer. The hidden and the input layers, where the outputs are generated, determine how well the output layers perform.

Artificial neural networks (ANNs) have been a prominent approach for gas and steam turbine power plants in recent years. Important ANN steps include building the model and dividing the data into training, testing, and cross-validation (CV). While the training set is used to optimize model variables, the test set is utilized for CV during training to avoid overlaps. Single-layer perception and multi-layer perception are the two types of ANNs. Multi-layer perception is a feedforward neural system in which impulses consistently continue to flow toward the output layer. There is only one concealed layer in a multi-layer perception. When compared to single-layer networks, multi-layer neural networks have greater processing power[28, 34].

The hyperparameters for each models were tuned based on preliminary analysis and best practices, see Table 3.

[Table 3. about here]

4. Model Training and Evaluation Strategy

Time-aware data splitting was used to train all models, with the last 20% data set for testing and the first 80% of the data used for training. This method reflects actual deployment conditions while maintaining the temporal framework of the plant operation. Standard metrics were used to evaluate the model's performance:

Mean Absolute Error (MAE) in MW

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \check{y}_i\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Where n is the number of data, y_i is the actual value, and $\check{y}_i$ is the predicted value.

Root Mean Squared Error (RMSE) in MW

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \check{y}_i\right)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \check{y}_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

Where $\bar{y}$ is the actual value

Preprocessing the operational dataset, training machine learning models utilizing time-aware data splitting, producing forecasts for predicted power production, and assessing prediction performance using statistical error measures make up the whole data analysis pipeline. To find the most reliable and accurate digital twin model for industrial deployment, a comparative analysis of the implemented models was carried out.

IV. RESULTS AND DISCUSSIONS

The performance of the proposed data-driven digital twin models was evaluated using a time-aware test dataset comprising the final 20% of the available operational data. This evaluation strategy reflects realistic deployment conditions, where the digital twin is trained on historical data and applied to later operating periods.

Table 4 recaps the prediction performance of the four investigated machine-learning techniques in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). Since the target variable is net electrical power output, all error metrics are reported in megawatts (MW).

[Table 4 about here]

The proposed model is evaluated as a performance digital twin through deviation-based monitoring behavior. The discrepancy between the measured and anticipated net power output is continuously represented by the deviation signal ΔMW. When the gas turbine is running at a steady state, ΔMW remains close to zero, which means that the digital twin accurately models the steady-state behavior of a gas turbine. The usefulness of the proposed methodology when it comes to real-time performance tracking and trends in degradation is confirmed by the existence of persistent deviation patterns, which demonstrate operating regimes and potential changes in performance that cannot be identified based on the raw power alone.

Although initially, XGBoost was the best in overall performance with the highest coefficient of determination (R 2 = 0.9125), lowest RMSE (1.4056), lowest MAE (1.0813) and highest R 2 (0.8918), SVR also demonstrated a high level of performance that revealed that it minimized error well on average. Random Forest model, however, resulted in much poor performance (R2 = 0.8864) with higher values of MAE and RMSE, showing a modest predictive power.

The ANN (MLP) model had the lowest explained variance (R^2 = 0.8381) and the most errors (MAE = 1.3825, RMSE = 1.9123). This can mean that the time series of the data in the specified structure cannot be easily generalized.

The relatively weaker performance of the ANN model may be associated with the sensitivity of neural networks to training data distribution, hyperparameter selection, and dataset size. Unlike boosting-based approaches, the implemented ANN architecture may not have been sufficiently deep or temporally structured to fully capture the operational variability of the gas turbine system. Furthermore, industrial operational datasets often contain noise, non-stationary behavior, and uneven operating distributions, which can affect neural network convergence and stability.

The anticipated expected net MW and the measured net MW for each digital twin model are directly compared in Figure 5. The scatter plots indicate that some of the operational points are near the optimal diagonal line, which indicates that the digital twin models are reflective of the steady-state performance of the real plant.

[Figure 5. about here]

The XGBoost model exhibits reduced dispersion across the operating range and performs better in terms of generalization with the tightest clustering around the diagonal. Strong agreement is also shown by the SVR and ANN models, but there is a little increase in scatter, particularly in challenging operating situations. The Random Forest model captures the overall trend while exhibiting localized step-like behavior due to its tree-based structure. These results show that the proposed digital twin approach can accurately estimate the projected net MW without the need for particular physical performance maps.

The performance deviation was calculated by comparing the measured net MW with the expected net MW estimated by the digital twin to facilitate performance monitoring:
The temporal history of the performance deviation for the assessed models is shown in Figure 6.

[Figure 6. about here]

Operating regimes and patterns that are not easily discernible from raw power measurements alone are revealed by the deviation signal. Measurement noise and temporary operating circumstances are the main causes of short-term variations in ΔMW. Persistent positive or negative deviations, on the other hand, point to modifications in plant performance. Positive deviations might indicate advantageous environmental circumstances or operational modifications, whereas persistent negative deviations point to performance losses possibly connected to compressor fouling, inlet pressure losses, or thermal efficiency degradation. The XGBoost-based digital twin is especially well-suited for long-term performance monitoring and trend analysis since it generates the most consistent and smooth deviation signal.

The superior performance of XGBoost can be theoretically attributed to its gradient boosting structure, which sequentially minimizes residual prediction errors while effectively capturing nonlinear interactions among operational and environmental variables. Gas turbine thermodynamic behavior is inherently nonlinear due to the coupled influence of ambient conditions, compressor dynamics, combustion characteristics, and turbine efficiency. Therefore, ensemble boosting techniques such as XGBoost are better suited to model these complex relationships compared to conventional regression approaches. In addition, the regularization mechanisms embedded within XGBoost improve generalization capability and reduce overfitting under varying operating conditions.

For the tree-based models, feature significance analysis was carried out to obtain insight into the physical determinants of net power output. The relative significance of the input variables as determined by the Random Forest and XGBoost models is shown in Figure 7a.

[Figure 7(a-b), about here]

The findings show that the most significant factors influencing net MW are compressor discharge pressure, fuel flow, and exhaust gas temperature, followed by ambient temperature and pressure. The identified feature importance rankings are physically consistent with gas turbine thermodynamic principles. Compressor discharge pressure directly reflects the compressor pressure ratio and air mass flow entering the combustion chamber, both of which strongly influence turbine output power. Similarly, fuel flow determines the thermal energy supplied during combustion, while exhaust gas temperature reflects the efficiency of the expansion process and combustion conditions. The significant influence of ambient temperature is also expected, since increased inlet air temperature reduces air density and consequently decreases compressor mass flow rate and turbine power generation capability. These observations confirm that the digital twin successfully captures the underlying thermodynamic behavior of the gas turbine

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

system.

The physical plausibility of the digital twin is confirmed by these results, which are in line with known gas turbine thermodynamic behavior. Figure 7b shows that the Weight magnitude-based approximations were used to evaluate input influence for the SVR and ANN models. The fact that these models are less interpretable than ensemble-based methods, even while they achieve competitive accuracy, further supports the benefit of XGBoost for industrial digital twin deployment.

The outcomes confirm that using data from industrial historians to create an accurate performance digital twin is feasible. Among the evaluated methods, the XGBoost-based digital twin offers the greatest overall performance since it combines strong deviation behavior, important feature information, and high prediction accuracy. It is therefore ideal for ongoing performance monitoring in industrial settings.

The proposed digital twin can be easily integrated with current PI historical infrastructures and doesn't require any specific design information, in contrast to conventional performance correction curves or physics-based models. The digital twin offers operators practical insights to identify degradation, measure performance losses, and assist data-driven decision-making by concentrating on expected performance rather than future forecasts.

V. CONCLUSIONS

This study successfully achieved its primary objective by developing and validating a data-driven performance digital twin for an industrial gas turbine. the twin uses real operational data from a GE Mark VIe control system and archived in a PI historian infrastructure. To estimate expected power output and monitor performance deviation. four evaluated machine-learning methods: Support Vector Regression, Random Forest regression, Extreme Gradient Boosting, and an Artificial Neural Network. the XGBoost-based digital twin outperformed the others in terms of accuracy and stability, and generated the most reliable deviation signal for continuous monitoring. The obtained results confirm that the proposed objectives of estimating the expected net power output and enabling continuous performance monitoring using real operational data were successfully achieved. The developed digital twin framework demonstrated strong predictive capability and effectively identified operational performance deviations under varying operating conditions.

The introduction of a performance deviation indicator, which is the difference between measured and predicted net MW, is a significant contribution of this work. Continuous performance monitoring is made possible by this deviation signal, which also enables operators to discern between actual performance decline and anticipated ambient-driven fluctuations. With considerable contributions from compressor, fuel, and temperature variables in line with gas turbine thermodynamics, the assessment of the importance of the features of analysis further confirms that digital twin performance is physically significant. The proposed architecture is scalable, non-intrusive, and best for integration into current industrial data infrastructures.

VI. PRACTICAL IMPLICATIONS AND FUTURE WORK

1. Practical Implications

For engineers and operators of power plants, the proposed performance digital twin gives several benefits. The digital twin enables condition-based maintenance methods and pre- performance loss detection by presenting a continuously updated estimate of predicted net MW. The deviation-based monitoring framework is applicable for brownfield installations because it can be applied without modifying the current control logic. Moreover, the digital twin outputs—like predicted power and performance deviation—can be preserved, displayed, and examined alongside traditional plant tags thanks to the interface with PI historian systems. This promotes operational decision-making, long-term performance benchmarking, and adoption in industrial settings.

2. Future Work

Although the current study concentrates on the net power production of gas turbines, a number of extensions are envisaged for future research. First, the digital twin framework can be expanded to a full combined-cycle performance digital twin by including additional models for condenser performance, steam turbine power output, and HRSG steam production. This would enable a thorough assessment of combined-cycle efficiency and heat rate. Second, future research can use the enhanced explainability methods such as SHAP analysis to gain a deeper insight into the causal factors of performance variances. Third, if plant conditions change over time, adaptive or online learning techniques could be investigated to update the digital twin.

To verify the efficacy of the suggested digital twin in real-time operating situations and to facilitate automatic performance monitoring and alerting, real-time deployment utilizing PI system interfaces and streaming data will be examined.

Acknowledgement

The authors sincerely thank the management and staff of the Erbil Perdawood Power Plant for their unwavering support during the data collection phase of this study

We also acknowledge the Department of Electrical Engineering, College of Engineering at Salahaddin University-Erbil, for providing access to laboratories, and resources facilities.

Conflict of interest

"The authors declare that there is no conflict of interest regarding the publication of this paper. No financial or personal relationships have influenced the research, analysis, or interpretation of the results presented in this study.

REFERENCES

[1] D. P. Francis, S. Lazarova-Molnar, and N. Mohamed, "Towards data-driven digital twins for smart manufacturing," in *International Conference On Systems Engineering*, 2020, pp. 445-454: Springer.

[2] S. J. A. T. E. Kim, "Prediction-focused machine learning-based visualization of compressor flow in a gas turbine engine digital twin," p. 127154, 2025.

[3] Z. Liu and I. A. J. E. Karimi, "Gas turbine performance prediction via machine learning," vol. 192, p. 116627, 2020.

[4] M. P. Boyce, *Gas turbine engineering handbook*. Elsevier, 2011.

[5] A. S. H. Ameen, I. I. J. Z. J. o. P. Hamarash, and A. Sciences, "A Machine Learning Models for Predicting Power Output in Gas Turbines: A Case Study from Perdawood Power Station, Erbil, Iraqi Kurdistan," vol. 37, no. 5, pp. 115-124, 2025.

[6] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, M. J. B. Hoffmann, and i. s. engineering, "Industry 4.0," vol. 6, no. 4, pp. 239-242, 2014.

[7] J. Ullrich, A. G. Voyiatzis, and E. R. Weippl, "Secure cyber-physical production systems: Solid steps towards realization," in *2016 1st International workshop on cyber-physical production systems (CPPS)*, 2016, pp. 1-4: IEEE.

[8] T. Khan, U. Khan, A. Khan, C. Mollan, I. Morkvenaite-Vilkonciene, and V. J. M. Pandey, "Data-Driven Digital Twin Framework for Predictive Maintenance of Smart Manufacturing Systems," vol. 13, no. 6, p. 481, 2025.

[9] T. H.-J. Uhlemann, C. Lehmann, and R. J. P. C. Steinhilper, "The digital twin: Realizing the cyber-physical production system for industry 4.0," vol. 61, pp. 335-340, 2017.

[10] S. Alketbi, A. B. Nassif, M. A. Eddin, I. Shahin, and A. Elnagar, "Predicting the power of a combined cycle power plant using machine learning methods," in *2020 International conference on communications, computing, cybersecurity, and informatics (CCCI)*, 2020, pp. 1-5:

IEEE.

[11] C. A. Saleel, "Forecasting the energy output from a combined cycle thermal power plant using deep learning models," *Case Studies in Thermal Engineering,* vol. 28, p. 101693, 2021/12/01/ 2021.

[12] Y. Tunckaya and E. J. J. o. t. E. I. Koklukaya, "Comparative analysis and prediction study for effluent gas emissions in a coal-fired thermal power plant using artificial intelligence and statistical tools," *Journal of the Energy Institute,* vol. 88, no. 2, pp. 118-125, 2015.

[13] M. A. Saeed *et al.*, "Electrical power output prediction of combined cycle power plants using a recurrent neural network optimized by waterwheel plant algorithm," *Frontiers in Energy Research,* vol. 11, p. 1234624, 2023.

[14] Q. Yi, H. Xiong, and D. J. E. Wang, "Predicting Power Generation from a Combined Cycle Power Plant Using Transformer Encoders with DNN," vol. 12, no. 11, p. 2431, 2023.

[15] N. Pachauri and C. W. J. I. A. Ahn, "Electrical energy prediction of combined cycle power plant using gradient boosted generalized additive model," vol. 10, pp. 24566-24577, 2022.

[16] A. Asghar, T. A. H. Ratlamwala, K. Kamal, M. Alkahtani, E. Mohammad, and S. J. H. Mathavan, "Sustainable operations of a combined cycle power plant using artificial intelligence based power prediction," vol. 9, no. 9, 2023.

[17] J. L. Lobo, I. Ballesteros, I. Oregi, and J. J. a. p. a. Del Ser, "Real-time Electrical Power Prediction in a Combined Cycle Power Plant," 2019.

[18] N. S. Santarisi and S. S. J. E.-E. J. o. E. T. Faouri, "Prediction of combined cycle power plant electrical output power using machine learning regression algorithms," vol. 6, no. 8, p. 114, 2021.

[19] B. Xu *et al.*, "A case study of digital-twin-modelling analysis on power-plant-performance optimizations," vol. 3, no. 3, pp. 227-234, 2019.

[20] J. Friederich, D. P. Francis, S. Lazarova-Molnar, and N. J. C. i. I. Mohamed, "A framework for data-driven digital twins of smart manufacturing systems," vol. 136, p. 103586, 2022.

[21] J. Lim *et al.*, "The EPRI Gas Turbine Digital Twin–a Platform for Operator Focused Integrated Diagnostics and Performance Forecasting," in *Turbo Expo: Power for Land, Sea, and Air*, 2021, vol. 84966, p. V004T09A009: American Society of Mechanical Engineers.

[22] Y. Ma, X. Zhu, J. Lu, P. Yang, and J. J. S. Sun, "Construction of data-driven performance digital twin for a real-world gas turbine anomaly detection considering uncertainty," vol. 23, no. 15, p. 6660, 2023.

[23] D. Ariansyah *et al.*, "Digital twin development: a step by step guideline," 2020.

[24] A. Redelinghuys, A. Basson, and K. Kruger, "A six-layer digital twin architecture for a manufacturing cell," in *International Workshop on Service Orientation in Holonic and Multi-Agent Manufacturing*, 2018, pp. 412-423: Springer.

[25] J. T. Lim, A. Habibullah, and E. Y. K. J. E. Ng, "Towards a digital twin for gas turbines: Thermodynamic modeling, critical parameter estimation, and performance optimization using PINN and PSO," vol. 18, no. 14, p. 3721, 2025.

[26] K. Rawal and A. Ahmad, "A Comparative Analysis of Supervised Machine Learning Algorithms for Electricity Demand Forecasting," in *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, 2022, pp. 1-6.

[27] W.-C. Hong, "Hybrid evolutionary algorithms in a SVR-based electric load forecasting model," *International Journal of Electrical Power & Energy Systems*, vol. 31, no. 7, pp. 409-417, 2009/09/01/ 2009.

[28] L. A. Abad, S. M. Sarabia, J. M. Yuzon, and M. C. Pacis, "A Short-Term Load Forecasting Algorithm Using Support Vector Regression & Artificial Neural Network Method (SVR-ANN)," in *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, 2020, vol. 11th, pp. 138-143: IEEE.

[29] B. Mallala, A. I. U. Ahmed, S. V. Pamidi, M. O. Faruque, and R. J. R. i. E. Reddy, "Forecasting global sustainable energy from renewable sources using random forest algorithm," vol. 25, p. 103789, 2025.

[30] L. Bandić, M. Hasičić, and J. Kevrić, "Prediction of power output for combined cycle power plant using random decision tree algorithms and ANFIS," in *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies*, 2019, pp. 406-416: Springer.

[31] M. J. T. E. P. o. S. T. E. Cosar and Mathematics, "Utilizing the eXtreme gradient boosting algorithm for artificial intelligence-supported learning analytics application," vol. 28, pp. 277-285, 2024.

[32] T. M. Singh and K. Lippert, "Digital Twins in Human Activity Prediction on Gait Using Extreme Gradient Boosting Local Binary Pattern: Healthcare 6.0," in *Exploring the*

Advancements and Future Directions of Digital Twins in Healthcare 6.0: IGI Global, 2024, pp. 240-261.

[33] Z. A. Ali, Z. H. Abduljabbar, H. A. Tahir, A. B. Sallow, and S. M. J. A. J. o. N. U. Almufti, "eXtreme gradient boosting algorithm with machine learning: A review," vol. 12, no. 2, pp. 320-334, 2023.

[34] A. Wankhede and V. Ghate, "Prediction of power of a combine gas and steam turbine using artificial neural network," in *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2018, pp. 1103-1108: IEEE.

*Corespondent e-mail address engasaad86@gmail.com

Received: 2026-04-25
Accepted: 2026-04-26
Published: 2026-07-20

Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

DAFTAR TABEL

Tabel 1. Digital twin services provided by the proposed framework ....
Table 2. Description and abbreviations of input features..
Table 3. Hyperparameter settings for models used in the digital twin..
Table 4 predictive performance of the ML models..

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

Table 1. Digital twin services provided by the proposed framework

Digital twin service	Output	Purpose
Expected power estimation	MW_expected	Benchmark expected behavior under current conditions
Deviation monitoring	ΔMW = MW_actual − MW_expected	Detect performance shifts and abnormalities
Degradation trending	ΔMW trend vs time	Track long-term performance loss
Interpretability	Feature importance	Improve trust and physical understanding
Decision support	Threshold alarms (optional)	Support proactive maintenance planning

Table 2. Description and abbreviations of input features

No.	Feature Name	Abbreviation	Unit	Category
1	Ambient Temperature	AT	°C	Ambient
2	Ambient Pressure	AP	inHg	Ambient
3	Specific Humidity	CH	kg/kg (or %)	Ambient
4	Natural Gas Fuel Flow	GP	kg/s	Operational
5	Grid Frequency	DF	Hz	Operational
6	Compressor Pressure Ratio	CPR	Ratio %	Compressor
7	Compressor Discharge Pressure	CPD	bar	Compressor
8	Exhaust Gas Temperature	$TTXM$	°C	Turbine

Table 3. Hyperparameter settings for models used in the digital twin

Model	Key Hyperparameters
SVR	kernel=RBF, C=100, ε=0.1, γ=scale
Random Forest	n_estimators = 500, max_depth = 16, min_samples_leaf = 2, max_features = sqrt
XGBoost	n_estimators = 600, learning_rate = 0.05, max_depth = 5, subsample = 0.8, colsample_bytree = 0.8, reg_lambda = 2, min_child_weight = 3
ANN (MLP)	hidden_layer_sizes = (64,32), activation = ReLU, solver = Adam, alpha = 0.001, early_stopping = True, max_iter = 3000

Table 4. predictive performance of the ML models

Model	MAE	RMSE	R^2
SVR	1.081262	1.562917	0.891832
Random Forest	1.331115	1.601371	0.886443
XGBoost	1.098966	1.405583	0.912513
ANN (MLP)	1.382546	1.912344	0.838058

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

DAFTAR GAMBAR

Figure 1. AVEVA PI DATA ....
Figure 2. Proposed Data-Driven Digital Twin Layers ....
Figure 3. Workflow for Data-Driven Digital Twin...
Figure 4. sample of gas turbine digital twin with real time data...
Figure 5. actual vs digital twin power expectation ....
Figure 6. Digital twin performance deviation...
Figure 7. (a-b), the significance of all 8 features based on the application of 4 ML models ....

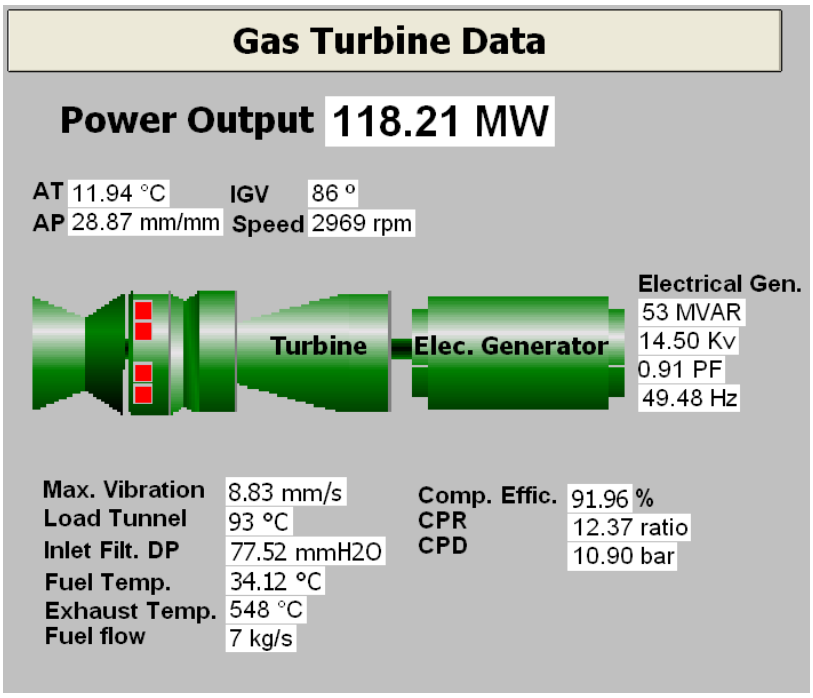

Figure 1. AVEVA PI Data

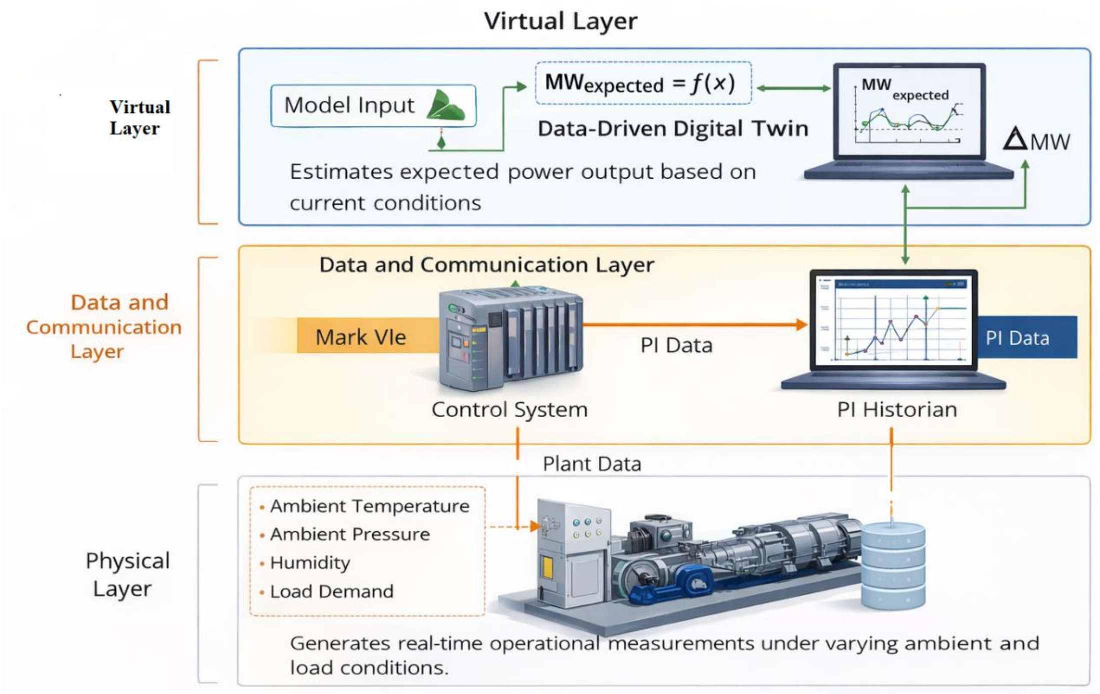

Figure 2. Proposed Data-Driven Digital Twin Layers.

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

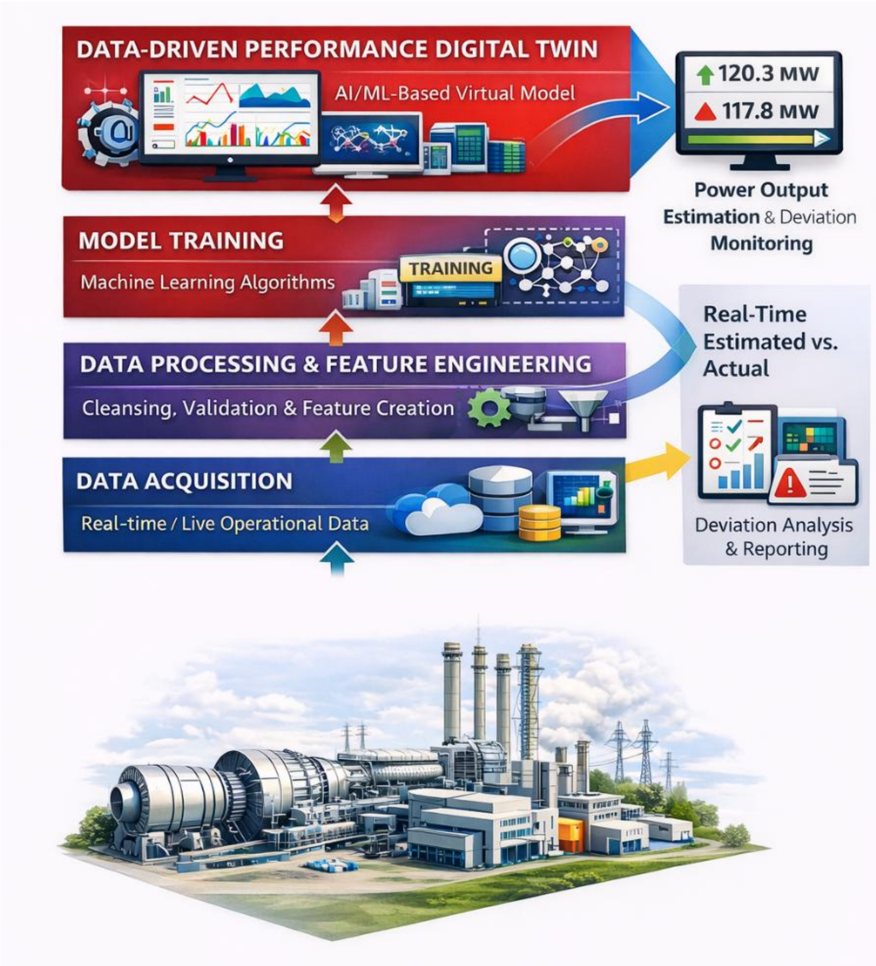

Figure 3. Workflow for Data-Driven Digital Twin

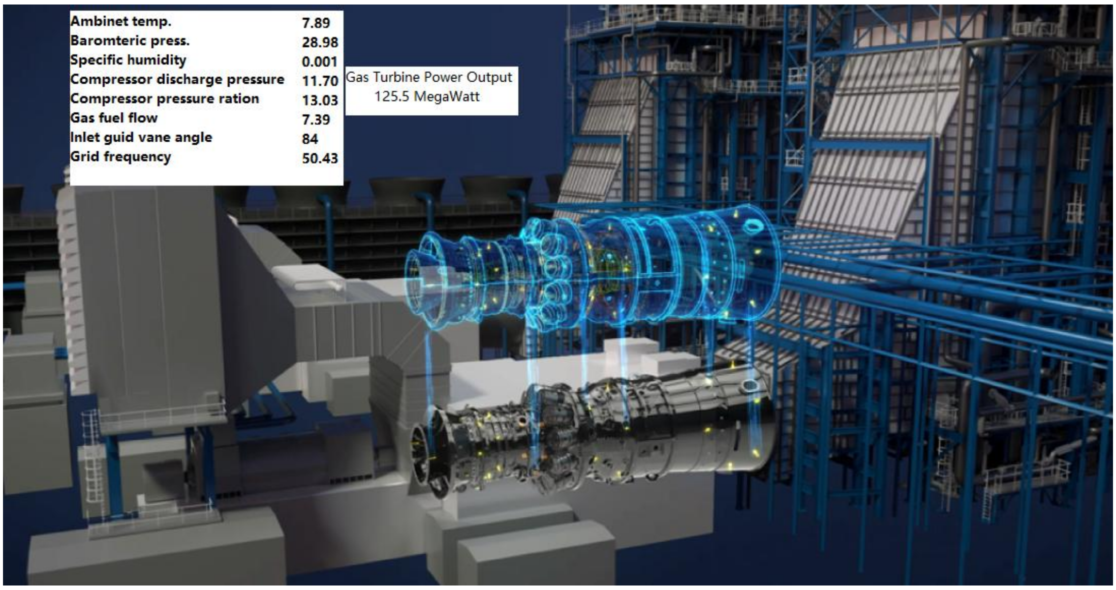

Figure 4. sample of gas turbine digital twin with real time data

JEEE-U Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

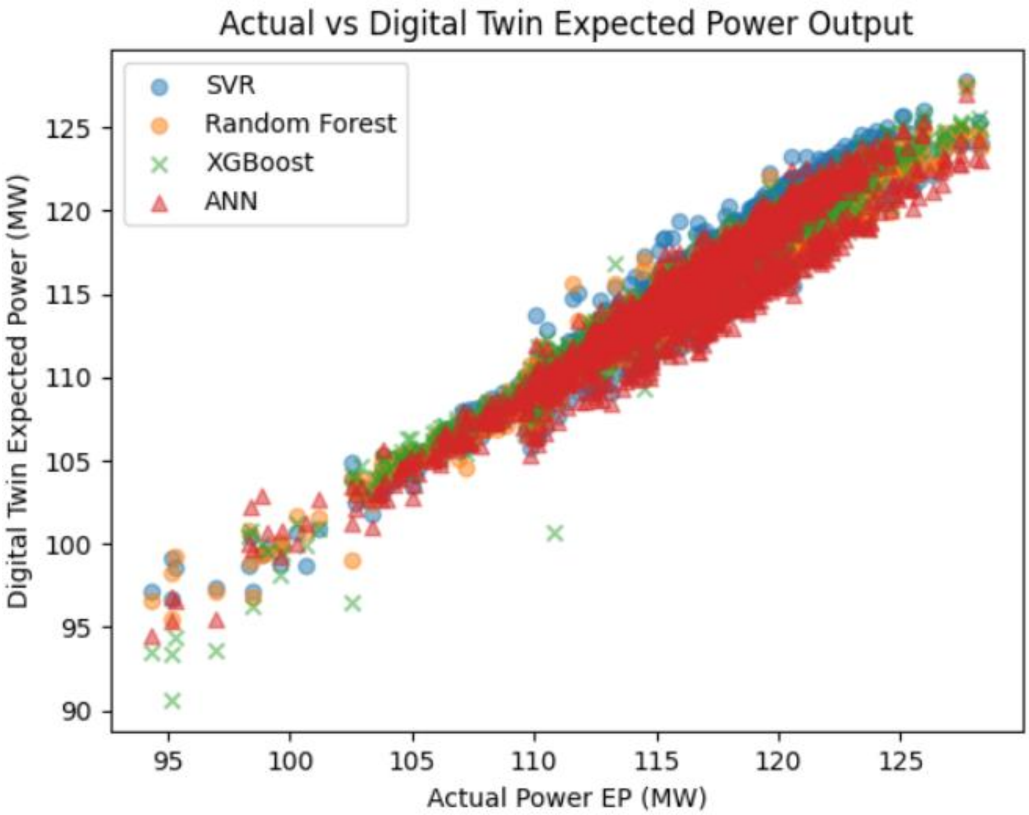

Figure 5. actual vs digital twin power expectation

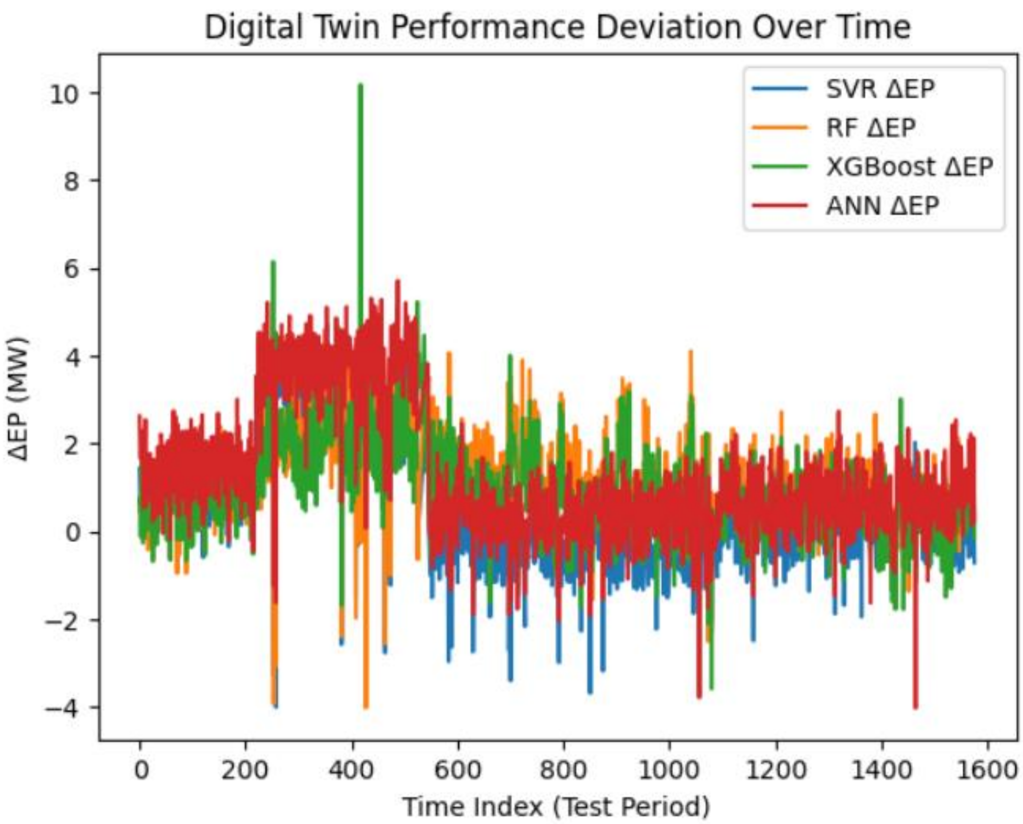

Figure 6. Digital twin performance deviation

JEEE-U
Journal of Electrical and Electronic Engineering-UMSIDA
ISSN 2460-9250 (print), ISSN 2540-8658 (online)
Vol. 10, No. 2, October 2026
10.21070/jeeeu.v10i1.1738

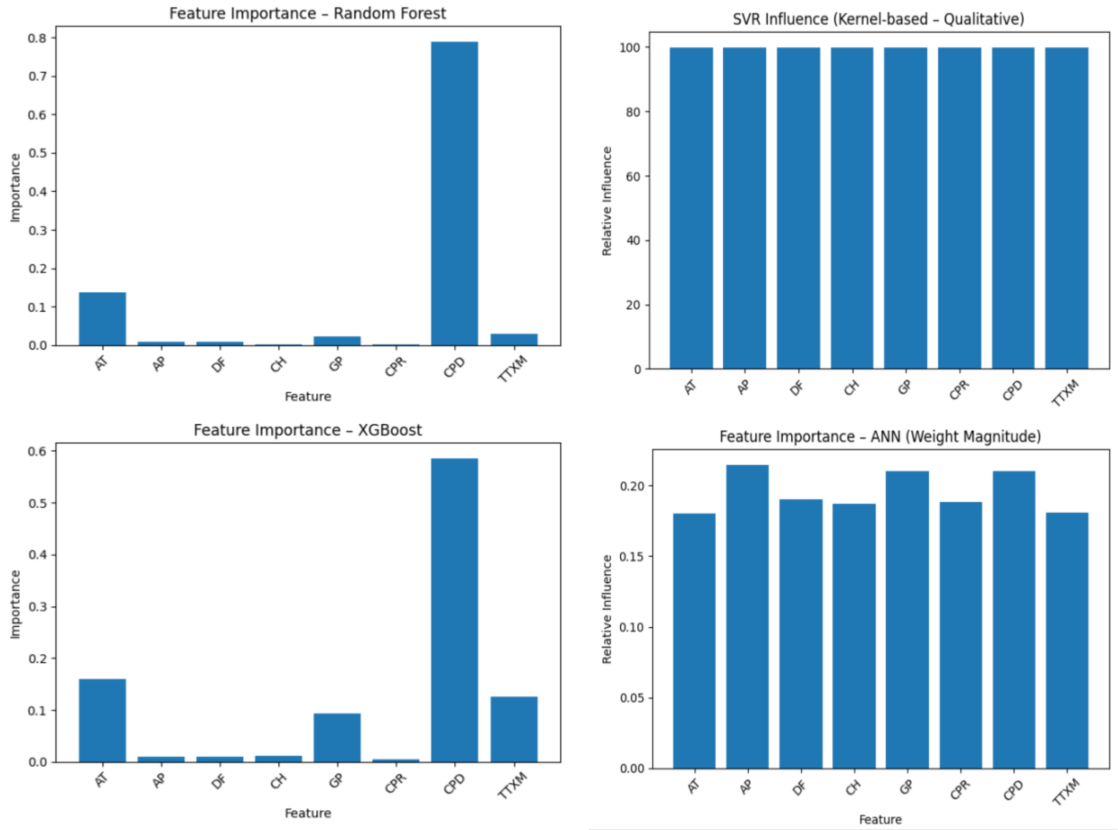

Figure 7. (a-b), the significance of all 8 features based on the application of 4 ML models.