

Benchmarking Linear vs. Non-Linear Predictive Models for Energy Demand Forecasting in Electrical Motor-Driven Compressors

Humam Nazir¹, Tarmizi Ahmad Izzuddin², Nur Izyan Zulkafli³, Mohamad Fani Bin Sulaima⁴, Mohd Hafiz Jali⁵, Haslenda Hashim⁶, Nur Saleha Jayiddin⁷, Azmi Md Lasin⁸, M Tarmidzi Iskandar⁹

^{1,4,5}Faculty of Electrical Technology and Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

²Faculty of Artificial Intelligence and Cybersecurity, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

³Faculty of Mechanical Technology and Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

⁶School of Chemical and Energy Engineering, Universiti Teknologi Malaysia, Skudai, 81310, Johor Bahru, Malaysia

^{7,8,9}Petronas Group Technical Solutions, Projects, Technology and Health, Safety, Security and Environment, 50088, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Jan 24, 2026

Revised Feb 28, 2026

Accepted Mar 14, 2026

Keywords:

Load Forecasting

Deep Learning

Multiple Linear -

Regression

Electrical Compressors

ABSTRACT

In the LNG regasification plants, among the most energy-demanding components are the electrical motor driven compressors. Artificial Neural Networks (ANNs) and Extreme Gradient Boosting (XGBoost) feature superior flexibility in modelling nonlinear and dynamic compressor behaviour, unlike the conventional methods such as Multiple Linear Regression (MLR), which is limited by its assumption of linearity. This study focuses on data-driven techniques for compressor power consumption prediction, in which MLR, ANN and XGBOOST are compared. Here, real operational data collected from two Boil-off Gas (BOG) and two Regasification Terminal Export Compressors (RGTEC) were considered to develop and evaluate the models. The findings suggest that non-linear machine learning models provide superior predictive performance in comparison to those of the traditional linear approach. The highest prediction accuracy for most compressors is achieved using the ANN model, with a R^2 values of 94.6%, 99.64% and 99.64% for BOG-A, BOG-C and RGTEC-A, respectively. Meanwhile, the best performance was found for XGBOOST for RGTEC-B, with a R^2 value of 94.17% and a significantly lower RMSE value. Contrary to that, MLR produced lower accuracy in several cases, particularly when subjected to complex operating conditions, in which linear assumptions were inadequate to capture system dynamics. In addition, this research features one of the first direct comparisons between linear and nonlinear models applied to real-time compressor data, unlike past research that depends on simulations or single-method analysis. It emphasizes the practical advantages of ANN and XGBoost for data-driven energy forecasting and operational optimizations in the gas industry.

Copyright © 2026 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Tarmizi Ahmad Izzuddin

Faculty of Artificial Intelligence and Cybersecurity, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia

Email: tarmizi@utem.edu.my

1. INTRODUCTION

In recent decades, demand for natural gas has steadily increased. Liquefied Natural Gas (LNG) technology converts natural gas into a denser form, making it easier to transport worldwide. After arrival, LNG can be stored and later converted back into gas. Since LNG is kept at very low temperatures, any heat entering the storage tank causes some of the liquid to evaporate, producing boil-off gas (BOG). If not managed, BOG can reach unsafe levels. BOG compressors collect and compress this evaporated gas so it can be handled or used safely [1]. Compressing BOG is a key function at regasification terminals and accounts for a significant part of the terminal's energy use [2]. At regassing terminals, compressing BOG is considered a main function and an important part of the terminal's energy use [2]. To transfer the gas into pipelines, the regassing terminal export compressor (RGTEC) further compresses the gas from BOG [3], highlighting the crucial role of compressors in the plant [4]. Moreover, compressors are necessary for processing and distributing gas efficiently through pipelines.

Nonetheless, operational efficiency is the focus for these systems. In a natural gas power plant, a substantial amount of energy is consumed by the electrical motor-driven compressor compared to any other major equipment [5]. Such higher consumption behaviour of compressors makes them an important indicator for the prediction of overall plant energy use. Additionally, 10% of the total industrial electricity is used by the industrial pneumatic systems, which use compressed air for automation [6]. Besides, the inefficient use of a compressor affects the environment. As an example, the carbon footprint is a byproduct of the gasification of LNG. On average, tonnes of carbon are emitted for every tonne of processed LNG [7]. Moreover, long-term health risks are also caused by the operational byproducts of these systems [8].

Due to this, it is essential to predict compressor power use to improve energy efficiency and performance in industry. Data-driven methods like machine learning and deep learning are now used in industrial systems to aid such predictive tasks. Using real-time compressor data, these approaches permit the prediction and optimization of performance [9].

Two main considerations for accurately predicting the power consumption of a compressor include operating conditions and performance characteristics. Suction and discharge temperatures, suction and discharge pressures and mass flow rate are among the important factors. These parameters directly influence the load and efficiency of compressors in industrial systems [10]. The compression ratio, gas density, and thermodynamic work required are affected by the change in these parameters [11]. Several studies have been conducted to find the relationship between these parameters and energy usage. As an example, mass flow and pressure dynamics were found to play a crucial role in compressor energy use within multiperiod operational models of long-distance gas networks [12]. It becomes difficult to accurately predict compressor power use when relying on measurable parameters and nonlinear system dynamics. This condition shows the need for better data-driven modelling methods.

MLR is an alternative approach to energy systems prediction because it is straightforward, understandable, and effective. It is an appropriate point to begin prediction because of its simplicity of implementation. However, the issue with such kinds of prediction models is that they suppose that there is a linear relationship between variables [13]. Such assumptions are not usually true in a real-world industrial system, like compressors. Complex, nonlinear interactions among variables determine the manner in which compressors operate [14]. The accuracy of MLR is therefore reduced in predicting the compressor loads in different conditions. To avoid these issues, Artificial Neural Networks (ANN) are now an essential data-driven tool to make a prediction. They are able to handle complicated relations and associations among various considerations. Unlike linear regression, ANNs do not require rigid rules regarding the input data and can adapt to changes in the way things work. In one case, a Gray RBF neural network was employed to estimate the performance of natural gas compressors when wet gas was added up to 30 percent [15]. The error of prediction made was 2.34 percent, which is better than the error that the linear interpolation would provide. These results indicate that neural networks are capable of controlling the complicated behaviour in compressor systems, and they work more efficiently in dynamic conditions. This advantage has given rise to the rising popularity of AI solutions within the industry. It is estimated that more than 60 percent of energy companies will utilize ANNs by 2025 in activities such as predicting energy usage, load forecasting, demand response, and predictive maintenance [16]. Along with neural networks, the ensemble machine learning approaches are demonstrated to be efficient in the modeling of complex, nonlinear industrial processes [17]. XGBoost is an effective tool that makes use of decision trees, and models are built in stages to enhance predictions. It is effective at finding non-trivial patterns, handling non-trivial features, and retaining accuracy even when the data is not arranged in order [18]. Unlike simple prediction algorithms, XGBoost takes additional measures to ensure its models are tougher and to prevent the errors that arise from overfitting on the data [19]. XGBoost is also typically employed in the prediction of energy demands, fault prediction, and maintenance planning due to these useful characteristics. These characteristics make it a preferred choice for compressor power consumption prediction, particularly when various factors are considered.

Despite these advances, the majority of earlier studies have utilized one way to predict the energy use of electric motor-driven compressors. Limited studies have integrated the use of MLR, ANN, and XGBoost for such predictions. In addition, very few studies have utilized real-time compressor data, and there are limited compressor-specific prediction models that are trained on real operating conditions using neural networks. Most studies also do not consider a detailed model performance comparison.

Therefore, in this study, real-time data from several compressors is used to test and compare three prediction methods: MLR, ANN, and XGBoost. Suction and discharge temperatures, suction and discharge pressures as well as mass flow rate are the input variables considered. Standard measures such as the coefficient of determination (R^2) and root-mean-square error (RMSE) are used to measure the accuracy. This paper aims to identify which method works best in real-world conditions.

2. RESEARCH METHOD

A. Compressor Plant Description

This study was conducted in a compressor plant in Melaka, Malaysia. The plant runs five compressors, which are categorized by function into two types. Here, the first three compressors are boil-off gas (BOG) compressors, which compress vaporized natural gas from the floating storage unit (FSU). The remaining two are regasification terminal export (RGTEC) compressors, positioned downstream of the BOG compressors, which further compress the gas before it enters the pipeline network. The other two are regasification terminal export (RGTEC) compressors, which are installed downstream of the BOG compressors and further compress the gas to pass through the pipeline network. Figure 1 shows the layout of the compressors, the data collection sensors, and the layout of the plant operation.

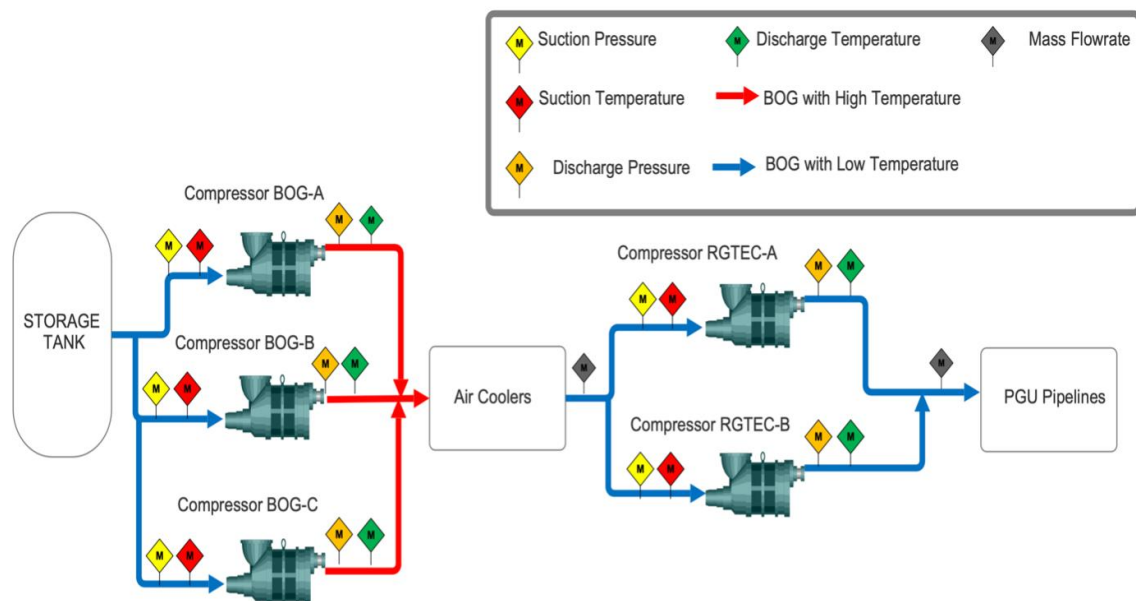


Figure 1. An illustration of the compressors' network arrangement in the plant

In this study, to store the Liquefied Natural Gas (LNG), the floating storage units (FSUs), which are equipped with specialized insulation that conserves cryogenic temperatures of approximately -160°C , are utilized. At this temperature, natural gas remains in liquid form at near atmospheric pressure, reducing its volume by about 600 times in comparison to its gaseous state and allowing efficient long-distance transport [20]. Irrespective of the insulation, part of the LNG unavoidably vaporizes, and if unmanaged, this vapor can accumulate to harmful levels. Therefore, to address this issue, the vaporized gas is directed to three BOG compressors, namely BOG A, BOG B and BOG C, for compression. As the compression process increase the gas temperature, air coolers are used to bring it down. In Figure 2, the electrical motor-driven compressor unit is shown. During minimal send-out mode, the cooled BOG discharge is distributed directly to the two RGTEC compressors, these being RGTEC A and RGTEC B, for further compression before being delivered to the Peninsular Gas Utilization (PPGU) pipelines (Figure 1). Contrary to that, during normal send-out mode, the discharge from the BOG compressors is routed to the regasification package unit. To ensure an efficient send-out while minimizing flaring during low-demand phases, the number of active compressors is regulated based on operational demand.

B. Dataset

Data for both compressor types were collected using sensors, as illustrated in Figure 1. This study considers eight independent variables for the BOG compressors and twelve independent variables for the RGTEC compressors, as listed in Table 1. The dependent variable for both compressor types is Power. The dataset was obtained from historical operational logs, recorded hourly over three weeks, yielding 534 time points. Data were available for BOG A, BOG C, RGTEC A, and RGTEC B, while BOG B remained shut down during the data collection period.



Figure 2. Electrical Motor-Driven Compressor [21]

Table 1. Dependent and Independent Variables

BOG-A & BOG-C	RGTEC-A & RGTEC-B	Dependent Variable
Flow rate, m ³ /h	Total flowrate (m ³ /hr)	Real Power (W)
Pressure input (barg)	Pressure input (barg)	
Pressure output (barg)	Pressure output (barg)	
Temperature input (degree C)	Temperature input (C)	
Temperature output (degree C)	Temperature output (C)	
Vibration data (Motor vibration) mm/s	Vibration data (Cross head 1st stage vibration)	
Vibration data (Crank case vibration) mm/s	Vibration data (Cross head 2st stage vibration)	
	Vibration data (Cross head 3st stage vibration)	
	Vibration data for each compressor RGTEC B (Crank case vibration)	
	Vibration data (Crank case vibration)	
	Vibration data (motor vibration)	

C. Data Preprocessing

2.C.1. General

In this study, the Python library [22] is used for data analysis. Pandas is one of the most sought-after Python libraries for analytics and data science. It is an open-source Python data analysis library that allows Python to work with data that mimics spreadsheets, making it easy to load, manipulate, align, merge and perform other tasks. It offers a variety of features and techniques to speed up pre-processing and data analysis. Data structures and data manipulation tools are incorporated to facilitate quick and simple data cleaning and analysis.

The power consumption over time, plotted on the X-axis, for each compressor type are given in Figure 3.

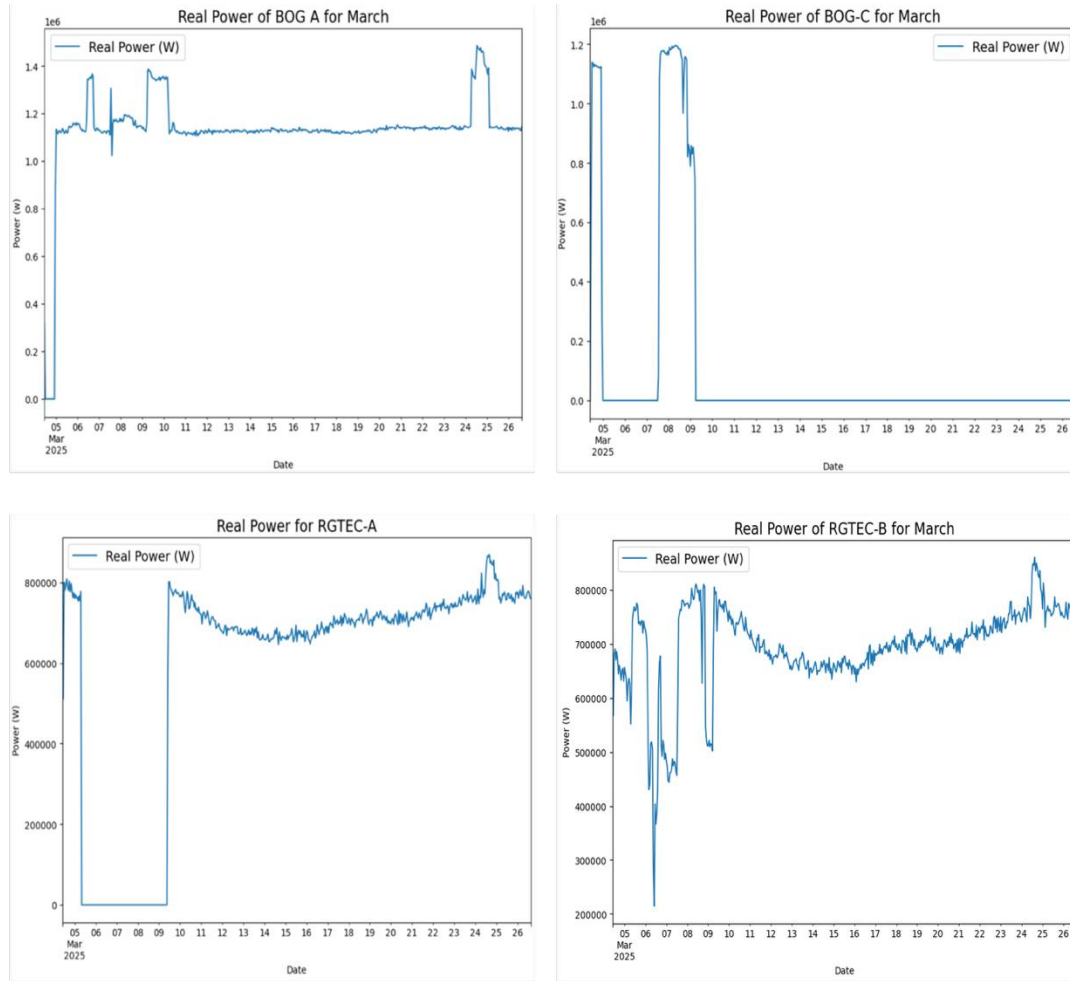


Figure 3. Real Power Consumption for BOG and RGTEC Compressors in March 2025

2.C.2. Standardization and Normalization

In this study, the input variables were measured in different units and scales (e.g. bar, kg/h, kW), which could bias model training. Therefore, normalization and standardization techniques were applied to rescale the features before model development. Normalization is a scaling technique that transforms a wide feature range into a standard range. It typically lies between 0 and 1. The normalization equation used in this study is given below.

$$X = \frac{x - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X represents the original value of a feature, X_{min} represents the minimum value of that feature, and X_{max} represents the maximum value of the feature [23].

Standardization is another scaling technique used in the data transformation step. It transforms the dataset so that the mean is zero and the standard deviation is one. Our study used the following formula for standardization:

$$Z = \frac{x - \mu}{\sigma} \quad (2)$$

Here, Z denotes the standardized value, and X denotes the original value of a specific feature. μ and σ stand for the mean and the standard deviation of the dataset, respectively [24]. In this study, both standardization and normalization were applied separately to the MLR and ANN models, and the preprocessing technique that yielded the best performance was retained for each case. The implementations were carried out using Python's StandardScaler [25] and MinMaxScaler [26] functions. For the MLR model, preprocessing had little to no effect on the results, whereas for the ANN model, standardization consistently produced better performance.

D. Predictive Models of Performance

2.D.1. Multiple Linear Regression (MLR)

MLR was the baseline model applied to make predictions of compressor power consumption depending on the variables related to its operation. The baseline model was initially chosen as MLR because it is highly interpretable and commonly is used as a starting point when performing energy forecasting in the industrial sector. MLR is a straightforward, clear, and popular regression approach, which approximates the linear relationship between a dependent variable and numerous independent variables, and thus it is a frequent benchmark in predictive modeling [27]. MLR is shown in Equation 3 [28].

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad (3)$$

Where y_i is the observed dependent variable, x_{ij} are the explanatory variables, β_j are the regression coefficients, β_0 is the intercept, and ε_i is the error term. In this study, MLR was implemented in Python using the Linear Regression function from Scikit Learn, which applies the Ordinary Least Squares (OLS) method to estimate regression coefficients by minimizing the residual sum of squares:

$$\min_w ||X_w - y||_2^2 \quad (4)$$

MLR is computationally efficient, but it is predicated on assumptions like non-multicollinearity, linearity, independence, and homoscedasticity. Multicollinearity, which inflates variance in coefficient estimates and diminishes model robustness, can result from correlated features in real-world operational data, such as suction and discharge pressures [29].

2.D.2. Artificial Neural Network (ANN)

ANNs were implemented to model nonlinear dependencies between compressor operational variables and power consumption. ANNs are made of input, hidden, and output layers of interconnected artificial neurons. To get a clear idea of the main elements used to construct an ANN, Figure 4 presents a general artificial neural network model that includes the main components of this type of model. The information (input) that the neuron receives from the external sensory system or from other neurons that it is connected to is represented by the symbols $x_1, \dots, x_p$. The vector of synaptic weights, $w = (w_1, \dots, w_p)$, alters the information received by simulating the synapse between biological neurons. These gains can amplify or attenuate the values they want to send to the neuron. The parameter b_j denotes the bias (intercept or threshold) of a neuron. Learning in this context refers to the process of altering the weights of connections among the nodes (neurons) of a given network. Because of their distinct advantages over conventional statistical techniques, Artificial Neural Networks (ANNs) were chosen for this investigation. ANNs can be used with both small and large datasets due to their capabilities to operate with an unlimited number of inputs and outputs. Complex and nonlinear relationships are necessary in the dynamics of the compressor operations since ANNs are able to model such relationships. ANNs are still capable of making important predictions even on incomplete or imperfect data, and, unlike most other methods, ANNs do not make strong assumptions about the distribution of the input variables.

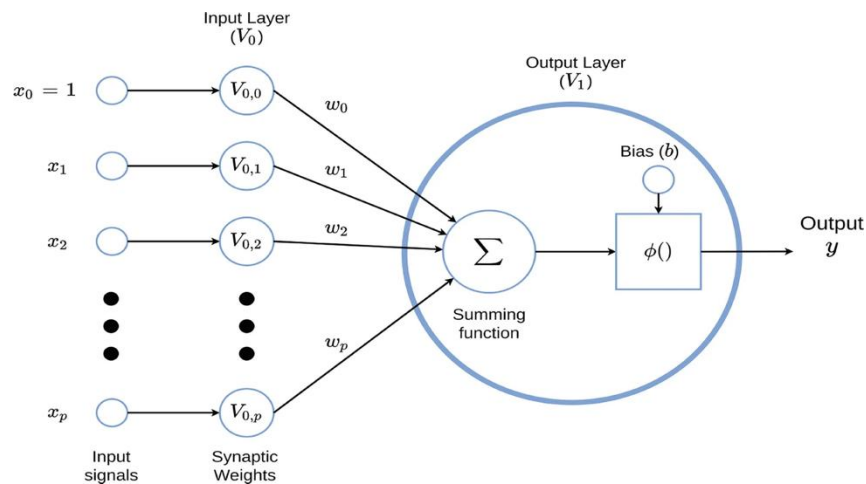
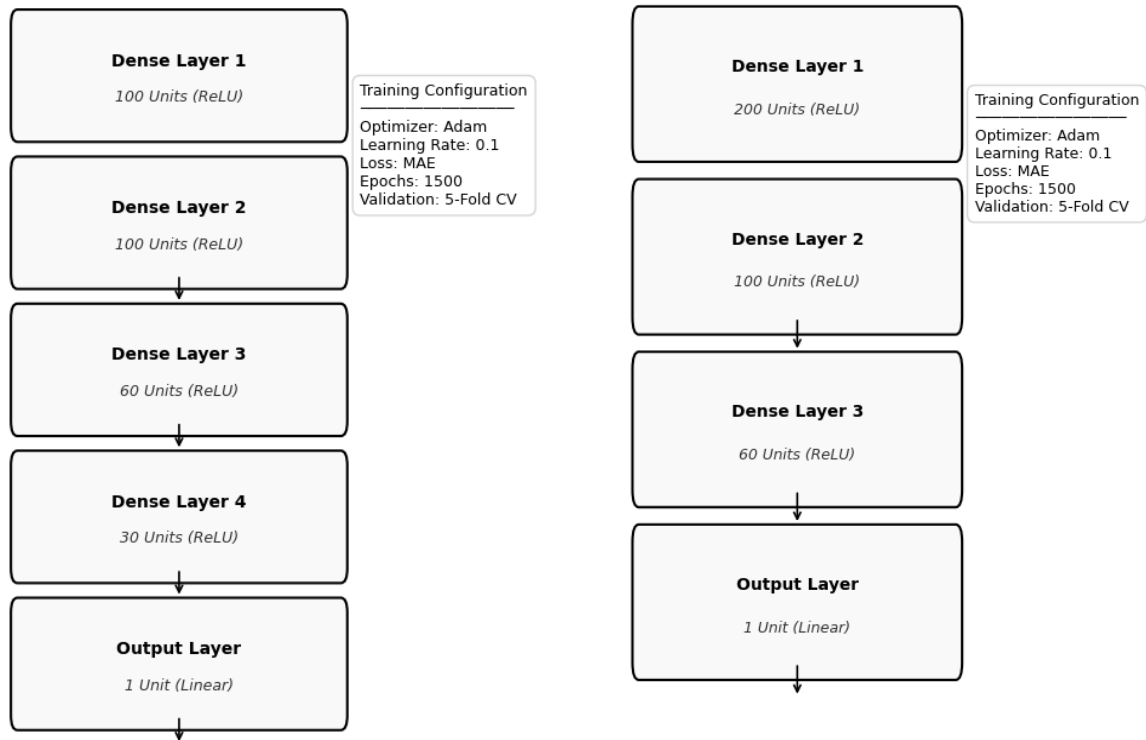
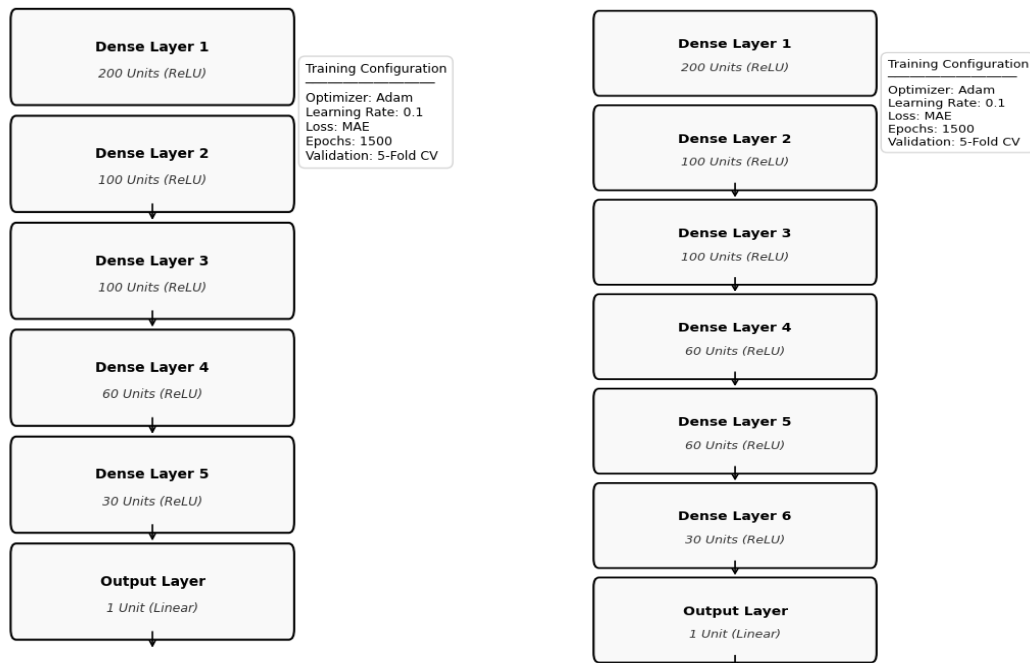


Figure 4. General Artificial Neural Network Model [30]



(a) ANN Architecture & Parameters for BOG A.

(b) ANN Architecture & Parameters for BOG C.



(c) ANN Architecture & Parameters for RGTEC A

(d) ANN Architecture & Parameters for RGTEC B

Figure 5. Structures of the best models for each compressor

Also, ANNs can multiprocessing, be fault-tolerant, and be highly generalized. Once trained on preliminary relationships, ANNs are able to predict outcomes in unknown operating conditions. Due to these properties, ANNs represent a powerful and versatile predictive model of industrial energy systems [31]. Figure 5 shows the structure of ANN generated with BOG A, BOG C, RGTEC A, AND RGTEC B, respectively.

The hidden layers and neurons were decided by trial and error until the most optimal topology had been reached. Even though a single hidden layer is sufficient to approximate complex functions in theory, more layers and neurons can increase the learning capacity in practice. Nevertheless, too large networks may decrease the generalization capacity and enhance the chances of overfitting. An over-parameterized model may fit the training data perfectly but fail to capture the true underlying relationships, thereby performing poorly on unseen data [32].

To introduce nonlinearity during training, the Rectified Linear Unit (ReLU) activation function was applied to each hidden layer. ReLU converts the weighted input from a node into an output by applying a simple piecewise linear transformation. It outputs the input directly if it is positive; otherwise, it outputs zero, as shown in equation [32].

$$(x) = x^+ = \max(0, x) \quad (5)$$

This is also known as a ramp function and is analogous to half-wave rectification in electrical engineering. During compilation, the Adam (Adaptive Moment Estimation) optimizer was employed to update the network weights. Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions that was proposed by [33]. Adaptive estimates of lower-order moments are the foundation of this algorithm. Through an iterative weight-updating procedure, Adam can be thought of as an extension of stochastic gradient descent for training machine learning models [34].

2.D.3. Extreme Gradient Boosting (XGBoost)

To provide a strong benchmark against the neural network models, XGBoost was implemented for compressor power consumption prediction. XGBoost is an efficient ensemble learning method based on gradient boosted decision trees. It builds multiple regression trees sequentially, where each new tree attempts to reduce the prediction error of the previous trees.

The final prediction of the model is obtained by summing the outputs of all individual trees:

$$\hat{Y}_i = \sum_{k=1}^K f_k(x_i) \quad (6)$$

Where, x_i represents the input feature vector, f_k denotes the prediction from the K^{th} regression tree, and K is the total number of trees.

XGBoost improves model Generalization by incorporating regularization, feature subsampling, and early stopping to prevent overfitting [19].

Table 2. Hyperparameter setting of XGBoost

Hyperparameter	Value	Description
n_estimators	2000	Maximum number of boosting trees
learning_rate	0.03	Controls the contribution of each tree
max_depth	7	Maximum depth of individual trees
subsample	0.8	Fraction of samples used for training each tree
colsample_bytree	0.8	Fraction of features randomly selected per tree
early_stopping_rounds	50	Stops training if validation error does not improve
eval_metric	RMSE	Evaluation metric used during training
random_state	42	Ensures reproducibility

The hyperparameters of the XGBoost model (Table 2) were selected through empirical tuning to balance model accuracy and computational efficiency.

2.D.4. Cross Validation and Training Strategy

Our 5-fold cross-validation model enabled us to test all predictive models using a reliable and fair evaluation procedure, including Multiple Linear Regression (MLR), Artificial Neural Network (ANN), and XGBoost [35]. The data was uniformly divided into five approximately equal portions and shuffled and set with a fixed random seed (random_state = 42) so that the results could be replicated. During every round, a training set consisting of four parts (80% of the data) was created and thus used, while the rest (20% of the data) was used as a test set [33] using the train-test-split function [36] of scikit-learn. The process was repeated

5 times in such a way that the parts served as the test set once. We averaged the assessment metrics of all the rounds to obtain the ultimate model performance. In the case of the MLR and ANN models, we directly trained the models using the training data and tested them using the test data. In the XGBoost model, we hold out 10 percent of the training data as a validation set to enable the program to early stop and prevent overfitting. This implied that in every round, approximately 72 percent of the data was used as training, 8 percent used as validation, and 20 percent as testing. Training was monitored on the validation set, and training was terminated when the validation error was not reduced within 50 rounds. That would enable us to compare all models in an equal manner and achieve a good approximation of the degree to which they would be suitable for new data.

2.D.5. Evaluation of Predictive Accuracy

The predictive accuracy of the MLR and ANN models was evaluated using R2 (coefficient of determination), MSE (mean squared error), and RMSE (root mean squared error).

Coefficient of Determination R2: The coefficient of determination can be interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^m (X_i - Y_i)^2}{\sum_{i=1}^m (\bar{Y} - Y_i)^2} \quad (7)$$

Where,

$$\bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i \quad (8)$$

(worst value = $-\infty$; best value = +1)

X_i is the predicted i th value, and the element Y_i is the actual i th value. The regression method predicts the X_i element corresponding to the Y_i element in the ground-truth dataset [37].

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): If outliers need to be found, MSE can be used. Because of the L2 norm, MSE is excellent at giving these points larger weights. It is evident that if the model ultimately produces a single extremely poor prediction, the squaring portion of the function will amplify

$$MSE = \frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2 \quad (9)$$

(best value = 0; worst value = $+\infty$)

Through the square root, there is a monotonic relationship between the two numbers, MSE and RMSE. Regression model orderings based on MSE and RMSE will be identical.

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2} \quad (10)$$

(best value = 0; worst value = $+\infty$)

3. RESULTS AND DISCUSSION

In the data analysis stage, the first step was to examine the relationships among the variables using a correlation matrix generated in Python. A correlation matrix is a table that displays the relationships between various variables [38]. The correlation coefficient, which indicates the strength of the relationship between two variables, is shown in each table cell. It facilitates the swift identification of trends, the understanding of connections, and the enhancement of data-driven decision-making. In Figure 6, the correlation matrices for BOG A, BOG B, RGTEC A, and RGTEC B are presented. Based on the BOG A compressor correlation matrix, an increasing outlet pressure results in a sharp increase in power consumption. A similar trend is observed for the crankcase vibration, which features a strong direct relationship with power, in which higher vibration levels correspond with increased power usage and vice versa. Besides, motor vibration also shows a significant direct influence on power consumption, with an effect comparable to that of the crankcase vibration.

These findings are in good agreement with earlier literature [39] in which it was argued that motor power consumption increases with increasing level of vibration. In contrast, it was found that a weak correlation was found between flow rate and real power, suggesting a non-linear relationship, emphasizing the limitation of MLR in capturing such patterns.

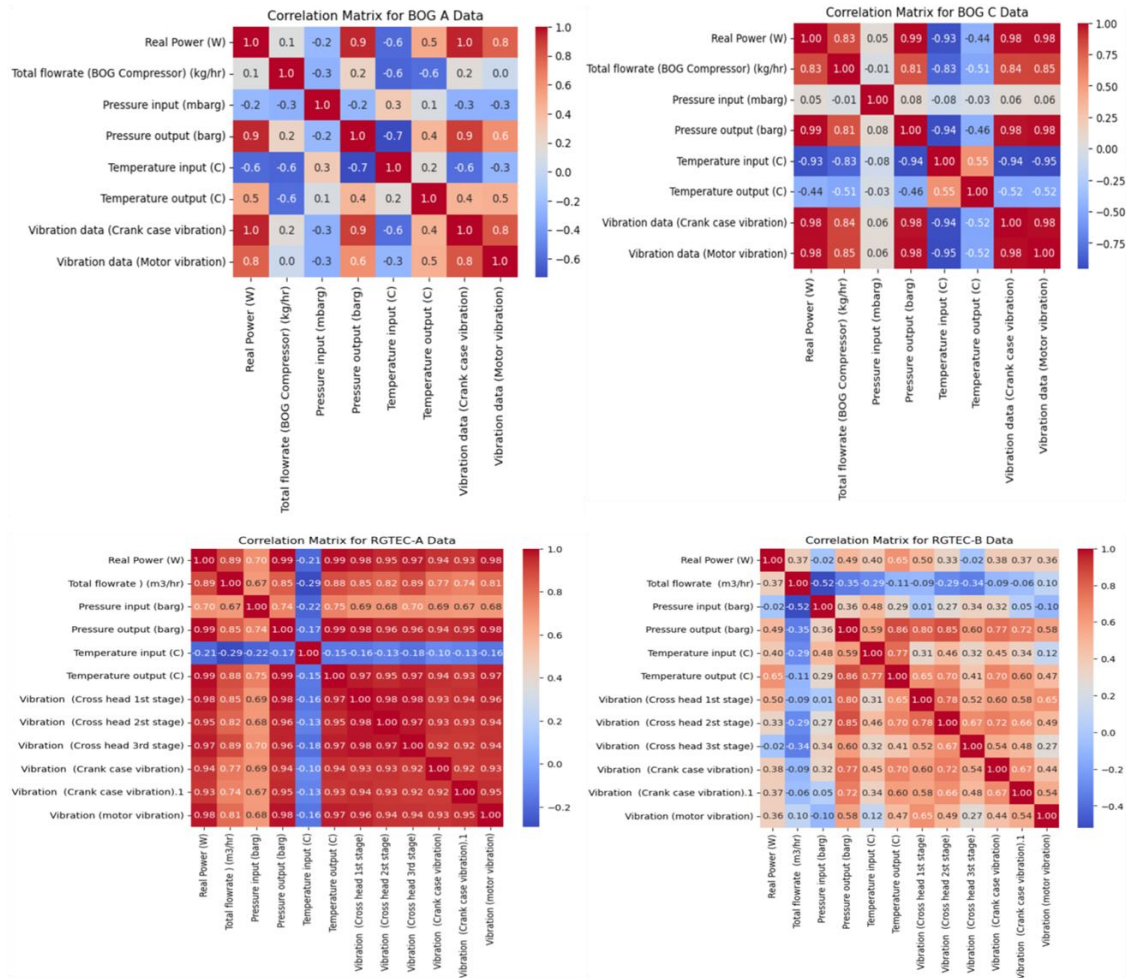


Figure 6. Correlation Matrices for BOG A, BOG C, RGTEC A, and RGTEC B

The BOG C compressor shows a similar observation to that of the BOG A compressor, in terms of the effect of outlet pressure, crankcase vibration, and motor vibration. In addition, for this case, the total flow rate also plays a strong role in the power consumption. Additionally, the indirect correlation between the compressor power and the inlet temperature is very high. Previous studies have validated that the volumetric efficiency, internal preheating, and valve behaviour depend greatly on inlet temperature, which in turn may have a bearing on the power demand, although the temperature in itself may not be directly proportional to the load [40]. This non-linear association also shows how the MLR was not able to fully describe the system dynamics.

The RGTEC A Compressor case demonstrates the strong direct influence of flow rate, outlet pressure, and motor as well as crankcase vibrations, on the power consumption, as indicated in the trend of BOG A and BOG C Compressors. In contrast, the results for RGTEC B compressor suggest a weaker effect of similar parameters.

Based on the relationship between operational variables and compressor power, the predictive models were developed and assessed. Three approaches were implemented: (1) the MLR, a conventional linear machine learning algorithm; (2) ANN, a nonlinear machine learning alternative; and (3) XGBoost. The results presented in the following section (Table 3) and the discussions presented are in terms of the average values from the 5-fold cross-validation described in the methodology section. This approach provides a reliable approximation of how well each model generalizes.

Initially, MLR was used as the baseline model for predicting compressor power consumption due to its ease of use and interpretability. Nonetheless, the intrinsically nonlinear dynamics of compressor systems restricted the MLR's predictive accuracy, although it captured the broad linear trends between operational parameters and power. Referring to the results presented in Table 3, it appears that the MLR model performed reasonably well in capturing linear relationships, but it was unable to sufficiently learn the underlying non-linear patterns in compressor power consumption.

The TensorFlow library of Python in Google Colab was used to train the ANN. The dataset was split into training and test sets, with 80% allocated to training and 20% to testing. A trial-and-error process was used to determine the best ANN model with a high R^2 and a low RMSE on the data. For this purpose, model tuning was done multiple times by varying hyperparameters, including the number of neurons, the number of hidden layers, and the optimizer's learning rate. Figure 5 summarizes the details of the selected best-performing model for each compressor, including the optimizer, loss function, number of epochs, and validation strategy.

Figure 7 illustrates the convergence dynamics of the training models and displays the final learning curves of the models. The number of training epochs to be used was well justified by the analysis of the loss curve in

Figure 7, which produced the most optimal performance indicators such as RMSE and R^2 , and this was used to set the number of training epochs to 1500 across all compressors. It is important to mention that the curves of validation loss did not change significantly during the longer training process, which indicates that the selected number of epochs was sufficient to ensure proper convergence without influencing the overall performance of generalization. The generalization ability of the models was tested on an independent test data that showed that the selected number of epochs was not causing overfitting. The Adam optimizer was found to be more predictive and convergent compared to other algorithms; thus, it was employed in all final models. It was also particularly effective at controlling the non-linear correlations within compressor datasets due to its momentum-based updates and adjustable learning rate procedure. This choice of Adam was further justified by the fact that Stochastic Gradient Descent (SGD), which was also tried during the optimization phase, was slower to converge and did not show any noticeable reduction in error figures. The learning rate was 0.01 with most compressors, but in the case of RGTEC B, the rate was increased to 0.1 to obtain the best performance and converge more quickly; this is because the unit is nonlinear and more complex than BOG compressors. The ANN models were found to be superior to the baseline MLR models in all cases, in terms of noticeable improvement in accuracy and generalization. This demonstrates that, in contrast to the linear statistical tools, neural networks provide a better framework in which to capture the non-linear characteristics of compressor load that are dynamic. Table 3 shows the performance comparison of Multiple Linear Regression (MLR), Artificial Neural Network (ANN), and XGBoost models on each of the four compressors, in terms of Root Mean Square Error (RMSE) and coefficient of determination (R^2). The ANN model was constantly better than MLR in all cases. ANN was able to predict BOG A compressor with an R^2 of 94.66% versus 92.85% of MLR. This is evidence that ANN was more precise in modeling the load variations. Likewise, ANN greatly reduced the magnitude of the errors over the MLR with BOG C and RGTEC A compressors, with almost perfect accuracy, where R^2 was 99.64% and 99.64, respectively. This result shows the extent to which ANN replicates nonlinear interactions in more complicated operating conditions, where MLR was unable to generalize.

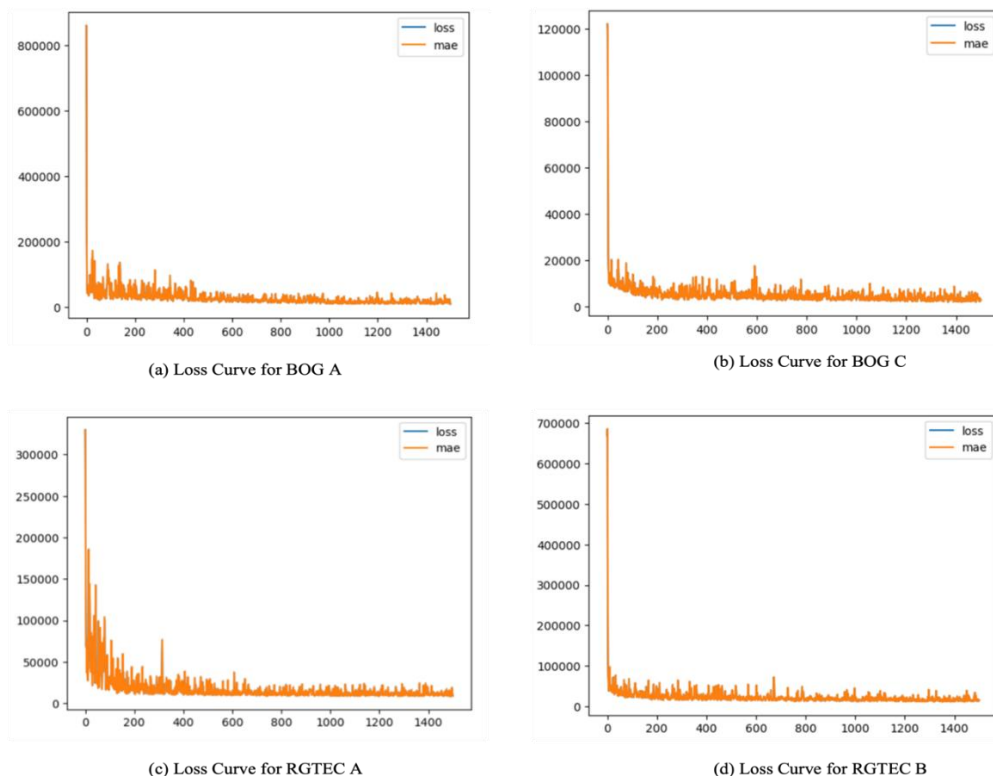


Figure 7. Loss Curves of ANN Model for Each Compressor Type

After the implementation of ANN, Extreme Gradient Boosting (XGBoost) was assumed as a second non-linear baseline. XGBoost is a distributed and optimized gradient-boosting library that is highly efficient and scalable, and is presented as a machine learning algorithm based on the Gradient Boosting framework. The model was set to use a learning rate of 0.03, 2000 estimators, and an internal validation set with early stopping (50 rounds) to avoid overfitting in the 5-fold cross-validation process.

Table 3 presents the performance comparison of MLR, ANN, and XGBoost models of each of the four compressors. Although the ANN model was always doing better than the baseline MLR, the emergence of XGBoost became a serious baseline on high precision energy prediction. In the case of the BOG A, BOG C, and RGTEC A compressors, ANN, as well as XGBoost, had almost flawless accuracy, with figures are 94% and 99%, respectively.

Table 3. Performance Comparison of MLR, ANN, and XGBoost Predictive Model

Compressor Type	Models	Max. Value of Power (W)	Average Value of Power (W)	RMSE (W)	Rr2-Score (%)	Best Model
BOG A	MLR			36553.69	92.85%	
	ANN	1485435.00	1131094.37	32780.62	94.66%	ANN/XGB
	XGBOOST			53907.32	85.45%	
BOG C	MLR			42780.00	98.15%	
	ANN	1195888.33	6641.19	15440.71	99.64%	ANN/XGB
	XGBOOST			35941.06	98.28%	
RGTEC A	MLR			16065.80	99.67%	
	ANN	868898.33	588911.96	15619.62	99.64%	ANN/XGB
	XGBOOST			18067.09	99.35%	
RGTEC B	MLR			42822.27	70.26%	
	ANN	860515.00	694412.06	28121.27	84.26%	XGB
	XGBOOST			19063.47	94.17%	

The largest divergence in performance was observed with the RGTEC B compressor, which was found to have the most complicated nonlinear behavior. The XGBoost model was chosen in this case because it is the best prediction model with an accuracy of 94.17 percent and significantly lower error rates compared to ANN and MLR. This finding means that in large-stochasticity industrial data sets, ensemble tree-based architectures, such as the XGBoost, can be stronger than traditional feed-forward neural networks. The result shows that ANN and XGBoost are capable of addressing non-linear correlations as well as dynamic shifts effectively and are more accurate in prediction compared to the various sorts of compressors. In spite of the fact that the ANN is near perfection when operating with BOG units, XGBoost is a more robust element when dealing with high-stochasticity intelligible context, and as such, it works better with the RGTEC B compressor, where it performed better than other models. In real-world energy prediction of LNG compressor systems, these non-linear models are more reliable since they are generalizable to unobserved data, whereas MLR can be useful in making quick linear approximations.

Graphical plots were used to compare predicted values with the actual test data of the intended models, such as MLR, ANN, and XGBoost, to evaluate the performance of the model further (Figure 8). It is completely evident in the comparison that whilst the predictions of MLR tend to follow the overall trend of the data, they vary much more where non-linear behavior is also high. ANN and XGBoost predictions, in their turn, are able to appropriately reflect linear and non-linear variations and demonstrate significantly stronger correspondence with the real test data.

Due to changes in the conditions of industrial plants over time, due to mechanical wear or season, it is necessary to refresh the models periodically. The suggested models do not need a lot of computing power. The example of a 6-layer ANN and XGBoost ensemble should also take no more than 90 to 120 seconds to train on an ordinary computer with a 4-core CPU and 8GB of memory. The inference phase is also extremely quick, with only under 10 milliseconds, thus readily able to utilize these models in real time with current SCADA or Distributed Control Systems (DCS). This implies that the predictive framework is precise, scalable, and simple to maintain within a normal IT environment of an LNG facility.

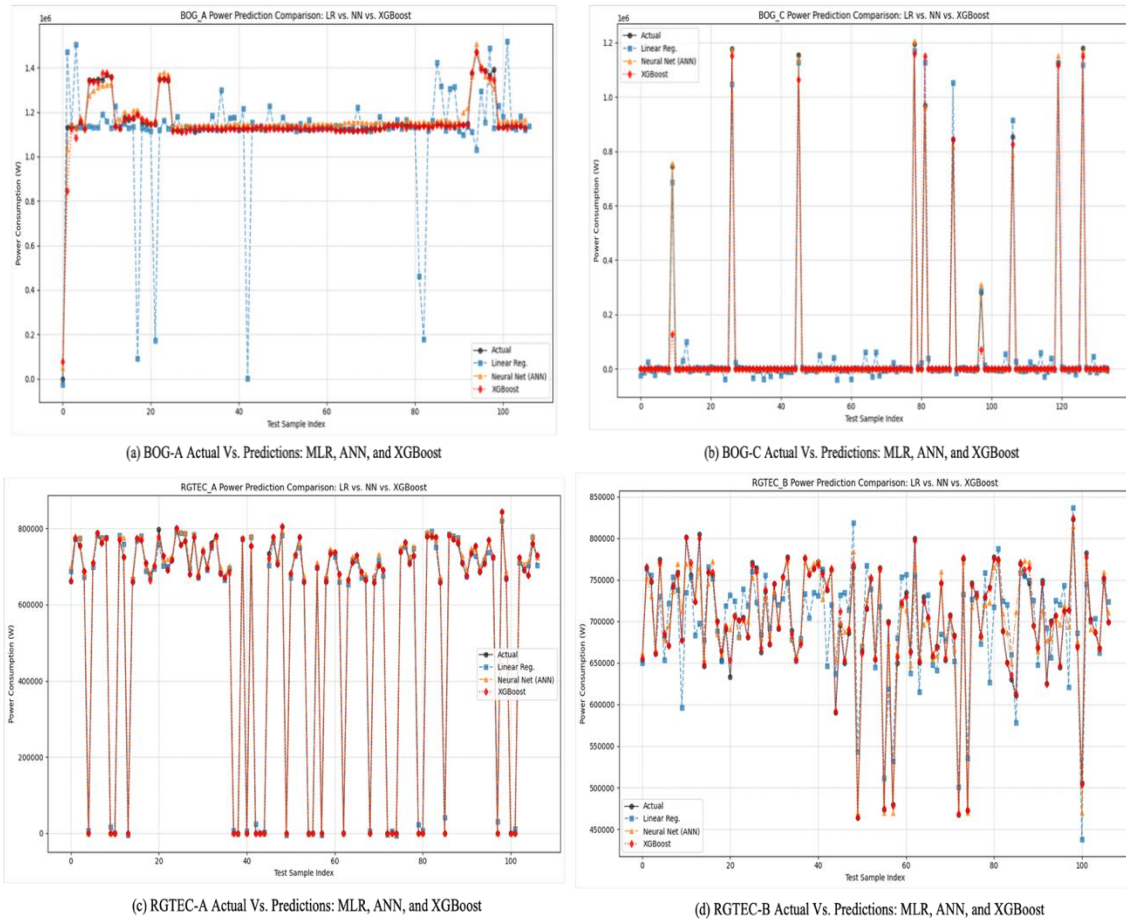


Figure 8. Actual vs. Predicted comparison of Power Consumption of Various Compressors using MLR, ANN, and XGBoost.

A. Model Explainability via SHAP Analysis

SHAP (Shapley Additive exPlanations) was adopted to enhance model interpretability to deal with the issues of the black box commonly seen with complex machine-learning models in critical infrastructure. This method enables one to decompose the decisions made by the prediction and make sure that the logic of the model is in accordance with thermodynamic characteristics[41].

3.A.1. Model Interpretability: ANN Feature Attribution

The ANN models were developed into SHAP summary plots of all types of compressors. As Figure 9 shows, the model has strong physical foundations. The SHAP plots indicate that these models are not only black boxes but are also concerned with parameters that are physically sound. As an illustration, the key factors influencing power consumption in all four compressors include Pressure output (barg) and Vibration data (Motor vibration). The values of SHAP demonstrate a definite tendency. The red dots of the values of Pressure output and Vibration, where the SHAP value is high, indicate that they are associated with an increase in the power demand. Conversely, the negative SHAP values tend to occur at lower values (blue dots), which means that when the loads are light, less electrical work is needed. The fact that the results of this model match the basic principles of thermodynamics, where increased discharge pressure and more vibration require more power, reinforces the validity of the model in real-time monitoring of energy.

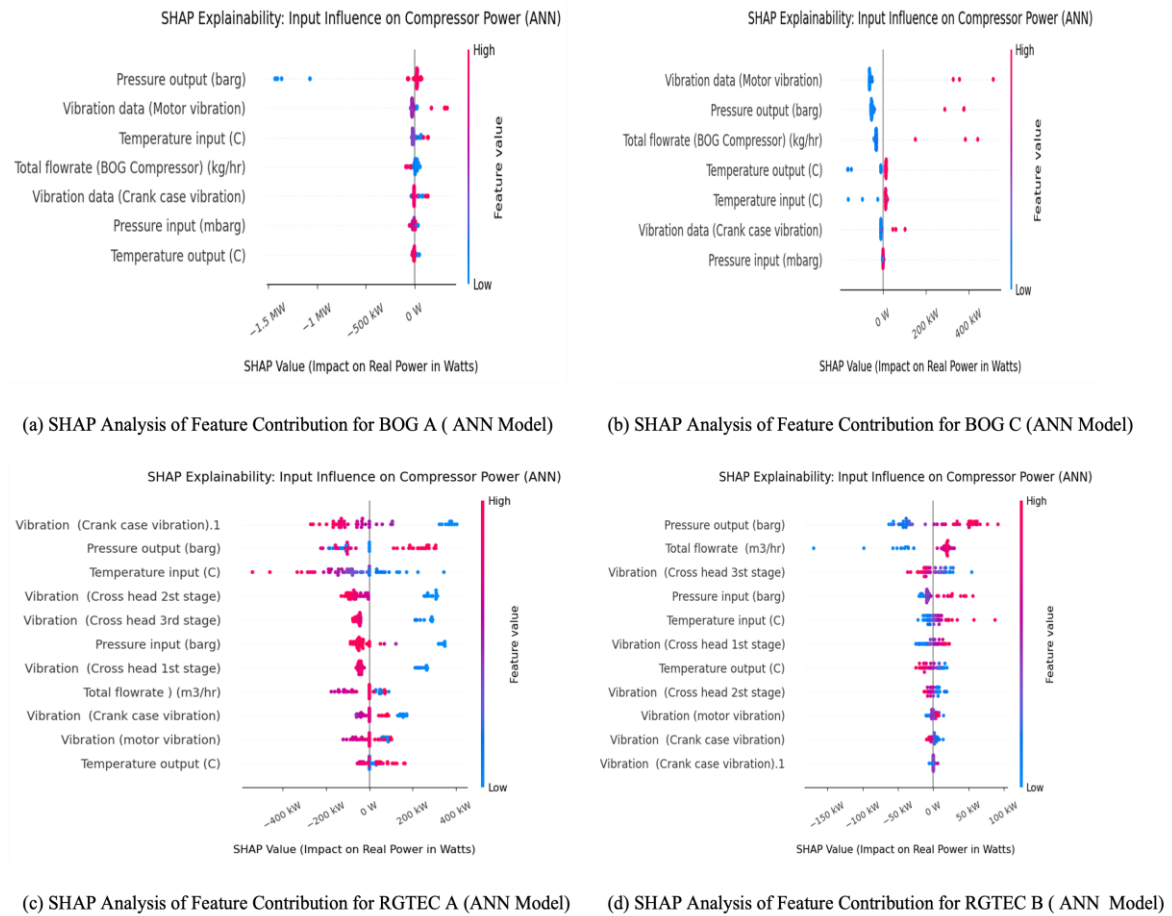


Figure 9. Features influence and Attribution using SHAP for ANN-based Power Prediction

3.A.2. Model Interpretability: XGBoost Feature Attribution

To make the ensemble-based approach more transparent, we used SHAP (SHapley Additive exPlanations) analysis for the XGBoost models (see Figure 10). XGBoost uses an ensemble of decision trees to find nonlinear patterns, and SHAP helps show how these trees make predictions about power demand. The summary plots of the four compressor units indicate that the models are the most sensitive in Pressure output (barg) and Total flowrate which are the most important features in each setup. The SHAP value distributions confirm that the model correctly identifies a positive relationship between Pressure output and motor power use, with higher values (red dots) pushing the prediction further to the right on the x-axis. For the RGTEC B unit, where the model performed best, SHAP analysis also shows that the model can identify detailed interactions between mechanical vibration signals and real-time load changes. This feature attribution helps solve the "black-box" issue when applying the model in crucial LNG infrastructure by providing engineers with an easy-to-measure method to confirm that the model's predictions are based on actual thermodynamic factors.

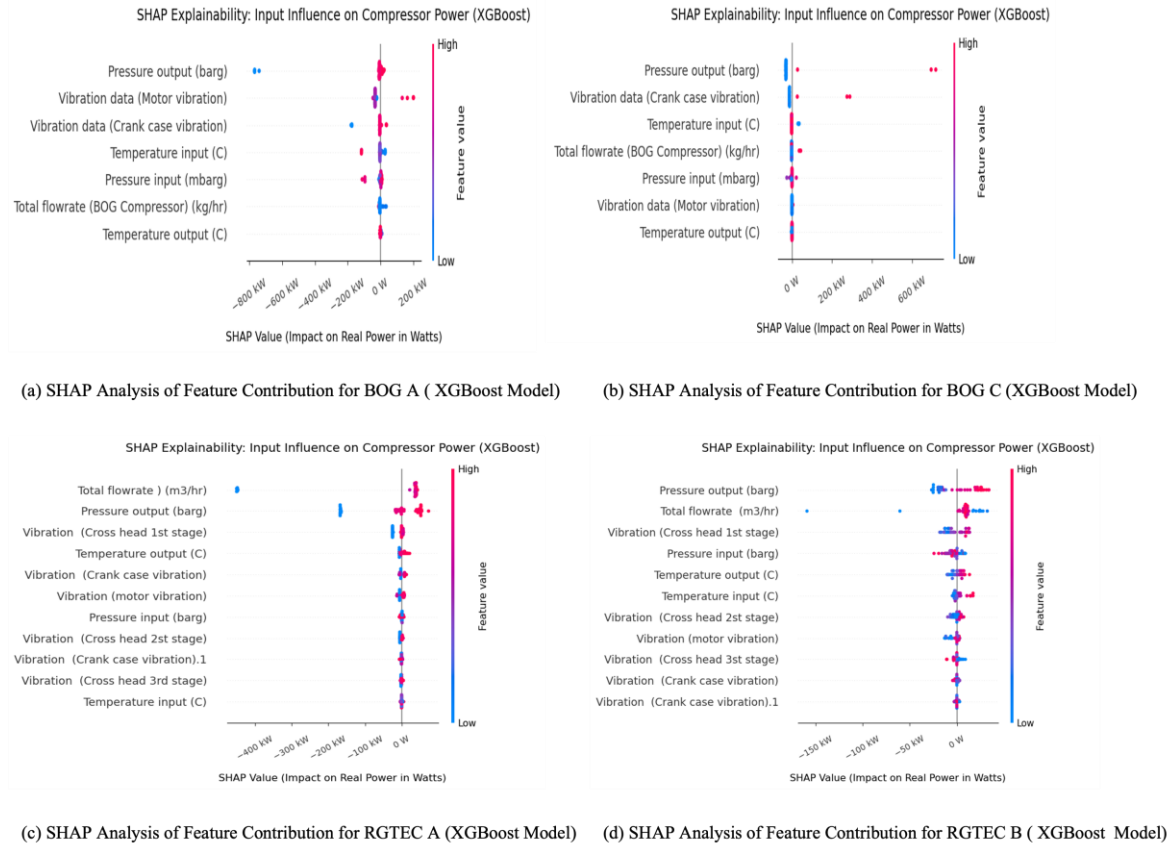


Figure 10. Features influence and Attribution using SHAP for XGBoost-based Power Prediction

4. CONCLUSION

Based on the key findings from this research, the following conclusions can be drawn:

- The main factors that influence the compressor load are discharge pressure and flow rate, and non-linear interactions limit the predictive performance of conventional linear models such as MLR.
- Across all compressors, both ANN and XGBoost models consistently outperformed the MLR model. More specifically, for the RGBET-B unit, with the most complex and stochastic operating environment, XGBoost showed the most effective predictive framework. Moreover, the Shapley Additive exPlanations (SHAP) analysis achieved enhanced transparency, suggesting that both ANN and XGBoost architectures prioritize physically useful parameters, such as discharge pressure and mechanical vibration, as per thermodynamic principles.
- A stronger connection between high-accuracy machine learning and operational trust is achieved due to the dynamics between model feature attribution and the physical system.
- Some of the beneficial features that these models can offer to the industry include an improvement in energy management, a reduction in operating costs, and environmental performance enhancement in LNG regasification plants.
- From an academic point of view, this study provides a thorough benchmarking framework and shows that nonlinear, ensemble, and neural network models are more reliable, interpretable, and scalable for energy forecasting, unlike conventional statistical methods.

More advanced deep learning models, such as hybrid systems or recurrent neural networks coupled with these prediction tools with optimization platforms to support real-time decision-making in large LNG operations, should be considered for future research.

ACKNOWLEDGMENTS

The authors are grateful to Petronas Academia Collaboration Dialogue 2023 for providing financial support, as indicated by the grant number INDUSTRI/PRSB/2024/FTKM/I00099. The authors would also like to acknowledge the Centre of Robotics, Instrumentation and Automation (CeRIA), Universiti Teknikal Malaysia Melaka (UTeM), for providing research facilities and support.

REFERENCES

- [1] M. S. Khan, M. A. Qyyum, W. Ali, A. Wazwaz, K. B. Ansari, and M. Lee, "Energy Saving through Efficient BOG Prediction and Impact of Static Boil-off-Rate in Full Containment-Type LNG Storage Tank," *Energies* 2020, Vol. 13, Page 5578, vol. 13, no. 21, p. 5578, Oct. 2020, doi: 10.3390/EN13215578.
- [2] G. Kang, J. Im, and C.-J. Lee, "Operational Strategy to Minimize Operating Cost in Lng Terminal Using a Comprehensive Boil-Off Gas Model," 2023, doi: 10.2139/SSRN.4642375.
- [3] Z. Bouabidi *et al.*, "Study on boil-off gas (Bog) minimization and recovery strategies from actual baseload lng export terminal: Towards sustainable lng chains," *Energies (Basel)*, vol. 14, no. 12, Jun. 2021, doi: 10.3390/EN14123478.
- [4] U. Humme, P. Radgen, S. Ülker, and R. Schelle, "Findings from Measurements of the Electric Power Demand of Air Compressors," *Energies* 2021, Vol. 14, Page 8395, vol. 14, no. 24, p. 8395, Dec. 2021, doi: 10.3390/EN14248395.
- [5] P. Sommani, S. Yoomak, and S. Ananwattanaporn, "Estimation of Power Consumption in Screw Gas Compressor Installed on Natural Gas Power Plant," pp. 101–110, 2024, doi: 10.1007/978-3-031-67987-2_10.
- [6] D. Gryboś and J. S. Leszczyński, "A Review of Energy Overconsumption Reduction Methods in the Utilization Stage in Compressed Air Systems," *Energies* 2024, Vol. 17, Page 1495, vol. 17, no. 6, p. 1495, Mar. 2024, doi: 10.3390/EN17061495.
- [7] M. Mukherjee *et al.*, "Greenhouse gas emissions from the US liquefied natural gas operations and shipping through process model based life cycle assessment," *Communications Earth & Environment* 2025 6:1, vol. 6, no. 1, pp. 1–14, Jan. 2025, doi: 10.1038/s43247-024-01988-2.
- [8] C. D. Davis *et al.*, "Community Health Impacts From Natural Gas Pipeline Compressor Stations," *Geohealth*, vol. 7, no. 11, p. e2023GH000874, Nov. 2023, doi: 10.1029/2023GH000874.
- [9] A. K. Chaudhari, "Integrating machine learning with industrial automation for enhanced predictive maintenance," *Using Computational Intelligence for Sustainable Manufacturing of Advanced Materials*, pp. 487–509, Apr. 2025, doi: 10.4018/979-8-3693-7974-5.CH021.
- [10] L. Zhao *et al.*, "Multi-Objective Optimization of Gas Storage Compressor Units Based on NSGA-II," *Energies* 2025, Vol. 18, Page 3377, vol. 18, no. 13, p. 3377, Jun. 2025, doi: 10.3390/EN18133377.
- [11] M. Gambelli *et al.*, "The Influence of Variable Operating Conditions and Components on the Performance of Centrifugal Compressors in Natural Gas Storage Reservoirs," *Energies* 2025, Vol. 18, Page 3930, vol. 18, no. 15, p. 3930, Jul. 2025, doi: 10.3390/EN18153930.
- [12] W. Zhang, Q. Tang, X. Gao, and R. Zhou, "Multi-Period Operational Modelling and Optimization for Large-Scale Natural Gas Networks Considering Linepack Functions in Long-Distance Transmission Pipelines," *Journal of Marine Science and Engineering* 2025, Vol. 13, Page 201, vol. 13, no. 2, p. 201, Jan. 2025, doi: 10.3390/JMSE13020201.
- [13] Y. Liu, Y. Li, G. Li, Y. Lin, R. Wang, and Y. Fan, "Review of multiple load forecasting method for integrated energy system," *Front. Energy Res.*, vol. 11, p. 1296800, Nov. 2023, doi: 10.3389/FENRG.2023.1296800/BIBTEX.
- [14] L. Zhong, R. Liu, X. Miao, Y. Chen, S. Li, and H. Ji, "Compressor Performance Prediction Based on the Interpolation Method and Support Vector Machine," *Aerospace* 2023, Vol. 10, Page 558, vol. 10, no. 6, p. 558, Jun. 2023, doi: 10.3390/AEROSPACE10060558.
- [15] Q. Peng *et al.*, "Centrifugal compressor performance prediction and dynamic simulation of natural gas hydrogen blended," *Int. J. Hydrogen Energy*, vol. 52, pp. 872–893, Jan. 2024, doi: 10.1016/J.IJHYDENE.2023.10.023.
- [16] "Ai In The Energy Industry Statistics: Reports 2025." Accessed: Sep. 08, 2025. [Online]. Available: <https://wifitalents.com/ai-in-the-energy-industry-statistics/>
- [17] D. Belayadi, D. Bendouda, N. Benali, and N. Hamouche, "Selecting an Appropriate Ensemble Machine Learning Algorithm for Intelligent Quality Management in Production," *Communications in Computer and Information Science*, vol. 2606 CCIS, pp. 211–225, 2025, doi: 10.1007/978-3-032-01922-6_18.
- [18] Z. Rais, S. Muhammad Fahmuddin, Saida, and A. T. Utomo, "Implementation of Machine Learning Algorithm with Extreme Gradient Boosting (XGBoost) Method in Hypertension Level Classification," *Journal of Applied Science, Engineering, Technology, and Education*, vol. 7, no. 1, pp. 126–136, Apr. 2025, doi: 10.35877/454RI.ASCI4191.
- [19] P. Jyotheeswari, N. Basyal, S. P. Gaire, and C. Patel, "Evaluating the impact of gamma (γ) and lambda (λ) on regression using XGBoost regressor," *Digital Transformation and Sustainability of Business*, pp. 404–407, Jan. 2025, doi: 10.1201/9781003606185-93.
- [20] "What is regasification? | Clarksons." Accessed: Aug. 27, 2025. [Online]. Available: <https://www.clarksons.com/glossary/regasification/>
- [21] "High Pressure BOG Compressor | Mitsui E&S Group." Accessed: Aug. 28, 2025. [Online]. Available: <https://www.mes.co.jp/en/products/bog/>
- [22] "pandas - Python Data Analysis Library." Accessed: Aug. 28, 2025. [Online]. Available: <https://pandas.pydata.org/>
- [23] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/J.ASOC.2019.105524.

- [24] K. Mahmud Sujon, R. Binti Hassan, Z. Tusnia Towshi, M. A. Othman, M. Abdus Samad, and K. Choi, "When to Use Standardization and Normalization: Empirical Evidence from Machine Learning Models and XAI," *IEEE Access*, vol. 12, pp. 135300–135314, 2024, doi: 10.1109/ACCESS.2024.3462434.
- [25] "StandardScaler — scikit-learn 1.7.1 documentation." Accessed: Aug. 28, 2025. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
- [26] "MinMaxScaler — scikit-learn 1.7.1 documentation." Accessed: Aug. 28, 2025. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>
- [27] G. Ciulla and A. D'Amico, "Building energy performance forecasting: A multiple linear regression approach," *Appl. Energy*, vol. 253, p. 113500, Nov. 2019, doi: 10.1016/J.APENERGY.2019.113500.
- [28] M. Lai, D. Zhang, Y. Li, X. Wu, and X. Li, "Application of Multiple Linear Regression and Artificial Neural Networks in Analyses and Predictions of the Thermoelectric Performance of Solid Oxide Fuel Cell Systems," *Energies* 2024, Vol. 17, Page 4084, vol. 17, no. 16, p. 4084, Aug. 2024, doi: 10.3390/EN17164084.
- [29] "1.1. Linear Models — scikit-learn 1.7.1 documentation." Accessed: Aug. 29, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares
- [30] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of Artificial Neural Networks and Deep Learning," *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, pp. 379–425, 2022, doi: 10.1007/978-3-030-89010-0_10.
- [31] M. G. M. Abdolrasol *et al.*, "Artificial Neural Networks Based Optimization Techniques: A Review," *Electronics* 2021, Vol. 10, Page 2689, vol. 10, no. 21, p. 2689, Nov. 2021, doi: 10.3390/ELECTRONICS10212689.
- [32] T. Szandala, "Review and comparison of commonly used activation functions for deep neural networks," *Studies in Computational Intelligence*, vol. 903, pp. 203–224, 2021, doi: 10.1007/978-981-15-5495-7_11/TABLES/6.
- [33] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Dec. 2014, Accessed: Aug. 31, 2025. [Online]. Available: <https://arxiv.org/abs/1412.6980v9>
- [34] N. D. Hoang, "Image Processing-Based Spall Object Detection Using Gabor Filter, Texture Analysis, and Adaptive Moment Estimation (Adam) Optimized Logistic Regression Models," *Advances in Civil Engineering*, vol. 2020, no. 1, p. 8829715, Jan. 2020, doi: 10.1155/2020/8829715.
- [35] "3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.7.1 documentation." Accessed: Aug. 31, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation
- [36] "train_test_split — scikit-learn 1.7.1 documentation." Accessed: Aug. 31, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
- [37] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, Jul. 2021, doi: 10.7717/PEERJ-CS.623/SUPP-1.
- [38] J. Hunt, "Pandas and Data Analytics," pp. 611–627, 2023, doi: 10.1007/978-3-031-40336-1_54.
- [39] M. Rigacci, R. Sato, and K. Shirase, "Power Consumption Simulation of Servo Motors Focusing on the Influence of Mechanical Vibration on Motor Efficiency," doi: 10.20965/ijat.2022.p0104.
- [40] D. Hu, J. Liu, F. Yi, Q. Yang, and J. Zhou, "Enhancing heat dissipation to improve efficiency of two-stage electric air compressor for fuel cell vehicle," *Energy Convers. Manag.*, vol. 251, p. 115007, Jan. 2022, doi: 10.1016/J.ENCONMAN.2021.115007.
- [41] Y. Gebreyesus, D. Dalton, D. De Chiara, M. Chinnici, and A. Chinnici, "AI for Automating Data Center Operations: Model Explainability in the Data Centre Context Using Shapley Additive Explanations (SHAP)," *Electronics (Switzerland)*, vol. 13, no. 9, May 2024, doi: 10.3390/ELECTRONICS13091628.

BIOGRAPHY OF AUTHORS



Humam Nazir is a postgraduate researcher in the Faculty of Electrical Technology and Engineering at Universiti Teknikal Malaysia Melaka (UTeM). He received his bachelor of Science in Industrial Engineering from the University of Engineering and Technology (UET) Peshawar, Pakistan. His research interests include industrial engineering optimization and machine learning for energy forecasting. He has worked as an Industrial Engineer Executive in the manufacturing industry and has contributed to multiple projects involving process improvement, value stream mapping, and deep learning based energy prediction for compressor networks. He is currently involved in research on the development of AI-driven forecasting of electrical compressor networks. He can be contacted at humamnazirengr@gmail.com



Tarmizi Ahmad Izzuddin    is a senior lecturer and researcher in the Faculty of Artificial Intelligence and Cybersecurity (FAIX) of Universiti Teknikal Malaysia Melaka (UTeM), where he has been promoting AI education and research since 2010. Dr. His international educational experience (connecting his graduate education in Shimane University in Japan with his Ph.D. education in Universiti Teknologi Malaysia (UTM)) is the core of the global perspective, which Tarmizi uses in his technical work. He also has a good industry contact with professional certifications in the fields of IBM, Skymind, and Nvidia which places him at the border of the scholarly and the hands-on. His research interests include the use of artificial neural networks, robotics, and AI in vision and signal processing in linking theoretical development with practical technological application in artificial intelligence. Dr. Tarmizi will be accessed through tarmizi@utem.edu.my.



Dr. Nur Izyan Zulkafli has her B.Eng. and M.Sc. in 2009 and 2011 respectively in Universiti Teknologi Petronas, Malaysia in the course Chemical Engineering. In 2019, she received her Ph.D. at Cranfield University, United Kingdom in Energy and Power. She teaches as a Senior Lecturer in the Faculty of Mechanical Technology & Engineering, Universiti Teknikal Malaysia Melaka (UTeM). At the moment, she is working as the Coordinator of Energy Management at the Centre of Smart Environment. She is a researcher who is involved with optimization modelling in the optimization of production and energy systems in buildings and process industries. Her research topics are energy and power efficiency, operations and maintenance planning, supply chain optimization and demand response programs. Dr. Nur Izyan may be contacted through nurizyan@utem.edu.my.