Comparative Analysis of Filter, Wrapper, and Embedded Feature Selection Methods to Predict Load Progression in Combined Cycle Power Plant (CCPP) Startup Across Diverse Startup Classes

Mogana Vadhna Suntrakumar
Faculty of Information and Communication Technology (FTMK)
Universiti Teknikal Malaysia Melaka (UTeM) Melaka, Malaysia
moganavadhna@gmail.com

Sharifah Sakinah Binti Syed Ahmad
Faculty of Information and Communication Technology (FTMK)
Universiti Teknikal Malaysia Melaka (UTeM) Melaka, Malaysia
sakinah@utem.edu.my

Mohamad Lutfi Samsudin
Department of Engineering
Malakoff Power Berhad
Kuala Lumpur, Malaysia
mlutfi.samsudin@malakoff.com.my

Muhamad Ridzhuan Othman
Department of Engineering
Malakoff Power Berhad
Kuala Lumpur, Malaysia
ridzhuan.othman@malakoff.com.my

***Abstract*— In the realm of big data, the utilization of advanced analytics techniques, particularly artificial intelligence (AI), holds paramount importance in extracting valuable insights from data generated by combined cycle power plants (CCPPs). This study meticulously compares three sophisticated feature selection methods: Pearson Correlation Coefficient (PCC) filtering, Random Forest-based Recursive Feature Elimination (RF-RFE) wrapper-based selection, and LASSO regularization for embedded selection, aimed at predicting load progression (Generator Watts) during startup using LSTM models. Leveraging a comprehensive dataset consisting of 240 startup processes, each comprising 14,000 observations and categorized into four startup classes (HOT, WARM 1, WARM 2, COLD), this research carefully selects features for each class independently through the aforementioned techniques. Subsequently, distinct LSTM models are constructed for each startup class based on the selected features, and their predictive capabilities are rigorously evaluated using performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The comparison highlights variations in the performance of feature selection techniques across different startup classes. Specifically, RF-RFE demonstrates superior efficacy in the Hot and Warm 1 classes, while LASSO regularization outperforms other methods in the Warm 2 and Cold classes. This study offers valuable insights into the importance of experimentation and empirical validation in selecting the optimal feature selection technique for predictive modeling tasks. While some research suggests that certain methods may outperform others, this study emphasizes the importance of tailoring feature selection approaches to dataset characteristics, guiding practitioners and researchers toward more accurate predictive models and improved real-world outcomes.**

***Keywords*— CCPP, load progression, big data, feature selection, predictive modeling, wrapper, filter, embedded.**

I. INTRODUCTION

In recent years, a profound transformation has been witnessed in the energy sector, driven by technological advancements and the increasing demand for sustainable energy solutions. As part of this evolution, combined cycle power plants (CCPPs) have emerged as a prominent technology for electricity generation due to their high efficiency and relatively low environmental impact. These plants utilize a combination of gas and steam turbines to generate electricity, making them an integral component of modern energy infrastructure. Such a combination of two or more thermodynamic cycles increases efficiency and reduces fuel costs [1]. CCPPs are considerably more efficient than other types of power plants, offering advantages such as quick start capability, lower specific emissions, and high efficiency. Operating them at maximum capacity is recognized to result in a favourable increase in its thermal efficiency [2]. Moreover, they are more conducive in environmental and economic aspects, leading to their adoption in many countries to satisfy the ever-increasing demand for electricity.

CCPPs are characterized by frequent start-up and shutdown cycles, often operating during the day and shutting down at night. Additionally, variations in seasonal demand necessitate proper distribution of heat and electricity in combined thermal power plants, requiring the implementation of various and complex control logic [3].

In parallel, the proliferation of big data and advanced analytics techniques has opened new avenues for optimizing the performance and operation of CCPPs. Leveraging vast amounts of data generated during plant operation provides valuable insights into plant dynamics, opportunities for efficiency improvements, and overall operational reliability. Moreover, the integration of artificial intelligence (AI) into data analytics processes further enhances predictive capabilities, enabling more accurate forecasting of key performance indicators. However, despite the potential benefits offered by big data analytics and AI in CCPP optimization, challenges remain in effectively harnessing the wealth of data available. One such challenge is the selection of relevant features from the dataset, which is critical for building accurate predictive models. With numerous feature selection techniques available, ranging from traditional statistical methods to more advanced machine learning algorithms, determining the most effective approach can be a daunting task.

In this context, this study aims to address this challenge by meticulously comparing three advanced feature selection techniques: filtering using Pearson Correlation Coefficient (PCC), wrapper-based selection with Random Forest based Recursive Feature Elimination (RF-RFE), and embedded selection through LASSO regularization. These techniques are evaluated in the context of predicting generator wattage during startup processes in CCPPs, utilizing Long Short-Term Memory (LSTM) models. By conducting a comprehensive analysis of these techniques, the aim is to identify the most effective approach for feature selection in the context of CCPP optimization. Additionally, the study aims to investigate the factors influencing the progression of generator watts in four startup classes (Hot, Warm 1, Warm 2, and Cold) and determine whether the same factors are selected for all classes or if a distinct set of factors emerges.

Furthermore, this study extends beyond technical considerations to explore the broader implications of its findings. By shedding light on the dynamics governing generator wattage progression during startup sequences in CCPPs, valuable insights are provided for optimizing startup durations, enhancing operational efficiency, and reducing downtimes. Ultimately, this research contributes to the ongoing discourse surrounding feature selection methodologies within the realm of CCPP optimization, paving the way for future advancements in this critical area.

II. FEATURE SELECTION

Feature selection is a crucial step in machine learning and data analysis, where the goal is to choose a subset of relevant features (or variables) from the original set of features without transforming the data. Selecting the right features is vital for improving machine learning models. It helps mitigate issues caused by large datasets, such as increased complexity and computational costs. Overfitting, where the model learns noise instead of patterns, is also addressed through feature selection [4][5]. By focusing on the most relevant features, data dimensionality is reduced, leading to better model accuracy and effectiveness. In the context of time series prediction, feature selection becomes particularly important as future values are forecasted based on past observations and historical data. This process helps identify the most important time periods to consider, improving the accuracy of predictions by capturing how data changes over time. [6].

A. Feature Selection Techniques

Choosing the right feature selection technique is paramount for optimizing machine learning models. These techniques can be broadly categorized into three groups: filter, wrapper, and embedded methods. The subsequent sections delve into each method, providing detailed explanations of their methodologies and practical applications, thereby establishing a cohesive understanding of their significance in model optimization.

- **Filter**

Filter methods, a prominent category in feature selection, operate by independently evaluating the relevance of individual features based on certain criteria, without considering the interaction between features or the underlying learning algorithm. These methods utilize various measures based on information theory, correlation, distance, consistency, fuzzy-set, and rough-set to evaluate features [7]. Initially, features are selected and ranked independently in the univariate case and collectively in the multivariate case to handle redundancies effectively [7]. Subsequently, a performance criterion is applied to select features with the highest ranking. Filter methods in feature selection offer several advantages, including their simplicity and efficiency in evaluating features based on data characteristics. These methods are computationally less demanding compared to wrapper methods since they do not involve training the model iteratively. Moreover, filter methods are generally applicable across various types of datasets and can effectively handle high-dimensional data [8]. However, filter methods have limitations, notably in their inability to consider feature dependencies and interactions. They tend to overlook the intrinsic relationships between features, potentially leading to suboptimal feature subsets [9]. Additionally, filter methods may struggle with feature selection tasks where the predictive performance of the model heavily depends on complex feature interactions.

- **Wrapper**

Wrapper methods, another prominent category in feature selection, take a different approach by assessing feature subsets based on the performance of a specific learning algorithm. Unlike filter methods, wrapper methods evaluate feature subsets by incorporating the learning algorithm's performance as the criterion for selection [10]. This iterative process involves training the model on different feature subsets to determine the optimal set that maximizes the model's performance [11]. One advantage of wrapper methods is their ability to consider feature dependencies and interactions, which filter methods often overlook. By evaluating feature subsets based on the learning algorithm's performance, wrapper methods can identify combinations of features that collectively contribute to better model performance. Additionally, wrapper methods are more suitable for feature selection tasks where the predictive performance of the model relies heavily on complex feature interactions. However, wrapper methods come with increased computational complexity compared to filter methods [12]. This is because wrapper methods involve training the learning algorithm iteratively on different feature subsets, which can be resource-intensive, especially for large datasets with numerous features. Moreover, wrapper methods may be prone to overfitting [13], as they optimize feature subsets based on the specific learning algorithm used, potentially resulting in reduced generalization performance on unseen data. Examples of wrapper methods encompass Forward Feature Selection (FFS), Backward Feature Elimination (BFE), Recursive Feature Elimination (RFE), and Bidirectional Search, etc. [14].

- **Embedded**

Embedded methods, the third category in feature selection techniques, integrate the feature selection process directly into the model building process. Unlike filter and wrapper methods, embedded methods treat feature selection as a subproblem of the main model-building process, optimizing the feature subset while constructing the learning algorithm. Regularization techniques, such as LASSO (Least Absolute Shrinkage and Selection Operator), Elastic Net, and Ridge Regression, are key aspects of embedded methods, functioning to simultaneously construct the learning model and automatically select features [13, 14]. During the implementation of the mining algorithm, embedded feature selection methods select features, considering their relevance to the learning algorithm's objective, while discarding noisy or redundant ones. By integrating the feature selection process into the model building, embedded methods effectively account for the dependencies between features, enhancing computational efficiency by avoiding repeated executions of the classifier and the examination of every feature subset. Embedded methods are adept at handling data with complex structures such as nonlinearity, further enhancing their suitability for various types of datasets [15]. Additionally, they tend to have lower computational costs compared to wrapper methods, as they avoid the need for iterative training on different feature subsets. However, one limitation of embedded methods is their close association with the learning algorithm used. Consequently, they may not perform optimally with other algorithms, limiting their versatility across different modelling scenarios.

III. METHODOLOGY

A. Data Understanding

The startup load progression data utilized in this study was obtained from a CCPP (Combined Cycle Power Plant) situated in Malaysia. The dataset comprises records of load progression during startup processes, captured at regular intervals. It encompasses records categorized by startup class (HOT, WARM1, WARM2, COLD) and spans the years 2017 to 2021. The details of the dataset are further illustrated in Table I, which outlines the distribution of data across different startup classes and years. With a total of 240 Excel files, each containing 14,401 rows and 92 columns, the dataset presents a diverse range of information. The variables (columns) are categorized into distinct groups, including load (Generator Watts) turbine speed, control and feedback, pressure, temperature, steam flow, loading rate, and steam conductivity. These variables collectively represent the factors influencing the load progression captured in the dataset. Table II provides an overview of the number of variables present in each category. Data types within the dataset include integers, floats, datetime values, and Boolean indicators, reflecting the various parameters and attributes captured. This comprehensive dataset overview lays the foundation for subsequent analysis and exploration.

TABLE I. DISTRIBUTION OF DATA ACROSS DIFFERENT STARTUP CLASSES AND YEARS.

Year	**2017**	**2018**	**2019**	**2020**	**2021**	**TOTAL**
HOT	7	27	25	4	0	**63**
WARM1	4	13	11	8	2	**38**
WARM2	7	23	20	14	8	**72**
COLD	2	11	8	28	18	**67**
TOTAL	**20**	**74**	**64**	**54**	**28**	**240**

TABLE II. NUMBER OF VARIABLES PRESENT IN EACH CATEGORY.

Variable Category	**Count of Variables**
GENERATOR WATTS (LOAD)	3
TURBINE SPEED	2
CONTROL AND FEEDBACK	39
PRESSURE	8
TEMPERATURE	28
STEAM FLOW	2
LOADING RATE	3
STEAM CONDUCTIVITY	3

B. Data Preprocessing

Data preprocessing begins with the extraction of relevant data from the original Excel files obtained from the above-mentioned data source. Unnecessary rows are removed from the data to streamline the dataset. Columns are then renamed to ensure clarity and consistency, facilitated by defining a mapping of old column names to new ones. Missing values are identified and handled appropriately, with imputation techniques such as filling with 0 values or applying linear interpolation where suitable. Outliers are identified using Z-Score and retained to preserve valuable insights into the industrial process. Feature engineering steps include adding new columns like "Reference" and "Class" to provide additional context to the data and removing unnecessary columns like constant and duplicated ones, resulting in 67 variables. Finally, the individual Data Frames from each Excel file are merged into a unified dataset, resulting in a comprehensive CSV file named 'COMBINED_DATA.csv', containing 3,312,227 rows, ready for further analysis. Fig. 1. illustrates the preprocessing steps in a structured and organized form.

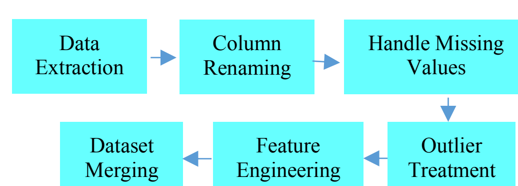

Fig. 1. Preprocessing steps

C. Data Selection for Analysis

The filtration process for data selection was critical in honing the analysis to focus on the most pertinent startup processes. By meticulously selecting the top 10 startup processes with the least total startup time (LST) and the top 10 with the largest total startup time (LGT) from each startup class (HOT, WARM1, WARM2, COLD), a total of 80 startup process files were carefully curated. This methodological approach was instrumental in pinpointing startup durations across various classes, shedding light on the dynamics of load progression. The identification of the 10 files with the shortest total startup times underscores the efficiency of certain startup procedures, while recognizing the 10 files with the longest total startup times highlights potential complexities within specific classes. The subsequent consolidation of selected startup processes into two datasets, COMBINED_LST.csv and COMBINED_LGT.csv, further refined the analysis, providing a focused lens to examine influential variables affecting load progression. By analysing the two datasets separately for each class, we can delve into the variables (parameters) that have a high influence on both favourable and less favourable scenarios, offering valuable insights for predictive modelling of load progression.

D. Feature Selection

In this section, the implementation of three advanced feature selection techniques—filtering using Pearson Correlation Coefficient (PCC), wrapper-based selection with Random Forest based Recursive Feature Elimination (RF-RFE), and embedded selection through LASSO regularization—is undertaken to identify the most relevant features contributing to the predictive performance of the model. These techniques are deemed crucial in machine learning model development, as they focus on selecting the optimal set of features to enhance predictive accuracy. The goal is to predict load progression (Generator Watts) during startup using LSTM models. Each technique's application will be explained in detail, highlighting their effectiveness in selecting influential features for the predictive task. By selecting these features, insights into the factors that impact the load progression of CCPP power plants in different startup classes can also be gained.

- **Feature Selection Algorithms**

1) PCC

The Pearson Correlation Coefficient (PCC) formula (1) provides a statistical measure of the linear relationship between two continuous variables [16]. It is calculated by dividing the covariance of the variables by the product of their standard deviations [17]. Essentially, the formula assesses how much two variables change together, relative to their respective means. A positive coefficient indicates a positive linear relationship, meaning that as one variable increases, the other tends to increase as well, while a negative coefficient signifies an inverse relationship, where one variable increases as the other decreases. The magnitude of the coefficient indicates the strength of the relationship, with values closer to +1 or -1 representing stronger correlations, while values closer to 0 suggest weaker or no linear relationship [18]. Overall, the formula offers insights into the degree and direction of association between two variables, aiding in the analysis of their interdependence. In feature selection, highly correlated features may introduce redundancy into the dataset, impacting model accuracy. Therefore, identifying and eliminating highly correlated features can enhance model performance by reducing overfitting and computational complexity. In this project, the selection criteria for using PCC involved assessing the correlation of each variable with the target variable, Generator Watts. A threshold of 0.7 was set to select the variables, indicating that only variables with a PCC magnitude greater than or equal to + 0.7/-0.7 were considered relevant for inclusion in the analysis. This threshold value ensured that only highly correlated features, demonstrating a strong linear relationship with the target variable, were retained for further investigation.

$$\rho(x, y) = \frac{cov(x,y)}{\sigma(x)\times\sigma(y)} \tag{1}$$

2) RF-RFE

Recursive Feature Elimination (RFE) is a wrapper-based feature selection [21] technique widely employed in machine learning to identify the most significant features contributing to the predictive performance of a model. Unlike filter methods such as Pearson Correlation Coefficient (PCC), which assess the correlation between individual features and the target variable independently, RFE evaluates the collective importance of features by recursively removing the least significant ones from the dataset. The process begins by training the model on the entire feature set and subsequently ranking the features based on their importance scores. Then, the least important feature(s) are eliminated, and the model is retrained on the reduced feature set [22]. This iterative process continues until the desired number of features is reached or until the model's performance stabilizes. RFE leverages the notion that removing irrelevant features can enhance model generalization by reducing noise and overfitting. Additionally, RFE is particularly useful when dealing with high-dimensional datasets where the number of features exceeds the number of samples, as it helps identify a subset of features that maximizes predictive performance while minimizing computational complexity. By iteratively assessing feature importance and pruning the least informative ones within the wrapper of model training and evaluation, RFE facilitates the creation of simpler and more interpretable models without sacrificing predictive accuracy. In this project, Recursive Feature Elimination (RFE) is employed with a Random Forest Regressor model for feature selection, leveraging Random Forest, a popular ensemble learning technique in machine learning that utilizes multiple decision trees to make predictions [23], where each tree is trained on a random subset of the data and features. RFE iteratively ranks and eliminates the least important features based on their impact on the model's predictive performance. By leveraging Random Forest's ability to assess feature importance, RFE identifies and retains the most relevant features for predicting load progression during startup.

3) LASSO

LASSO, which stands for Least Absolute Shrinkage and Selection Operator, is a feature selection technique, proposed by Tibshirani [19] commonly used in machine learning and regression analysis. It incorporates regularization into the model fitting process to prevent overfitting and to promote sparsity in the resulting feature set. In LASSO, the objective function includes a penalty term that penalizes the absolute magnitude of the coefficients of the features, in addition to the usual loss function. This penalty encourages smaller coefficient values and can drive some coefficients to exactly zero, effectively removing corresponding features from the model. If a group of features exhibits high correlation, LASSO selects only one from the group and shrinks the rest to zero [20]. The formula for the LASSO objective function is given by formula (2). Here n is the number of observations, p is the number of features, yi is the target variable for observation i, xij is the value of feature j for observation i, $\beta 0$ is the intercept term, βj is the coefficient for feature j, and λ is the regularization parameter. The regularization parameter (λ) controls the strength of the regularization penalty. A larger value of λ results in greater regularization, leading to more coefficients being shrunk towards zero and more features being eliminated from the model. Conversely, a smaller value of λ reduces the effect of regularization, allowing more features to contribute to the model. One of the key advantages of LASSO is its ability to perform feature selection and regularization simultaneously. By shrinking certain coefficients to zero, LASSO effectively selects the most relevant features while discarding the less important ones, thus producing a sparse model that is easier to interpret and less prone to overfitting.

$$\min\beta\left\{\frac{1}{2n}\sum_{i=1}^{n}(yi-\beta 0-\sum_{j=1}^{p}\beta jxij\,)^2\;+\lambda\sum_{j=1}^{p}\mid\beta j\mid\right\} \tag{2}$$

- **Feature Selection Criteria**

The selection criteria for identifying influential features involved the application of all three feature selection techniques—PCC, RF-RFE, and LASSO—on both the COMBINED_LST and COMBINED_LGT datasets separately for each startup class (HOT, WARM 1, WARM 2, COLD). By independently selecting features for each class, unique sets of influential features were obtained. The most influential variables were determined as those common to both LST and LGT scenarios, demonstrating high influence across different startup conditions.

E. Prediction Model

The prediction model employed in this project utilizes Long Short-Term Memory (LSTM) neural networks, a type of recurrent neural network (RNN) known for its ability to learn and remember long sequences of data. For each of the four startup classes (HOT, WARM1, WARM2, COLD), an LSTM model is trained using distinct sets of influential variables selected by the three feature selection techniques employed, namely PCC, RF-RFE, and LASSO, resulting in a total of 12 models. These models are trained and evaluated individually on sequences of influential variables specific to each startup class, allowing for a comprehensive analysis of the impact of feature selection on predictive performance across different operational scenarios in the CCPP power plant. The LSTM architecture for each model comprises three layers, each containing 50 units, with dropout regularization layers incorporated to prevent overfitting. Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are computed to assess the performance of each model, providing valuable insights into the load progression during startup processes in the CCPP power plant.

F. Evaluation Materics

The evaluation of the prediction model involves several key metrics to assess its performance.

1) Mean Squared Error (MSE)

Mean Squared Error (MSE) quantifies the average squared difference between predicted and actual values. It provides a measure of the overall accuracy of the model's predictions, with lower values indicating better performance in minimizing prediction errors. Formula (3) presents the calculation for MSE. In the equation, y_i represents the actual value, $\hat{y}_i$ denotes the predicted value, and n is the total number of samples.

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{3}$$

2) Mean Absolute Error (MAE)

Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values. It offers a straightforward assessment of prediction errors, providing insight into the magnitude of discrepancies between predicted and observed values. Formula (4) presents the calculation for MAE. In the equation, $y_i$represents the actual value, $\hat{y}_i$ denotes the predicted value, and n is the total number of samples.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{4}$$

3) Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is the square root of the average squared differences between predicted and actual values. It offers an interpretable measure of prediction error in the same units as the target variable, providing insight into the magnitude of errors made by the model in predicting the target variable. Formula (5) presents the calculation for RMSE. In the equation, y_i represents the actual value, $\hat{y}_i$denotes the predicted value, and n is the total number of samples.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{5}$$

IV. RESULT AND DISCUSSION

This section presents the outcomes of applying various feature selection techniques to predict load progression across different startup classes (HOT, WARM1, WARM2, COLD) in Combined Cycle Power Plants (CCPPs). The selected feature categories encompass a range of variables, including pressure, temperature, control and feedback parameters, and steam flow. Each feature selection technique, namely PCC, RF-RFE, and LASSO, is evaluated based on performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Table III and Figures (2-4) illustrate the results in detail, providing a comprehensive overview of the performance of each technique across all classes. The following sections will delve into the results for each class, analyzing the effectiveness of the feature selection methods in capturing the nuances of load progression dynamics during startup operations in CCPPs.

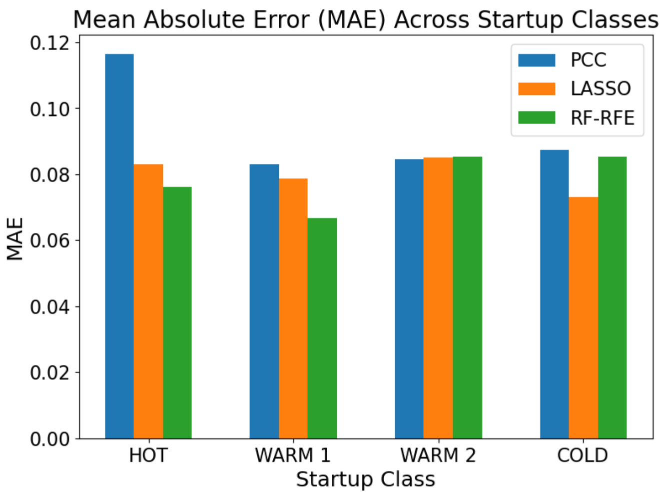

Fig. 2. Mean Absolute Error (MAE) comparison across startup classes.

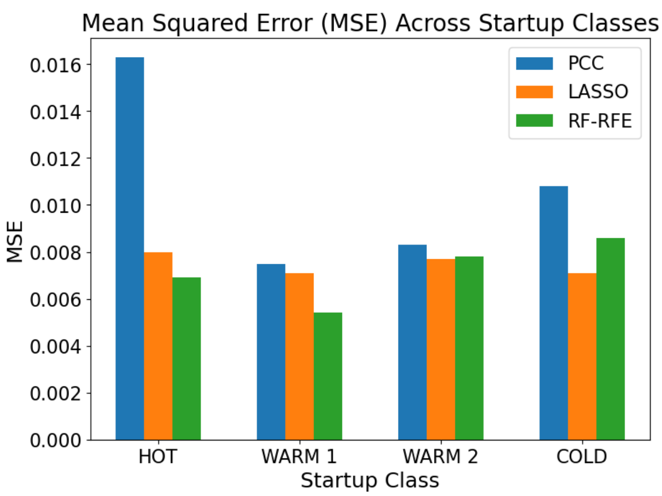

Fig. 3. Mean Squared Error (MSE) comparison across startup classes.

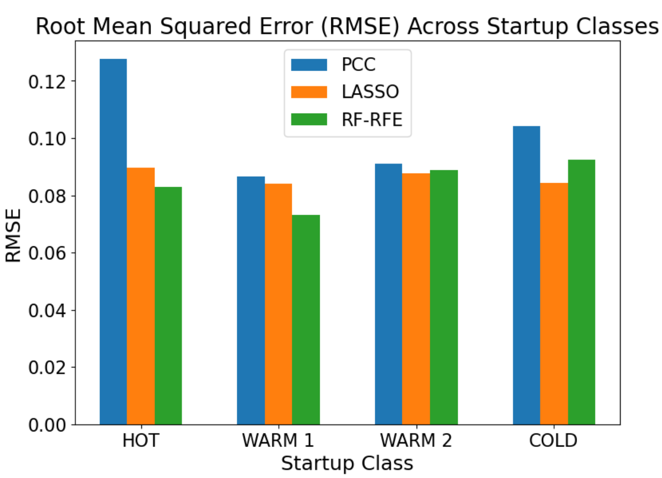

Fig. 4. Root Mean Squared Error (RMSE) comparison across startup classes.

- **Hot Start-up Class**

The results for the Hot startup class, obtained using different feature selection techniques, reveal significant variations in the selected variable categories and the performance metrics across methods. For the PCC method, the selected variable categories include pressure, temperature, control & feedback, and steam flow. The model trained on these features achieved a Mean Absolute Error (MAE) of 0.1164, a Mean Squared Error (MSE) of 0.0163, and a Root Mean Squared Error (RMSE) of 0.1277. These metrics provide insights into the predictive accuracy and the model's ability to capture the dynamics of the startup process when utilizing the PCC-selected features. In contrast, additional features, including turbine speed, were identified by the RF-RFE technique, alongside pressure, temperature, control & feedback, and steam flow. Improved performance metrics were observed with a lower MAE of 0.0760, a decreased MSE of 0.008, and a reduced RMSE of 0.0829. The inclusion of turbine speed likely contributed to the enhanced predictive capabilities of the model. Similarly, pressure, temperature, control & feedback, and turbine speed were identified as influential variables for predicting the load progression during startup in the Hot class using the LASSO feature selection method. Despite a slightly higher MAE of 0.0830 compared to RF-RFE, the LASSO-selected features achieved competitive performance metrics, with an MSE of 0.008 and an RMSE of 0.0896. The consistency in performance across different feature selection techniques underscores the robustness of the identified variable categories in capturing the underlying dynamics of the startup process.

- **Warm 1 Start-up Class**

The PCC method identified essential variable categories, comprising Pressure, Temperature, Control & Feedback, and Steam flow. The resulting performance metrics revealed a mean absolute error (MAE) of 0.0831, a mean squared error (MSE) of 0.0075, and a root mean squared error (RMSE) of 0.0865. When comparing the RF-RFE and LASSO regularization techniques in the context of the WARM 1 startup class, both methods demonstrated notable improvements over the baseline PCC approach. RF-RFE, by extending the feature set to include Turbine Speed alongside the categories identified by PCC, exhibited superior predictive accuracy. With an MAE of 0.0668, MSE of 0.0054, and RMSE of 0.0732, RF-RFE showcased enhanced model performance compared to PCC alone. On the other hand, LASSO regularization also selected Turbine Speed as an influential variable category, along with Pressure, Temperature, Control & Feedback, and Steam Flow. While LASSO yielded slightly different performance metrics, including an MAE of 0.0786, MSE of 0.0071, and RMSE of 0.0842, the overall predictive accuracy was competitive with RF-RFE. However, it's worth noting that despite the similarities in performance, RF-RFE exhibited a marginally superior fit to the data compared to LASSO. This nuanced difference suggests that while both techniques are effective in enhancing predictive accuracy by incorporating additional variable categories, RF-RFE may offer a slight advantage in capturing the underlying dynamics of the startup process in the WARM 1 class.

TABLE III. FEATURE SELECTION PERFORMANCE SUMMARY FOR START-UP CLASSES (PCC, RF- RFE, LASSO).

Start-up class	Feature Selection Technique	Selected Variable Categories	MAE	MSE	RMSE
HOT	PCC	Pressure, Temperature, Control & Feedback, Steam flow	0.1164	0.0163	0.1277
	RF-RFE	Pressure, Temperature, Control & Feedback, Steam Flow, Turbine Speed	0.0760	0.008	0.0829
	LASSO	Pressure, Temperature, Control & Feedback, Turbine Speed	0.0830	0.008	0.0896
WARM 1	PCC	Pressure, Temperature, Control & Feedback, Steam flow	0.0831	0.0075	0.0865
	RF-RFE	Pressure, Temperature, Control & Feedback, Steam Flow, Turbine Speed	0.0668	0.0054	0.0732
	LASSO	Pressure, Temperature, Control & Feedback, Steam Flow, Turbine Speed	0.0786	0.0071	0.0842
WARM 2	PCC	Pressure, Temperature, Control & Feedback, Steam Flow	0.0845	0.0083	0.0912
	RF-RFE	Pressure, Temperature, Control & Feedback, Steam Flow, Turbine Speed.	0.0854	0.0078	0.0888
	LASSO	Pressure, Temperature, Control & Feedback, Steam Flow, Turbine Speed.	0.0850	0.0078	0.0877
COLD	PCC	Pressure, Temperature, Steam Flow	0.0873	0.0108	0.1041
	RF-RFE	Pressure, Temperature, Control & Feedback, Steam Flow, Turbine Speed.	0.0854	0.0086	0.0925
	LASSO	Pressure, Temperature, Control & Feedback, Steam Flow, Turbine Speed.	0.0731	0.0071	0.0843

- **Warm 2 Start-up Class**

For the WARM 2 class, when utilizing the PCC method, the selected variable categories consisted of Pressure, Temperature, Control & Feedback, and Steam Flow. The resulting model performance metrics included a mean absolute error (MAE) of 0.0845, a mean squared error (MSE) of 0.0083, and a root mean squared error (RMSE) of 0.0912. Similarly, RF-RFE identified the same variable categories as PCC, with the addition of Turbine Speed. Despite this expanded feature set, the model's performance remained comparable to the PCC selected features, with an MAE of 0.0854, an MSE of 0.0078, and an RMSE of 0.0888. In comparing LASSO and RF-RFE for the WARM 2 class, both RF-RFE and LASSO identified the same variable categories: Pressure, Temperature, Control & Feedback, Steam Flow, and Turbine Speed. Despite this similarity, there were nuanced differences in the model performance metrics obtained through each technique. LASSO yielded a MAE of 0.0850, MSE of 0.0078, and RMSE of 0.0877. While the variations in performance metrics between RF-RFE and LASSO are subtle, they signify a nuanced difference in predictive accuracy. LASSO achieved slightly better results in terms of RMSE, indicating potentially improved fit to the data.

- **Cold Start-up Class**

For the COLD startup class, each feature selection technique revealed unique sets of influential variable categories. When employing the PCC, the selected variables included Pressure, Temperature, and Steam Flow, resulting in an MAE of 0.0873, MSE of 0.0108, and RMSE of 0.1041. In contrast, RF-RFE identified additional variables, such as Control & Feedback and Turbine Speed, alongside the PCC-selected ones. The RF-RFE based model exhibited slightly improved performance with an MAE of 0.0854, MSE of 0.0086, and RMSE of 0.0925. Similarly, LASSO regularization also expanded the feature set to include Control & Feedback and Turbine Speed, resulting in the lowest MAE of 0.0731, MSE of 0.0071, and RMSE of 0.0843 among the three techniques. This comprehensive comparison highlights the effectiveness of RF-RFE and LASSO in capturing additional variability compared to PCC. Additionally, while both RF-RFE and LASSO selected similar sets of influential variables, LASSO exhibited slightly better predictive accuracy and model performance compared to RF-RFE, suggesting its superiority in capturing the underlying dynamics of the COLD startup process.

A. *Comparative Performance of Feature Selection Techniques Across Startup Classes*

The analysis of model performance across various startup classes highlights significant differences in the effectiveness of feature selection techniques. Notably, RF-RFE demonstrated superior performance in the Hot and Warm 1 classes, while LASSO regularization excelled in the Warm 2 and Cold classes. In the Hot and Warm 1 classes, RF-RFE consistently identified a subset of influential features, resulting in models with lower Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) compared to other feature selection techniques. This indicates its effectiveness in capturing the underlying dynamics of the startup processes. Conversely, in the Warm 2 and Cold classes, LASSO regularization emerged as the top-performing feature selection method, identifying relevant features and mitigating overfitting. Notably, both RF-RFE and LASSO selected turbine speed across all startup classes, indicating its crucial role in predicting load progression despite its indirect relationship with generator watts. In contrast, the PCC method did not select turbine speed across all classes, highlighting its limitation in capturing complex interactions and non-linear relationships among variables. While PCC prioritizes features based on their correlations with the target variable, it may overlook variables like turbine speed, which contribute to predictive accuracy through non-linear interactions with other features. This underscores the importance of considering complementary feature selection techniques like RF-RFE and LASSO, which can capture such nuanced relationships and improve the effectiveness of predictive models in capturing the underlying dynamics of startup processes.

B. *Influence of Selected Influential Variables on Performance*

The selection of influential variables significantly impacts the predictive accuracy and performance of the LSTM model in predicting the load progression of the Combined Cycle Power Plant (CCPP) during startup. Initially, a comprehensive set of variables, including Turbine Speed, Control and Feedback, Pressure, Temperature, Steam Flow, Loading Rate, and Steam Conductivity, were considered for feature selection. However, through various feature selection techniques such as PCC, RF-RFE, and LASSO regularization, a refined subset of variables emerged as the most influential for predicting load progression during startup processes. Among the originally considered variables, Turbine Speed, Control and Feedback, Pressure, Temperature, and Steam Flow consistently emerged as the most influential features across different feature selection techniques. Notably, these variables exhibit better predictive performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) compared to other variables. Therefore, including these variables enables more accurate predictions of load progression and facilitates informed decision-making in power plant operations.

V. CONCLUSION

In this study, the application of advanced machine learning techniques was utilized to predict load progression during startup processes in CCPPs, employing Long Short-Term Memory (LSTM) neural networks and three feature selection techniques: Pearson Correlation Coefficient (PCC) filtering, Random Forest-based Recursive Feature Elimination (RF-RFE) wrapper-based selection, and LASSO regularization for embedded selection. The critical role of feature selection in enhancing predictive accuracy and model performance across different startup classes was revealed. The comparison of feature selection methods showcased variations in their efficacy, with wrapper-based RF-RFE demonstrating superior performance in capturing the dynamics of startup processes in the Hot and Warm 1 classes, while embedded methods (LASSO regularization) outperformed others in the Warm 2 and Cold classes. These findings underscore the importance of tailoring feature selection approaches to dataset characteristics for optimal predictive modeling outcomes. Furthermore, our study identified influential variable categories such as Turbine Speed, Control and Feedback, Pressure, Temperature, and Steam Flow in predicting load progression during startup. Incorporating these variables into predictive models improved accuracy and performance, providing valuable insights for decision-making in power plant operations. However, it is important to acknowledge certain limitations, such as the lack of additional contextual factors in the dataset, which may limit the generalizability of our findings. Future research could explore the integration of additional variables and feature selection techniques to refine predictive models further. Additionally, the application of ensemble learning methods and hybrid approaches may offer avenues for improving model robustness and generalization capabilities in real-world scenarios. In conclusion, this study contributes to the understanding of feature selection techniques and their impact on predictive modeling in CCPP startup processes. By leveraging advanced machine learning methodologies and empirical validation, the importance of data-driven approaches in enhancing predictive accuracy and facilitating informed decision-making in complex industrial settings is emphasized.

ACKNOWLEDGMENT

This work was supported by Malakoff Energy Sdn Bhd under the Industrial Grant Vote No: INDUSTRI/MALAKOFF/FTMK/2023/I00077. The authors express their appreciation for this support.

REFERENCES

[1] R. A. Joy, "An Interpretable Catboost Model to Predict the Power of Combined Cycle Power Plants," in International Conference on Information Technology (ICIT), Amman, Jordan, 2021.

[2] Ahmad SS, Pedrycz W. Data and feature reduction in fuzzy modeling through particle swarm optimization. Applied Computational Intelligence and Soft Computing. 2012.

[3] P. Kaewprapha, P. Prempaneerach, V. Singh, T. Tinikul and N. Intarangsi, "Machine Learning Approaches for Estimating the Efficiency of Combined Cycle Power Plant," in International Electrical Engineering Congress (iEECON2022), Khon Kaen, THAILAND, 2022.

[4] A. Kole, "A Review and Study on Advanced Control and Automation Functions and Future Control for a Modern Combined Cycle Power Plant," in nternational Conference on Intelligent Control Power and Instrumentation (ICICPI), Kolkata, India, 2016.

[5] A. M. yesaswini and d. K.Annapurna, " A Hybrid Approach for Intrusion Detection System to Enhance Feature," in Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS 2023), Trichy, India, 2023.

[6] F. Marzbani, A. Osman and M. S. Hassan, "A Hybrid Multi-model Ensemble Feature Selection and SVR Prediction Approach for Accurate Electric Vehicle Demand Prediction: A US Case Study," in 2023 IEEE International Conference on Environment and Electrical Engineering and 2023 IEEE Industrial and Commercial Power Systems Europe (EEEIC), Madrid, Spain, 2023.

[7] M. Cherrington, T. Fadi, L. Joan and X. Qiang, "Feature Selection: Filter Methods Performance Challenges," in 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019.

[8] Y. Bouchlagem, Y. Akhiat and S. Amjad, "Feature Selection: A Review and Comparative Study," in ICIES'22, Morocco, 2022.

[9] T. S. Binita Kumari, " Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review," International Journal of Computer Science and Information Technologies, vol. 2(3), pp. 1048-1053, 2011.

[10] P. H. Prastyo, I. Ardiyanto and R. Hidayat, "A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods," in 2020 6th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2020.

[11] D. Vidanagama, T. Silva and A. Karunananda, "Hybrid Filter-Wrapper Approach for Feature Selection in Deceptive Consumer Review Classification," in 2021 5th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI), Sri Lanka, 2021.

[12] S. E. M. M. Johannes Grohmann, " Detecting Parametric Dependencies for Performance Models Using Feature Selection Techniques," in 2019

IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication, Rennes, France, 2019.

[13] B. K. Abhigya Mahajan, "A Review of Machine Learning Algorithms and Feature Selection Techniques for Cardiovascular Disease Prediction: Insights and Implications," in 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2023.

[14] X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang and H.-S. Ye, "A Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection," IEEE ACCESS, vol. 6, pp. 22863 - 22874, 2018.

[15] Y. T. Naing, M. Raheem and N. K. Batcha, "Feature Selection for Customer Churn Prediction: A Review on the Methods & Techniques applied in the Telecom Industry," in 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 2022.

[16] P. Saha, S. Patikar and S. Neogy, "A Correlation - Sequential Forward Selection Based Feature Selection Method for Healthcare Data Analysis," in IEEE International Conference on Computing, Power and Communication Technologies (GUCON), New Delhi, India, 2020.

[17] Y. Zheng, Y. Li, G. Wang, Y. Chen, Q. Xu, J. Fan and X. Cui, "A Novel Hybrid Algorithm for Feature Selection Based on Whale Optimization Algorithm," IEEE Access, vol. 7, pp. 14908 - 14923, 2018.

[18] B. W. Maulana, S. B. Utomo, K. Anam and I. Bilfaqih, " Analysis of Movement Patterns Based on Electroencephalograph Signals Using Pearson Correlation Coefficient," in 2023 International Seminar on Intelligent Technology and Its Applications (ISITIA), Surabaya, Indonesia, 2023.

[19] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, pp. 267-288, 1996.

[20] M. Hamada, J. J. Tanimu, M. Hassan, H. A. Kakudi and P. Robert, "Evaluation of Recursive Feature Elimination and LASSO Regularization-based optimized feature selection approaches for cervical cancer prediction," in 2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), Singapore, 2021.

[21] M. G. Ismail, M. A. E. Ghany and M. A.-M. Salem, "Enhanced Recursive Feature Elimination for IoT Intrusion Detection Systems," in 2022 International Conference on Microelectronics (ICM), Casablanca, Morocco, 2022.

[22] R.K.Jeyauthmigha and R.C.Suganthe, "Recursive Feature Elimination and Clustering Technique for Network Anomaly Detection," in IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India, 2018.

[23] K. Sudhakar, S. M. R. Mallireddy, P. Tallapureddy, S. Vuddanti, B. Katragadda and S. R. Manda, "Water Level Prediction using Random Forest Algorithm," in 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA), oimbatore, India, 2023.