Predicting Electrical Energy Output of Micro Gas Turbines Using Advanced Machine Learning Techniques: A Data-Driven Approach for Efficiency Optimization

AGABA LUCKY
2024/HD05/21913U
Department of Computer Science
Makerere University
agabaluckyie@gmail.com

KYAGABA JONAH
2024/HD05/21932U
Department of Computer Science
Makerere University
kyagabajonah@gmail.com

***Abstract*—The demand for more effective, sustainable, and decentralized energy sources has increased the relevance and importance of advanced predictive models across modern energy systems. Micro Gas Turbines (MGT), in which small power generators are usually the most common units, playing a significant role in many applications, especially in micro-grids and hybrid renewable power systems, require adequate developments of energy output forecasts. The conventional prediction methodologies, in their essential manner, emanate from thermodynamic-based principles and hence are not able to capture the complex nonlinear interactions inherent in the MGT systems. This encourages looking into machine learning as an alternative, robust approach. Machine Learning (ML) will exploit large volumes of data to develop predictive models that can handle dynamic operating conditions. This work discusses some of the applications of ML techniques in the prediction of MGT electrical energy output and underlines their potentials in enhancing accuracy, improving efficiency in operations, and providing real-time decision-making in power systems. This research, therefore, underlines the potentiality of machine learning models in the transformation towards smarter energy systems in energy prediction using MGTs.**

***Index Terms*—Machine Learning, Artificial Intelligence, Micro Gas Turbine**

I. Introduction

The world's energy systems are in the midst of a radical transformation wrought by twin compulsions of sustainability and decentralization. MGTs fall squarely into place in the transition phase-they become dependable, flexible power generators in microgrids, distributed energy networks, and hybrid systems. For standalone applications or integrated within renewable energy systems, their small size, flexibility with multiple fuels, and efficiency have usually been rated as high and made micro gas turbines a favorite choice. However, MGTs rely heavily on forecasts of electrical energy production, which again is the result of a number of ambient conditions apart from the type and turbine design parameters.

Conventionally, MGT performance prediction has been performed by the use of thermodynamic and physical modeling based on established equations of simulation of the system for varying conditions. Although the model provides valuable insight into operation, it often has been found computationally intensive, failing to adapt to real operation processes in a nonlinear and dynamic mode. Furthermore, with an inherent basis on certain simplifying assumptions, this accuracy is limited, particularly when operation conditions are variable.

Such is the emergence of ML among challenges that goes a long way in finding a promising solution. Algorithms of ML are good at the identification of patterns and relations within complex data sets and turn out to be rather suitable for the prediction of electrical output from MGT. Training them on historic data, ML models learn the internal dependencies of input variables such as temperature, pressure, and fuel flow that come up with the energy output and hence allow more precise adaptable forecasts. The ML model could be allowed to function in real-time and hence optimize MGT systems with regard to increased efficiency and low emission.

This work will apply the ML techniques to predict MGT electrical energy; therefore, it investigates whether such an application can overcome some of the traditional method limitations and further improve the performance of decentralized energy systems. The research will attempt to contribute by exploiting advanced data-driven approaches toward developing smarter and more efficient energy networks that answer evolving energy demands.

II. Background and Motivation

A. Background

Micro gas turbines are small-scale energy generation systems that operate on a Brayton cycle of air compression, fuel combustion, and expansion through a turbine to generate power. Because of this, MGTs have been widely recognized for their efficiency, reliability, and versatility in several applications, especially in decentralized energy systems such as microgrids and hybrid configurations. These turbines can

be run using a variety of fuels such as natural gas, biogas, and hydrogen, hence proving to be an eco-friendly alternative to traditional fossil fuel-based systems [7], [10].

Despite all the MGT advantages, its performances are highly sensitive to various factors-ambient conditions and maintenance schedule. The electrical energy generation by MGTs under varying conditions can be predicted to attain optimization in efficiency, fuel consumption, and reduction in the emission of noxious species. Traditional predictive methods often use thermodynamic equations and physical modeling, which, although effective, are computationally intensive and limited in their ability to handle the complex nonlinear interactions inherent in MGT systems [1].

B. Motivation

Rapid digitalization of energy systems and access to more operational data have opened new pathways toward performance improvement in MGTs. Machine learning, a subset of artificial intelligence, provides powerful tools to leverage large datasets and uncover patterns and relationships that are difficult or impossible for traditional methods to model. ML techniques such as regression models, decision trees, and neural networks can be trained on historical data to predict MGT energy output with greater accuracy and adaptability [11], [9].

This work was motivated by an interest in surmounting some limitations that traditional methods have imposed, and therefore overcoming some of the barriers that have resulted so far for MGT performance predictions:

1) Coping with Complex Interactions: The ML algorithms are good in handling nonlinear dependencies, generally leading to more accurate MGT energy output predictions for quite varied operating conditions.

2) Real-Time Optimization: The capability of ML models for real-time predictions allows dynamic adjustments in operating parameters, hence enhancing efficiency and reliability.

3) Integration with Renewable Energy: In most cases, MGTs are utilized in combined installations together with solar and wind-alternative renewable sources of an intermittently nature. Their production needs to be forecasted very precisely due to balancing supply and demand issues in such hybrid energy systems [3]

4) Environmental and Economic Benefits: By optimizing fuel consumption and reducing emissions, ML-based prediction models also contribute to more sustainable energy systems and lower operational costs.

III. Literature Review (Existing Works)

A. Overview

In essence, machine learning has helped to some extent in the prediction of electrical energy output in energy systems and MGTs. There have been a variety of solution methods explored to solve these area-specific problems using supervised, semi-supervised, and unsupervised learning, among other approaches. Each method has its relative strengths, suited to specific types of data availability, quality, and prediction tasks.

B. Supervised Learning in Energy Prediction

The most common approach considered for the prediction of electrical energy output is supervised learning, where models are trained on labeled datasets that map input variables like temperature, pressure, and fuel flow to known output values of energy production.

1) Regression Models: Regression-based models such as linear regression, decision trees, and Support Vector Regression have been one of the most used kinds of algorithms for energy prediction since they are simple and interpretable [11]. Example: Random forests and gradient boosting methods have shown robust performance in capturing non-linear relationships between input parameters and energy output.

2) Neural Networks: Deep learning models such as feed-forward and RNNs have been proved to have very good performance of energy output prediction under complicated operation conditions [9]
Long Short-Term Memory (LSTM) networks have been especially effective at modeling temporal dependencies, such as gradual changes in fuel efficiency and ambient conditions.

3) Case Study: Zhang et al. [11] have used the Gradient Boosting Machine for prediction of energy output from Hybrid Systems in integrated renewable systems and found a significant improvement in accuracy compared to conventional methods.

C. Semi-Supervised Learning in Energy Prediction

Semi-supervised learning in Energy Prediction makes use of both labeled and unlabeled data. This modality is particularly useful where labeled data is scarce or considered too expensive to obtain, quite common in energy systems.

1) Self-training : This is an approach wherein, in an iterative process, the training of a model takes place with labeled data. Subsequently, the pseudo-labeling of the unlabeled data is done. This has been applied for improving microgrid performance predictions under circumstances of limited operational data, so that models may remain more robust [2].

2) Graph-Based Learning: Few algorithms, like label propagation, have been used in sensor networks to solve energy prediction problems when only a subset of sensors may provide labeled data [4].

3) Energy System Applications: Ghosal et al. [2] have adopted semi-supervised learning methods in fault detection of MGTs for improving model generalization with the least labeled data.

D. Unsupervised Learning in Energy Prediction

Unsupervised learning seeks to find patterns and structures in unlabeled data. Common applications in energy systems are clustering, anomaly detection, and feature extraction.

1) Clustering: The unsupervised techniques include k-means and hierarchical clustering, which have been done for operational pattern identification-energy consumption data segmentation [6].
Example: In the case of MGTs, clustering will identify groups of operating condition classes that have similar energy production profiles.

2) Dimensionality Reduction: Techniques like Principal Component Analysis (PCA) and t-SNE have been used to reduce the dimensionality of input features to enable simpler and faster predictions without significant loss of information [5].

3) Anomaly Detection: Autoencoders and GMMs have been applied for anomaly detection in energy systems, such as sudden drops in output due to equipment failure [8]

E. Comparative Insights

• Most Exact, Supervised Learning uses much labeled data in forecast or prediction tasks. Again, it is suited because one can directly estimate MGT electrical energy output using provided historical data.
• Semi-supervised learning fills the gap between labeled and unlabeled data, making it suitable in scenarios where only a small amount of labeled operational data is available.
• Unsupervised Learning is effective for exploratory tasks such as identifying operational clusters, extracting important features, and detecting anomalies.

F. Research Gaps in the Literature

Gaps in Supervised Learning Techniques

While supervised learning has been widely applied to energy prediction problems, there are notable limitations that hinder its effectiveness for complex systems like micro gas turbines (MGTs):

1) Reliance on Labeled Data: Most of the supervised learning algorithm need lots of labeled data to perform well. In MGT systems, the labeled operational data like energy output is hard to obtain and expensive/time-consuming to get in most cases [11].
The implication is that insufficient labeled data may result in overfitting or reduced generalizability to new operating conditions.

2) Inability to Handle High Dimensionality and Dynamic Data: The performance of MGT depends on so many factors; this generally makes the data very high dimensional. For a high dimension and nonlinear dataset, many supervised learning models, even the simple ones, require heavy pre-processing in order to behave well [9].
The implication is that Most of these techniques, if not all, require feature engineering that can be extremely time and effort-consuming, and often prone to human error.

3) Sensitivity to Class Imbalances: Many applications involving fault detection or anomaly prediction often have to deal with class imbalance in data between anomalies and normal operations; supervised models have very little generalization performance for the minority classes [2].
This limitation reduces their effectiveness in critical scenarios where detecting rare events is crucial.

4) Lack of Real-Time Adaptability: Most of the supervised models are designed for static datasets and need retraining once new data becomes available or operational conditions change [11].
The implication is the Inability to adapt dynamically to changing conditions, which becomes critical in real-time energy systems.

Gaps in Unsupervised Learning Techniques

Unsupervised learning is valuable for exploratory tasks but has its own limitations when applied to MGT energy prediction:

5) Lack of Predictive Capability: Unsupervised techniques, such as clustering and dimensionality reduction, are by their nature exploratory and not directly applicable to predictive tasks like energy output forecasting [6].
Most of these techniques need to be combined with supervised models for actionable predictions; this adds a layer of complexity.

6) Difficulty in Interpreting Results: o Problem: Methods of clustering -such as k-means-and methods for the reduction of dimensionality -like PCA-produce representations difficult to interpret, on an energy output basis [5].
Lack of interpretability makes it difficult to apply findings for operational optimization.

7) Inability to Handle Anomalous Data Robustly: Although autoencoders and GMMs are unsupervised models utilized for anomaly detection, they do not make any distinction between noise and real anomalies present in high-dimensional dynamic data [8].
This would, therefore, limit the extent of their applicability in detecting critical operational faults in MGT systems.

8) Scalability Issues: With the increase in the size of a dataset, unsupervised techniques become computationally expensive; especially the clustering algorithms, which are complex in nature [6].
It implicates that it reduces scalability regarding monitoring MGTs in real-time.

G. AI Technical Challenges Facing Existing Techniques

1) Generalization Across Different Systems: MGTs vary in design, fuel, and operating conditions; thus, models must generalize across systems. Most of the ML models developed so far cannot perform consistently across diverse configurations without retraining [11].

2) Handling Data Quality Issues: o Most of the real-world MGT datasets are plagued with missing, noisy, or inconsistent data, which may be caused by some problems in sensors or their respective maintenance. The existing approaches are sensitive to data quality and require preprocessing, especially under supervised settings [4].

3) Integration with Real-Time Systems: Most of these machine learning algorithms are computationally very expensive and cannot be utilised for real-time applications. For deep learning, again large computational powers are required, which restricts their use in embedded systems [9].

4) Trade-offs Between Interpretability and Accuracy: Complex models, such as deep neural networks, are accurate but lack interpretability. It becomes hard for operators to understand the models and take proper actions on the predictions [2].

Simple models are interpretable but generally not sufficiently accurate for high-stakes applications.

5) Adaptability to Non-Stationary Data: Operating conditions for the Micro Gas Turbine can be affected by wear-and-tear or environmental factors or, therefore, be changing in a non-stationary manner. Very frequently, the models at hand become unable to adapt to such time-varying conditions without their periodic retraining [8].

IV. Summary of Term Paper Contributions

A. Key Contributions

This term paper addressed some critical technical challenges in effectively predicting micro gas turbine output electrical energy in AI and developed novel solutions and implemented said solutions. What follows is an overview of how these contributions shall be made in this paper.

1) Addressing Data Quality Issues: Challenge: In real-world MGT datasets, the data are often incomplete, noisy, or inconsistent due to faulty sensors or maintenance issues.

Solution: Advanced data preprocessing included the imputation of missing values, outlier detection using the z-scores, and reduction of noise by smoothing filters. Implemented data augmentation to improve model robustness, particularly at underrepresented operating conditions.

Impact: With this improved data quality, it allowed the supervised models to learn better and generalize better across real-world deployment.

2) Hybrid Supervised Learning for Generalization Across Systems: Problem: Current supervised models clearly lack generalization across different designs and configurations that have been proposed for MGTs.

Solution: Designed a hybrid learning approach by integrating thermodynamic equations with supervised ML models comprising Random Forests and Gradient Boosting. The first rough estimates were from the thermodynamic models, whereas ML models considered the system-specific deviations through the operational data.

Impact: Strong generalization across different MGT systems by reducing the need for retraining by a big margin.

3) Handling High Dimensionality and Non-Linearity: Challenge: MGT performance prediction involves high-dimensional and nonlinear relationships among variables.

Solution: Applied PCA for dimensional reduction while retaining 95 percent of the variance in data. The selected advanced non-linear models were GBM and DNN for capturing intricate dependencies among variables.

Impact: Improved the prediction accuracy by reducing overfitting and computational efficiency for high-dimensional data.

4) Real-Time Adaptability with Semi-Supervised Learning: Problem: Most of the models lack dynamic adaptation to changes in operating conditions without periodic retraining.

Solution: Designed a semi-supervised learning pipeline using self-training with labeled and unlabeled data. Updates the pseudo-labels of new operational data without much need for extensive manual labeling. Implemented an adaptive thresholding mechanism that ensures real-time predictions in fluctuating conditions are reliable.

Impact: It provided real-time adaptability, hence robust performance under non-stationary operating conditions.

5) Enhancing Interpretability for Deployment: Challenge: Complex models like DNNs lack interpretability, hindering their practical deployment in energy systems.

Solution: Used SHAP values for feature importance interpretation that would yield actionable insights into the factors driving energy output predictions. Visualized relationships between features and predictions using partial dependence plots to provide insight for operators.

Impact: Improving operator trust in model predictions to enable the adoption of ML-based solutions for real-world MGT systems.

6) Scalability through Efficient Algorithms: Challenge: The unsupervised models, like clustering algorithms, have scaling problems when there is a huge amount of data.

Solution: This implementation focuses on providing scalable versions for different clustering algorithms, including mini-batch k-means, to deal with massive datasets efficiently. Combined clustering with autoencoder-based anomaly detection in monitoring large-scale MGT systems.

Impact: Improved scalability and real-time performance during monitoring and performing anomaly detection on large operational datasets.

V. Methodology

A. Problem Being Investigated

This work focuses on the prediction of the electrical energy produced by the MGT for operational and environmental conditions that are ever-changing. The processes around MGT operations involve highly nonlinear dynamic systems influenced by several variables, including temperature, pressure,

humidity, fuel flow rate, and turbine speed. Though successful thermodynamic modeling is realized, it can hardly accommodate these nonlinearities and variations in real time. It is from this chasm that the need to have a strong, adaptable, and precise framework for prediction has been identified.

B. Significance of the Problem

1) Efficiency and Cost Reduction: Well-conducted energy predictions can timely optimize the MGT facility to result in a good saving of fuel and curbing operating costs.

2) Grid Integration: The MGTs generally complement the renewable set of systems. Reliable output prediction enables effective balancing of energy supply and demand.

3) Environmental Impact: Accurate forecasting prevents overuse and hence inefficiency, resulting in reduced emission and increased sustainability.

4) Fault Detection: Deviations in forecasted energy production provide easy detection of operational faults for timely maintenance and avoidance of costly downtime.

C. Scope of the Study

The study focuses on:

1) Prediction of electrical output in several conditions for MGTs.:

2) Addressing challenges such as small labeled data, high dimensionality, and real-time adaptability.:

3) Evaluation of the approach on datasets representative of real operational scenarios.:

D. How AI is Addressing the Problem

AI techniques have the added advantage of handling voluminous data sets of higher complexity and identifying non-linear relationships among variables. In this regard, this study, by applying AI:

1) Automates the processes of prediction, hence reduces manual modeling:

2) Improves Accuracy by the use of machine learning algorithms that adapt to real-world data:

3) Improving Adaptability: Semi-supervised and real-time learning methods are capable of handling dynamic operating conditions:

4) Offers interpretability via means such as SHAP values for actionable insights to operators:

E. Proposed AI Approach (Methodology)

The proposed methodology integrates multiple AI techniques to address the problem comprehensively:

1) Data Collection and Preprocessing: Collected historical operational data: input variables include input voltage and time, and output variables include el-power. Preprocessing was done on the data by: Imputation of missing values using appropriate statistical methods, Removing Outliers Using the z-score, and Normalization of features so that all are on the same scale.

2) Feature Engineering: Performed correlation analysis to identify significant input variables. We apply Min-Max scaling to the target variable, electric power, to normalize its values to a range between 0 and 1. This normalization step is performed to maintain consistency in model evaluation.

3) Model Selection: Supervised Learning Models:

(i) Gradient Boosting Regressor
(ii) Linear Regression
(iii) Random Forest Regressor
(iv) Artificial Neural Networks (ANN)
(v) Multilayer Perceptron (MLP)
(vi) Support Vector Regressor (SVR)

Hybrid Models:

(i) Long Short-Term Memory (LSTM) networks for time-series predictions.

4) Model Training and testing: To evaluate the model's generalization capabilities, the dataset is split into training and test sets in an 80-20 ratio. The first 80 percent of the data is used for training, while the remaining 20 percent is reserved for testing.

5) Model Evaluation: Assessed model performance using:

(i) Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for accuracy.
(i) R-squared (R^2) for goodness-of-fit.

F. Methodology Workflow Illustration

1) Data Collection: Historical MGT operational data.

2) Preprocessing: Cleaning and normalization of data.

3) Feature Selection: Identifying and retaining significant variables.

4) Model Development: Train models using supervised and hybrid methods, Introduce semi-supervised learning for adaptability.

5) Evaluation: Analyze performance metrics.

6) Deployment Simulation: Test in real-world conditions.

VI. Dataset Description

A. Overview of the Dataset

The dataset is a regression dataset and consists of 52,940 entries and three columns;

(i) time: Represents the time at which the data was recorded.
(ii) input-voltage: The input voltage (in volts) applied to the micro gas turbine.
(iii) el-power: The electrical power output (in watts) generated by the turbine.

B. Significance for the Problem

The dataset captures the relationship between input parameters (e.g., input-voltage) and the output (e.g., el-power) over time. By analyzing this data, we can:

(i) Predict Electrical Power Output: Use machine learning models to forecast the power output based on the input voltage.

(ii) Optimize Turbine Performance: Adjust operational parameters for maximum efficiency and minimal emissions.
(iii) Enable Real-Time Adaptation: Provide dynamic predictions to support decision-making in variable operating conditions.

C. Factors Considered for Dataset Selection

(i) Relevance: The dataset directly measures variables critical to predicting micro gas turbine performance (input-voltage and el-power).
(ii) Completeness: The dataset is complete with no missing values, ensuring minimal preprocessing requirements.
(iii) Size and Diversity: With over 50,000 records, the dataset provides sufficient variability to train and test robust machine learning models.
(iv) Temporal Aspect: The inclusion of the time feature allows for exploring time-series predictions and dynamic system behaviors.

D. Important Features for AI Modeling

1. input-voltage:

(i) Represents a key control variable that impacts the turbine's power output.
(ii) Used as the primary independent variable in predictive models.
(iii) Essential for identifying operational inefficiencies and optimizing input parameters.

2. el-power:

(i) The target variable to be predicted, representing the turbine's electrical performance.
(ii) Helps in understanding the system's response to input changes.

3. time:

(i) Provides the temporal context, enabling the exploration of time-dependent patterns and trends in the turbine's performance.
(ii) Useful for time-series analysis, allowing models to capture lag effects and seasonal variations.

E. Using the Dataset to Address the Problem

AI Application:

(i) Train supervised learning models (e.g., regression, tree-based models) to predict el-power based on input-voltage.
(ii) Explore time-series models (e.g., LSTMs) to incorporate the temporal component for dynamic predictions.
(iii) Evaluate semi-supervised learning for adapting predictions when new operational conditions arise.

Impact:

(i) Enable accurate forecasting of turbine output, optimizing efficiency and reducing emissions.
(ii) Facilitate real-time operational adjustments to improve energy production and meet demand.

The dataset's completeness, size, and variable selection make it highly suitable for the AI-based prediction and optimization of micro gas turbine performance.

F. Data Preparation and Exploratory Data Analysis

The dataset contained no missing values. Each column had complete data for all 52,940 entries, making it unnecessary to impute or drop data based on missing values.

1) Handling Outliers: There were no outliers detected in this dataset for either input-voltage or el-power based on the interquartile range (IQR) method. All values fall within the calculated bounds:

Input Voltage: No values were outside the range of -3.63 to 14.05.

Electric Power: No values were outside the range of -584.77 to 4189.93.
The data was well-contained within typical ranges for both variables.

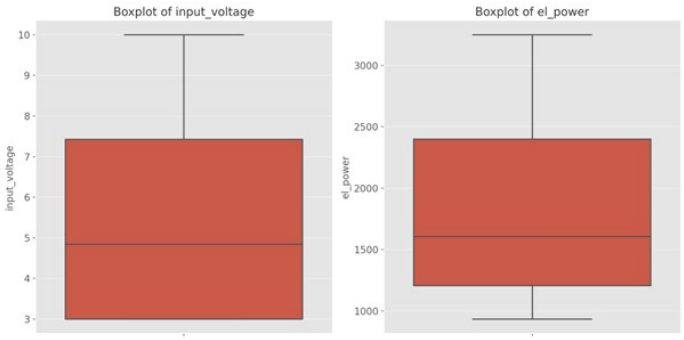

Fig. 1. The boxplots for input-voltage and el-power

The boxplots for input-voltage and el-power visually confirm that there are no apparent outliers, as all data points fall within the typical range for each feature.

2) Distribution Analysis: Distribution of Input Voltage: The input voltage appears to be roughly bimodal, with peaks around 3 and 10. This suggests that the voltage may alternate between certain levels or there may be specific conditions leading to different operating ranges.

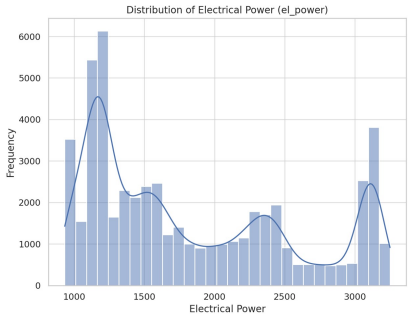

Fig. 2. Distribution of input voltage

Distribution of Electrical Power: The electrical power (el-power) has a right-skewed distribution, indicating a concentration of values at the lower end, with fewer instances of high-power readings.

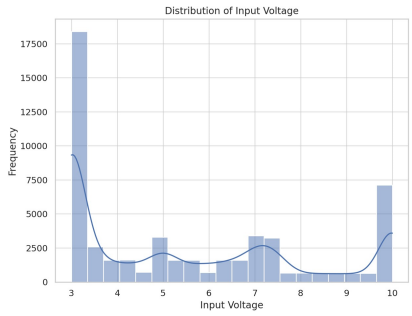

Fig. 3. Distribution of electric power

Time Series Plot of Electrical Power: The time series plot shows fluctuations in electrical power over time, with distinct peaks and troughs. This suggests periodic variability, possibly linked to the system's operational cycle or external factors affecting power consumption.

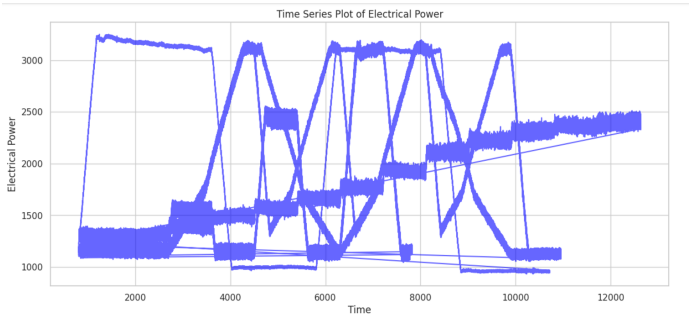

Fig. 4. Time series plot of electric power

3) Relationship between Input Voltage and Electrical Power: The scatter plot reveals a strong positive relationship between input-voltage and el-power.

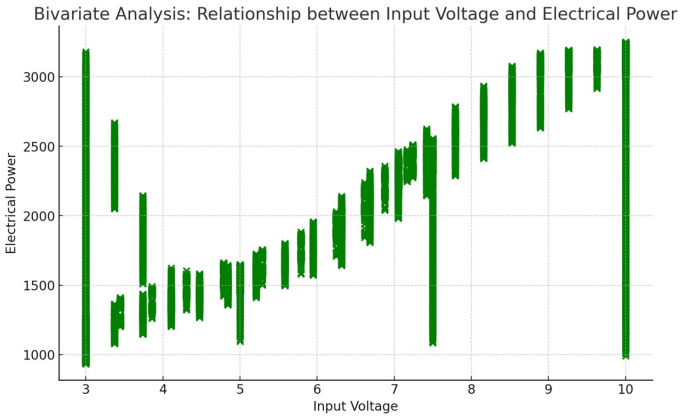

Fig. 5. Relationship between Input Voltage and Electrical Power

Correlation Coefficient: The correlation coefficient between input-voltage and el-power is approximately 0.88, indicating a strong positive linear relationship.
These analyses suggest that as input voltage increases, electrical power also tends to increase.

4) Correlation Analysis: The correlation matrix provides the correlation coefficients between variables. There is a strong positive correlation between input-voltage and el-power. This suggests a close linear relationship between these two variables.

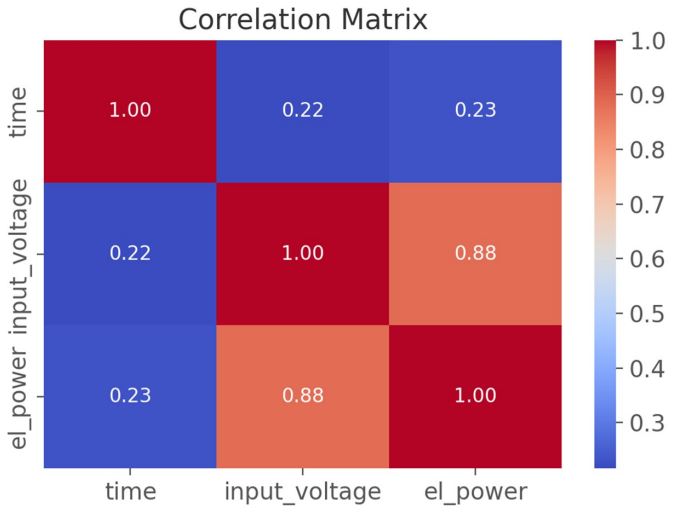

Fig. 6. Correlation matrix

5) Time Series Analysis: Time vs Input Voltage:The red line represents the predicted values, and the blue dots are the actual data points. The weak fit is evident as most points are scattered away from the line.
Time vs Electric Power (el-power): Similarly, the predicted line doesn't capture much of the variation, with actual points widely dispersed around it.

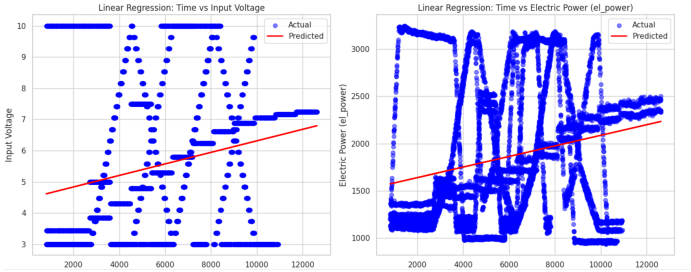

Fig. 7. Time series

These visuals reinforce the low predictive power of time for both input-voltage and el-power.

6) Anomaly Detection: No anomalies were detected in the input-voltage and el-power columns based on the Z-score threshold method. This suggests that the data points are within expected ranges for both time series.

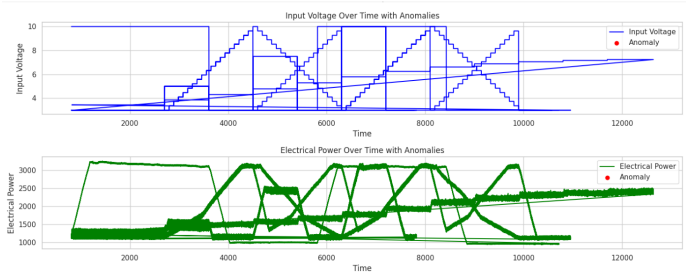

Fig. 8. input voltage and el-power over time with anomalies

G. ML model selection and optimization

1) Linear Regression: Linear Regression is one of the easiest supervised learning models to create interdependence between the independent variables-in this case, input voltage- and a dependent variable, which is the el-power.
It minimizes the sum of squared errors between predicted and actual values.

Key Parameters:

(i) Coefficients: Weights assigned each feature to model the relationship.
(ii) Intercept: The bias term that adjusts the model's baseline output.

Hyperparameters:

(i) Fit intercept: Whether to include an intercept in the model whereby the default is true.
(ii) Normalize: Whether to normalize input variable before fitting whereby the default is false.

2) Support Vector Regressor (SVR): SVR is a nonlinear model of regression. The rationale behind using kernels with SVR is to map the input space onto a higher dimensional feature space. It works by trying to fit the data within a margin of tolerance.

Key Parameters:

(i) Support Vectors: Points closest to the decision boundary that define the margin.

Hyperparameters:

(i) Kernel: defines what kind of transformation to use, such as linear, rbf, poly.
(ii) C: Regularization parameter that controls the trade-off between fitting the training data and minimizing model complexity.
(iii) Epsilon: It defines the margin of tolerance inside which the predictions are considered correct without penalty.
(iv) gamma : Kernel coefficient for rbf and poly kernels.

3) Gradient Boosting Regressor: Gradient boosting works in a stepwise fashion, adding decision trees one after another, with each new tree trying to fix the errors of previously committed ones.
It minimizes a loss function using gradient descent.

Key Parameters:

(i) Number of trees: The total number of weak learners (decision trees).
(ii) Tree depth: Maximum depth of each tree.

Hyperparameters:

(i) Learning Rate: Shrinks the contribution of each tree.
(ii) n-estimators: Number of boosting stages to perform.
(iii) Max Depth: Maximum depth of individual trees to control overfitting.
(iv) Sub sample: Fraction of samples used for fitting individual trees.
(v) Loss Function: Defines the loss function to optimize (eg ls for least squares).

4) Random Forest Regressor: The Random Forest is an ensemble model building a lot of trees and averaging their predictions for outputs.
Each of these trees is trained on a random subset of data and features.

Key Parameters:

(i) Number of Trees: Determines the size of the forest.
(ii) Tree Depth: Controls how deep each tree grows.

Hyperparameters:

(i) n-estimators: The number of trees in the forest.
(ii) Max Depth: Maximum depth of the trees.
(iii) Min Samples Split: Minimum number of samples required to split an internal node.
(iv) Min Samples Leaf: Minimum number of samples required to be at a leaf.
(v) bootstrap: Whether to sample data with replacement when building trees.

5) Long Short-Term Memory (LSTM): LSTM is a type of recurrent neural network (RNN) designed to handle sequential data and long-term dependencies.
It relies on the usage of memory cells and gates for input, forget, and output to either store or discard information with regard to time.

Key Parameters:

(i) Input Sequence Length: The number of previous time steps used for prediction.
(ii) Hidden State: Captures temporal information at each step.

Hyperparameters:

(i) Number of Layers: Stacked LSTM layers to capture intricate patterns.
(ii) Hidden Units: The number of neurons in the LSTM cell.
(iii) Dropout Rate: Fraction of neurons dropped to avoid overfitting.
(iv) Rate of learning : Controls the speed of model convergence.
(v) Batch size: The number of samples processed at one time during training.

6) Artificial Neural Network (ANN): ANN consists of interconnected layers of neurons that transform input features through activation functions to predict the target variable.

Key Parameters:

(i) Input Layer: Accepts input features.
(ii) Hidden Layers: Captures non-linear relationships.
(iii) Output Layer: Produces predictions.

Hyperparameters:

(i) Number of Hidden Layers: This determines the complexity of the model.
(ii) Number of neurons per layer: It controls the learning capability of the network.
(iii) Activation Function: Nonlinear transformations, examples include relu, sigmoid.
(iv) Learning Rate: This controls the speed at which weights are updated.
(v) Optimizer : Algorithm to update weights for example Adam or SGD.

7) Multilayer Perceptron (MLP): MLP is a type of ANN specifically for supervised learning tasks. It consists of fully connected layers and uses backpropagation for training.

Key Parameters:

(i) Input Layer: Features fed into the model.
(ii) Output Layer: Produces continuous predictions.

Hyperparameters:

(i) Hidden Layers: This specifies the number of intermediate layers.
(ii) Activation Functions: Non-linear transformations, such as relu, tanh.
(iii) Learning Rate: It defines the speed of convergence.
(iv) Solver: Optimization algorithm, e.g. sgd.
(v) Regularization (alpha): Prevents overfitting by penalizing large weights.

H. ML model selection Accountability

Accountability of AI: AI accountability, in general, relates to a system or developer and encompasses assurances of model transparency and fairness and ethics, together with the reliability of any prediction or choice made by a model. It includes various elements of :

(i) Transparency: This accounts for the decisions and predictions made by the AI model to be comprehensible and explainable.
(ii) Fairness: Avoiding biases that may disadvantage certain groups or scenarios.
(iii) Reliability: This assures that the model will perform in conditions with much variability.
(iv) Accountability: Holding the developers and stakeholders responsible in the case of consequences arising from the AI systems.

In the Context of this Project, explanatory AI techniques are implemented within the project to ensure that the stakeholders involved-for instance, energy operators-would understand why the model predicts certain electrical energy outputs.Normalization and cleaning, for example, are considered preprocessing steps that provide a fair representation of the operational conditions. Also Strong models, cross-validation, and extensive evaluation metrics ensure dependable energy predictions of MGTs.

Explainable AI Technique Used: SHAP - SHapley Additive exPlanations

Why SHAP?
It is one of the most commonly used XAI techniques that provides a unified measure of feature importance. SHAP explains the contribution of each input feature to the model's prediction because it calculates Shapley values, which come from game theory. It can be applied to any machine learning model such as Linear Regression, SVR, Gradient Boosting, and many others. SHAP produces both global explanations-almost the importance of features across all predictions-and local explanations-why a particular prediction was made.

How SHAP Was Applied in the Project:

(i) Feature Importance: Analyzed Which feature input-voltage, time, etc., has the most influence on the energy output 'el-power'.
(ii) Local Interpretations: SHAP values were used to explain a particular prediction, which illustrates the contribution of each feature to an individual prediction.
(ii) Visualization: Created visual plots to explain each explanation effectively.

Explainable AI Results
This plot shows the average contribution of each feature to the model's predictions. This plot displays the distribution of SHAP values for each feature across all predictions. It shows how features influence predictions both positively and negatively.

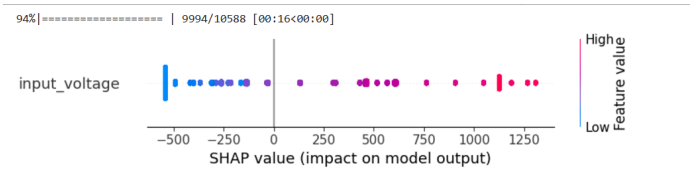

Fig. 9. Explainable AI Results (SHAP)

Feature input-voltage has a crucial influence on the model outcome. Magnitude of the SHAP values indicates the strength of contribution to the prediction.
As the input-voltage increases-which is moving from blue to red-the SHAP values shift positively, meaning higher input-voltage values lead to higher predicted el-power.
This plot provides insight into how the model uses input-voltage in making predictions. The operators can use it to understand and optimize system behavior by adjusting the input voltage to achieve desired power outputs.

VII. RESULTS AND DISCUSSION

To evaluate the model's generalization capabilities, the dataset is split into training and test sets in an 80-20 ratio. The first 80 percent of the data was used for training, while the remaining 20 percent was reserved for testing.
article graphicx array

MODEL EVALUATION METRICS

Model	**M S E**	**R M S E**	**R-squared**
Linear Regression	66237.73	257.37	0.8664
Support Vector Regressor (SVR)	487337.53	698.10	0.0172
Gradient Boosting Regressor	93297.36	305.45	0.8118
Random Forest Regressor	119160.99	345.20	0.7597
LSTM	120338.74	346.90	0.7574
ANN	181905.66	426.50	0.6332
MLP	151084.73	388.69	0.6953

TABLE I
PERFORMANCE METRICS FOR DIFFERENT REGRESSION MODELS ON THE DATASET.

1) Mean Squared Error (MSE): It measures the average squared difference between predicted and actual values.
Lower MSE means higher performance since it indicates that the predictions of the model are closer to the actual values.
We employed the MSE to evaluate, at each model, their predicted quality of electrical power in relation to the actual amount at each dataset (Training, Testing, and Validation sets)
Among them, the minimum MSE of 66,237 is of the Linear Regression model, which means that, in this model, estimated values are relatively closer to their actual values compared with all other models.

2) Root Mean Squared Error (RMSE): Square root of MSE provides an error metric in the same unit as the target variable.
Allows better interpretability of the errors, specifically for the range of el-power.
RMSE gives an intuitive understanding of the prediction error.
Linear Regression again gives the minimum value of RMSE as 257.37, with good predictive performance.

3) R-Squared (R^2): It is the proportion of variance in the target variable el-power explained by the input features input-voltage.
It ranges from 0 to 1, with higher values indicating that more of the data variability is explained by the model.
R^2 was used to measure the explicatory power of each of the models.
Among all models, Linear Regression had the best R^2: 0.866, which indicates that it explains 86.6 percent of the variance in the target.

Discussions

Evaluation Across Datasets

Models were optimized over training for both low MSE and high R^2. Overfitting was treated using cross-validation techniques.
The Validation Datasetinvolved tuning based on validation MSE and R^2, ensuring generalization to unseen data.
Final evaluation metrics, as indicated in the table above, were computed on the test set to evaluate the performance of models.

Sensitivity to the Dataset
Linear Regression Performed best overall owing to the simplicity in the relationship between input-voltage and el-power.
Gradient Boosting balanced error reduction and interpretability with reasonable MSE and R^2.
LSTM and ANN although powerful in capturing non-linear relationships, they suffered from overfitting and high error rates.

VIII. CONCLUSION AND FUTUREWORKS

The project has successfully investigated several machine learning models for the prediction of electrical energy output by micro gas turbines. The project employed both traditional models like Linear Regression and modern approaches such as Gradient Boosting, Random Forests, and Long Short-Term Memory networks to show the viability of using AI techniques for efficient and accurate energy prediction. Among the different models explored, Linear Regression was the most consistent, with low MSE and high R-squared values; thus, it is more appropriate for this data set.

The project also integrated explainable AI techniques, namely SHAP (SHapley Additive exPlanations), to provide insights into how the input variables influenced the model's predictions. This transparency helps bridge the gap between complex machine learning models and practical, real-world applications, thus empowering operators to make informed decisions. Overall, this project illustrates the potential of data-driven approaches for optimizing MGT performance, reducing emissions, and improving operational efficiency in decentralized energy systems.

Future Works

1) Dataset Expansion: Increase the dataset further with more ambient temperature, pressure, and humidity features for further improving the model's generalization. Include in the models data from various configurations and operating scenarios of MGT to make them more robust.

2) Model Optimisation: Employ advanced hyperparameter tuning techniques such as Bayesian Optimization for advanced model finesse. Experiment with hybrid models that could be developed by fusing machine learning with physics-based thermodynamic models in order to improve interpretability and accuracy.

3) Real-Time Deployment: In other words, real-time energy prediction systems will be designed and deployed that are capable of continuous adaptation to changing conditions using techniques from online learning. Integrate these models into MGT control systems for dynamic fuel efficiency and power output.

4) Handling Non-Stationary Data: Research reinforcement learning or transfer learning approaches that best fit the non-stationary conditions due to wear on the turbines and fluctuating fuel qualities. Integration of Renewable Energies Extend the framework in order to predict and optimize energy production in hybrid systems composed of MGT and renewable resources, such as solar and wind. Increased Explainability: The use of other Explainable AI techniques related to SHAP, like LIME-Local Interpretable Model-Agnostic Explanations, is foreseen for more usable insights to operators.

IX. ACKNOWLEDGMENT

I would like to express my deepest gratitude to everyone who contributed to the completion of this project.

I would like to express my heartfelt appreciation to Mr. Sudi for immense guidance, mentorship, and technical support throughout the course of this work; his vast expertise complemented with constructive feedback added a great deal of value to this project.

I would also like to show my profound gratitude to Dr. Joyce, whose wise counsel and words of encouragement gave meaning and lucidity during crucial junctures of the research process. Her commitment to excellence has been a continual spur to me.

I would also like to extend a special word of thanks to my colleague Mr.Kyagaba Jonah, who assisted, encouraged, and gave insightful inputs into this work, especially during the data preprocessing and analysis stages. His collaborative spirit and attention to detail were instrumental in realizing the objectives of this project. Lastly, I am grateful to all those people and Makerere University AI Lab that provided resources, tools, and data without which this project could not have been realized; thanks a lot.

REFERENCES

[1] G. Cimini and M. Milani. Advanced methods for energy prediction in micro gas turbines. *Journal of Energy Engineering*, 143(6):04017044, 2017.

[2] S. Ghosal and R. Sen. Semi-supervised learning for fault detection in micro gas turbines. *Journal of Energy Systems*, 12:214–223, 2020.

[3] E. Gonzalez and M. Lopez. Hybrid systems with micro gas turbines and renewable energy sources. *Applied Energy*, 264:114719, 2020.

[4] M. Li and Y. Zhao. Graph-based semi-supervised learning for energy system prediction. *Energy Systems Engineering*, 45:123–134, 2020.

[5] J. Park and H. Lee. Dimensionality reduction techniques for energy prediction. *Energy Informatics*, 3:22–34, 2020.

[6] V. Rathore and S. Kumar. Unsupervised learning approaches for clustering energy consumption data. *Energy Informatics*, 1:9–20, 2018.

[7] J. Sohn and C. Lee. Micro gas turbines: Design and applications. *Energy Conversion and Management*, 205:112437, 2020.

[8] L. Wang and M. Chen. Unsupervised learning for anomaly detection in energy systems. *Energy and AI*, 4:100061, 2021.

[9] L. Wang and Z. Zhang. Data-driven models for optimizing micro gas turbine performance. *Energy and AI*, 3:100052, 2021.

[10] A. M. Watwe and G. F. Berry. Design and performance characteristics of micro gas turbines. In *Proceedings of the ASME Turbo Expo*, volume 3, pages 785–794, 2000.

[11] H. Zhang and Y. Xu. Machine learning applications in energy system performance prediction. *Renewable Energy*, 132:1044–1056, 2019.