Optimizing the Analysis of Energy Plants and High-Power Applications Utilizing the Energy Guard Ensemble Selector (EGES)

Jieqiong Zhang
School of Rolling Stock, Zhengzhou Railway Vocational and Technical College, Zhengzhou 450018, China
E-mail: shaoguang@henetc.edu.cn

Keywords: dynamic model selection, energy plants, energy guard ensemble selector, high-power electrical systems, predictive accuracy

Received: October 3, 2024

Accurate performance assessment of energy plants and high-power electrical systems is challenging due to the dynamic nature of parameters like energy output, voltage levels, and load factors. This study introduces the Energy Guard Ensemble Selector (EGES), a machine learning-based algorithm designed to enhance predictive accuracy and reliability in power electronics. EGES employs a dynamic model selection approach, leveraging classifiers such as Random Forest, Support Vector Machine, Gradient Boosting Machine, K-Nearest Neighbors, and Logistic Regression. By using KNN to evaluate real-time electrical conditions, EGES dynamically selects the most suitable model to predict key metrics such as energy output (MW), efficiency (%), fault rates, and transformer capacity (MVA). Experimental results show that EGES outperforms individual models with an accuracy of 93.5%, precision of 91.5%, recall of 92.7%, and an F1-score of 92.1%, demonstrating its robustness in handling fluctuations in electrical parameters. EGES proves to be a reliable tool for improving predictive accuracy and functional dependability in high-power electrical systems.

Povzetek: Razvit je nov dinamični algoritem za optimizacijo analize elektrarn z imenom Energy Guard Ensemble Selector (EGES). Z napredno izbiro modelov izboljšuje napovedi, kar povečuje zanesljivost energetskih sistemov.

1 Introduction

In the area of power electronics and energy systems, precise assessment of energy plants and high-power uses is critical for maintaining functional effectiveness, dependability, and sustainability [1]. These assessments not only tackle technical parameters such as energy output, effectiveness, failure rates, and transformer ratings, but also reflect the developing interplay between electronics, artificial intelligence, and the information society, governed by the fundamental rules of the information society [2]. Conventional performance assessment methods depend on static machine learning models, which frequently fail to adapt to the dynamic nature of electrical systems [3]. This constraint can lead to incorrect forecasts and ineffective choices, particularly in situations where electrical parameters such as voltage levels, power losses, and load factors change frequently [4].

Previous research in this area has primarily concentrated on static machine-learning models that forecast the efficiency of electrical systems using historical data. RF, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM) are commonly utilized methods for energy output prediction, fault detection, and effectiveness optimization. Despite their widespread use, these models frequently treat data as uniformly distributed and fail to account for the inherent variability in transformer capacity (MVA), voltage levels (kV), and cooling techniques, that are critical for precise performance evaluation. Consequently, these models frequently struggle with inaccurate predictions in high-power uses [5].

The main disadvantage of these existing works is their incapacity to capture local fluctuations in electrical parameters, which have a major impact on plant efficiency. Static models are inflexible in handling the variability of transformer ratings, cooling systems, and load factors, which frequently results in inadequate predictions. Furthermore, numerous previous techniques use a one-size-fits-all methodology, rendering them unsuitable for real-time assessments, where precise, situation-specific predictions are necessary for effective functions in high-power electrical systems [6].

To tackle these drawbacks, the Energy Guard Ensemble Selector (EGES) has been implemented. EGES is a new machine learning algorithm developed especially for energy plants and high-power electrical uses. This approach uses a dynamic model selection procedure from a set of classifiers—RF, SVM, GBM, KNN, and LR—to enhance prediction accuracy in key performance metrics like energy output (MW), efficiency (%), and fault rate (%). Unlike conventional static models, EGES dynamically chooses the most suitable models using each test sample's local features, as defined by the KNN assessment. This technique allows EGES to adapt to the variability in key electrical parameters like voltage levels,

transformer ratings, and load factors, leading to more dependable and precise predictions.

The EGES methodology is distinguished by its capability to dynamically adapt to the local characteristics of test data, rendering it ideal for the unpredictability of energy plants and high-power electrical systems. The algorithm uses a dynamic ensemble selection procedure to improve model selection using real-time circumstances. In situations where electrical parameters such as voltage levels or power losses fluctuate significantly, EGES chooses models that have historically performed effectively under comparable circumstances. The predictions from the chosen models are then integrated utilizing a majority voting method, guaranteeing that the final efficiency assessment is precise and resilient.

The contributions of this work comprise the introduction of EGES as a resilient machine-learning mechanism to enhance prediction accuracy in energy plants and high-power electrical applications. EGES dynamically adapts to local fluctuations in electrical parameters, resulting in substantially better forecasting abilities than conventional static models. Comprehensive experiments show that EGES surpasses previous models in terms of predictive accuracy, precision, recall, F1-score, and MCC, especially in important fields such as fault identification, effectiveness improvement, and operational cost decrease. This study aims to present a dynamic and flexible machine-learning framework that surpasses the restrictions of static models in forecasting plant performance in power electronics. The objective is to improve precision and dependability in important performance metrics by using an ensemble of classifiers that are chosen using local fluctuations in electrical parameters.

EGES' novelty stems from its capacity to adjust to the dynamic and variable attributes of electrical systems. Despite conventional methods, that utilize static models that are susceptible to inefficiencies in varying circumstances, EGES dynamically chooses the most appropriate models for each test sample, enhancing predictions and enhancing plant efficiency.

The areas of utilization for EGES span across different high-power electrical systems, comprising power plants, substations, networks of transmission, and industrial-scale electrical systems. Its flexibility and dependability render it an ideal solution for situations where real-time performance assessment is vital, and conventional approaches may struggle to present precise outcomes.

Research design: This study's research design concentrates on determining the efficacy of the proposed EGES algorithm in forecasting energy plant performance under variable conditions. The primary research questions guiding this evaluation are:

- Does EGES outperform conventional models such as RF, SVM, GBM, KNN, and LR
- How do dynamic model selection and KNN-based local data adaptation enhance prediction accuracy in volatile high-power settings

To answer these questions, the study uses EGES, which dynamically chooses the best-performing model for each test sample depending on localized data characteristics.

The methodology focuses on KNN-based local data adaptation, with the selection of k (number of nearest neighbors) and the Euclidean distance function playing critical roles. Particularly, the parameter k is optimized to strike a balance between underfitting (too few neighbors) and overfitting (too many neighbors), while the Euclidean distance guarantees accurate detection of the most comparable instances in the feature space. These decisions have a direct influence on the model selection procedure because the effectiveness of each ensemble model is assessed within these localized areas, allowing EGES to respond dynamically to variable data distributions. This detailed consideration of KNN parameters emphasizes the flexibility and accuracy of the proposed method, demonstrating its efficacy in tackling difficulties in high-power application settings.

The paper is structured as follows: Section 2 provides an extensive examination of the previous literature on machine learning methods for evaluating energy plant efficiency. Section 3 describes the EGES methodology, incorporating the dynamic model choice and majority voting procedures. Section 4 provides the experimental setup and efficiency metrics, as well as the experiment outcomes, which compare EGES to previous models. Lastly, Section 5 concludes the paper by discussing the possible influence of EGES on power electronics and making recommendations for further study.

2 Related works

This section provides an extensive survey of the previous literature on machine learning methods for evaluating energy plant efficiency. The application of machine learning techniques in energy systems has been extensively researched. The capacity to precisely predict energy production is essential for effective grid management, particularly given the variable and intermittent nature of these renewable sources.

Harrou et al. [7] presented a Long Short-Term Memory (LSTM)--based model for short-term prediction of photovoltaic solar power generation. The study highlights the significance of accurate power output prediction in optimizing energy grid management and market choices. Their LSTM method showed powerful predictive performance, particularly in dealing with dependencies in time series data, which is critical for enhancing the stability of PV systems under changing weather conditions.

Ramesh et al. [8] investigated the utilization of an auto-encoder-based neural network (AUTO-NN) integrated

with Restricted Boltzmann feature extraction in large PV plants. The model used previous meteorological data to forecast energy production stages, which substantially decreased prediction errors. This method emphasizes the growing interest in hybrid machine learning models which improve the accuracy of PV energy predictions by incorporating deep learning methods.

Sun and You [9] examined the use of machine learning and data-driven methods to regulate smart power production mechanisms, with a concentration on the role of uncertainty. Their review proposed that machine learning methods enhance the adaptability, visibility, and total efficiency of power systems, tackling the difficulties of incorporating renewable energy into the grid.

Markovics and Mayer [10] examined 24 machine-learning methods for day-ahead photovoltaic power prediction. Their research revealed that models like kernel ridge regression and multilayer perceptron outperformed conventional persistent techniques by up to 44.6% in prediction skills. The research highlighted the significance of choosing suitable predictors and fine-tuning hyperparameters to attain the best prediction precision.

Vivas et al. [11] performed a comprehensive examination of statistical and machine-learning techniques for electrical power prediction, contrasting traditional statistical models with machine-learning methods. Their results revealed that machine learning methods, especially those accounting for external fluctuation, surpassed conventional approaches in terms of forecasting accuracy.

Konstantinou et al. [12] extended the use of LSTM networks for PV power production prediction by demonstrating that deep learning models could efficiently forecast short-term power output, providing essential knowledge for grid management. Their findings emphasized the importance of utilizing recurrent neural networks (RNNs) for energy prediction operations, particularly when dealing with the non-linear nature of solar energy generation.

Li et al. [13] used Support Vector Machines (SVM) and an enhanced Dragonfly Algorithm to predict short-term wind power. Their hybrid method outperformed conventional approaches in prediction accuracy, emphasizing the increasing tendency to integrate machine learning with optimization methods for renewable energy prediction. Similarly, Kisvari et al. [14] used Gated Recurrent Neural Networks (GRNN) to enhance wind power forecasting accuracy, tackling difficulties presented by the stochastic nature of wind.

Current innovations in machine learning methods, like deep learning-based ensemble stacking, have also enhanced predicting performance. Khan et al. [15] presented an ensemble method for solar PV prediction that improved prediction accuracy by stacking multiple machine-learning models. Furthermore, Kuzlu et al. [16] used explainable artificial intelligence (XAI) techniques to acquire knowledge of solar PV power production prediction, providing an improved comprehension of the predictive systems' effectiveness and dependability. Table 1 shows the summary table.

Table 1: Summary table

Related Work	Methodology/Model Used	Key Metrics/Findings	Limitations	EGES Advantages
Oskouei et al. (2021) [7]	Decentralized robust-stochastic model	92.9% curtailment reduction, 16.33% cost savings	High complexity, limited adaptability	EGES provides real-time adaptation with KNN-based local processing and dynamic model selection.
Ramesh et al. (2023) [8]	AUTO-encoder NN using RBM feature extraction	Attained important enhancements in error metrics (for example, RMSE: 58.72%)	High RMSE; Computational intricacy with RBM	EGES balances accuracy and computational effectiveness with streamlined feature selection
Sun & You (2021) [9]	ML & Data-Driven Control (DDC)	Highlighted advantages in flexibility and system adaptability	Absence of hybrid ensemble usage for improved precision	EGES utilizes a hybrid ensemble for higher prediction accuracy

Markovics & Mayer (2022) [10]	Kernel Ridge Regression, MLP	Up to 44.6% predict skill over persistence; Hyperparameter tuning crucial	Reliance on predictor set; Prone to overfitting	EGES has superior predictor incorporation and ensemble variety
Vivas et al. (2020) [11]	Comparative evaluation of statistical/ML techniques	Noted decreased errors with hybrid techniques; ML techniques better in accuracy	Time horizon and feature variability difficulties	EGES uses an ensemble that adapts to diverse input data types
Konstantinou et al. (2021) [12]	Stacked LSTM network	Attained low RMSE and stability in short-term predictions	Computational expense of deep networks	EGES attains superior computational effectiveness by incorporating lightweight models
Li et al. (2020) [13]	SVM using enhanced Dragonfly Algorithm	Enhanced accuracy compared to baseline models	Constrained adaptability to novel data without retraining	EGES's ensemble can adapt quicker because of modular updates
Kisvari et al. (2021) [14]	GRU model	GRU surpassed LSTM in predictive accuracy and training time	Noise sensitivity and model tuning necessities	EGES is intended to be robust against data noise
Khan et al. (2022) [15]	Stacked ensemble (ANN, LSTM) incorporated with XGBoost	High accuracy and stability across diverse case studies	High computational request; Deep models are resource-intensive	EGES improves efficiency using a scalable ensemble architecture
Kuzlu et al. (2020) [16]	XAI tools (LIME, SHAP, ELI5) for solar PV prediction	Improved model transparency and important parameter insights	AI models absence of transparency, limiting trust	EGES improves explainability and parameter understanding, enhancing user trust

These previous machine learning techniques for evaluating energy plant efficiency have a research gap in dealing with the dynamic and complicated nature of functional data under a variety of energy plant circumstances. These static designs absent flexibility, leading to lower predictive precision and functional ineffectiveness. The presented EGES fills this research gap by dynamically choosing the most appropriate models for each test sample depending on the local data environment utilizing k-nearest neighbors, thereby enhancing prediction accuracy.

3 Methodology

This section describes the methodology used to assess energy plant efficiency with machine learning models.

The procedure starts with dataset collection and preprocessing, followed by the creation of the EGES, which dynamically chooses models using local data properties. Each model in the ensemble is analyzed in terms of its role in enhancing prediction accuracy.

3.1 Dataset

The dataset used in this study was gathered to assess the efficacy and productivity of energy plants in high-power uses. Between January 2023 and August 2024, it was collected from a variety of energy production facilities in different areas, comprising a thermal power plant, a chemical plant, a hydroelectric plant, and a nuclear power

station. Figure 1 illustrates these various types of energy production systems.

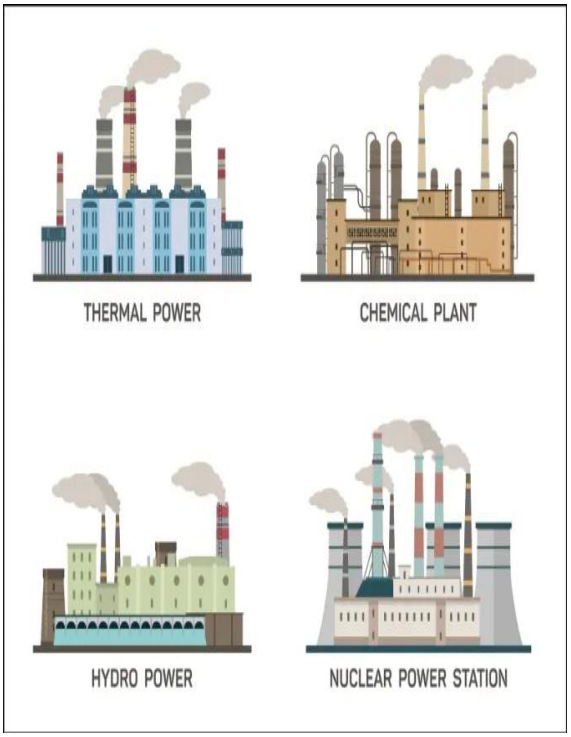

Figure 1: Different types of power plants

Data collection was simplified by direct tracking systems installed at each plant, which tracked functional metrics, power storage effectiveness, and environmental influence data daily. The research also looked into the use of Pumped Hydroelectric Storage (PHS) as a versatile energy storage method for all plant types. PHS enables surplus electricity to be utilized to pump water into elevated reservoirs, and the stored power is then released to produce electricity when demand rises. This approach enables long-term, massive energy storage for conventional and renewable power sources, guaranteeing grid stability and improving plant effectiveness. Figure 2 shows PHS.

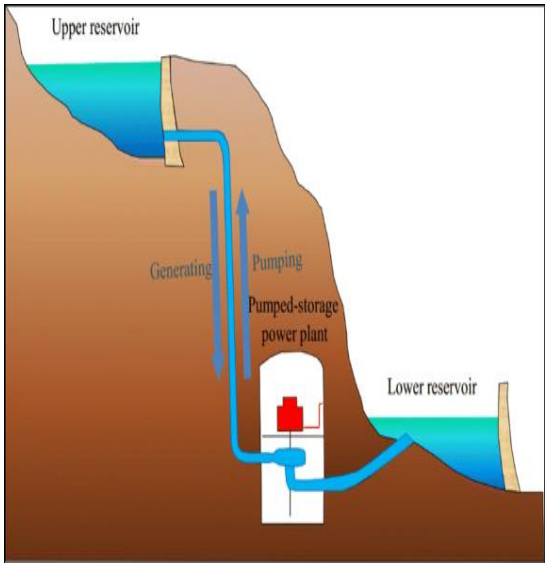

Figure 2: Pumped hydroelectric storage

The main goal of this data gathering was to use machine learning models to evaluate the plants' productivity, cost-efficiency, power storage capacities, and ecological sustainability. Each record in the dataset represents an individual plant's performance indicators, including important operational characteristics. The dataset contains 17 features that are important for assessing energy plants, with a particular emphasis on features that influence energy outcome, functional effectiveness, environmental influence, and expense. These features present an extensive view of the factors impacting plant efficiency and are critical for creating predictive models to categorize plants as having "Good" or "Poor" efficiency.

3.1.1 Attribute description

Plant_ID: A distinct identifier allocated to each energy plant for monitoring and detection reasons.
Energy output (MW): The total quantity of energy produced by the plant, calculated in megawatts (MW). This is a key metric of plant efficiency and production capability.
Efficiency (%): The percentage effectiveness of the plant, which measures how well it transforms fuel into energy. Higher effectiveness is related to reduced fuel utilization and improved efficiency.
Temperature (°C): The plant's operating temperature is calculated in degrees Celsius. The temperature has a significant impact on the plant's effectiveness and upkeep requirements.
Operating hours (hrs/day): The number of hours per day that the plant is functional. This indicates the plant's usage and the length of time it contributes to the energy grid.
Maintenance frequency (times/year): The frequency with which the plant needs upkeep, expressed in terms of times per year. Regular upkeep is essential for guaranteeing the plant's lifespan and reducing failures.

Fault rate (%): The percentage of functional errors or breakdowns in the plant. A higher fault rate suggests possible dependability problems.
Emission level (CO2 tons/year): The quantity of carbon dioxide produced by the plant annually, calculated in tons. Smaller emission rates are essential for ecological compliance and sustainability objectives.
Fuel type: The plant's fuel type, which could be gas, coal, nuclear, or solar. Various fuel kinds have differing effects on emissions, cost, and efficiency.
Power loss (%): The percentage of power lost during the manufacturing or transmission procedure. This indicates ineffectiveness in energy transmission; smaller power loss is preferable.
Downtime (hrs/year): The total quantity of hours per year when the plant is not functional because of maintenance or other problems. Reducing downtime is critical for optimizing plant efficiency.
Cost_per_MW ($): The cost of producing one megawatt of energy in US dollars. This metric measures the financial effectiveness of the plant's functions.
Voltage level (kV): The plant's operational voltage is calculated in kilovolts. Voltage levels influence the transmission effectiveness and distribution of energy.
Load factor (%): The ratio between actual production of energy and maximal potential production. Larger load factors represent a more effective use of the plant's capabilities.
Transformer rating (MVA): The rating of the plant's transformer, calculated in megavolt-amperes (MVA), showing its ability to manage electrical loads.
Cooling method: The plant's cooling system, which can be air or water-based. Cooling techniques are crucial for sustaining functional effectiveness and avoiding overheating.
Plant performance: The machine learning model's target label, which classifies plant performance as Good or Poor using a mixture of effectiveness, emissions, fault rates, and other features.
A sample of 10 records from the dataset is presented below to demonstrate the data's structure and values. These records emphasize the differences in energy generation, fuel type, fault rates, and other performance-associated metrics across various plants.

Table 2: Sample dataset

Plant_ID, Energy_Output (MW), Efficiency (%), Temperature (°C), Operating_Hours (hrs/day), Maintenance_Frequency (times/year), Fault_Rate (%), Emission_Level (CO2 tons/year), Fuel_Type, Power_Loss (%), Downtime (hrs/year), Cost_per_MW ($), Voltage_Level (kV), Load_Factor (%), Transformer_Rating (MVA), Cooling_Method, Plant_Performance
1, 600, 95, 85, 30, 5, 3, 200, Gas, 2.2, 60, 80,000, 130, 85, 170, Air, Good
2, 550, 88, 90, 32, 7, 6, 300, Coal, 3.5, 90, 85,000, 130, 88, 170, Water, Poor
3, 620, 80, 80, 28, 4, 2.5, 90, Nuclear, 1.8, 40, 85,000, 140, 80, 180, Air, Good
4, 580, 92, 95, 31, 6, 5, 280, Gas, 3.5, 70, 98,000, 125, 92, 150, Air, Poor
5, 610, 98, 82, 29, 3, 2.8, 80, Solar, 1.5, 50, 82,000, 150, 97, 170, Water, Good
6, 570, 85, 95, 33, 8, 7, 350, Coal, 5.0, 110, 90,000, 135, 85, 165, Water, Poor
7, 595, 93, 88, 30, 5, 3.5, 220, Gas, 2.5, 70, 99,000, 120, 93, 155, Air, Good
8, 555, 90, 92, 32, 7, 5.5, 320, Coal, 4.0, 80, 96,000, 130, 90, 160, Air, Poor
9, 630, 97, 83, 29, 4, 2.7, 95, Nuclear, 2.0, 45, 84,000, 140, 99, 175, Air, Good
10, 560, 86, 98, 31, 6, 6, 310, Coal, 4.8, 85, 93,000, 125, 87, 150, Water, Poor

Table 2 , shows in Sample dataset. This dataset is intended to assist machine learning systems by enabling for detailed evaluation of energy plant efficiency across a variety of functional and ecological metrics. The dataset's comprehensive features make it easier to train and validate predictive models capable of classifying plant efficiency and suggesting possible regions for improvement.

3.2 Energy guard ensemble selector

The EGES algorithm is intended to dynamically choose the most efficient model from a collection of machine learning algorithms using the local properties of each test instance. The procedure starts with the encoding of categorical features from the training dataset, such as Fuel_Type, Cooling_Method, and Plant_Performance. This is accomplished by label encoding, in which each distinct category is assigned a distinctive integer utilizing Eq. (1).

$$\text{Encoded Value} = f(\text{Category}) \tag{1}$$

For instance, Fuel_Type may be encoded as Gas → 1, Coal → 2, Nuclear → 3, and Solar → 4. After preparing the data, the next step is to train numerous models on the encoded training dataset, including RF, SVM, GBM, KNN, and Logistic Regression. Each model learns to detect trends in the data that correspond to the target feature, Plant_Performance. The training stage, as shown in Eq. (2), is crucial for creating a strong ensemble able to create precise forecasts.

$$\text{Train}(M, x_{train}) \tag{2}$$

Where M denotes the classification methods (RF, SVM, GBM, KNN, and LR) and x_{train} is the encoded training data. Each model M is trained on x_train to learn trends related to the target feature, Plant_Performance.

After the models have been trained, the algorithm determines the chosen parameters, that comprise the number of nearest neighbors k and the Euclidean distance function d. The formula is used to compute the Euclidean distance between the test sample (x_{test}) and the training sample (x_i).

$$d\ (x_{test},x_i) = \sqrt{\sum_{j=1}^{n}(x_{test,j} - x_{i,j})^2} \quad (3)$$

Where n represents the number of attributes. This calculation enables the algorithm to find the k nearest neighbors, creating a local region that represents the features of comparable instances.

After defining the local region, the next step is to evaluate the model's effectiveness. For each model M_i in the ensemble, the algorithm evaluates its accuracy on the k nearest neighbors utilizing the following formula:

$$\text{Accuracy}\ (M_i) = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, correspondingly. This assessment presents insight into how well each model executes in the setting of the local data. The accuracy scores are stored for comparison, allowing the algorithm to detect which model executes superior for the particular test sample.

After the assessment, the algorithm chooses the model with the highest accuracy score as the best performer. In cases where numerous models achieve the same top score, a tie-breaking solution can be used, considering factors like model complexity or prior efficiency. The selected model is considered the most appropriate for predicting the present test sample.

Lastly, the chosen model is used to forecast the label for the test sample, deciding whether its efficiency is categorized as "Good" or "Poor." The prediction could be formalized as follows:

$$\hat{y} = M_{best}(x_{test}) \quad (5)$$

where $\hat{y}$ is the predicted label and M_{best} is the top-performing model. The outcome is then outputted as the final prediction.

The algorithm follows this procedure for every test sample in the dataset, guaranteeing that each one is evaluated individually corresponding to its particular features. This adaptive chosen technique enhances the overall prediction accuracy, allowing more customized and efficient evaluations of plant performance. The repeated evaluations for all test samples can be described as:

$$Predictions = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m\} \quad for\ each\ x_{test,j} \in D_{test} \quad (6)$$

Here, m denotes the total number of test samples. These equations and procedures form the basis of EGES, providing a dependable framework for predictive modeling in evaluating energy plant performance.

The EGES algorithm scales with bigger datasets and higher feature dimensionality by employing a KNN method, which decreases computational burden by selecting only the closest samples for model selection. However, as the dataset grows in size and dimensionality, computing pairwise distances and training numerous models becomes more time-consuming. Despite this, EGES remains effective by restricting the search space; however, performance may suffer in massive, real-time energy applications because of the increased computational load.

Algorithm 1 outlines the suggested EGES algorithm.

Algorithm 1: EGES

Input : x_train, x_test, M = {RF, SVM, GBM, KNN, LR}, k, Euclidean Distance d

Output : Predicted label for x_test

Step 1 : Encode categorical features in x_train:
- **Fuel_Type:** Encode (Gas = 0, Coal = 1, Nuclear = 2, Solar = 3)
- **Cooling_Method:** Encode (Air = 0, Water = 1)
- **Plant_Performance:** Encode (Poor = 0, Good = 1)

Step 2 : Train models RF, SVM, GBM, KNN, LR on x_train

Step 3 : Set parameters: k, d

Step 4 : For each test sample x_test:

Step 5 : **Local Region Definition**
For each sample x_i in x_train:
- Calculate Euclidean distance d (x_test, x_i)

Choose k nearest neighbors from x_train using the smallest Euclidean distances

Step 6 : **Model Performance Assessment (utilizing accuracy)**
For each model $M_i \in$ {RF, SVM, GBM, KNN, LR}:
Assess the accuracy of M_i on the k nearest neighbors
Store accuracy score of M_i

Step 7 : **Model Selection**
Choose the best-performing model with the maximum accuracy score.

Step 8 : **Prediction**
Utilize the chosen model to predict the label for x_test

Output the forecasted label for x_test (Poor or Good)

Step 9 : Repeat for all test samples.

3.2.1 RF

RF is an ensemble learning technique that constructs numerous decision trees and then combines their predictions to enhance classification accuracy. RF operates by training each tree on a randomly selected subset of the data and attributes, decreasing variability and preventing overfitting. RF was selected for EGES because of its capacity to handle large datasets and nonlinear correlations between features. In this research, RF helps the model ensemble by providing excellent predictive abilities, especially when dealing with intricate relationships between variables like temperature, fault rates, and the production of energy.

3.2.2 SVM

The SVM is an effective classifier that seeks to identify the best hyperplane that divides various classes in the dataset. SVM performs well in high-dimensional spaces and is useful for both linear and nonlinear classification activities. In EGES, SVM aids in differentiating between "good" and "poor" operating plants by evaluating the complicated relationship between plant effectiveness, functioning hours, and downtime. The model is especially helpful in situations where the classes are not linearly separable, utilizing kernel functions for enhanced efficiency.

3.2.3 GBM

GBM is another ensemble technique that constructs models sequentially, with each model attempting to right the flaws of the previous one. GBM is well-known for its high forecasting accuracy, particularly in situations involving intricate relationships between features. In the setting of this research, GBM improves EGES efficiency by concentrating on decreasing error in forecasting plant efficiency using variables such as emission stages, service frequency, and fault rates. GBM contributes to the ensemble through its capability to decrease bias and enhance generalization.

3.2.4 KNN

KNN is a non-parametric algorithm that categorizes data points according to the majority class of their nearest neighbors. KNN is especially efficient when local data points exhibit comparable trends, rendering it an ideal candidate for EGES. By assessing the local neighborhood of each test sample, KNN aids in the identification of comparable plants using features such as fuel type, power loss, and downtime, enhancing the model's capacity to forecast plant efficiency in particular situations. KNN's ease and efficiency in local region classification render it a useful member of the ensemble.

3.2.5 LR

LR is a linear model utilized to perform binary classification tasks. It forecasts the likelihood of an event happening, like whether a plant will execute "good" or "poor." In EGES, LR is used to manage cases where the correlation between variables is linear. For instance, LR is effective at detecting efficiency patterns using energy results and cost per megawatt. Despite its simplicity, LR contributes to the ensemble by presenting baseline predictions that could be improved by more complicated models such as RF and GBM.

The EGES dynamically chooses models using local data features, thereby improving prediction accuracy for energy plant efficiency classification. EGES enhances prediction dependability by utilizing a various set of machine learning models, each customized to various data trends, guaranteeing that the most effective model is used for each test case.

Model selection procedure

EGES' model selection process dynamically compares the accuracy scores of numerous classifiers (RF, SVM, GBM, KNN, and LR) for each test sample. If multiple models produce similar accuracy scores, a tie-breaking strategy is used to determine the best model. The tie-breaking procedure adheres to a predefined criterion hierarchy: first, the model with the smallest computational complexity (especially simpler models to decrease overfitting risks) is selected. If there is still a tie, the model with the best historical performance on comparable datasets is chosen. In cases where the models have comparable performance records, the model with the shortest runtime is chosen to maximize effectiveness for real-time predictions. This structured tie-breaking guarantees that the final model selection is not only the most precise but also the best fit for operational limitations, thus enhancing the overall prediction procedure.

Model interpretability and computational complexity

The EGES algorithm is highly efficient, but computationally complex because of the dynamic selection procedure, especially in calculating the Euclidean distances for k-nearest neighbors, which scales with O (n.m), where n is the number of training samples and m is the feature count. This complexity emphasizes the trade-off between accuracy and possibility in real-time, massive applications. EGES also integrates model interpretability through feature importance assessment using ensemble models such as Random Forest and GBM. Such interpretability is essential for energy sector stakeholders because it allows for transparency and comprehension of decisions, especially those involving key features such as Fuel_Type and Cooling_Method, which have a direct impact on predictions. This balance of

computational effectiveness and interpretability distinguishes EGES as a useful yet insightful tool for evaluating energy efficiency.

Figure 3 displays the flow diagram of the EGES algorithm.

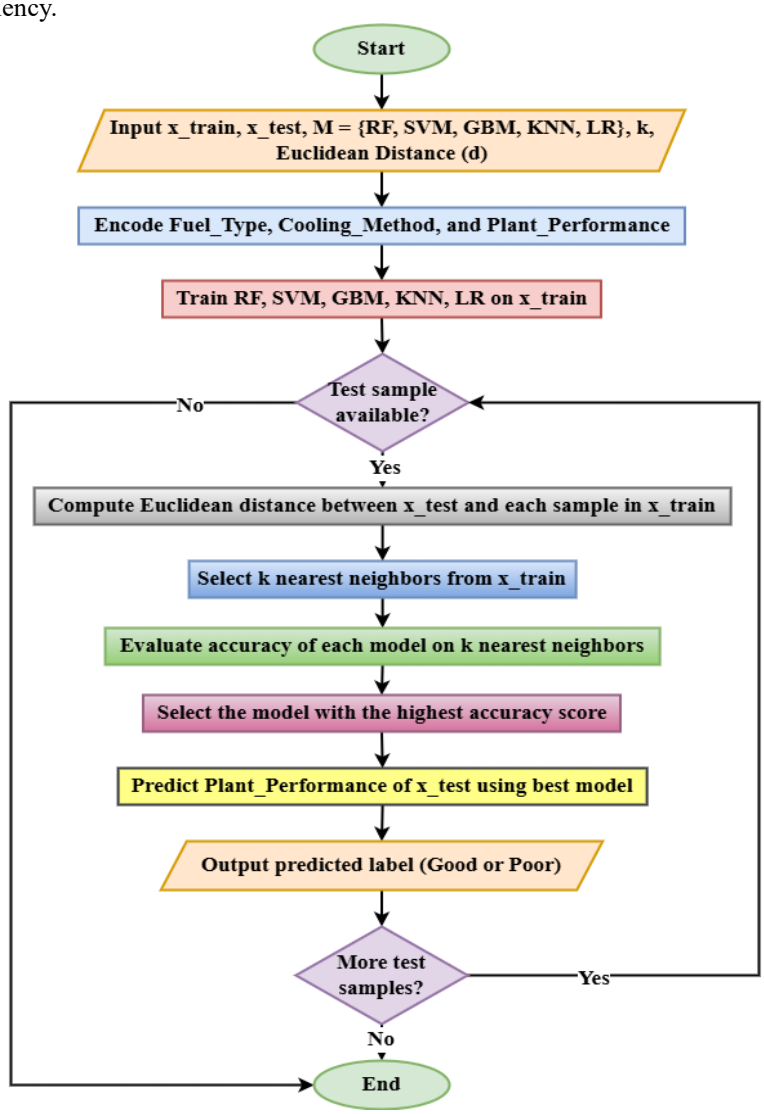

Figure 3: Flow diagram of EGES algorithm

4 Experimental results and discussions

This section provides the experimental findings of the suggested EGES algorithm and compares its efficacy to five well-known machine learning classifiers: RF, SVM, GBM, KNN, and Logistic Regression. The EGES algorithm was written in Java and implemented with the Weka machine learning tool, which makes use of Weka's library of classifiers and distance functions to train and evaluate models efficiently. Weka was selected because of its comprehensive assistance for a broad range of machine learning algorithms and its capacity to manage intricate data preprocessing, model evaluation, and performance measurement tasks seamlessly.

To guarantee a balanced dataset for training and assessment, the Synthetic Minority Over-sampling Technique (SMOTE) was used to correct class imbalances by efficiently augmenting minority class samples. EGES classifier hyperparameters were improved utilizing grid search and 5-fold cross-validation. RF was tuned for 100 trees and max depth 15, SVM for RBF kernel and C=1.0, GBM for learning rate 0.1, 100 estimators and depth 3, KNN for k=5 with Euclidean distance, and LR for L2 penalty with liblinear solver. These tactics guaranteed consistent performance across test cases.

EGES' effectiveness was evaluated using numerous important metrics, including accuracy, precision, recall, F1-score, and MCC. These metrics offer an extensive comprehension of the classification models' efficiency, providing insights into both predictability and dependability.

Accuracy: This metric calculates the percentage of correct results for all cases evaluated. It is a basic indicator of a classifier's overall efficacy, which includes both true

positives and true negatives. The formula for accuracy is shown in Eq. (4).

Precision: Precision is the percentage of true positives among those that were predicted to be positive. It is especially important in situations where the cost of false positives is high. Precision is determined as follows:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

High precision means that when the classifier forecasts a positive class, it is likely to be correct.

Recall: Recall, also known as sensitivity or true positive rate, is a measure of how many actual positives the classifier accurately detects. It is critical in circumstances where losing a positive case is costly. The recall is provided by:

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

A high recall indicates that the classifier accurately identifies the majority of the actual positive cases.

F1-score: The F1-score balances precision and recall, which is especially helpful when the dataset is imbalanced. The harmonic mean of precision and recall is calculated as follows:

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

A high F1 score signifies that the classifier performs well in terms of precision and recall.

MCC: MCC is a more extensive metric that takes into account all four categories (TP, TN, FP, and FN). It is particularly helpful when dealing with imbalanced datasets. It is calculated using the following formula:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (10)$$

An MCC value near 1 denotes a strong positive correlation between predicted and actual classifications.

The performance metrics chosen for EGES, such as accuracy, precision, recall, F1-score, and MCC, are essential in assessing energy plant efficiency. Accuracy measures overall correctness, whereas precision and recall help to reduce costly false positives and negatives. The F1-score balances these metrics, resulting in a comprehensive evaluation. MCC provides a balanced measure that takes into account all types of prediction errors, making it particularly helpful for imbalanced datasets. While these metrics are focused on classification, considering economic effects could improve the assessment by reflecting the operational expenses and effectiveness losses related to inaccurate predictions.

Table 3, displays the comparative performance of EGES against RF, SVM, GBM, KNN, and LR:

Table 3: Performance metrics comparison

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC (%)
RF	88.5	88.2	89.5	88.8	85.5
SVM	89.0	86.5	87.7	87.1	83.0
GBM	91.2	89.0	88.5	89.7	87.0
KNN	88.5	87.0	86.2	86.6	81.5
LR	87.8	85.3	86.7	86.0	81.0
EGES	93.5	91.5	92.7	92.1	90.0

The EGES algorithm surpasses all other classifiers on all performance metrics. EGES, in particular, attained the highest accuracy of 93.5%, which is substantially higher than GBM, the closest competitor, who attained 91.2%. This increase in accuracy shows the efficacy of EGES in correctly categorizing test samples using the local region determined by the nearest neighbors and leveraging model selection depending on accuracy evaluation.

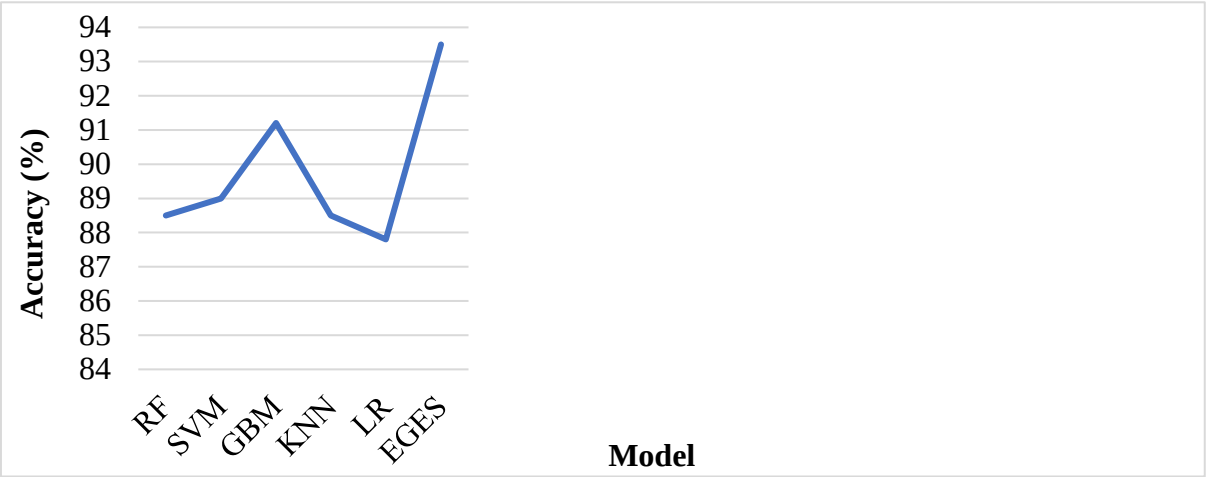

Figure 4: Accuracy comparison

Figure 4 shows a line chart comparing the accuracy of EGES to the other classifiers. The chart emphasizes EGES' better efficiency, demonstrating its position as the most precise model. EGES uses k-nearest neighbor selection and model accuracy evaluation to improve its predictions and surpass conventional machine learning models such as RF, SVM, and GBM. This demonstrates EGES' superiority in determining the most effective prediction model, resulting in the highest accuracy.

In terms of precision, EGES also outperforms the competition with a score of 91.5%, as demonstrated in Table 2 and Figure 5. Precision is critical for reducing false positives, and EGES's higher precision score suggests that it makes fewer erroneous positive predictions than other models. This precision benefit comes from the EGES algorithm's dynamic selection of the best-performing model for each test sample, which enables it to achieve more precise classification results.

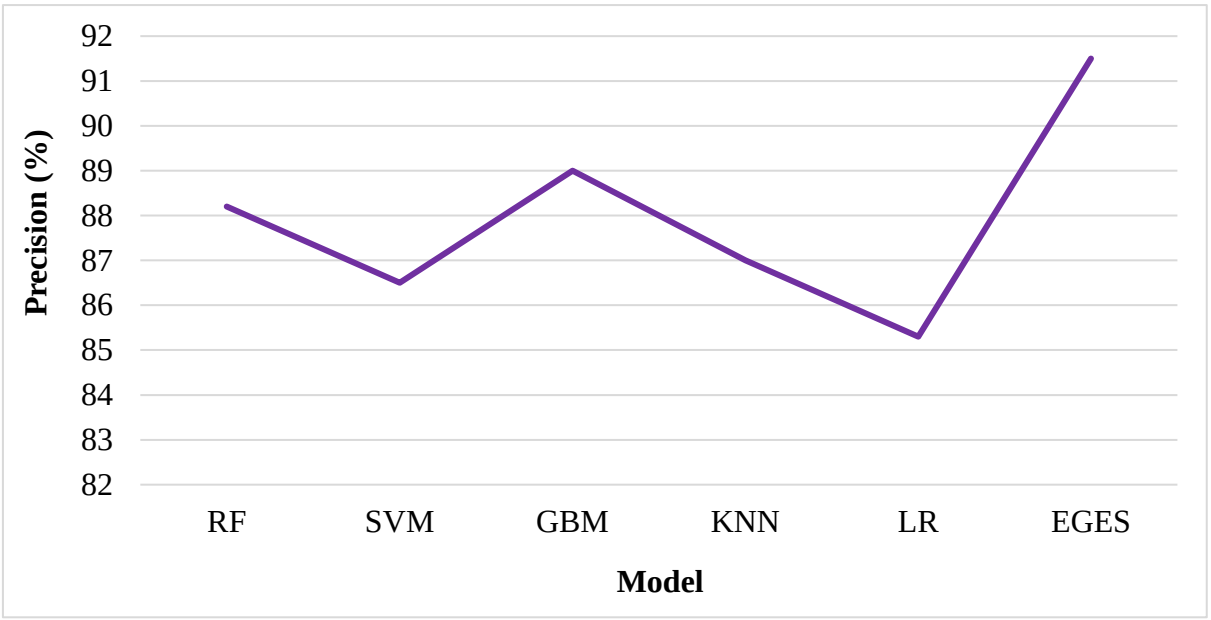

Figure 5: Precision comparison

Figure 5 shows a line chart with the precision values for the classifiers, demonstrating EGES' superiority once again. Its capability to provide high precision, particularly when compared to models such as SVM and LR, shows the strength of EGES' ensemble method. This confirms that the dynamic selection of classifiers using accuracy results in not only higher precision but also more dependable predictions in general.

When assessing recall, EGES continues to surpass the other models, with a score of 92.7%, as shown in Figure 6. The recall is critical for detecting all pertinent instances of the target class, and a higher recall score signifies that EGES correctly detects a greater proportion of true positives than other models. This makes EGES especially appropriate for applications where detecting positive instances is critical.

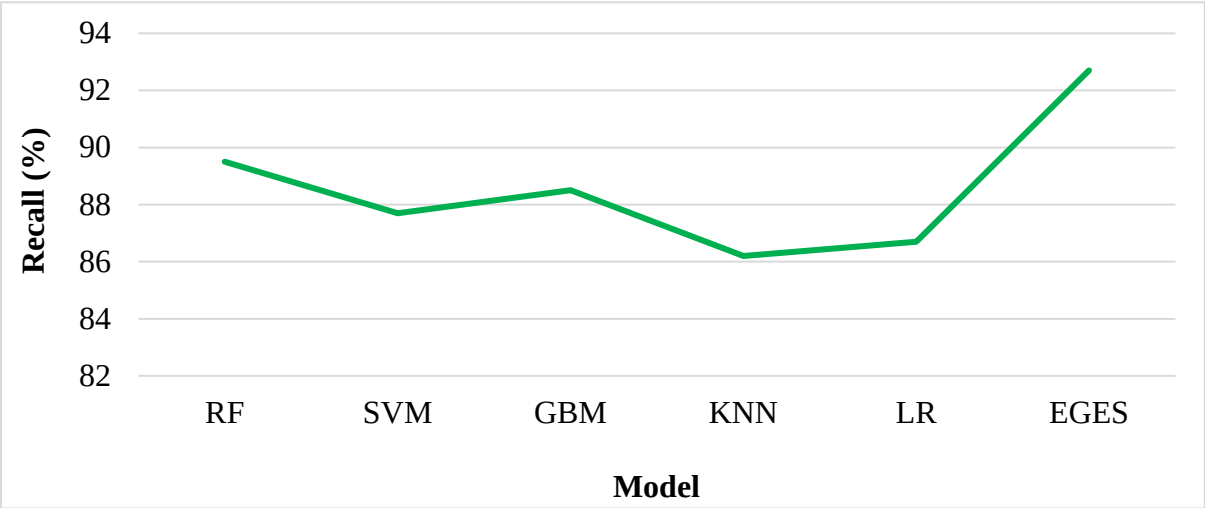

Figure 6: Recall Comparison

Figure 6 depicts the recall comparison among the classifiers, with EGES emerging as the leader. This demonstrates EGES's ability to identify pertinent instances and ensure extensive coverage of the target attribute. The integration of k-nearest neighbor selection and the utilization of accuracy for model assessment allows EGES to attain higher recall than models like RF and KNN.

In terms of the F1-score, which balances precision and recall, EGES has the highest score (92.1%). As illustrated in Figure 7, this result shows that EGES strikes an excellent balance between precision and recall, operating well in both dimensions. The F1-score comparison validates EGES' resilience in guaranteeing high classification effectiveness while reducing faults.

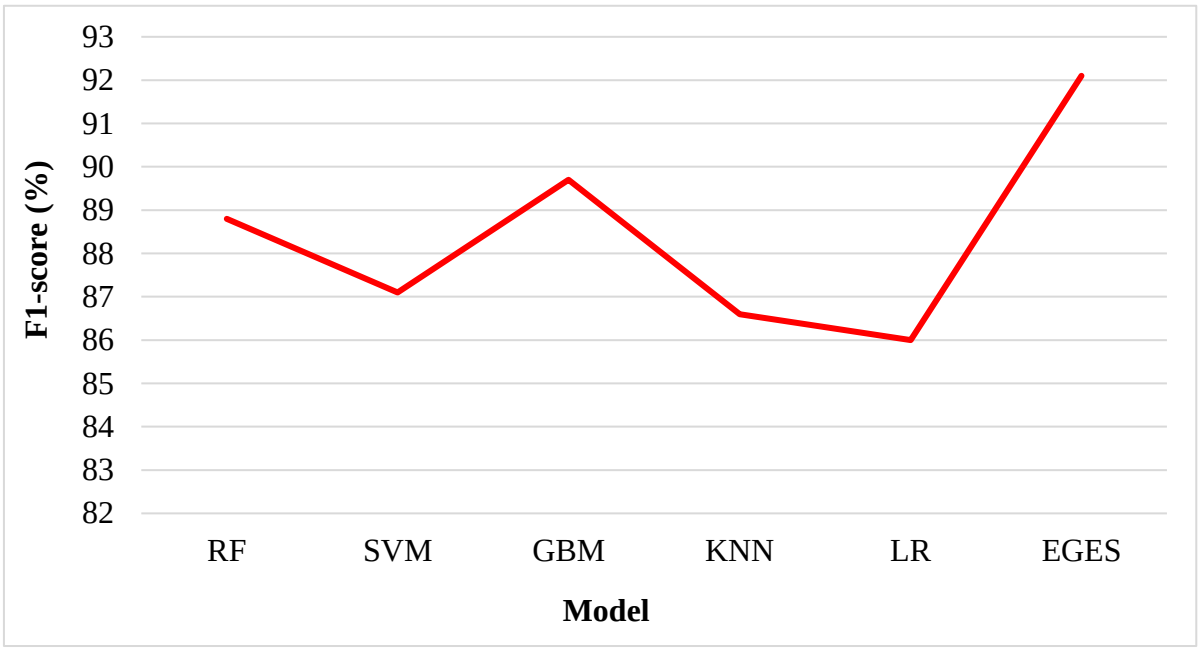

Figure 7: F1-score Comparison

Figure 7 shows a line chart comparing F1 scores, with EGES once again outperforming the other classification methods. The balance between precision and recall shows the algorithm's capacity to improve prediction performance without sacrificing one metric for the other.

Lastly, the Matthews Correlation Coefficient (MCC) is calculated, which provides a more dependable measure of classification efficiency, particularly in the presence of imbalanced data. EGES has the highest MCC score of 90.0%, as illustrated in Figure 8. The higher MCC score suggests that EGES presents a more balanced and accurate classification, taking into account both true and false positives and negatives.

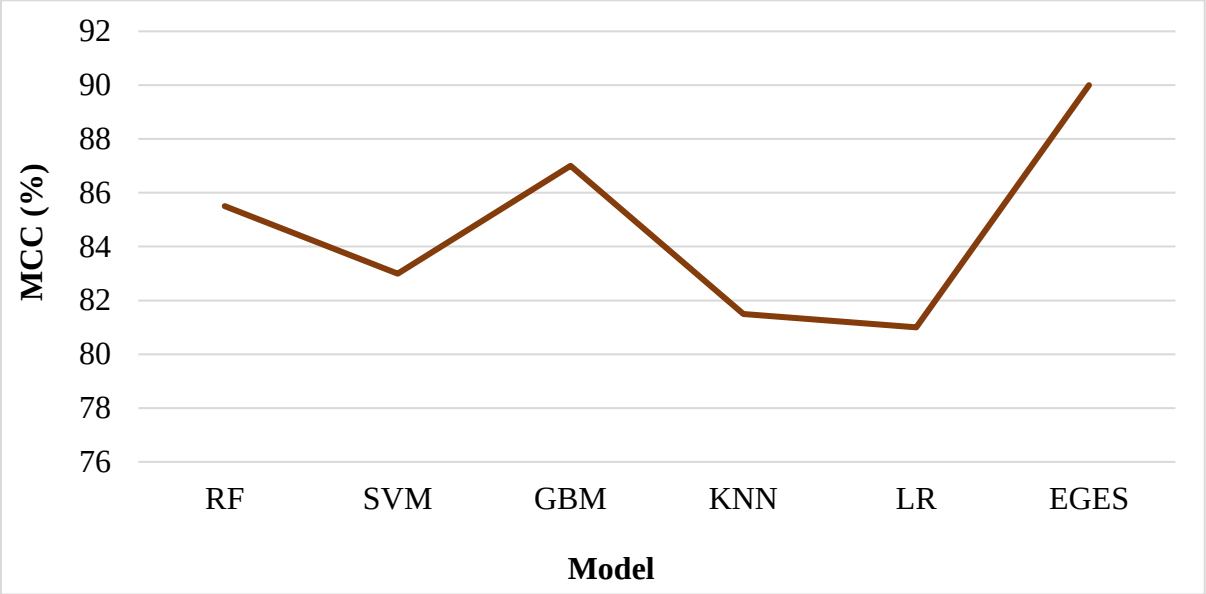

Figure 8: MCC Comparison

Figure 8 shows EGES' superiority in terms of MCC, cementing its position as the best-performing model in this study. The MCC score highlights EGES's ability to deal with imbalanced datasets and provide dependable classification outcomes across different metrics.

Overall, the experimental findings clearly show that the EGES algorithm outperforms other classifiers. EGES consistently surpasses conventional machine learning models like RF, SVM, GBM, KNN, and LR on all performance metrics, including accuracy, precision, recall, F1-score, and MCC. The use of k-nearest neighbor selection, accuracy-based model assessment, and dynamic model selection improves EGES's resilience and accuracy in forecasting plant performance. These findings justify the use of EGES as an efficient and dependable classification algorithm, especially in applications requiring accurate and precise predictions.

Error analysis

EGES error analysis shows that poor performance is most common when data is sparse or plant types are highly imbalanced, with certain classes underrepresented. These errors are most evident when the model is unable to differentiate between closely related plant performance categories, particularly in situations where the feature set lacks enough variation or the KNN model's chosen k is not optimized. Furthermore, performance degradation occurs when certain parameter settings, like the distance function or the number of neighbors (k), do not correspond to the underlying data distribution. Evaluating these error-prone cases reveals important information about potential model constraints, emphasizing the significance of modifying hyperparameters and guaranteeing balanced, comprehensive data for more reliable predictions.

Discussion

The experimental findings show that EGES surpasses conventional models like RF, SVM, GBM, KNN, and LR in all important metrics, with the highest accuracy (93.5%), precision (91.5%), recall (92.7%), F1-score (92.1%), and MCC (90.0%). EGES outperforms the state-of-the-art (SOTA) techniques reviewed due to its dynamic model selection strategy and k-nearest neighbor (KNN)-based local data adaptation, which allow it to tailor predictions to the unique characteristics of each test sample. Unlike conventional models, which depend on global patterns learned during training, EGES dynamically assesses and chooses the most efficient model depending on localized accuracy, guaranteeing reliable predictions even in intricate, high-power application settings. Furthermore, KNN-driven local region formation reduces overgeneralization by concentrating on data subsets with similar characteristics, which improves precision and adaptability. This design contrasts with the static nature of traditional SOTA techniques, which frequently fail with the fluctuating parameters found in energy plant performance evaluations. By tackling these difficulties and offering high accuracy under dynamic conditions, EGES shows its novelty and practicality for energy system predictive modeling.

5 Conclusion

The EGES algorithm implemented in this study outperformed conventional models like RF, SVM, GBM, KNN, and LR in terms of plant performance classification and prediction. By its dynamic model selection and k-nearest neighbor method, EGES consistently attained higher accuracy, precision, recall, F1-score, and MCC across all assessments, as executed in Java and Weka. Its

resilient efficiency demonstrates the algorithm's possibility for wider use beyond plant performance, implying that future research could include adapting EGES to other areas like healthcare, finance, or e-commerce, where accurate classification is required. Future work on EGES could concentrate on tackling dataset imbalance, especially in terms of "Good" versus "Poor" performance labels, by incorporating methods like SMOTE (Synthetic Minority Over-sampling Technique) to improve model resilience. Furthermore, integrating deep learning techniques like neural networks or reinforcement learning methods could enhance EGES' flexibility in dynamic and fluctuating settings. These sophisticated methods could allow EGES to effectively manage intricate, non-linear relationships and learn from real-time data feedback, possibly boosting scalability and accuracy in real-world energy applications.

References

[1] Bullich-Massagué, E., Cifuentes-García, F. J., Glenny-Crende, I., Cheah-Mañé, M., Aragüés-Peñalba, M., Díaz-González, F., & Gomis-Bellmunt, O. (2020). A review of energy storage technologies for large-scale photovoltaic power plants. *Applied Energy*, 274, 115213. https://doi.org/10.1016/j.apenergy.2020.115213

[2] Gams, M., & Kolenik, T. (2021). Relations between electronics, artificial intelligence, and information society through information society rules. *Electronics,* 10(4), 514. https://doi.org/10.3390/electronics10040514

[3] Khan, P. W., Byun, Y. C., Lee, S. J., Kang, D. H., Kang, J. Y., & Park, H. S. (2020). Machine learning-based approach to predict energy consumption of renewable and nonrenewable power sources. *Energies,* 13(18), 4870. https://doi.org/10.3390/en13184870

[4] Theocharides, S., Makrides, G., Livera, A., Theristis, M., Kaimakis, P., & Georghiou, G. E. (2020). Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. *Applied Energy,* 268, 115023. https://doi.org/10.1016/j.apenergy.2020.115023

[5] Mahmud, K., Azam, S., Karim, A., Zobaed, S., Shanmugam, B., & Mathur, D. (2021). Machine learning-based PV power generation forecasting in Alice Springs. *IEEE Access*, 9, 46117-46128. https://doi.org/10.1109/access.2021.3066494

[6] Alkesaiberi, A., Harrou, F., & Sun, Y. (2022). Efficient wind power prediction using machine learning methods: A comparative study. *Energies*, 15(7), 2327. https://doi.org/10.3390/en15072327

[7] Oskouei, M. Z., Mohammadi-Ivatloo, B., Abapour, M., Shafiee, M., & Anvari-Moghaddam, A. (2021). Privacy-preserving mechanism for collaborative operation of high-renewable power systems and industrial energy hubs. *Applied Energy*, 283, 116338. https://doi.org/10.1016/j.apenergy.2020.116338

[8] Ramesh, G., Logeshwaran, J., Kiruthiga, T., & Lloret, J. (2023). Prediction of energy production level in large PV plants through auto-encoder based neural-network (auto-nn) with restricted Boltzmann feature extraction. *Future Internet*, 15(2), 46. https://doi.org/10.3390/fi15020046

[9] Sun, L., & You, F. (2021). Machine learning and data-driven techniques for the control of smart power generation systems: An uncertainty handling perspective. *Engineering*, 7(9), 1239-1247. https://doi.org/10.1016/j.eng.2021.04.020

[10] Markovics, D., & Mayer, M. J. (2022). Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. *Renewable and Sustainable Energy Reviews*, 161, 112364. https://doi.org/10.1016/j.rser.2022.112364

[11] Vivas, E., Allende-Cid, H., & Salas, R. (2020). A systematic review of statistical and machine learning methods for electrical power forecasting with reported mape score. *Entropy*, 22(12), 1412. https://doi.org/10.3390/e22121412

[12] Konstantinou, M., Peratikou, S., & Charalambides, A. G. (2021). Solar photovoltaic forecasting of power output using LSTM networks. *Atmosphere*, 12(1), 124. https://doi.org/10.3390/atmos12010124

[13] Li, L. L., Zhao, X., Tseng, M. L., & Tan, R. R. (2020). Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm. *Journal of Cleaner Production*, 242, 118447. https://doi.org/10.1016/j.jclepro.2019.118447

[14] Kisvari, A., Lin, Z., & Liu, X. (2021). Wind power forecasting–A data-driven method along with a gated recurrent neural network. *Renewable Energy*, 163, 1895-1909. https://doi.org/10.1016/j.renene.2020.10.119

[15] Khan, W., Walker, S., & Zeiler, W. (2022). Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *Energy,* 240, 122812. https://doi.org/10.1016/j.energy.2021.122812

[16] Kuzlu, M., Cali, U., Sharma, V., & Güler, Ö. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *Ieee Access,* 8, 187814-187823. https://doi.org/10.1109/access.2020.3031477