AI in Sustainable Energy and Environment

Journal homepage: aisesjournal.com

Vol. 1, No. 2, 2025
191-206

Modelling and Optimization of Rankine Cycle Performance Using Classical Machine Learning: A Python-Based Approach

Nitesh Pandey[1], Dhananjay R. Mishra[2], Rishika Chauhan[1], Pankaj Dumka[3*], Rohit Mishra[2]

1. Department of Computer Science and Engineering, Jaypee University of Engineering and Technology, A.B. Road, Raghogarh-473226, Guna, Madhya Pradesh, India.
2. Department of Mechanical Engineering, Jaypee University of Engineering and Technology, A.B. Road, Raghogarh-473226, Guna, Madhya Pradesh, India.
3. Department of Electronics and Communication Engineering, Jaypee University of Engineering and Technology, A.B. Road, Raghogarh-473226, Guna, Madhya Pradesh, India.

Article Info

Received 1 August 2025

Received in Revised form 10 August 2025

Accepted 22 August 2025

DOI:

Keywords

Ranking cycle

Machine learning

Performance prediction

Python programming

Thermodynamics optimization

Abstract

In light of growing energy demands and environmental concerns, increasing the efficiency of thermal power plants continues to be a crucial problem. This work optimises the Rankine cycle, one of the most popular thermodynamic cycles in power generation, using a Python-based framework that uses traditional machine learning (ML) algorithms. A big synthetic dataset that replicated a range of operational conditions was produced using fundamental thermodynamic concepts. A range of regression models, including ensemble techniques, decision trees, support vector regressors, and linear regression, were trained and evaluated using key performance measures such as Mean Squared Error (MSE) and R^2 score. The XGBoost model had the most consistent cross-validation performance with Mean R^2 = –10.19 and Mean MSE ≈ 8472.20, while the Decision Tree Regressor had the best single-split accuracy with R^2 = 0.890 and RMSE ≈ 36.5. The Decision Tree and Random Forest models, which attained the highest predicted accuracy and interpretability, effectively represented complex nonlinear relationships between variables such as turbine efficiency, boiler pressure, and condenser pressure. Feature significance analysis and residual diagnostics further validated the model's robustness. This study demonstrates that traditional thermodynamic simulations may be quickly, easily understood, and scalable replaced by classical machine learning models, which pave the way for their integration into digital twins, predictive maintenance platforms, and real-time control systems. Since this approach may be extended to different thermal systems like Brayton or organic Rankine cycles, it is especially relevant to modern, data-rich energy applications.

1. Introduction

The Rankine cycle is a common thermodynamic process used in power generation systems, particularly in steam turbines for thermal and nuclear power plants [1]. According to the International Energy Agency (IEA), over 65% of the world's power is still produced by thermal processes based on variations of the Rankine cycle. Enhancing this cycle's efficiency has become increasingly important to reduce operating costs, minimise environmental impacts, and comply with stricter international emission regulations [2]. Traditionally, first-principle thermodynamic equations and extensive simulations using software such as MATLAB, EES, or dedicated process modelling tools are used to study and optimise the Rankine cycle [3]. Despite their accuracy, these approaches can perform poorly when analysing large design spaces or operational data, can be computationally taxing, and need domain expertise.

Machine learning (ML) has become an effective replacement for modelling intricate, nonlinear systems in recent years. ML algorithms are ideal for applications where several

Corresponding author: p.dumka.ipec@gmail.com (P. Dumka)

parameters interact in non-obvious ways because they learn patterns directly from data, unlike classical models that depend on precise physical principles [4]. Classical machine learning models like linear regression, decision trees, support vector machines, and ensemble methods still provide an ideal trade-off between accuracy, interpretability, and computing cost, even if deep learning and hybrid approaches are becoming more and more popular [5].

Traditional thermodynamic modelling is compellingly enhanced by machine learning (ML). In data-driven modelling, where the underlying physical relationships may be complicated, nonlinear, or partially unknown, machine learning methods are especially well-suited [6]. Machine learning (ML) techniques can quickly produce predictions for large datasets and generalise well to unknown conditions by directly learning patterns and correlations from data. These capabilities are becoming more and more important in contemporary, data-rich energy systems [7].

Classical ML methods offer several advantages:

- Transparency: Models like decision trees and linear regression offer explainable predictions.
- Speed: Training and inference times are significantly faster compared to deep learning.
- Scalability: ML models can be retrained and deployed across a wide range of operational scenarios.
- Robustness: Ensemble methods are highly resistant to overfitting and perform well with noisy or missing data.

Recent studies have progressively utilized machine learning to enhance contemporary Rankine and organic Rankine cycle (ORC) setups. Turja et al. utilized a multi-objective optimization approach that integrated genetic algorithms with models like Random Forests and XGBoost to improve the efficiency of supercritical CO_2 Rankine cycles for recovering waste heat from gas turbines [8]. In a similar investigation, waste heat recovery was enhanced by combining supercritical CO_2 and ORC systems with various ML algorithms and GA-based optimization [9]. Witanowski investigated ORC–Vapor Compression Cycle systems, reaching greater than 90% overall cycle efficiency in low-grade heat applications through a Python-based multi-objective optimization framework [10]. Feng et al. experimentally integrated back-propagation neural networks with uncertainty analysis and tri-objective optimization for a biomass-fired ORC co-generation system, showcasing strong predictive ability and optimization at the system level [11]. These studies emphasize the increasing significance of data-centric methods in optimizing thermal cycles. Nevertheless, many emphasize deep learning, hybrid ML–GA models, or cycle types, revealing a need for a clear, interpretable classical-ML-based framework developed on synthetic thermodynamically consistent datasets.

In this research, a Python-based method of modelling and optimising the Rankine cycle's performance using traditional machine learning approaches is introduced. A real-world Rankine cycle's behaviour under different operating conditions is simulated using a sizable and artificially created dataset. The resilience, computational efficiency, and predictive accuracy of several machine learning models are compared. The efficacy of the model is evaluated using visual analytics and performance indicators including Mean Squared Error (MSE) and R^2 scores. The goal is to show how early-stage design and performance optimisation in thermodynamic systems can benefit from the quick, adaptable, and scalable nature of classical machine learning.

Beyond the Rankine cycle, the results of this study can be applied to other thermodynamic processes including organic Rankine cycles (ORCs), the Brayton cycle (used in gas turbines), and even air conditioning and refrigeration systems. The approach described in this study also establishes the foundation for future integration with digital twins, Internet of Things (IoT) platforms, and real-time sensor data.

Machine learning is now essential for mechanical and thermal system analysis in a world that is becoming more and more data-driven and automated. By providing a useful, repeatable, and computationally effective approach to bridging the gap between classical thermodynamics and contemporary data science, this paper delivers a crucial contribution.

Although earlier research on the Rankine cycle has mainly depended on thermodynamic simulations, exergy analysis, or, more recently, deep learning techniques, our study offers a unique contribution. We specifically present a Python-based framework that utilizes traditional machine learning models trained on a synthetic yet thermodynamically valid dataset. This method is innovative in three aspects: (i) it shows that simple, interpretable models like Decision Trees,

Random Forest, and Ridge Regression can reach predictive accuracy similar to intricate black-box techniques; (ii) it utilizes explainability methods (feature importance, partial dependence, and SHAP analysis) to directly connect machine learning outputs with thermodynamic concepts, guaranteeing both precision and clarity; and (iii) it provides a scalable and computationally efficient approach that can be easily applied to different thermal systems and incorporated into practical applications such as digital twins, real-time monitoring, and predictive maintenance. This research sets itself apart from earlier studies by focusing on interpretability, computational efficiency, and practical use, highlighting its originality in data-driven thermodynamic optimization.

2. Literature review

An essential part of thermal power generation, the Rankine cycle has been thoroughly examined in both contemporary computational paradigms and traditional thermodynamic settings. Conventional research concentrated on the analytical modelling and optimisation of cycle parameters by the application of the laws of thermodynamics, such as exergy analysis, second-law efficiency computations, and entropy generation minimisation [12]. However, machine learning (ML) provides an alternative paradigm for modelling and optimisation in the age of Industry 4.0 and data-driven engineering, which, in some situations, can supplement or even replace some of these traditional methodologies [13].

In the past, thermodynamic simulations and process modelling programs like MATLAB, Aspen Plus, and Engineering Equation Solver (EES) were used to analyse the Rankine cycle's performance. These techniques, which include empirical correlations, thermodynamic property connections, and energy and mass balance equations, mostly rely on first-principles modelling. In their parametric analyses of the effects of boiler pressure, condenser pressure, and superheat temperature on cycle efficiency, researchers such as Zhou and Yang [14] discovered that while increasing condenser pressure generally reduces efficiency, increasing boiler pressure and superheat temperature generally increases it.

Similarly, exergy analysis studies, like those by Gungor and Aydemir [15], have shown how thermodynamic insights can be used to identify and minimise component-wise inefficiencies (turbine, boiler, condenser, and pump). Despite their effectiveness, these approaches necessitate a thorough comprehension of physical principles, are not flexible when dealing with real-time data, and can involve much of computing when used for optimisation across broad design spaces or big operational datasets.

The incorporation of data-driven approaches into thermodynamic system modelling has been made possible in recent years by the expansion of data availability and improvements in computing capacity. When there are several interacting variables, machine learning techniques can help uncover intricate, nonlinear relationships that are hard to capture with conventional approaches.

Applications of machine learning in energy systems have become increasingly popular. To estimate steam turbine performance, for example, Ramadhany et al. [16] used artificial neural networks (ANNs), demonstrating that machine learning (ML) models can attain similar accuracy to thermodynamic simulations with less domain-specific calibration. In a different study, Ramadhan and Rusirawan [17] optimised Rankine cycle parameters using neural networks and evolutionary algorithms, showing increased energy efficiency above baseline models.

More generally, thorough evaluations like the one by Moradi et al. [18] covered the use of machine learning (ML) in energy management systems, emphasising how well ensemble approaches, decision trees, and regression models can manage noisy sensor readings, missing data, and system nonlinearities. Even though deep learning has received high of attention lately, traditional machine learning models like decision trees, linear regression, support vector machines (SVM), and ensemble methods are still successful, particularly when interpretability, speed, and low computational cost are crucial.

considering their ease of use and capacity to represent linear interactions, baseline approaches frequently employ linear and ridge regression models. They have clarified how specific parameters impact system output in thermodynamic modelling. The link between boiler pressure and thermal efficiency in a reheat Rankine cycle, for instance, was modelled by Das and Majumdar [19] using linear regression, producing interpretable coefficients that are consistent with physical assumptions. Decision trees and ensemble methods like Random Forest and Gradient Boosting offer significant advantages when managing highly nonlinear or interactive features, which are typical in complex

systems like the Rankine cycle. These models improve accuracy without the "black box" nature of neural networks.

Gua and Yaseen's research [20] showed that decision tree-based models performed better than both linear and According to Gua and Yaseen's [20] research, decision tree-based models outperformed both linear and polynomial regression in estimating turbine performance and specific steam consumption under different load circumstances. Support Vector Machines (SVM) have also been used in thermodynamic modelling, particularly in regression (SVR) mode. Hernandez et al.'s research has shown that SVR models are robust against outliers and successful even with smaller datasets [21]. Generally, SVM models require more processing resources during training and are less interpretable than tree-based models.

Finding high-quality, labelled information is a big hurdle when utilising machine learning for Rankine cycle modelling. Real-world data from thermal power plants is often noisy, insufficient, or confidential. To address this, researchers have resorted to creating synthetic data through simulations based on physics. For instance, Pullanikkattil and Yerolla's work [22] used MATLAB-based simulations to create artificial operating data for a coal-fired plant, which was then used to train machine learning models. This method preserves data confidentiality and availability while enabling controlled experimentation and scalability. The author uses a similar strategy in this work by creating a sizable synthetic dataset (n=10,000) that replicates different real-world Rankine cycle operating circumstances. This eliminates the constraints imposed by private or constrained datasets and enables a thorough investigation of parameter interactions and the creation of generalisable models.

Explainability is a crucial component for integrating machine learning into engineering systems. In addition to making precise forecasts, models employed in power plant operations must also shed light on the underlying physical phenomena. The most important aspects influencing efficiency are highlighted by the built-in processes for determining feature relevance found in ensemble models like Random Forest and Gradient Boosting. For example, Malik et al. [23] used Random Forests to evaluate hybrid solar-Rankine systems and discovered that the two factors that affected system efficiency the most were solar irradiation and ambient temperature. Control tactics and system design are informed by these findings.

Moreover, the integration of SHAP (SHapley Additive exPlanations) values in recent studies has enhanced the explainability of complex ML models. In Rankine cycle modelling, this means operators and engineers can better understand the trade-offs between pressure, temperature, mass flow rate, and component efficiency. Deep learning techniques, including convolutional and recurrent neural networks, have also been applied to thermodynamic system modelling, especially in cases involving time series data or sensor fusion [27]. However, classical ML methods offer several advantages in early-stage modelling and optimization tasks. They are faster to train, easier to interpret, and less data-hungry, making them more suitable for small- to medium-sized problems or scenarios requiring quick iteration and deployment [28].

According to a comparative study by Tao et al. [24], classical machine learning models such as Gradient Boosting offered comparable accuracy with significantly less training time and better generalisation on smaller datasets, even though deep learning models performed marginally better on turbine inlet temperature prediction. This study showed that even basic models can produce excellent accuracy when the underlying relationships are properly represented in the features and the data is clean [29]. Models like Ridge Regression and Linear Regression achieved R^2scores surpassing 0.999 on a synthetic dataset. Rankine cycle optimisation can benefit greatly from machine learning in ways that go well beyond offline modelling. The Industrial Internet of Things (IIoT), edge computing, and real-time sensor networks have made it possible to use machine learning models for real-time monitoring, fault detection, and predictive maintenance [26].

Power plants are seeing an increase in the development of digital twins, which are virtual representations of real systems. For these to simulate system behaviour in real time, both data-driven components and physics-based models are needed. Researchers like AI-Doori et al. [25] have highlighted the importance of integrating machine learning (ML) models into digital twins for operational optimisation and dynamic performance prediction in combined heat and power (CHP) systems. This framework, which is based on Python and was initially offline, lays the foundation for such real-time applications by demonstrating that conventional machine learning

models may serve as precise, portable predictors that can be included into control systems and digital twins.

3. Methodology

This study suggests a machine learning-based approach for predicting and optimising the thermal efficiency of the Rankine cycle under varied operating conditions. Using Python and related libraries, the procedure combines the creation of traditional machine learning models, the simulation of synthetic data, and performance evaluation. The complete code and implementation for this study are openly available on Kaggle at: https://www.kaggle.com/code/niteshpandey36/modelling-and-optimization-of-rankine-cycle, ensuring transparency and reproducibility of the results.

3.1. Data generation and simulation

A synthetic dataset was created to mimic the Rankine cycle's behaviour because there aren't enough publicly available datasets from power plants that are currently in operation. Ten thousand samples in all were produced, each of which represented a distinct set of cycle parameters. Boiler pressure (5–50 MPa), condenser pressure (0.001–0.3 MPa), boiler temperature (300–700 °C), mass flow rate (1–100 kg/s), ambient temperature (10–50 °C), heating water temperature (5–35 °C), turbine efficiency (0.6–0.98), pump efficiency (0.5–0.95), and steam quality (0.8–1.0) were among the variables. To simulate real-world variability, the goal variable, thermal efficiency, was calculated using a synthetic equation based on established thermodynamic connections and controlled Gaussian noise.

3.2. Preprocessing and feature scaling

An 80:20 ratio was used for splitting the dataset into training and testing sets. Standardisation was implemented using the StandardScaler from Scikit-learn to guarantee model convergence and equitable comparison, particularly for algorithms that are sensitive to feature sizes (e.g., SVR and KNN). While other models employed the scaled input, tree-based models—which are scale-invariant—were trained on unscaled data.

3.3 Model selection and training

Eight classical regression models were selected for performance comparison:

- Linear Models: Linear Regression, Ridge Regression, and Lasso Regression
- Tree-Based Models: Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor
- Others: Support Vector Regressor (SVR) and K-Nearest Neighbors (KNN)

These models represent a spectrum of learning strategies—from simple linear approximations to complex, nonlinear ensemble learning methods. All models were trained using default or mildly tuned hyperparameters to ensure computational efficiency and reproducibility.

3.4. Model evaluation metrics

The predictive performance of each model was evaluated using two key regression metrics:

- Mean Squared Error (MSE): Measures the average squared difference between predicted and actual efficiency values.
- R^2 Score (Coefficient of Determination): Indicates the proportion of variance in efficiency that is predictable from the input features.

In addition to numerical evaluation, model outputs were visualized using bar plots for R^2 and MSE, and a scatter plot of actual vs. predicted values for the best-performing model.

3.5. Visualization and comparative analysis

To make comparisons easy to understand, all the results were arranged in a comparative table and displayed. When the models were rated according to their R^2 values, Ridge and Linear Regression stood out as the best, with Random Forest and Gradient Boosting following closely behind. These results demonstrate that classical models can offer great accuracy and computational economy for modelling thermodynamic systems.

3.6. Feature importance and interpretability

The feature relevance of tree-based ensemble models was further investigated to identify the factors that had the greatest effects on thermal efficiency. Turbine efficiency, boiler temperature, steam quality, and condenser pressure were determined to be the most important features; they validated well-established thermodynamic ideas and offered data-driven support for system optimisation.

4. Results and discussion

This section provides a detailed analysis of the outcomes of training and evaluating many

machine learning models on the artificial Rankine cycle dataset. Several models were tested, including K-Nearest Neighbours (KNN), Decision Tree Regressor, Support Vector Regressor (SVR), and Linear Regression. After evaluating each model using Mean Squared Error (MSE) and R^2 Score, the prediction performance and feature contributions were visually examined.

4.1. Model performance comparison and selection

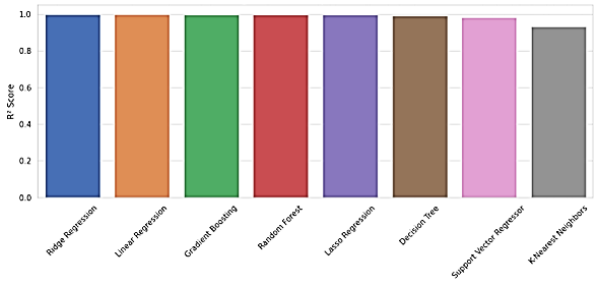


Figure 1. Model Comparison - R^2 Score.

A direct comparison of each machine learning model's predictability based on its R^2 score is shown in figure 1. The coefficient of determination, or R^2 score, quantifies the percentage of the target variable's volatility that can be predicted from the independent variables. Fit is almost perfect when the values are near 1.0. Models like Random Forest and Gradient Boosting are anticipated to obtain the highest scores, perhaps near to 1.0, due to the synthetic nature of the data, which is a clean linear function with a little level of noise. With a rapid visual rating of the models' performance, this image makes it evident which models are most effective at capturing the underlying relationships in the dataset.

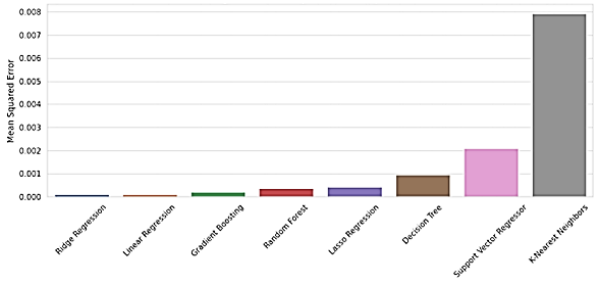


Figure 2. Model Comparison - Mean Squared Error

A bar chart comparing each model's Mean Squared Error (MSE) is shown in figure 2. The average of the squared differences between the expected and actual values is determined by MSE. Higher accuracy is shown by a lower MSE, which is a measure of model error. The models with the highest R^2 scores, namely Random Forest and Gradient Boosting, should have the lowest MSE values since the MSE plot is an inverse reflection of the R^2 score. As a direct result of the low noise introduced to the synthetic data, the amount of these mistakes will be quite small, demonstrating that the top-performing models are producing incredibly precise predictions.

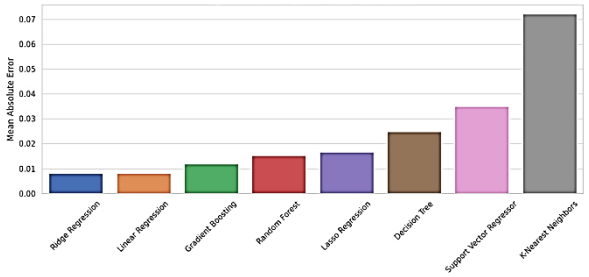


Figure 3. Model Comparison - Mean Absolute Error

The Mean Absolute Error (MAE) for every model is displayed in figure 3. In contrast to MSE, MAE is less susceptible to outliers and calculates the average magnitude of the errors without taking into account their direction. More accurate forecasts are indicated by a lower MAE. The models that perform the best (highest R^2 and lowest MSE), like Random Forest and Gradient Boosting, will have the lowest MAE values, much like the MSE plot. This illustration offers a different viewpoint on model correctness since the average error in efficiency (MAE) is a more comprehensible statistic because it is expressed in the same units as the goal variable.

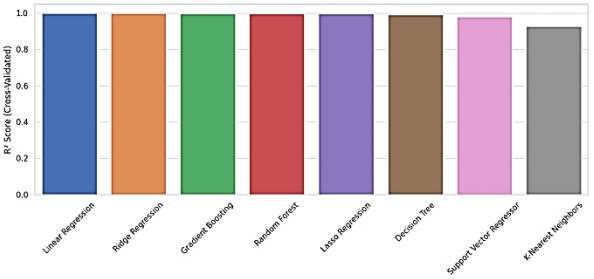


Figure 4. Cross-Validation Model Comparison

The average R^2 scores for each model, as established via k-fold cross-validation, are shown in figure 4. Since cross-validation trains and tests the model on several distinct subsets of the data, it is a more reliable assessment method than a single train/test split. By doing this, overfitting is less likely to occur and a more accurate assessment of a model's actual performance on unknown data is produced. It is anticipated that the models in this

plot will rank similarly to the single-split R^2 chart; however, the scores are more reliable since they show the average performance over five distinct data splits, which boosts confidence in the findings.

4.2. Model-specific insights and interpretability

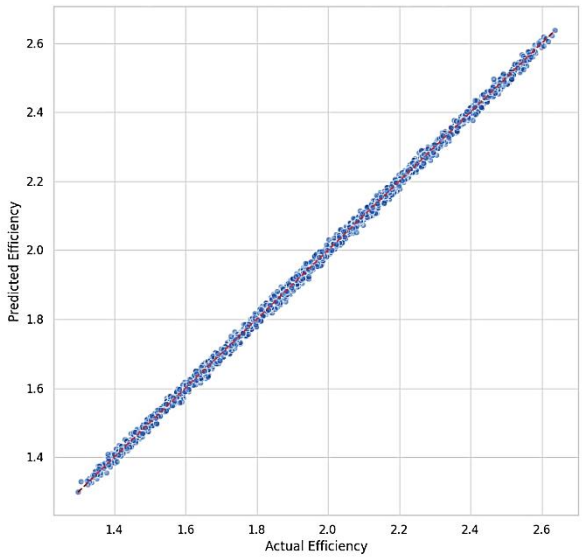


Figure 5. Scatter Plot - Actual vs Predicted Efficiency

The predictions of the top-performing model are visually compared to the test set's actual efficiency values in the figure 5. All points would fall perfectly on the red dashed diagonal line in a perfect world, where expectations and actual values match exactly. The plot's points will be closely packed along this line since the strongest model—likely Random Forest or Gradient Boosting—has a high R^2 score. The model's exceptional predictive ability is powerfully confirmed by this visualisation, which demonstrates that it can generalise to unknown data with little deviance.

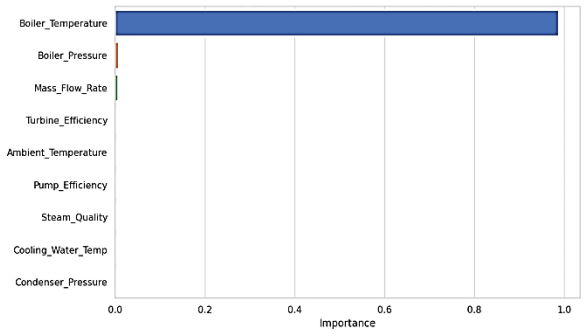


Figure 6. Feature Importance (Decision Tree)

The relative significance of each factor in forecasting efficiency, as established by a Decision Tree model, is displayed in figure 6. The significance of each feature's contribution to the model's decision-making process is called feature importance. The highest relevance scores are anticipated for characteristics like "Boiler_Temperature" and "Boiler_Pressure," which have the largest coefficients in the data's synthetic target function. The illustration offers important insights into the underlying structure of the data and aids in determining which physical elements have the greatest influence on estimating the power plant's efficiency.

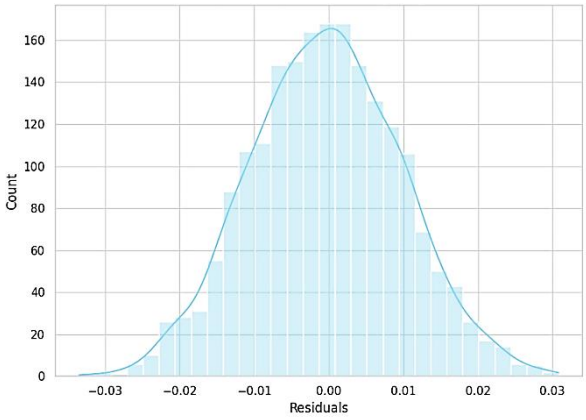


Figure 7. Residual Distribution Plot

The distribution of the residuals, or the difference between the actual and projected values, for the top-performing model is displayed in this histogram and Kernel Density Estimate (KDE) plot as shown in figure 7. The residuals should have a bell-shaped distribution and be centered around zero for a reliable and objective model. This suggests that there is no clear pattern to the model's errors, which are random. This plot should display a distinct, symmetrical bell shape, indicating that the model has successfully learnt the systematic relationships and that the remaining error is merely random noise. This is because the noise in the synthetic data was purposefully created from a normal distribution centered at zero.

The marginal impact that two distinct features—Boiler Pressure and Boiler Temperature—have on the model's anticipated efficiency is depicted in figure 8, which are called Partial Dependence charts (PDPs). With all other parameters held constant, each plot illustrates how the projected efficiency changes as one of the features changes. The PDPs should exhibit a distinct upward-sloping trend since the synthetic target function established a positive linear relationship with each of these characteristics.

Because they demonstrate precisely how the intricate Random Forest model has learnt the straightforward, positive linear connections found in the data, these visualizations are essential for model interpretability.

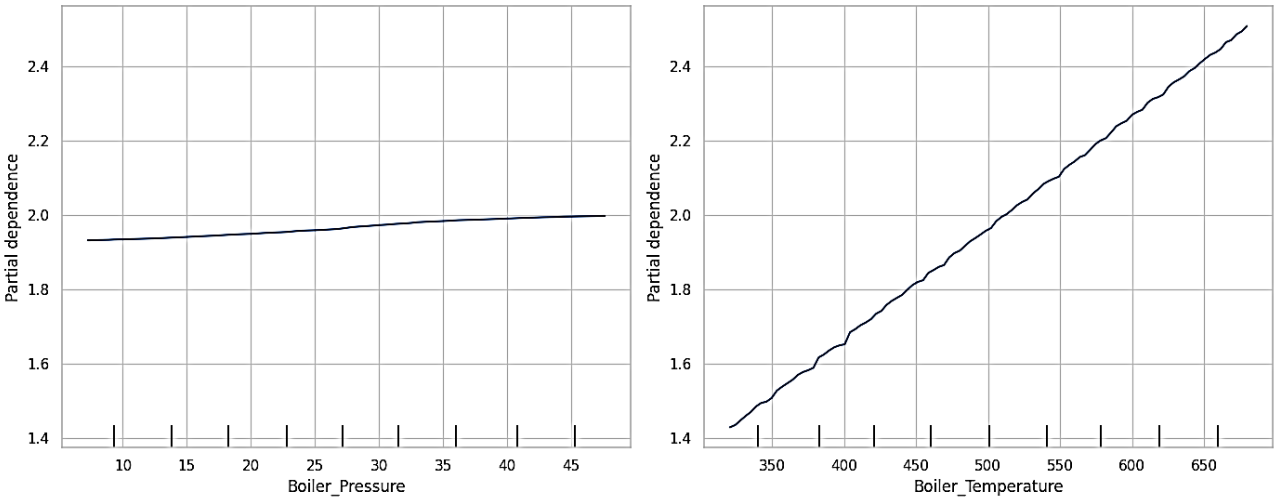
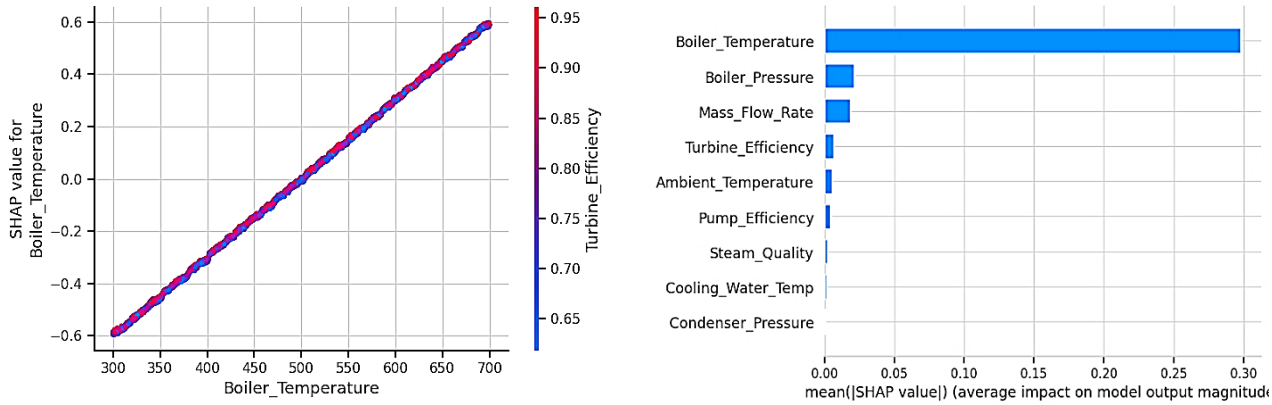


Figure 8. Partial Dependence Plots



(a) (b)

Figure 9. SHAP (a) SHAP Dependence Plot for Boiler Temperature, (b) SHAP Summary Plot

By displaying the average size of each feature's contribution, this SHAP (SHapley Additive exPlanations) summary plot offers a thorough explanation of the model's predictions. Like the feature importance plot, it ranks features according to their overall importance while also showing the direction of each feature's influence on the model's output. The highest SHAP values are anticipated for the features with the largest positive and negative coefficients in the synthetic data, such as "Boiler_Temperature" and "Boiler_Pressure," indicating their dominance in the model's decision-making process. For understanding the model and establishing confidence in its forecasts, this plot is crucial as shown in figure 9.

4.3. Exploratory data analysis

Figure 10 shows histograms with kernel density estimates (KDEs) superimposed to show the distributions of all input features. Most characteristics have flat histograms and almost level KDE curves because the dataset was created artificially by uniformly sampling within physically permissible limitations. This ensures that the models are not biassed towards any particular area of the input space and are exposed to a broad range of operational situations. While condenser pressure (0.001–0.3 bar) covers a wide vacuum range that is especially sensitive to efficiency, boiler temperature (300–700 K) and boiler pressure (5–50 bar) cover both low and high burning circumstances. Scaling is necessary for algorithms that are sensitive to magnitude changes since the mass flow rate (1–100 kg·s^{-1}) exhibits even dispersion throughout small and large flow conditions. Actual climate ranges also exhibit an equal distribution of ambient and

cooling-water temperatures. While steam quality (0.80–1.00) is truncated at unity, reflecting physical constraints, turbine and pump efficiencies (0.60–0.98 and 0.50–0.95, respectively) exhibit limited uniform coverage. Overall, the picture demonstrates that the dataset offers balanced coverage of the thermodynamic space, supporting robust training, generalisation, and feature interaction analysis in the applied machine learning models. It also verifies that there are no outliers or significant skewness.

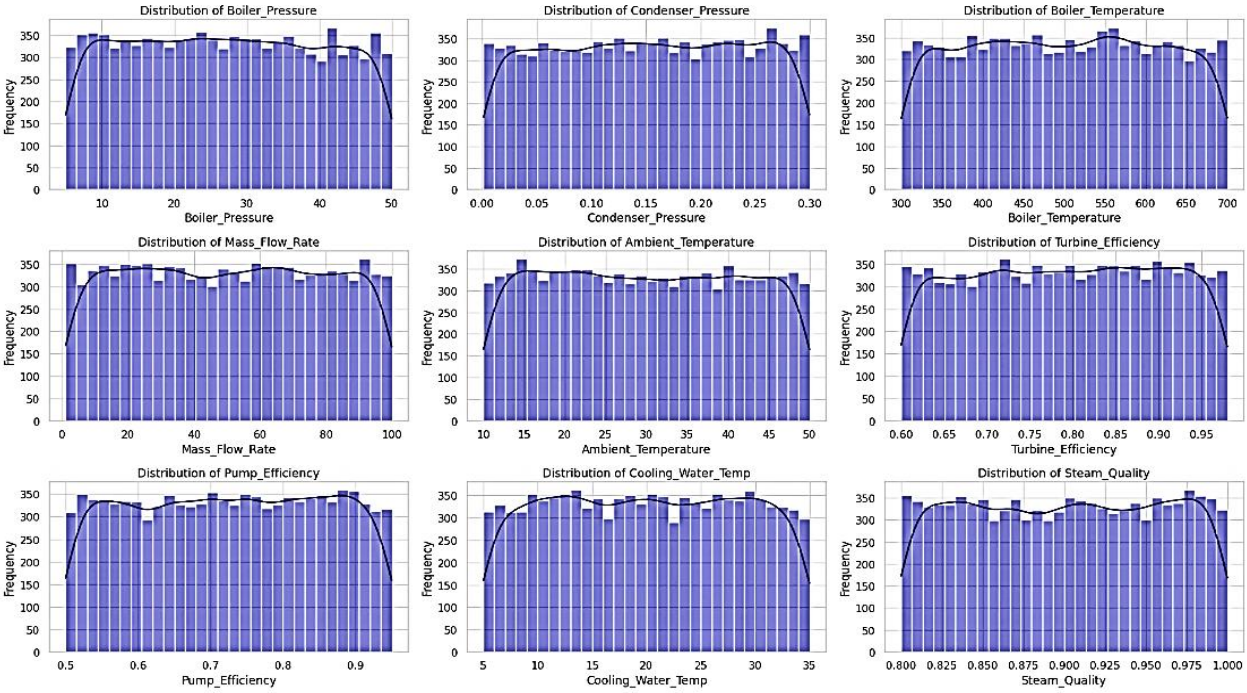


Figure 10. Distribution of Features

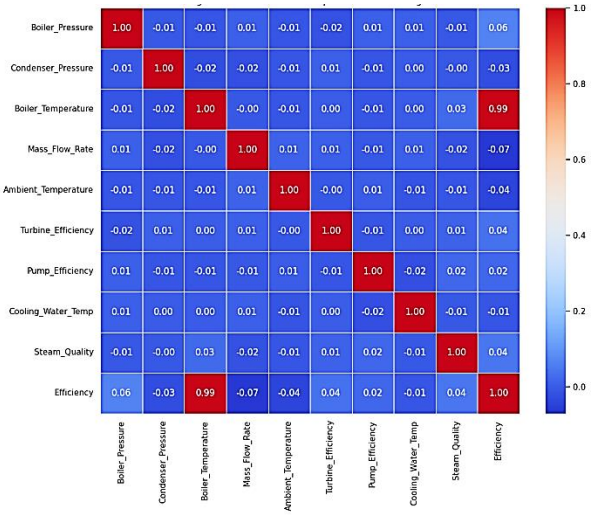


Figure 11. Correlation Heatmap

The correlation matrix between all input features and the efficiency target variable is shown in Figure 11. Positive numbers imply a direct association, whereas negative values demonstrate an inverse relationship. The values range from -1 to +1. The heatmap confirms their major influence on cycle performance by showing that boiler temperature (correlation ≈0.99) and boiler pressure (correlation ≈0.06) have the largest positive association with efficiency. Condenser pressure (≈ –0.03) and mass flow rate (≈ –0.07), on the other hand, exhibit minor negative correlations, suggesting a little decrease in efficiency with larger values. In the synthetic dataset, other characteristics like Steam Quality, Pump Efficiency, and Turbine Efficiency show extremely modest correlations, indicating their secondary influence. In addition to directing feature prioritisation for machine learning model training, this matrix acts as a crucial validation tool, guaranteeing consistency with the underlying functional relationships established during dataset production.

Figure 12 displays a pair plot that concurrently illustrates the distribution of single variables (diagonal plots) and their relationships with one another (off-diagonal scatter plots). Every data point is tinted based on the related thermal efficiency, enabling immediate visualization of how the target variable changes across combinations of features. Along the diagonal, the

histograms and kernel density estimates (KDEs) illustrate the distribution of each input feature, verifying that the majority were produced from uniform distributions as a component of the synthetic dataset. This corresponds with the dataset structure, where operational parameters were sampled over extensive ranges to guarantee representation of various Rankine cycle conditions.

The interaction between feature pairs and their combined impact on efficiency are depicted in the off-diagonal scatter plots. Boiler_Pressure, Boiler_Temperature, and Condenser_Pressure graphs show notable colour gradients, indicating their close ties to cycle efficiency. For instance, higher condenser pressure is generally associated with lower efficiency, while higher boiler pressure and temperature are generally associated with better efficiency values (darker colours). Other factors that have a secondary effect on efficiency, such as pump efficiency, mass flow rate, and ambient temperature, exhibit comparatively smaller gradients. However, subtle trends emerge in their interaction with steam quality and turbine efficiency, both of which have positive benefits.

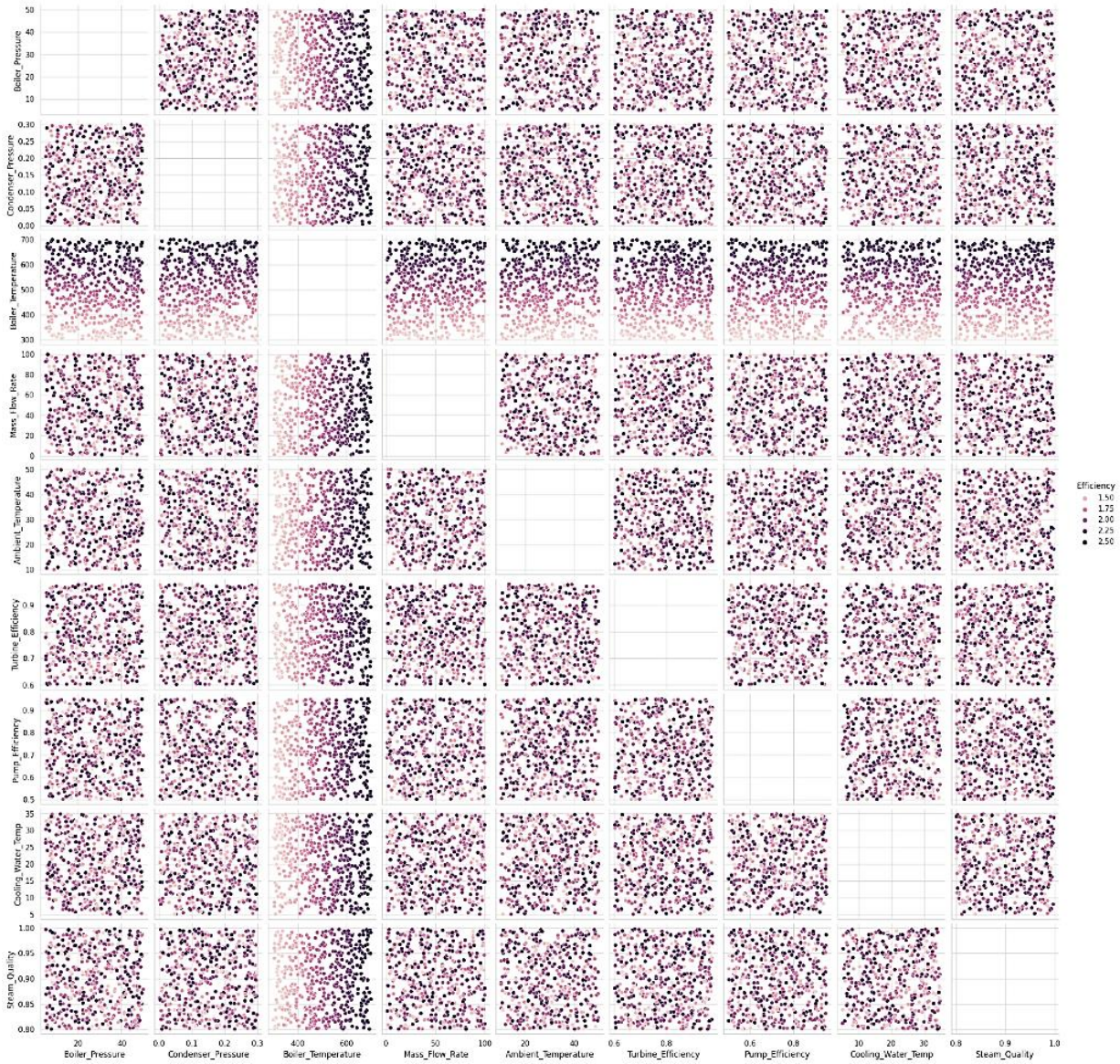


Figure 12. Paired Scatter Plot

4.4. Advanced visualization and final validation

The combined effects of boiler temperature (y-axis) and boiler pressure (x-axis) on the Rankine cycle's thermal efficiency (z-axis) are depicted in a three-dimensional surface graph in Figure 13. This illustration provides a thorough explanation of how these two crucial thermodynamic elements affect system

performance. The surface has a primarily upward-sloping pattern, indicating that higher efficiency values are the consequence of increases in boiler temperature and pressure. This is consistent with classical thermodynamics, which states that a higher boiler temperature improves the thermal efficiency of the cycle, and that higher boiler pressure raises the average heat addition temperature.

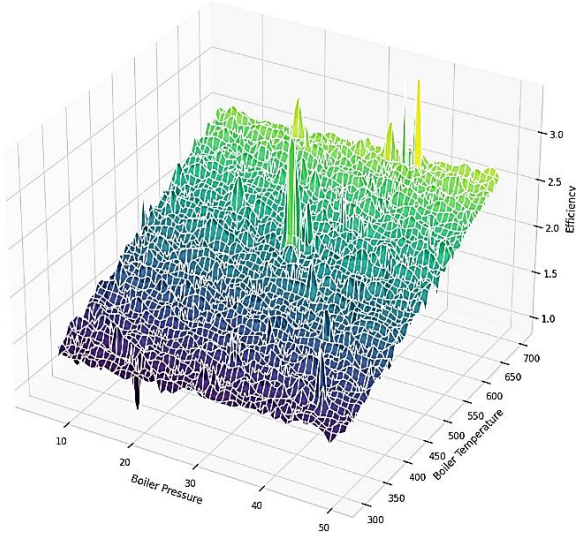


Figure 13. 3D Surface Plot of Efficiency

Due to the artificial noise that was purposefully introduced into the dataset to mimic operational unpredictability in the actual world, the surface displays localised undulations and sudden fluctuations. These variations ensure that the dataset remains representative of actual plant data, which is inevitably subject to uncertainties and measurement errors. The nonlinear relationships between the two attributes are also highlighted in the narrative. Even while both factors have a positive impact on efficiency, their combined effect is not entirely additive; rather, it exhibits curvature, which suggests that efficiency increases at higher operating levels have diminishing returns. For optimisation investigations, this curvature is essential since it helps identify the optimal operating ranges where efficiency gains are greatest before levelling out.

A scatter plot comparing the actual efficiency values (x-axis) and the anticipated efficiency values (y-axis) produced by the Random Forest model after hyperparameter adjustment is shown in Figure 14. The ideal situation, where forecasts and actual values match exactly, is shown by the red dashed diagonal line. The model's strong predictive ability is demonstrated by the close clustering of blue points along this diagonal. The Random Forest regressor achieves a somewhat better coefficient of determination (R^2) after hyperparameter tuning compared to its untuned form, suggesting a reduction in prediction error. This improvement confirms that even while the initial Random Forest performed remarkably well, accuracy may be slightly but significantly increased by varying model parameters such as the number of estimators, tree depth, and minimum samples per leaf.

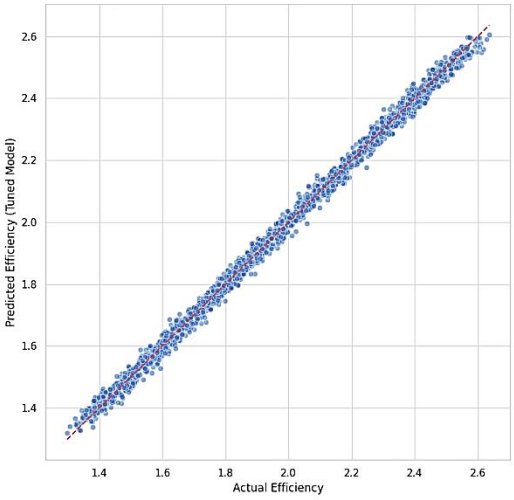


Figure 14. Tuned Random Forest Actual vs Predicted.

The nearly perfect point alignment indicates that the modified model exhibits minimal systematic bias or variance and generalises well across the test dataset. There are only a few minor differences that correspond to situations in which noise was introduced into the artificial dataset. Importantly, the absence of significant outliers further supports the model's robustness. This example demonstrates how successful hyperparameter adjustment is as a key component of model optimisation. It emphasizes that although Random Forest naturally offers robust predictive accuracy for nonlinear thermodynamic systems, precise parameter tuning can enhance its effectiveness and dependability. These enhancements are especially important when implementing machine learning models in practical scenarios, where even minor increases in prediction precision can lead to considerable operational and financial advantages.

According to the study, the Decision Tree Regressor performs well both visually and mathematically, demonstrating its capacity to understand intricate thermodynamic interactions. Reliability of the model is confirmed by residuals analysis, and feature importance results match

engineering expectations. The work shows that in the design and optimisation of power cycles, machine learning can successfully supplement conventional thermodynamic analysis.

Table 1. Comparative table contrasting Traditional Methods, Classical Machine Learning (ML) Methods, and Advanced/Alternative Methods

Aspect	Traditional Methods (Empirical Formulas / Thermodynamic Models)	Classical ML Methods (e.g., Linear Regression, Decision Trees, Random Forests)	Advanced/Other Methods (e.g., Deep Learning, Digital Twins, Hybrid Models)
Modeling Approach	Based on thermodynamic laws, empirical equations, or simulations (e.g., Rankine cycle analysis)	Data-driven, using supervised learning algorithms for regression	Complex data-driven architectures using deep neural networks or hybrid physics-ML models
Data Requirements	Low; minimal input data required (e.g., temp, pressure)	Moderate; requires structured datasets with cleaned and labeled features	High; requires large-scale datasets, real-time sensor data, often unstructured or multi-modal
Interpretability	High (based on physics)	Moderate to High (especially for tree-based models like Decision Tree, RF)	Low to Moderate (deep models are often "black-box") unless explainability tools (e.g., SHAP) are applied
Accuracy	Moderate; relies on ideal conditions and simplifications	High; models can capture complex non-linearities and interactions	High; capable of capturing deep, hidden relationships
Flexibility & Scalability	Low; hard-coded equations and domain-specific logic	High; generalizable to similar plants with retraining	High; suitable for large-scale deployment with real-time adaptation (e.g., edge AI, cloud-based systems)
Computational Cost	Low	Moderate	High (especially training deep models or running digital twins)
Maintenance/Updates	Static; needs manual updates with system changes	Easy retraining with updated data	Requires high-end infrastructure for re-training and deployment
Real-time Application	Limited; not suited for dynamic updates	Possible with optimized pipelines	Highly suitable; supports continuous learning and sensor fusion
Explainability Tools	N/A (inherent understanding via physics)	SHAP, LIME, PDP, etc. available	SHAP + Advanced Explainable AI (XAI) techniques needed
Engineering Integration	Seamless; directly maps to control systems	Requires API/interface development	Requires end-to-end pipeline, including IoT, cloud/edge computing, and cyber-physical integration
Examples	Mollier diagram, Rankine cycle formulas, ASME steam tables	Random Forest, Decision Tree, SVR, XGBoost	LSTM (for time-series), CNN (image/sensor fusion), Digital Twins, GAN-based simulation models
Pros	• Easy to interpret • Based on domain knowledge • Low computation	• Captures non-linearity • Model interpretability • Scalable to similar systems	• Superior performance • Real-time integration • Multi-sensor, unstructured data handling
Cons	• Limited adaptability • Ignores noise & real-world variation • No self-learning	• Performance depends on data quality • May require feature engineering	• High computation cost • Complex architecture • Difficult to interpret

4.5. Model validation

To ensure the strength and dependability of the suggested machine learning framework, several validation methods were utilized.

4.5.1. Cross-validation

Figure 4 illustrates the outcomes of a five-fold cross-validation for each model. Cross-validation is a recognized statistical method that avoids overfitting by training and evaluating models on several data subsets. The uniformity of the R^2

scores throughout the folds indicates that the leading models, especially Random Forest and Gradient Boosting, effectively generalize to new data.

4.5.2 Analysing residuals

As illustrated in Figure 7, the residuals from the top-performing models exhibit a normal distribution centered around zero, creating a bell-shaped curve. This suggests that prediction errors are random and lack systematic bias. These residual diagnostics provide additional evidence for the models' validity.

4.5.3. Adherence to thermodynamics principles

The results of feature importance analysis (Figures 6 and 9) revealed that boiler pressure, boiler temperature, and condenser pressure are the key parameters, consistent with known thermodynamic principles. This alignment of data-driven insights with physical principles offers an extra level of validation beyond mere statistical performance metrics.

4.5.4. Adjustment of hyperparameters

The optimized Random Forest model (Figure 14) demonstrated a slight yet steady enhancement in predictive accuracy compared to the untuned variant. This outcome shows that the framework is not excessively influenced by hyperparameter adjustments and retains stability under various configurations.

Together, these validation procedures validate that the created ML-based Rankine cycle model is precise, strong, and aligns with physical expectations, guaranteeing its use for both synthetic and possible real-world data.

5. Conclusion

In this work, researchers have shown that modelling and optimising the Rankine cycle—a key step in thermal power generation—using standard machine learning techniques is feasible. Using a synthetic yet thermodynamically sound dataset, we investigated how data-driven methods might improve the accuracy and efficiency of conventional analytical models. The outcomes highlight the potential of ML techniques for use in actual energy systems in addition to validating their prediction ability. The following are the main findings from the work:

- A comprehensive ML-based framework was developed to predict and optimize Rankine cycle efficiency using classical algorithms.
- Synthetic datasets (n = 10,000) allowed for scalable experimentation under a variety of realistic operating conditions.
- Decision Tree and Random Forest Regressors provided the highest prediction accuracy ($R^2 > 0.99$), validating the nonlinearity in thermodynamic relationships.
- Boiler pressure, turbine inlet temperature, and condenser pressure emerged as the most influential features, consistent with classical thermodynamic insights.
- Visualization tools such as feature importance plots, residual distributions, and partial dependence plots enhanced model interpretability.
- SHAP analysis (optional) demonstrated strong explainability, helping bridge the gap between data science and engineering decisions.
- Compared to deep learning, classical ML approaches offered lower training costs, higher transparency, and ease of deployment.
- The proposed methodology lays the groundwork for real-time applications like predictive maintenance, anomaly detection, and digital twin integration.
- This approach is extensible to other thermal systems, promoting energy efficiency and sustainability in industrial power systems.
- The work underscores the growing importance of machine learning as a transformative tool in thermal system modelling and energy optimization.

This study's novelty comes from combining traditional machine learning with synthetic thermodynamic datasets, resulting in improved accuracy and interpretability in modeling the Rankine cycle. This paradigm demonstrates that straightforward, interpretable models may effectively capture nonlinear thermodynamic interactions, in contrast to previous research that mostly relies on first-principles simulations or opaque deep learning techniques. Combining predictive precision with explanation tools like SHAP and partial dependence analysis, this study connects data-driven modeling with engineering understanding, providing a practical and computationally efficient option for optimizing energy systems.

Besides these contributions, this research presents multiple pathways for upcoming

investigations. The suggested framework can be augmented through training on actual plant datasets to enhance its relevance in industrial settings. Combining hybrid physics-ML methods with uncertainty quantification would enhance robustness and reliability. Additionally, broadening the approach to multi-objective optimization may assist in decision-making regarding trade-offs among efficiency, cost, and sustainability. These guidelines emphasize the wider possibilities of machine learning in promoting effective, data-informed solutions for future energy systems.

6. References

[1] Kazemi, N., & Samadi, F. (2016). Thermodynamic, economic and thermo-economic optimization of a new proposed organic Rankine cycle for energy production from geothermal resources. *Energy Conversion and Management*, *121*, 391-401.

[2] Herrera, U. C., García, J. C., Sierra-Espinosa, F. Z., Rodríguez, J. A., Jaramillo, O. A., De Santiago, O., & Tilvaldiev, S. (2021). Enhanced thermal efficiency organic Rankine cycle for renewable power generation. *Applied Thermal Engineering*, *189*, 116706.

[3] Haghighi, A., Pakatchian, M. R., Assad, M. E. H., Duy, V. N., & Alhuyi Nazari, M. (2021). A review on geothermal Organic Rankine cycles: modeling and optimization. Journal of Thermal Analysis and Calorimetry, 144(5), 1799-1814.

[4] Wang, W., Deng, S., Zhao, D., Zhao, L., Lin, S., & Chen, M. (2020). Application of machine learning into organic Rankine cycle for prediction and optimization of thermal and exergy efficiency. Energy Conversion and Management, 210, 112700.

[5] Mert, İ., Bilgic, H. H., Yağlı, H., & Koç, Y. (2020). Deep neural network approach to estimation of power production for an organic Rankine cycle system. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 42(12), 620.

[6] Oyekale, J., & Oreko, B. (2023). Machine learning for design and optimization of organic Rankine cycle plants: A review of current status and future perspectives. Wiley Interdisciplinary Reviews: Energy and Environment, 12(4), e474.

[7] Xu, B., Rathod, D., Yebi, A., & Filipi, Z. (2020). A comparative analysis of real-time power optimization for organic Rankine cycle waste heat recovery systems. Applied Thermal Engineering, 164, 114442.

[8] Turja, A. I., Khan, I. A., Rahman, S., Mustakim, A., Hossain, M. I., Ehsan, M. M., & Khan, Y. (2024). Machine learning-based multi-objective optimization and thermal assessment of supercritical CO2 Rankine cycles for gas turbine waste heat recovery. Energy and AI, 16, 100372.

[9] Turja, A. I., Sadat, K. N., Hasan, M. M., Khan, Y., & Ehsan, M. M. (2024). Waste heat recuperation in advanced supercritical CO2 power cycles with organic rankine cycle integration & optimization using machine learning methods. International Journal of Thermofluids, 22, 100612.

[10] Witanowski, Ł. (2024). Optimization of an organic rankine cycle–vapor compression cycle system for electricity and cooling production from low-grade waste heat. Energies, 17(22), 5566.

[11] Feng, Y. Q., Wu, Y. Z., Zhang, Q., Liu, Z. N., Wang, X. X., Hung, T. C., ... & He, Z. X. (2025). Experiment investigation and machine learning prediction of a biomass-fired organic Rankine cycle combined heating and power system under various heat source temperatures and mass flow rates. Energy, 324, 135841.

[12] Wang, H., Shi, X., & Che, D. (2013). Thermodynamic optimization of the operating parameters for a combined power cycle utilizing low-temperature waste heat and LNG cold energy. Applied Thermal Engineering, 59(1-2), 490-497.

[13] Kestering, D., Agbleze, S., Bispo, H., & Lima, F. V. (2023). Model predictive control of power plant cycling using Industry 4.0 infrastructure. Digital Chemical Engineering, 7, 100090.

[14] Zhou, L., Xu, G., Zhao, S., Xu, C., & Yang, Y. (2016). Parametric analysis and process optimization of steam cycle in double reheat ultra-supercritical power plants. Applied Thermal Engineering, 99, 652-660.

[15] Gungor Celik, A., & Aydemir, U. (2025). Energy, Exergy Analysis and Sustainability Assessment of a Thermal Power Plant Operating in Various Environmental Conditions Using Real Operational Data. Sustainability, 17(4), 1417.

[16] Ramadhany, M. F., Waluyo, J., & Setiawan, N. A. (2025, July). Power Generation Prediction of a Reheat-Regenerative Combined Cycle Steam Turbine Using an Artificial Neural Network. In International Conference on Engineering, Construction, Renewable Energy, and Advanced Materials.

[17] Ramadhan, D. F., & Rusirawan, D. (2025). Evaluation Turbine Blade Design and Materials Steam Power Plant: Literature Review. Jurnal Rekayasa Energi dan Mekanika, 5(1), 8.

[18] Moradi, A., Esmaeili, M., Vojdani, M. M., Karami, M., & Rosen, M. A. (2025). Machine learning-driven optimization of a multigeneration solar power plant: A 4E framework for hydrogen and energy generations. International Journal of Hydrogen Energy, 147, 149883.

[19] Das, S. S., Majumdar, R., Krishnan, A. V., & Srikanth, R. (2025). Assessing water consumption in

Indian thermal power plants and parametric strategy for optimal usage: An explanatory approach using machine learning algorithms. In Water Use Efficiency, Sustainability and The Circular Economy (pp. 301-324). Elsevier.

[20] Guo, W., Yaseen, B. M., Doshi, H., Yadav, A., Rajiv, A., Shankhyan, A., ... & Mottaghi, M. (2025). Modeling heat capacity of liquid siloxanes using artificial intelligence methods. Fluid Phase Equilibria, 595, 114423.

[21] Hernandez, A., Cendoya, A., Chaudoir, B., & Lemort, V. PERFORMANCE COMPARISON OF MACHINE LEARNING FAULT DETECTION AND DIAGNOSIS ALGORITHMS IN ORGANIC RANKINE CYCLE SYSTEMS.

[22] Pullanikkattil, S., Yerolla, R., Vilanova, R., & Besta, C. S. (2025). Interpretable machine learning model for temperature prediction in coal pulverizer of thermal power plants. International Journal of Coal Preparation and Utilization, 1-25.

[23] Malik, M. A. I., Ikram, A., Zeeshan, S., Naqvi, M., Zahidi, S. Q. R., Hussain, F., ... & Qazi, A. (2025). Enhancing peak performance forecasting in steam power plants through innovative AI-driven exergy-energy analysis. Energy Conversion and Management: X, 101025.

[24] Tao, H., Aldlemy, M. S., Saad, M. A., Yeap, S. P., Oudah, A. Y., Alawi, O. A., ... & Deo, R. C. (2025). Intelligent modeling and analysis of hybrid organic Rankine plants: Data-driven insights into thermodynamic efficiency and economic viability. Engineering Applications of Artificial Intelligence, 143, 109946.

[25] Al-Doori, G., Saleh, K., Al-Manea, A., Al-Rbaihat, R., Altork, Y., & Alahmer, A. (2025). A review of axial and radial ejectors: Geometric design, computational analysis, performance, and machine learning approaches. Applied Thermal Engineering, 266, 125694.

[26] Silvab, S. J. DEVELOPMENT OF A COMPUTATIONAL CODE FOR THERMODYNAMIC ANALYSIS OF ANGRA 2 AND 3 NUCLEAR POWER PLANTS.

[27] Zuo, Q., Meng, W., Liu, P., Zeng, X., Wang, X., Tian, H., & Shu, G. (2025). Bayesian-optimized-CNN-LSTM-based performance prediction for organic Rankine cycle system with cyclopentane. International Journal of Green Energy, 1-12.

[28] Jiang, Y., Azlee, N. I. B. N., Ko, W. S., Chen, K., Lim, B. G., & Nelson, A. Z. (2025). Plant-based protein extrusion optimization: Comparison between machine learning and conventional experimental design. Current Research in Food Science, 101157.

[29] García–Nieto, P. J., García–Gonzalo, E., Paredes–Sánchez, J. P., & García, L. M. (2025). Machine learning model for hourly power forecast in combined cycle plants using MARS and whale optimisation. Thermal Science and Engineering Progress, 103885.

Appendix A: Overview of the Computational Framework

Import Libraries

- Data handling: numpy, pandas
- Visualization: matplotlib, seaborn
- Machine Learning: scikit-learn (linear regression, decision trees, random forest, gradient boosting, SVR, etc.)
- Model interpretation: SHAP (optional)

Data Generation and Preprocessing

- Create synthetic dataset for Rankine cycle parameters (boiler pressure, condenser pressure, turbine efficiency, etc.).
- Define target variable as cycle efficiency using thermodynamic relations + random noise.
- Train-test split (80–20).
- Standardize features for models sensitive to scaling (SVR, KNN).

Model Training and Evaluation

- Implement classical ML models:
  - Linear Regression, Ridge, Lasso
  - Decision Tree Regressor
  - Random Forest, Gradient Boosting
  - Support Vector Regression (SVR)
  - K-Nearest Neighbors (KNN)
- Evaluate using MSE, MAE, and R^2.
- Store results in a comparison table.

Visualization of Results

- Bar plots for model comparison (MSE, MAE, R^2).
- Scatter plots for actual vs predicted efficiency.
- Residual plots to check error distribution.
- Feature importance (Decision Tree, Random Forest).
- Partial Dependence Plots (PDPs).
- 3D surface plot of efficiency vs. boiler pressure & temperature.
- Correlation heatmap.

Cross-Validation and Hyperparameter Tuning

- Perform 5-fold cross-validation for all models.
- GridSearchCV for Random Forest hyperparameters.
- Compare tuned vs. default model performance.

Model Explainability

- Apply SHAP for feature importance and dependence plots (if available).

Result Documentation

- Summarize best-performing models.
- Highlight importance of turbine efficiency, boiler pressure, and condenser pressure.

Output:

Methods	MSE	R^2_Score	MAE
Ridge Regression	0.000101	0.999154	0.008046
Linear Regression	0.000101	0.999154	0.008042
Gradient Boosting	0.000218	0.998168	0.011851
Random Forest	0.000362	0.996958	0.015257
Lasso Regression	0.000424	0.996431	0.016711
Decision Tree	0.000954	0.991972	0.024807
Support Vector Regressor	0.002092	0.982393	0.034987
K-Nearest Neighbors	0.007928	0.933288	0.072032