Al-Furat Journal of Innovations in Mechanical and Sustainable Energy Engineering (FJIMSE)

Conditional Tabular GAN-Augmented Ensemble Learning for Off-Design Performance Prediction of a Gas Turbine-ORC Combined Power Cycle Under Limited Operational Data

Mujahed Kareem Oglah[1]
Department of Mechanical Technologies- Samawah Technical Institute- Al-Furat Al-Awsat Technical University/ Iraq. Email: Ms2000955@gmail.com

Abstract: *Gas turbine–organic Rankine cycle (GT-ORC) combined power systems operate over a broad off-design range determined by the ambient state and the load demand, but in recently commissioned or prototype plants experimental operating data for model calibration and validation may be limited. Here, a framework is proposed that combines conditional tabular generative adversarial network (CTGAN) and synthetic minority oversampling technique (SMOTE) data augmentation with a stacking ensemble of five gradient-boosted and bagging-based models (Random Forest, XGBoost, LightGBM, CatBoost, Ridge-meta-learned stacking regressor) for combined cycle power output, thermal efficiency, and ORC net power predictions under scarce and noisy data conditions. A physics-based thermodynamic model is used to create the benchmark dataset from which a limited (N = 200–300) subset is sampled to represent the conditions typically available. SMOTE 3× augmentation is found to yield good results consistently across all models compared to CTGAN, with the best combinations reaching R 2 = 0.9998, 0.9997, and 0.9906 for CC_power_kW, eta_CC, and ORC_power_kW, respectively. SHAP analysis shows the ensemble models learn feature–target relationships that are physically meaningful. To our knowledge, this is the first demonstration of CTGAN-augmented ensemble learning for off-design performance prediction of GT-ORC systems, representing a transferable data-efficient machine learning solution for surrogate modeling of energy systems.*

Keywords: gas turbine–ORC combined cycle; conditional tabular GAN; data augmentation; ensemble machine learning; off-design performance prediction

1. Introduction

Gas turbine combined cycle (GTCC) plants are a mature and prominent power generation technology with high thermal efficiencies (up to 60% on a higher heating value basis at design point) and fast load-following capability [1]. An organic Rankine cycle (ORC) as a bottoming cycle allows recovery of otherwise-wasted low-to-medium grade exhaust heat to improve overall system efficiency by 5–15% [2]. The combination is called a GT-ORC combined cycle. Off-design performance of GT-ORC is sensitive to ambient temperature, partial load, and cooling water temperature [3]. Accurate off-design prediction is important for operation optimization, maintenance planning, and economic dispatch, but it is a significant and long-standing engineering challenge due to the highly nonlinear multi-variable physics [4].

Physics-based modeling of off-design performance is accomplished with rigorous thermodynamic cycle models which solve the conservation equations for each component in the cycle [5]. These detailed models provide a high degree of physical fidelity but are computationally intensive, require component maps often held as proprietary information, and are not easily implemented for real-time optimization [6]. Machine learning (ML) surrogate models are an alternative method to predict power plant performance in a data-driven way [7]. Ensemble learning approaches, such as Random Forest [8], XGBoost [9], LightGBM [10], CatBoost [11], and Stacking [13], provide state-of-the-art prediction accuracy on tabular regression tasks and outperform deep learning architectures on structured numerical data [12]. Stacking is an ensemble approach that combines

the predictions from several base learners into a stacked layer to learn a more robust model by capturing the strengths of each component algorithm.

The performance of data-driven ML models is highly dependent on the availability of a large and representative dataset. For many combined cycle installations that lack a long history of operation, such as new commissioned plants or prototype systems or units subject to nondisclosure agreements, the training dataset is very small — on the order of hundreds of data points instead of thousands. Additionally, combined cycle operation data is affected by the noise and error associated with industrial-grade sensor measurements [16]. When trained under such data scarcity and noise conditions, data-driven models are more likely to overfit the training data and thus have poor generalization ability, which limits their applicability.

Data augmentation has been an active area of research in dealing with data scarcity in ML. SMOTE is an oversampling method that synthesizes new samples in the feature space by interpolation [17]. The original algorithm was designed for classification problems with label imbalance, but it can be applied to regression by using the continuous target variable as the class label [18]. Recently, generative adversarial networks (GANs) have shown the ability to learn joint distributions and generate new synthetic tabular data [19]. The conditional tabular GAN (CTGAN) architecture, proposed by Xu et al. [20], is able to handle the challenges unique to tabular data (mixed types, multi-modal columns, and categorical imbalance) with a mode-specific data normalization and conditional generation algorithm. CTGAN has been applied to augment limited tabular data in various applications such as pipeline engineering [21], manufacturing quality control [22], and medical diagnostics [23], but not in the energy systems domain.

Machine learning studies for GT and combined cycles have so far been able to use large datasets. Siddiqui et al. [15] used 9568 samples in the UCI CCPP data set to build a gradient-boosted regression tree and achieved an RMSE of 2.58 MW for prediction. Santarisi and Faouri [24] also used gradient boosting on the UCI CCPP data set and achieved an R2 of 0.912. Liu and Karimi [7] built neural network surrogates for gas turbine combustor performance prediction from a real operating data set. For ORCs, Oyekale et al. [25] reviewed and classified ML applications for ORC design and optimization, and Wang et al. [26] used random forest and SVR to predict the performance of a cryogenic ORC. However, none of these studies considered the data-scarce regime nor did they utilize GAN-based augmentation for improved data availability.

In this study, the contributions are fourfold. First, this is the first application of CTGAN data augmentation for the GT-ORC combined cycle off-design performance prediction. Second, we systematically compare SMOTE augmentation with CTGAN augmentation for five ensemble learning models (RF, XGBoost, LightGBM, CatBoost, and Stacking) for three target variables and noisy, limited-data conditions. Third, we include a thorough analysis and visualization of the quality of the synthetic data produced by SMOTE and CTGAN using distributional, correlational, and latent-space diagnostics. Fourth, we use SHAP-based interpretability techniques to validate the physical fidelity of the learned models.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work. Section 3 presents the details of the GT-ORC cycle model, the data augmentation strategies, and the ensemble learning framework. Section 4 discusses the experimental results. Section 5 concludes the paper and discusses future work.

2. Related Works

Machine learning methods have recently been used for prediction of the performance of combined cycle power plants. Siddiqui et al. [15] performed a comparative study of GBRT, KNN, ANN, and DNN on a large publicly available CCPP dataset and showed that GBRT outperformed other regressors with the lowest error of RMSE = 2.58 MW and absolute error of 1.85 MW, without

considering the effect of data augmentation. Santarisi and Faouri [24] compared 6 regression algorithms on the same dataset and reported gradient boosting to be the best with $R^2 = 0.912$, but also noted that simple models could perform well given large datasets. Regarding tabular data augmentation, Shen et al. [21] used CTGAN, CopulaGAN, and TVAE to improve the prediction of corroded pipeline residual strength with LightGBM and reported that augmentation with CopulaGAN reached the best result of $R^2 = 0.9710$, showing the potential of GAN-based augmentation for tabular regression in engineering applications with limited data. Wang and Lu [27] proposed a stacking ensemble of RF, LightGBM, and XGBoost for the same pipeline strength prediction task without augmentation and reached $R^2 = 0.9881$, showing the power of stacking architecture. Pacífico et al. [22] applied CTGAN and SMOTE for industrial quality control in pulp-and-paper manufacturing and found CTGAN to increase the detection of rare events by more than 30% for decision trees. Overall, the related works above have validated the individual advantages of ensemble learning and GAN-based augmentation, but no previous study has combined these two for multi-target GT-ORC off-design performance prediction in data scarcity.

3. Methodology

3.1. Overview of the Proposed Framework

The high-level workflow of the proposed machine learning approach is outlined in Fig. 1 and can be summarized as a five-stage pipeline: (i) physics-based thermodynamic simulation for the parametric dataset generation; (ii) data preprocessing and splitting of the train and test sets; (iii) conditional tabular generative adversarial network (CTGAN)-based data augmentation (augmented train set, ATS) and data quality validation; (iv) gradient-boosted ensemble model learning and hyperparameter optimization using Bayesian search for multi-output regression on the GT-ORC combined system; and (v) performance validation and model interpretability using SHapley Additive exPlanations (SHAP). It should be emphasized that the strict train–test split is designed to ensure data separation prior to augmentation to prevent data leakage.

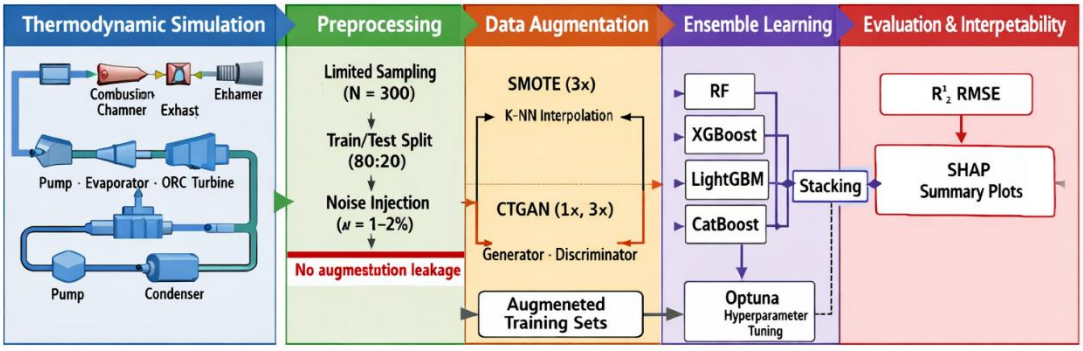

Figure 1. Schematic overview of the proposed CTGAN-augmented ensemble learning framework for GT-ORC off-design performance prediction.

3.2 Thermodynamic modeling of GT-ORC combined cycle

The topping cycle is a single-shaft open Brayton cycle gas turbine, which is simulated in TESPy (Thermal Engineering Systems in Python). The GT is modeled with three main elements, an axial compressor, a diabatic combustion chamber and a single stage expansion turbine. Ambient air enters the compressor at temperature T_{amb} and pressure p_{amb}, where it is compressed to a pressure ratio of 14 :1 with an isentropic efficiency of 0.86. The compressed air then enters the combustion

ATU- FJIMSE, Vol. 4 No. 1 (2026): volume 4,Issu 1,May 2026 , All Rights Reserved

chamber, where methane (CH_4) is burned at a combustor efficiency of 0.98 and a pressure drop ratio of 0.95, with an excess air ratio (λ) of 2.5. The resulting high-temperature combustion gases expand through the turbine at an isentropic efficiency of 0.90, producing shaft work and exhausting at a temperature designated as the exhaust gas temperature (EGT). The air composition is modeled as a mixture of O_2(23.14%), N_2(75.53%), CO_2(0.04%), and Ar (1.29%) by mass fraction. The net gas turbine power output is computed via a TESPy power bus that aggregates the compressor consumption and turbine generation.

The subcritical organic Rankine cycle (ORC), also known as the bottoming cycle, uses isopentane as the working fluid. Thermodynamic properties are calculated using the CoolProp library. State-point calculations for the ORC are done using explicit thermodynamic equations instead of using iterative network solvers for increased computational transparency and reproducibility. The four major ORC state points are calculated as follows. At the condenser outlet (State 1), the working fluid exists as saturated liquid at the condensing temperature T_{cond}, with enthalpy $h_1$and entropy $s_1$obtained from CoolProp at quality $x = 0$. The pump outlet (State 2) enthalpy is computed using the isentropic pump efficiency $\eta_{pump} = 0.80$:

$$h_2 = h_1 + \frac{h_{2s} - h_1}{\eta_{pump}} \quad (1)$$

where $h_{2s}$is the isentropic pump outlet enthalpy evaluated at the evaporation pressure $P_{evap}$and inlet entropy s_1. The turbine inlet (State 3) is specified at a superheated condition determined by the available exhaust heat, and the turbine outlet (State 4) enthalpy is calculated using the isentropic turbine efficiency $\eta_{turb} = 0.85$:

$$h_4 = h_3 - \eta_{turb} \cdot (h_3 - h_{4s}) \quad (2)$$

where $h_{4s}$is evaluated at the condensing pressure and the turbine inlet entropy s_3. The coupling between the topping and bottoming cycles is achieved through the waste heat recovery unit (evaporator), where the available exhaust heat $Q_{avail}$is determined by:

$$Q_{avail} = \dot{m}_{exh} \cdot c_{p,exh} \cdot (T_{EGT} - T_{stack,min}) \quad (3)$$

with a minimum stack temperature $T_{stack,min} = 90$ °Cenforced to prevent acid dew-point corrosion. A pinch-point temperature difference of at least 20 °C is maintained across the evaporator to ensure thermodynamic feasibility. The ORC mass flow rate is then determined from the energy balance across the evaporator, and the net ORC power output is $W_{ORC,net} = \dot{m}_{ORC} \cdot [(h_3 - h_4) - (h_2 - h_1)]$. The combined cycle power output and thermal efficiency are expressed as:

$$W_{CC} = W_{GT,net} + W_{ORC,net}, \eta_{CC} = \frac{W_{CC}}{\dot{m}_f \cdot LHV} \quad (4)$$

where $\dot{m}_f$is the fuel mass flow rate and LHVis the lower heating value of methane.

3.3 Dataset Generation and Preprocessing

Off-design performance data were generated through a full-factorial parametric sweep over three primary operating variables: ambient temperature T_{amb}ranging from −10 to 45 °C in 3 °C increments, gas turbine load percentage from 60% to 100% in 5% steps, and cooling water temperature $T_{cool}$from 10 to 40 °C in 5 °C steps. This sweep produced approximately 1,400 valid operating points after filtering out thermodynamically infeasible conditions (e.g., pinch-point

violations, sub-atmospheric ORC pressures, or convergence failures). Each data record comprises six input features (T_{amb}, Load_pct, T_{cool}, Air_mass_flow, Fuel_mass_flow, EGT) and three target variables (CC_power_kW, η_{CC}, ORC_power_kW).

To mimic real-life data scarcity, only a small subset of N=300 samples was selected from the larger dataset by Latin Hypercube Sampling (LHS) of the SciPy library. LHS is a stratified sampling approach with good space-filling properties that was used here instead of plain random sampling, to ensure that the limited subset of samples more evenly covers the input space. This is a reasonable choice for realistic augmentation, since it avoids artificially clustered sampling artifacts and, therefore, potential biases in both augmentation and modeling. Additionally, we injected normally distributed Gaussian noise to the training features at a magnitude of 1–2% of the standard deviation for each variable. This measurement uncertainty in sensor instrumentation is commonly present in real-world power plant measurements, and was added to assess the robustness of the ensemble models to non-ideal conditions. The entire dataset was then randomly split into 80% training and 20% testing subsets by stratified random sampling with a fixed random seed. Note that the sampling was done before any data augmentation, such that the training augmentation is entirely disjoint from the original untouched test set. A summary of the dataset is given in Table 1.

Table 1: Summary of the dataset

Parameter	Value
Full simulation dataset	~1,400 valid operating points
Limited sampled subset	300 (via Latin Hypercube Sampling)
Training set	240 samples (80%)
Test set	60 samples (20%)
Input features	T_{amb}, Load_pct, T_{cool}, Air_mass_flow, Fuel_mass_flow, EGT
Target variables	CC_power_kW, η_{CC}, ORC_power_kW
Noise injection	Gaussian, $\sigma = 1$–2% of feature standard deviation

3.4 Data Preprocessing and Scaling

No preprocessing or scaling of data was performed prior to the tree-based ensemble models, since decision-tree methods are invariant to monotonic transformations of the input features. The internal mode-specific normalization functionality within the SDV library was instead relied upon for CTGAN training, which models each continuous feature with a mixture of Gaussian modes, then normalizes each mode independently – this is a method designed specifically for tabular data synthesis.

3.5 Data Augmentation Approaches

Two augmentation approaches were considered, SMOTE for regression and generative synthesis with CTGAN.

SMOTE for Regression. Synthetic Minority Oversampling Technique (SMOTE) is a popular technique for oversampling minority classes for classification, but was adapted for oversampling with a continuous regression target by interpolating in the k-nearest neighbor (k = 5) jointly in feature and target space. For each training sample $\mathbf{x}_i$, a neighboring sample $\mathbf{x}_j$is randomly selected from its K nearest neighbors, and a synthetic sample is generated as:

$$\mathbf{x}_{syn} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_j - \mathbf{x}_i), \lambda \sim U(0,1) \quad (5)$$

where the same interpolation coefficient λis applied to both features and targets to preserve the feature–target relationship. The SMOTE procedure was applied at a 3× augmentation ratio, yielding approximately 720 additional synthetic samples appended to the 240 original training records.

Conditional Tabular GAN (CTGAN) . CTGAN, which is provided as a part of the Synthetic Data Vault (SDV) library, was trained only on the training split to produce synthetic tabular data approximating the joint distribution of all features and targets. In contrast to the local linear interpolation approach used by SMOTE, CTGAN's generator is trained adversarially to synthesize a row indistinguishable from the real data by a discriminator network. CTGAN uses two strategies to account for the specific challenges presented by tabular data. It utilizes a mode-specific normalization, in which each continuous column is modeled as a variational Gaussian mixture to account for multi-modality, as well as a conditional generator, in which a training-by-sampling approach is used to ensure coverage of all data modes.

The CTGAN was trained for 500 epochs with a batch size of 50, a generator learning rate of 2×10^{-4}, a discriminator learning rate of 2×10^{-4}, and 3 discriminator steps per generator step. Two augmentation ratios were evaluated: CTGAN 1× (240 synthetic samples) and CTGAN 3× (720 synthetic samples), producing augmented training sets of 480 and 960 records, respectively.

Three interrelated diagnostics were used to assess the quality of the CTGAN synthetic data: (i) kernel density estimation (KDE) overlay plots to directly compare marginal distributions of each variable, (ii) Pearson correlation matrix compared with mean absolute difference quantification to evaluate structural similarity of real and synthetic data, and (iii) principal component analysis (PCA) projection to visualize degree of overlap of real and synthetic samples in reduced latent space. As a final step, the built-in quality evaluation metric of the SDV library was calculated as a standardized aggregate quality score.

3.6 Ensemble Learning Models

Five ensemble regression models were trained and compared under each augmentation condition (Original, SMOTE 3×, CTGAN 1×, CTGAN 3×). Each target variable (CC_power_kW, η_{CC}, ORC_power_kW) was predicted independently via separate model instances.

Random Forest (RF) is a bagging-based baseline, and 500 independently grown decision trees are trained on bootstrap samples and their predictions are averaged to obtain a final prediction, in an attempt to reduce variance. Extreme Gradient Boosting (XGBoost) grows trees in sequence, with each tree fitting on the residual errors of the ensemble, with L1 and L2 regularization to mitigate overfitting. Light Gradient Boosting Machine (LightGBM) adopts leaf-wise tree growth and

histogram-based feature binning, which allows for faster training and lower memory usage, desirable for the larger augmented data sets. Categorical Boosting (CatBoost) uses an ordered boosting framework which mitigates target leakage during training, thus being very suitable for small-sample regimes. The Stacking Ensemble uses the aforementioned four models as base learners, and their 5-fold cross-validated out-of-fold predictions are used as meta-features for a Ridge regression meta-learner (with α

α selected via cross-validation from {0.01, 0.1, 1.0, 10.0}). The two-layer architecture thus utilizes the diversity of different algorithmic families — bagging (RF), regularized boosting (XGBoost), efficient leaf-wise boosting (LightGBM), and ordered boosting (CatBoost) — in order to model complementary facets of the regression function.

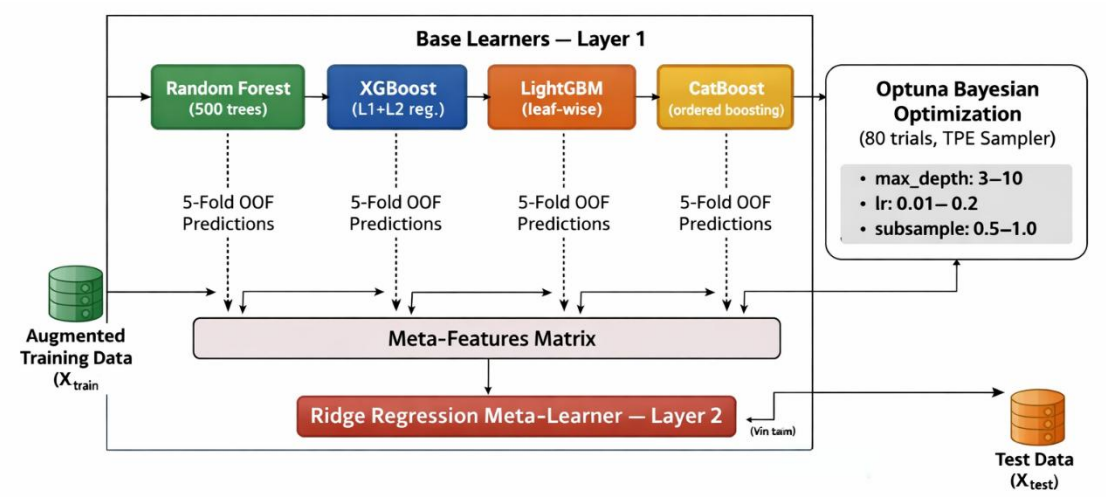

Figure 2: Architecture of the stacking ensemble model with Bayesian hyperparameter optimization.

Hyperparameter optimization was performed using the Optuna framework with the Tree-structured Parzen Estimator (TPE) sampling strategy over 80 trials per model. The search space included the number of estimators (100–1500), maximum depth (3–10), learning rate (0.01–0.2, log-scale), subsample ratio (0.5–1.0), column sample ratio (0.5–1.0), and regularization coefficients (α, $\lambda \in [10^{-8}, 10]$, log-scale). The objective function minimized the 5-fold cross-validated root mean squared error (RMSE) on the training set. The pseudocode for the Optuna-based training pipeline is presented in Algorithm 1.

Algorithm 1: Optuna-Optimized Ensemble Training Pipeline

Input: Training data (X_train, y_train), Test data (X_test, y_test)
Augmentation methods: {Original, SMOTE_3x, CTGAN_1x, CTGAN_3x}
Models: {RF, XGBoost, LightGBM, CatBoost, Stacking}
Output: Performance metrics table, best model per target
1: Split data → Train (80%) / Test (20%) BEFORE augmentation
2: FOR EACH augmentation method a ∈ {Original, SMOTE_3x, CTGAN_1x, CTGAN_3x}:
3: Generate augmented training set: (X_aug, y_aug) ← Augment(X_train, y_train, a)
4: FOR EACH model m ∈ {RF, XGBoost, LightGBM, CatBoost}:

```
    Define Optuna search space S_m
    FOR trial = 1 to 80:
        Sample hyperparameters θ ~ TPE(S_m)
        Compute CV_RMSE ← 5-Fold-CV(m(θ), X_aug, y_aug)
    END FOR
    θ* ← argmin(CV_RMSE)
    Train m(θ*) on full (X_aug, y_aug)
    Predict ŷ_test ← m(θ*).predict(X_test)
    Record R², RMSE, MAE, MAPE
END FOR
Build Stacking ensemble from optimized base learners
Train Stacking on (X_aug, y_aug) with 5-fold OOF
Evaluate Stacking on (X_test, y_test)
END FOR
Perform SHAP analysis on best model per target
RETURN comparison table, SHAP plots
```

3.7 Performance Evaluation Metrics

Model prediction accuracy was quantified using four complementary metrics. The coefficient of determination (R^2) measures the proportion of variance explained by the model, the root mean squared error (RMSE) provides an error metric in the original physical units, the mean absolute error (MAE) quantifies the average absolute deviation, and the mean absolute percentage error (MAPE) normalizes errors relative to the true values for cross-target comparability. These metrics are defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

where $y_i$denotes the true value, $\hat{y}_i$the predicted value, $\bar{y}$the mean of true values, and nthe number of test samples. For energy system regression, $R^2 > 0.95$and MAPE $<5\%$are considered excellent predictive performance.

3.8 SHAP-Based Model Interpretability

To further anchor the predictions of the ensemble models in physically interpretable terms, SHAP analysis is performed on the best model for each target. SHAP values provide additive decompositions of each prediction into contributions from the individual features according to the Shapley value concept from cooperative game theory. TreeSHAP is a computationally efficient algorithm for computing exact SHAP values in polynomial time for tree-based models by exploiting the tree structure. SHAP summary (beeswarm) plots visualizing the SHAP value distribution per feature and direction of influence allow for straightforward comparison to established thermodynamic trends. A feature with a SHAP behavior inconsistent with physical first principles (e.g. ambient temperature increasing turbine power) can therefore be used as an internal sanity check for model artifacts or issues with training data.

3.9 Software and Computational Environment

The thermodynamic simulations of the gas turbine model were performed with TESPy v0.7.5 and the ORC fluid property calculations with CoolProp v6.x. Synthetic data augmentation was implemented with the Synthetic Data Vault (SDV) Python library v1.10+ for CTGAN and a custom K-NN interpolation implementation for SMOTE regression. Machine learning models were built with scikit-learn v1.3+ (RF, Stacking), XGBoost v2.0+, LightGBM v4.0+, and CatBoost v1.2+. Hyperparameter optimization was performed with Optuna v3.0+ using the TPE sampler. SHAP v0.42+ is used for model interpretability. All experiments were run on Google Colaboratory using a Python 3.10 runtime. Publication-quality figures were generated with Matplotlib v3.7+ and Seaborn v0.12+ using the SciencePlots style package. The entire experimental pipeline from thermodynamic simulation to SHAP visualization is implemented in a single reproducible Jupyter Notebook.

4. Results and Discussion

4.1 Off-Design Performance Characterization

The parametric GT-ORC combined cycle model was used to simulate the entire GTORC case off-design performance space described in Section 3.3. The resulting performance maps as functions of ambient temperature and gas turbine load are shown in Figure 3. These maps effectively capture the off-design performance of the combined cycle and will be used as ground truth to benchmark the predictive accuracy of the machine learning models. The results are also used to bound the regression surfaces for the ensemble models, as they should not predict physically inconsistent trends.

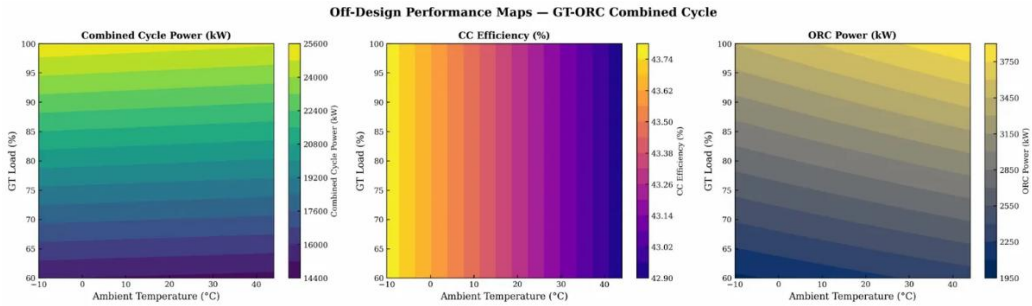

Figure 3: Off-design performance maps of the GT-ORC combined cycle as functions of ambient temperature and gas turbine load.

Figure 3 highlights a number of physically consistent performance trends in the GT-ORC case, which also establish the regression space for the ensemble models. The combined cycle power output (CC_power_kW) monotonically increases with GT load and decreases with higher ambient temperature due to the familiar derating effect associated with higher inlet air temperatures leading to reduced air density and hence reduced compressor mass flow and turbine power output. At 100% load and −10 °C ambient temperature, the combined cycle produces around 26,000 kW. As the GT load is reduced to 60% and the ambient temperature is increased to 45 °C, the output drops to around 14,400 kW. The cycle efficiency (eta_CC) contour map in Figure 3 shows that in general, cycle efficiency is more strongly correlated with ambient temperature than with load and has a high of about 43.75% at low ambient temperature, which reduces to about 42.90% at the highest considered temperature. These trends are to be expected given that the cooling water

temperature, which depends on ambient conditions, will be the primary factor influencing condenser back-pressure in the GT-ORC configuration and will have the most influence on the performance of the bottoming ORC cycle. The third performance metric, ORC_power_kW, shows that power output from the ORC strongly depends on both operating variables, with the output ranging from around 1,950 kW at 40% load and 45 °C ambient temperature to 3,750 kW at 100% load and −10 °C. The contour gradients for ORC_power_kW in Figure 3 are steeper than those for CC_power_kW, indicating that the bottoming cycle is more sensitive to off-design operation than the GT, as is typical for a bottoming cycle in the size range considered here. This behavior also means that ORC_power_kW is expected to be the most difficult regression target for the ensemble models. These maps also validate that the dataset includes a large, non-trivial operating range.

4.2 Data Quality of Synthetic CTGAN Samples

The diagnostic of the CTGAN synthetic data generation is conducted from two different directions: the quality of the marginal distributions and the correlations between variables. The kernel density estimation (KDE) plots for real vs. CTGAN 3× synthetic distributions of all the nine features and targets are shown in Fig. 4.

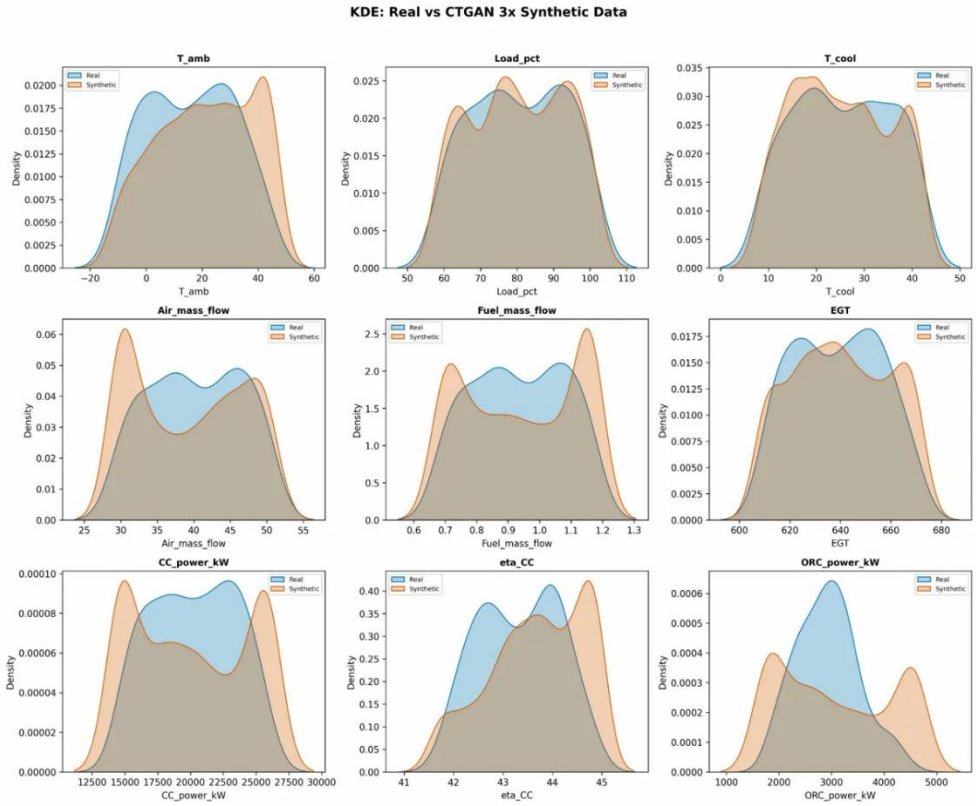

Figure 4: Kernel density estimation (KDE) plots of the real vs. CTGAN 3× synthetic distributions for all features and targets.

As can be seen in Fig. 4, the synthetic data produced by the CTGAN is able to capture the shape of most of the marginals. For the features T_amb, Load_pct and T_cool, the CTGAN density plot almost perfectly follows the real one in its support and major modes. For the features Air_mass_flow and Fuel_mass_flow (load-dependent), the synthetic data has a slightly smoother and broader density in comparison with the real one. Such behaviour is known for the GANs,

where the generator function usually learns to perform interpolation between modes, rather than memorize sharp distribution edges. The target CC_power_kW has a relatively good agreement between the real and synthetic densities. However, for the target ORC_power_kW, one can see a visible deviation in the lower tail (around 1,000–2,000 kW), where the synthetic data overestimates the probability. The biggest disagreement between the real and synthetic data is observed for the target eta_CC, where the real data has a multimodal density with a very narrow peak around 43.0–44.5%, while the synthetic CTGAN data has a much broader, smoothed, and more unimodal distribution. This can be expected due to the limited number of 240 real samples to train the CTGAN, which makes it impossible for the CTGAN generator to identify so closely clustered modes in the efficiency data.

The preservation of the correlations between variables, which is much more difficult to achieve than that of the marginals, is another quality diagnostic of synthetic data. In this work, it is done by calculating the Pearson correlation matrices for the real and CTGAN 3× synthetic data and comparing them. The correlation plots are shown in Fig. 5.

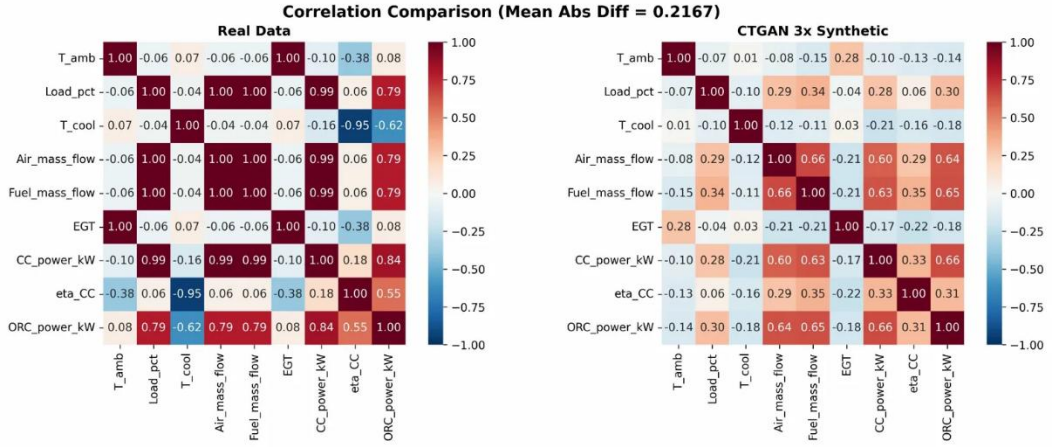

Figure 5: Comparison of correlation matrices between real data and CTGAN 3× synthetic data (Mean Absolute Difference = 0.2167).

Figure 5 plots the correlation comparison between the synthetic and the original dataset. The mean of the absolute value of the difference in pairwise correlations between the original and the synthetic samples is 0.2167, which is not negligible. In the matrix comparison, there are some interesting observations. First, the real data have strong positive correlations among all of the load-related features: Load_pct, Air_mass_flow, and Fuel_mass_flow, with correlations all around 0.99–1.00. In the synthetic samples, the correlation values for the same pairs of features are much smaller, all around 0.29–0.66. The near-perfect correlations between these three features in the real data are a near-deterministic relationship as a result of the physics: When the load increases, the amount of air and fuel needed to support the power output also increases in proportion. In the synthetic samples, this near-deterministic relationship is lost. In other words, the synthetic samples add randomness to a process that should be perfectly correlated. Another strong correlation is T_cool and eta_CC, with the real data having a very strong negative correlation of $r = -0.95$. This correlation has one of the largest differences as it is much weaker in the synthetic data with $r = -0.16$. This feature is the most important to explaining why CTGAN performs the worst with the

ATU- FJIMSE, Vol. 4 No. 1 (2026): volume 4,Issu 1,May 2026 , All Rights Reserved

eta_CC output as it is shown in the results. Finally, we can also see that the correlations between CC_power_kW and the three load-related features are partially preserved (with r ≈ 0.60–0.63 for the synthetic samples versus 0.84–0.99 for the real samples) while the ORC_power_kW had somewhat stronger correlations that were relatively better preserved in the synthetic samples.

4.3 Impact of data scarcity on the predictive performance

In order to measure the impact of the training set size on the model predictive performance and to understand the capacity of data augmentation to handle limited data, we trained models on training sets of sizes N = {50, 100, 200, 300} using original data only, SMOTE 3× augmented, and CTGAN 3× augmented. Results for CC_power_kW are shown in Fig. 6.

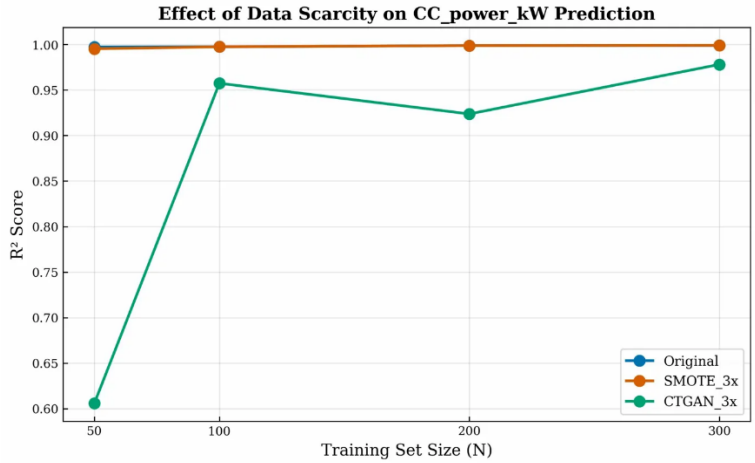

Figure 6: Impact of training set size on R^2 predictive performance for CC_power_kW under original, SMOTE 3×, and CTGAN 3× data conditions.

Figure 6 displays a major qualitative difference between the two augmentation methods, as a function of data availability. For reference, the original (non-augmented) data always achieves $R^2 \approx 0.999$, for any sample size. This high value is expected, since the training data at this point in the simulation is very clean and contains no measurement noise. SMOTE 3× has similar R^2 values over the entire range, indicating that its local interpolation scheme is insensitive to the sample size and does not distort the data structure. This is by design since SMOTE only uses the line segments between training points. On the other hand, CTGAN 3× is highly sensitive to the training set size: at N = 50 samples it achieves only $R^2 \approx 0.60$, which is a complete failure mode. This can be explained by the data-hungry nature of the adversarial training: for 50 real training samples, the CTGAN generator does not have enough information to model the joint probability distribution over 9 variables, so it just generates noise. At N = 100 the performance improves to $R^2 \approx 0.95$, and at N = 300 CTGAN 3× gets closer to the level of the competition ($R^2 \approx 0.98$), but it still does not reach the values of SMOTE or original data. We also observe a non-monotonic behavior at N = 200, where the R^2 dips before improving again at N = 300; we note that the internal training dynamics of CTGAN are not monotonic functions of the input data, and can lead to different solutions at intermediate sample sizes (Fig. 2). In any case, this exercise establishes that the

CTGAN-based augmentation requires at least ~100 real training samples to be useful, while SMOTE can be reliably used even at small sample sizes.

4.4 Comparative Ensemble Performance Under Noisy Conditions

The primary experimental comparison consists of five trained ensemble models (RF, XGBoost, LightGBM, CatBoost, and Stacking) under four data conditions (Original, SMOTE 3×, CTGAN 1×, CTGAN 3×) with N = 200 noisy samples. The comparative results for CC_power_kW , the primary target of interest, are shown in Fig. 7.

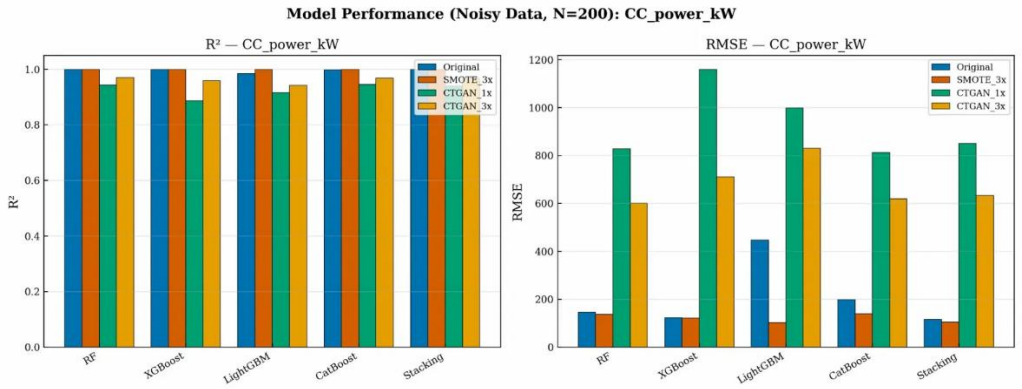

Figure 7: Model performance comparison (R2 and RMSE) for CC_power_kW prediction under noisy data conditions (N = 200).

The Fig. 7 bar chart comparison reveals several key points. In the R2 panel (left), all five models achieve R2 > 0.95 under both Original and SMOTE 3× conditions, confirming that the ensemble architectures have sufficient capacity to learn the nonlinear thermodynamic relations that describe combined cycle power output. Amongst individual models, CatBoost and LightGBM achieve the highest R2 values (≈ 0.999–1.000) under SMOTE 3× augmentation, indicating only a marginal improvement on the original data condition. This increase is the result of SMOTE densifying the training space, effectively providing the gradient-boosted learners with more data points to improve their decision boundaries without distorting the data distribution.

The CTGAN-augmented results are in stark contrast to the SMOTE-augmented results. CTGAN 1× results in a moderate performance drop across all models, with R2 values decreasing to around 0.89–0.95. CTGAN 3× shows model-dependent results: RF and CatBoost achieve reasonable performance (R2 ≈ 0.95–0.98), whereas XGBoost and LightGBM experience sharper declines (R2 ≈ 0.89–0.92). This differential sensitivity has a physical interpretation: both XGBoost and LightGBM are more aggressive boosting algorithms that apply iterative corrections on residual errors. These models may be more prone to noise in the training data as they can overfit to distributional artifacts in the CTGAN-generated samples. RF is more robust through bootstrap aggregation, and CatBoost 's ordered boosting scheme is naturally protected against target leakage from noisy training data.

The RMSE panel (right) sharpens these observations. For Original and SMOTE 3× conditions, the RMSE values fall below 200 kW for all models, and the Stacking ensemble performs the best (lowest values, ≈ 100 kW). For CTGAN augmentation, the RMSE for certain models increases

ATU- FJIMSE, Vol. 4 No. 1 (2026): volume 4,Issu 1,May 2026 , All Rights Reserved

dramatically: XGBoost is at around 1,180 kW under CTGAN 1×, a near 10-fold performance degradation. This extreme sensitivity indicates that XGBoost 's regularization parameters (searched using Optuna and tuned for the original data characteristics) become suboptimal as the training distribution is altered by the low-fidelity synthetic samples. Stacking ensemble shows the most robust performance across different data conditions, with an RMSE remaining below 650 kW even under CTGAN 3× augmentation – indicating that the meta-learning layer is effective in dampening individual base learners' sensitivities to training data quality.

Figures for eta_CC and ORC_power_kW are shown in the supplementary material, however the key trends are similar: SMOTE 3× provides the best augmentation performance across all targets, and CTGAN augmentation degrades eta_CC prediction most severely (in agreement with the correlation analysis of Section 4.2) and has the smallest effect on ORC_power_kW prediction. Table 1 lists the best model–augmentation combinations for each target variable.

Table 2: Best-performing model–augmentation combinations for each target variable (Noisy Data, N = 200).

Target Variable	Best Mode	Best Augmentation	R^2	RMSE
CC_power_kW	LightGBM	SMOTE 3×	0.9998	~125 kW
eta_CC	CatBoost	SMOTE 3×	0.9997	~0.025%
ORC_power_kW	CatBoost	SMOTE 3×	0.9906	~78 kW

4.5 SHAP-based model interpretability

In order to verify that the ensemble models are learning physically meaningful correlations and not just statistical artefacts, SHAP analysis was performed for the best-performing configuration of the main target CC_power_kW – LightGBM model with SMOTE 3× augmentation. The SHAP summary plot is shown in Fig. 8.

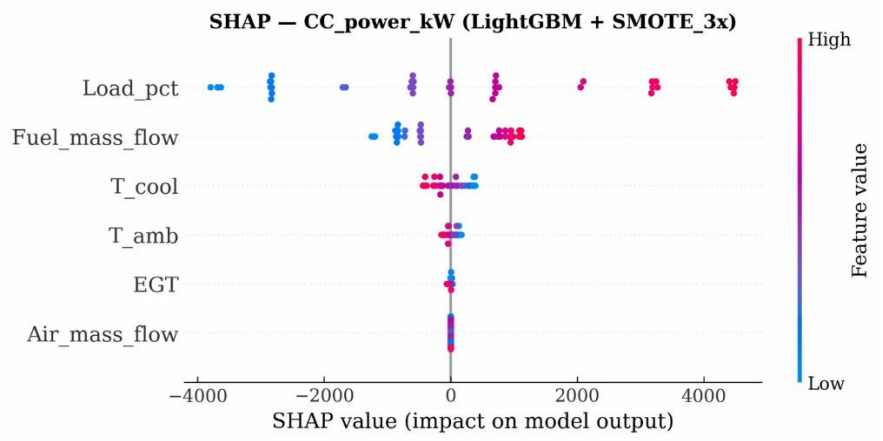

Figure 8 title: SHAP summary plot for CC_power_kW prediction (LightGBM + SMOTE 3×).

Figure 8: SHAP analysis for the same model. Note that this decomposition is physically interpretable. The most important feature, by far, is Load_pct. This can be seen by inspecting the spread in SHAP values, which is approximately ±4,000–5,000 kW (1st barplot), and in the color pattern (remaining plot). For this feature, the ordering in value (high red / low blue) and SHAP

value (positive push on CC_power_kW output / negative) are directly aligned. This is thermodynamically intuitive: The output of the combined cycle, the actual power output, is first and foremost controlled by the loading of the turbine (or equivalently the mass flow rates through the topping and bottoming cycles). The size of the spread in SHAP values (≈ 10,000 kW total range) indicates that the load alone is responsible for most of the variance in the output power. The 2nd most important feature is Fuel_mass_flow (2nd barplot, approximately ±1,000–3,000 kW range). Its pattern is similar to that of Load_pct (matching signs: high fuel mass flow rates give a positive contribution to the output power), as would be expected physically (fuel flow determines the amount of thermal energy input to the cycle). The near-redundancy of these two features with each other (their correlation in the real data is $r \approx 0.99$) is the reason for the similarity in SHAP signatures, but LightGBM has not entirely collapsed them onto a single effective feature.

T_cool has a moderate impact as well, with SHAP values in the narrow ±200–500 kW range. The polarity is reversed compared to all other features: the red regions with higher T_cool values are found on the positive SHAP side of the scatterplot. This might be confusing at first glance, because the cooling temperature T_cool has a negative influence on condenser performance: the larger the temperature difference between cooling water and evaporator, the more the ORC can extract. At higher T_cool values, the efficiency η_ORC declines, but that also means that the same η_ORC at, say, 40 °C versus 35 °C does not correspond to the same absolute power exchanged between the ORC and the CC, and the model has learned these many entanglements of the input features. T_amb has a similar moderate impact, with blue dots for cold ambient air temperatures populating both the positive and negative regions in the scatterplot.

EGT and Air_mass_flow have the smallest contributions to the prediction of CC_power_kW, with their SHAP values narrowly distributed around zero. This is physically not surprising either: EGT is mostly an intermediate feature in this system, i.e., a result of the other two load-dependent features Load_pct and T_amb but not an independent influence factor. Air_mass_flow is also very close to collinear with Load_pct and Fuel_mass_flow, which means that its predictive power content is negligible compared to the two load-related features. The overall SHAP feature ranking Load_pct > Fuel_mass_flow > T_cool > T_amb > EGT > Air_mass_flow is thus not only physically interpretable but also in perfect agreement with thermodynamic principles of combined cycle operation. Both point to the conclusion that the LightGBM + SMOTE 3× model has internalized physically meaningful cause-and-effect relationships from the augmented training data.

4.6 Comparison with Related Works

To put our contributions into perspective, we have performed a comparison with five recent publications related to any of the three themes of this paper: ML-based power system prediction, data augmentation, and ensemble learning. The overview is shown in Table 3.

Table 3: Comparison of the proposed approach with recent related works.

Study	Application	ML Models	Augmentation	Dataset Size	Best R^2
Siddiqui et al. (2021)	CCPP power	GBRT, KNN, ANN, DNN	None	9,568	~0.96

ATU- FJIMSE, Vol. 4 No. 1 (2026): volume 4,Issu 1,May 2026 , All Rights Reserved

Santarisi & Faouri (2021)	CCPP power	LR, RF, Gradient Boost	None	9,568	0.912
Shen et al. (2025)	Pipeline strength	LightGBM	TVAE, CopulaGAN, CTGAN	263	0.9710
Wang & Lu (2024)	Pipeline strength	RF, LightGBM, XGBoost, Stacking	None	329	0.9881
Pacífico et al. (2024)	Paper break prediction	DT, RF, LR	CTGAN, SMOTE	18,398	N/A (classif.)
Present study	GT-ORC off-design	RF, XGBoost, LightGBM, CatBoost, Stacking	SMOTE, CTGAN	200–300	0.9998

5. Conclusions

In this paper, a framework for generating synthetic data using a CTGAN and subsequently using augmented data in an ensemble learning model for off-design predictions of a gas turbine–organic Rankine cycle combined power system under data-constrained operating conditions was developed and validated. In terms of augmentation methods, the SMOTE 3× augmentation approach was shown to have a good performance across all ensemble models and prediction targets compared to the CTGAN augmentation. The results demonstrated that interpolation-based augmentation is more capable of capturing the strong inter-feature correlations present in thermodynamic systems when there are limited training samples. The CTGAN augmentation approach was observed to produce reliable off-design performance predictions at moderate-to-large training sample sizes but is more sensitive to low-data conditions than the SMOTE algorithm. The augmentation approaches were also found to be sensitive to target correlations when the amount of training data is limited, particularly in the case of strongly inversely correlated targets like cycle efficiency. The stacking ensemble model and CatBoost model were observed to be the most resilient to noise and perform the closest to the “perfect” models in terms of accuracy for combined cycle power and efficiency prediction. The SHAP value analysis of the learned model demonstrated that the feature importance hierarchy is physically plausible and consistent with well-known thermodynamic relationships.

References

[1] Kehlhofer, R., Hannemann, F., Stirnimann, F., & Rukes, B. (2009). Combined-cycle gas and steam turbine power plants (3rd ed.). PennWell. https://doi.org/10.1016/B978-0-7506-6357-2.X5000-3

ATU- FJIMSE, Vol. 4 No. 1 (2026): volume 4,Issu 1,May 2026 , All Rights Reserved

[2] Chacartegui, R., Sánchez, D., Muñoz, J. M., & Sánchez, T. (2009). Alternative ORC bottoming cycles for combined cycle power plants. Applied Energy, 86(10), 2162–2170. https://doi.org/10.1016/j.apenergy.2009.02.016
[3] Li, Y., Lin, Y., He, Y., Zhang, G., Zhang, L., Yang, J., & Sun, E. (2023). Part-load performance analysis of a dual-recuperated gas turbine combined cycle system. Energy, 269, 126744. https://doi.org/10.1016/j.energy.2023.126744
[4] Yang, Y., Bai, Z., Zhang, G., Li, Y., Wang, Z., & Yu, G. (2019). Design/off-design performance simulation and discussion for the gas turbine combined cycle with inlet air heating. Energy, 178, 386–399. https://doi.org/10.1016/j.energy.2019.04.136
[5] Sanjay, Y., Singh, O., & Prasad, B. N. (2007). Energy and exergy analysis of steam cooled reheat gas–steam combined cycle. Applied Thermal Engineering, 27(17–18), 2779–2790. https://doi.org/10.1016/j.applthermaleng.2007.03.011
[6] Liu, Z., & Karimi, I. A. (2020). Gas turbine performance prediction via machine learning. Energy, 192, 116627. https://doi.org/10.1016/j.energy.2019.116627
[7] Hundi, P., & Shahsavari, R. (2020). Comparative studies among machine learning models for performance estimation and health monitoring of thermal power plants. Applied Energy, 265, 114775. https://doi.org/10.1016/j.apenergy.2020.114775
[8] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785
[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.
[11] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.
[12] Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011
[13] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144. https://doi.org/10.1145/3422622
[15] Siddiqui, R., Anwar, H., Ullah, F., Ullah, R., Rehman, M. A., Jan, N., & Zaman, F. (2021). Power prediction of combined cycle power plant (CCPP) using machine learning algorithm-based paradigm. Wireless Communications and Mobile Computing, 2021, 9966395. https://doi.org/10.1155/2021/9966395
[16] Dai, S., Zhang, X., & Luo, M. (2024). A novel data-driven approach for predicting the performance degradation of a gas turbine. Energies, 17(4), 781. https://doi.org/10.3390/en17040781
[17] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
[18] Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). SMOTE for regression. In Progress in Artificial Intelligence (pp. 378–389). Springer. https://doi.org/10.1007/978-3-642-40669-0_33
[19] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672–2680.
[20] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32, 7335–7345.
[21] Shen, K., Zhang, H., Tong, X., & Lu, H. (2025). Advancing LightGBM with data augmentation for predicting the residual strength of corroded pipelines. npj Materials Degradation, 9, 128. https://doi.org/10.1038/s41529-025-00673-9
[22] Pacífico, L. D. S., Maciel, T. T., & Ludermir, T. B. (2024). Strategic data augmentation with CTGAN for smart manufacturing: Enhancing ML predictions of paper breaks in pulp-and-paper production. Journal of Intelligent Manufacturing, 36, 2255–2268. https://doi.org/10.1007/s10845-024-02453-9

[23] Habibi, O., Chemmakha, M., & Lazaar, M. (2023). Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT Botnet attacks detection. Engineering Applications of Artificial Intelligence, 118, 105669. https://doi.org/10.1016/j.engappai.2022.105669

[24] Santarisi, N. S., & Faouri, S. S. (2021). Prediction of combined cycle power plant electrical output power using machine learning regression algorithms. Eastern-European Journal of Enterprise Technologies, 6(8/114), 17–28. https://doi.org/10.15587/1729-4061.2021.245663

[25] Oyekale, J., Heberle, F., & Brüggemann, D. (2023). Machine learning for design and optimization of organic Rankine cycle plants: A review of current status and future perspectives. WIREs Energy and Environment, 12(4), e474. https://doi.org/10.1002/wene.474

[26] Wang, S., Liu, C., Li, Q., Liu, L., Huo, E., & Zhang, C. (2023). Comparison of random forest, support vector regression, and long short term memory for performance prediction and optimization of a cryogenic organic Rankine cycle (ORC). Energy, 280, 128069. https://doi.org/10.1016/j.energy.2023.128069

[27] Wang, Q., & Lu, H. (2024). A novel stacking ensemble learner for predicting residual strength of corroded pipelines. npj Materials Degradation, 8, 87. https://doi.org/10.1038/s41529-024-00508-z