IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 14, 2022, accepted February 16, 2022, date of publication February 22, 2022, date of current version March 9, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3153720

Electrical Energy Prediction of Combined Cycle Power Plant Using Gradient Boosted Generalized Additive Model

NIKHIL PACHAURI, (Member, IEEE), AND CHANG WOOK AHN, (Member, IEEE)
Artificial Intelligence Graduate School, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding author: Chang Wook Ahn (cwan@gist.ac.kr)

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) Grant funded by the Korea Government (MSIT) (Artificial Intelligence Graduate School Program, Gwangju Institute of Science and Technology) under Grant 2019-0-01842, and in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2021R1A2C3013687.

ABSTRACT A combined cycle power plant (CCPP) employs gas and steam turbines to generate 50% more power while utilizing the same fuel as a normal single cycle plant. The performance of a CCPP under full load is affected by a variety of factors such as weather, process interactions, and coupling, which makes it challenging to operate. Therefore, a reliable assessment of the maximum output power of a CCPP is required to improve plant reliability and monetary performance. In this paper, a predictive model based on a generalized additive model (GAM) is proposed for the electrical power prediction of a CCPP at full load. In GAM, a boosted tree and gradient boosting algorithm are considered as shape function and learning technique for modeling a non-linear relationship between input and output attributes. Furthermore, predictive models based on linear regression (LR), Gaussian process regression (GPR), multilayer perceptron neural network (MLP), support vector regression (SVR), decision tree (DT), and bootstrap-aggregated tree (BBT) are also designed for comparison purposes. Results reveal that GAM improves the RMSE by 74%, 68.8%, 70.3%, 54.8%, 21.2%, and 17.3% compared to LR, GPR, MLP, SVR, DT, and BBT, respectively. Furthermore, the results of the Man-Whitney U test and rank analysis also confirm the effectiveness of GAM for energy prediction of CCPP. Finally, it can be concluded that the proposed method is effective, robust, and accurate for the assessment of the maximum output power of a CCPP to improve plant consistency and financial performance.

INDEX TERMS Combined cycle power plant, electrical energy, generalized additive model, linear regression, decision tree, Man-Whitney U test.

I. INTRODUCTION

In order to analyze a thermodynamic system, various hypotheses are needed to compensate for the uncertainty in the solution. In real time applications, these hypotheses are impractical for analyzing complex systems. It involves solving hundreds of nonlinear equations, resulting in excessive computational requirements. To circumvent this constraint, machine and deep learning techniques are gaining popularity as a way to avoid thermodynamic-based techniques, discover counter-intuitive aspects, and provide performance efficiencies beyond design variables. These advances result from the discovery of diverse and complex correlations and interconnections between important input and output attributes [1]. A combined cycle power plant (CCPP) is a well-known example of a thermodynamic system. The performance of a power plant under full load is affected by a variety of factors such as weather, process interactions, coupling, and so on, which makes it challenging to create a reliable mathematical model for CCPP. The CCPP uses gas and steam turbines to produce 50% more energy using fuel similar to a standard simple cycle plant. However, accurate estimation of output power at maximum load is essential to enhance plant efficiency and financial operations [2]. Reliable energy generation assessment tools can help to conserve energy and maximize returns on existing megawatt-hours (MWh), which improves power

The associate editor coordinating the review of this manuscript and approving it for publication was Hiram Ponce.

plant efficiency, especially when facing the limits of raw material conservation and high profitability. Hence, precise power generation forecasting has great importance in enhancing the efficiency of power plants and improving environmental conditions [3]. In recent years, researchers have utilized various approaches based on machine learning (ML) algorithms to predict the output power at full load of CCPP. Previous studies on the prediction and control of CCP are reviewed in Table 1.

TABLE 1. Literature Survey on prediction and control of CCPP.

Authors	Methodology	Results obtained
Calvo & Corchado [4]	Rule-base and multilayer perceptron neural network (MLP) inspired tuning of the Proportional-Integral-Derivative control	MLP perform better in comparison to linear discriminator analysis (LDA) and decision trees (DT)
Akdemir [5]	Artificial neural network (ANN) is designed for the output power prediction	The mean squared error (MSE) decreases to 3.176 after introducing two fold cross-validations.
Ahn, & Hur [6]	Continuous Conditional Random Field Model (CCRF) is implemented for electrical load prediction	The root mean squared error (RMSE) for CCRF is reduced to 2.3% and 2.5% compared to regression tree (RT) and ANN.
Janoušek et.al. [7]	Random forest (RF) in the regression framework is used for the output power prediction	Up to 4% reduction in RMSE and MAE compared to K5star with parallel RF
Elfaki & Ahmed [8]	Bayesian Regularized ANN (BRNN) and Levenberg-Marquardt ANN (LMNN) are designed for output power prediction.	BRNN have a lower training and testing value for MSE compared to LMNN.
Lorencin et.al. [9]	Genetic algorithm optimized MLP (GMLP) is proposed for CCPP power output forecasting.	GMLP is better than linear regression (LR) and pace regression (PR) by 20.6 % and 6 %
Wood [10]	Memetic optimized transparent-open-box (MTBO) ML algorithm for power prediction	The RMSE achieved by MTBO is 2.89 MW, which shows the efficacy of proposed ML technique.
Hundi, & Shahsavari [11]	RF is implemented for the prediction of output power	The R^2 achieved by RF is 95.9 % which is higher than LR (92.4 %) and MLP (93.8 %)
Qu et.al. [12]	Grid search optimized stacked ensemble ML algorithm for electricity generation	Proposed method improves the RMSE by 19.3 % and 3.5 % compared to RF and vote ensemble
Karaçor et.al. [13]	Life prediction using fuzzy logic (FL) and ANN	The relative error varies between 0.59% and 3.54% for FL, which is higher than ANN (0.001% and 0.84%)
Moayedi & Mosavi [14].	Water cycle (WCA), Ant Lion (ALO) and Satin Bowerbird (SBO) optimized MLP	The results reveal that WCA is more efficient than ALO and SBO for MLP optimization
Arferiandi et.al. [15]	Different ANN configurations were implemented for different combinations of features (total features = 3) to accurately predict the heat rate.	The results show that the regression coefficient (0.994) is higher for the ANN applied to the first and third shape combinations.

In [16], the LR2,1 norm-based online sequential extreme learning algorithm (LR21OS-ELM) is designed for different prediction problems. The performance of LR21-OS-ELM is compared with that of ELM and LR21-ELM for electrical energy prediction. Results demonstrate that the proposed ML algorithm outperformed ELM and LR21-ELM in terms of RMSE. In [17], the Ridge and support vector regression models are designed and implemented for the energy prediction of CCPP. The regression coefficient for SVR (0.98) is higher than the ridge regression (0.92), which shows the higher predicting accuracy of SVR. In [18], principal component analysis (PCA)-based K-means and agglomerative clustering are used for CCPP energy prediction. The results show that the proposed algorithms have an accuracy of 80% compared to the support vector machine (SVM) and regression tree. A deep learning neural network (DNN) is designed for the CCPP energy forecasting [19]. The predicting performance of DNN is compared with that of sequential API and functional API based ANN. Results show the superior performance of deep learning neural networks.

Various ML techniques are used in the literature to predict the electrical energy output of CCPP. Each ML algorithm has its own pros and cons. For example, the number of neurons in each hidden layer, synaptic weights, learning rate, and bias values all have a significant impact on ANN performance. Fuzzy logic-based ML methods require an accurate estimation of the rule base, which is a time-consuming operation. On the other hand, the prediction accuracy of SVR and SVM is determined by the appropriate values of their corresponding hyper-parameters. From the viewpoint of the above discussion, the aim of this work is to investigate a competent energy forecasting model for CCPP based on a boosting based generalized additive model (GAM), which can provide energy experts with the necessary insight into CCPP energy generation. These predictive models are simple to interpret while enhancing forecast accuracy. It also generally outperforms most linear techniques, such as linear regression, while providing greater interpretability compared to other ML algorithms. These predictive models allow current knowledge to be integrated during the model construction process in order to improve the prediction performance. The key contributions of the article are as follows.

1. A boosting-based generalized additive model (GAM) is proposed for the output energy prediction of CCPP.
2. The predictive performance of GAM is compared with that of linear regression (LR), Gaussian process regression (GPR), multilayer perceptron neural network (MLP), support vector regression (SVR), decision tree (DT), and bootstrap-aggregated tree (BBT).
3. The Mann–Whitney U test, violin plots, and rank analysis are all utilized to perform a detailed assessment of the models' outcomes.

The following are the remaining sections of this work: Sections II and III provide a full discussion of the dataset as well as the proposed algorithm. Sections IV and V describe

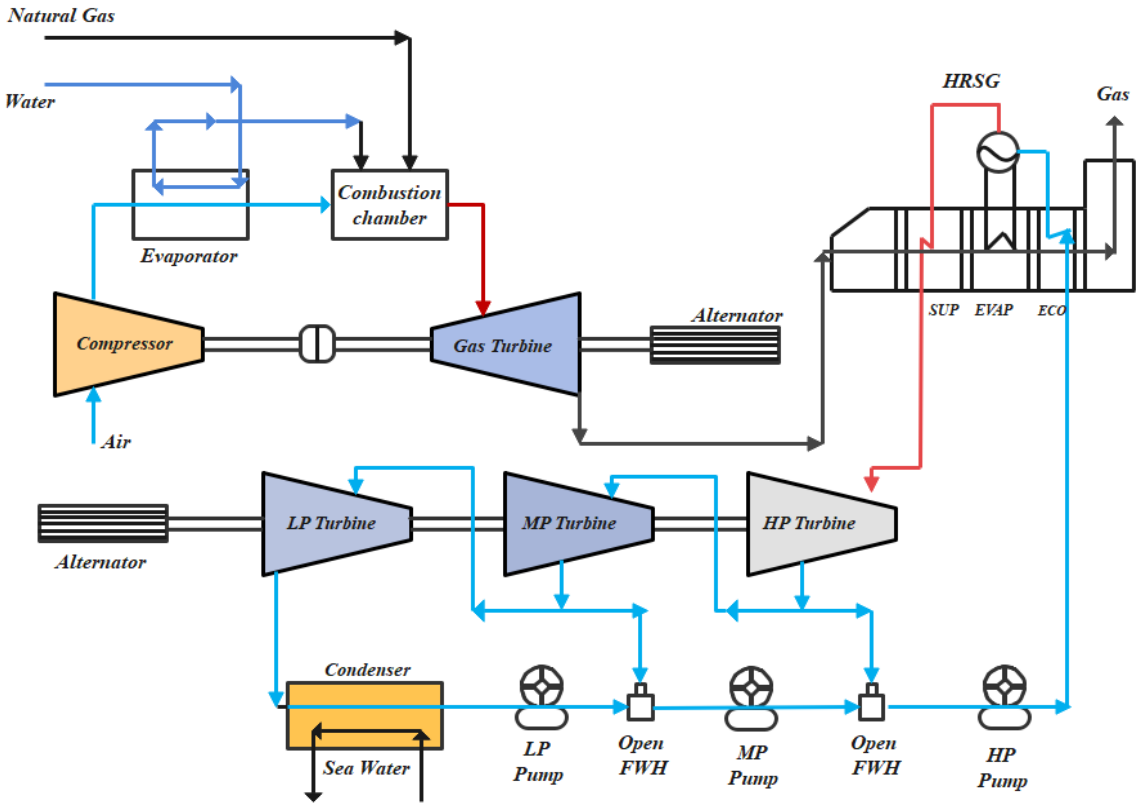


FIGURE 1. A schematic representation of combined cycle power plant [19].

the results and discussion, followed by the conclusion of the work.

II. COMBINED CYCLE POWER PLANT (CCPP) SYSTEM

A CCPP is a combination of steam and gas turbines (ST and GT) with heat recovery steam generators (HRSG). The power in a CCPP is produced using ST and GT, which are integrated in one cycle and transported from one turbine to the other [20]. In the CCPP, GT not only produced electric power but also hot emissions consisting of NOx and COx gases. These gases are passed through HRSG, where they are converted to steam and generate electricity due to coupled ST and generators. Thus, the GT generator generates energy, and the remaining heat from the exhaust gas is used to make steam, which in turn generates power via the ST generator [21]. For this study, the data set is from CCPP-1 [22] with a small production capacity of 480 MW, consisting of one 160 MW ABB ST, two dual HRSGs and two 160 MW ABB 13E2 GTs, as shown in Figure 1. The CCPP data set contains 47840 (9568 per year) data points collected between 2006 and 2011 while the plant was operating at full load. The power (PE) generated by the combination of GT and ST is primarily affected by four environmental variables: ambient temperature (AT), exhaust vacuum (V), relative humidity (RH), and ambient pressure (AP). Thus, AT, V, RH, and AP act as input attributes while PE acts as output attribute of the ML algorithm. A statistical description of the CCPP data set is presented in Table 2. For better understanding of the data set, all input and output attributes are described as histogram fit in Figure 2.

TABLE 2. Statistical description of CCPP dataset.

S.No	Attributes	Range	Median	Standard deviation
1	Ambient Temperature (AT)	1.81 -37.11°C	20.30	7.45
2	Exhaust Vacuum (V)	25.36-81.56 cm Hg	52.08	12.70
3	Relative Humidity (RH)	31.15 - 100.16%	75	14.51
4	Ambient Pressure (AP)	992.89-1033.30 mbar	1012.90	5.69
5	Electrical energy output (PE)	420.26-495.76 MW	451.55	17.06

A. DATA PREPROCESSING

Preprocessing of the data (DP) is regarded as the most crucial phase in any data-driven investigation. It provides information about the dataset's outliers, redundancies, and missing terms. The highly diverged data points from the other points are known as outliers and should be removed from the dataset [23]. In this work, a quartile-based outlier detection

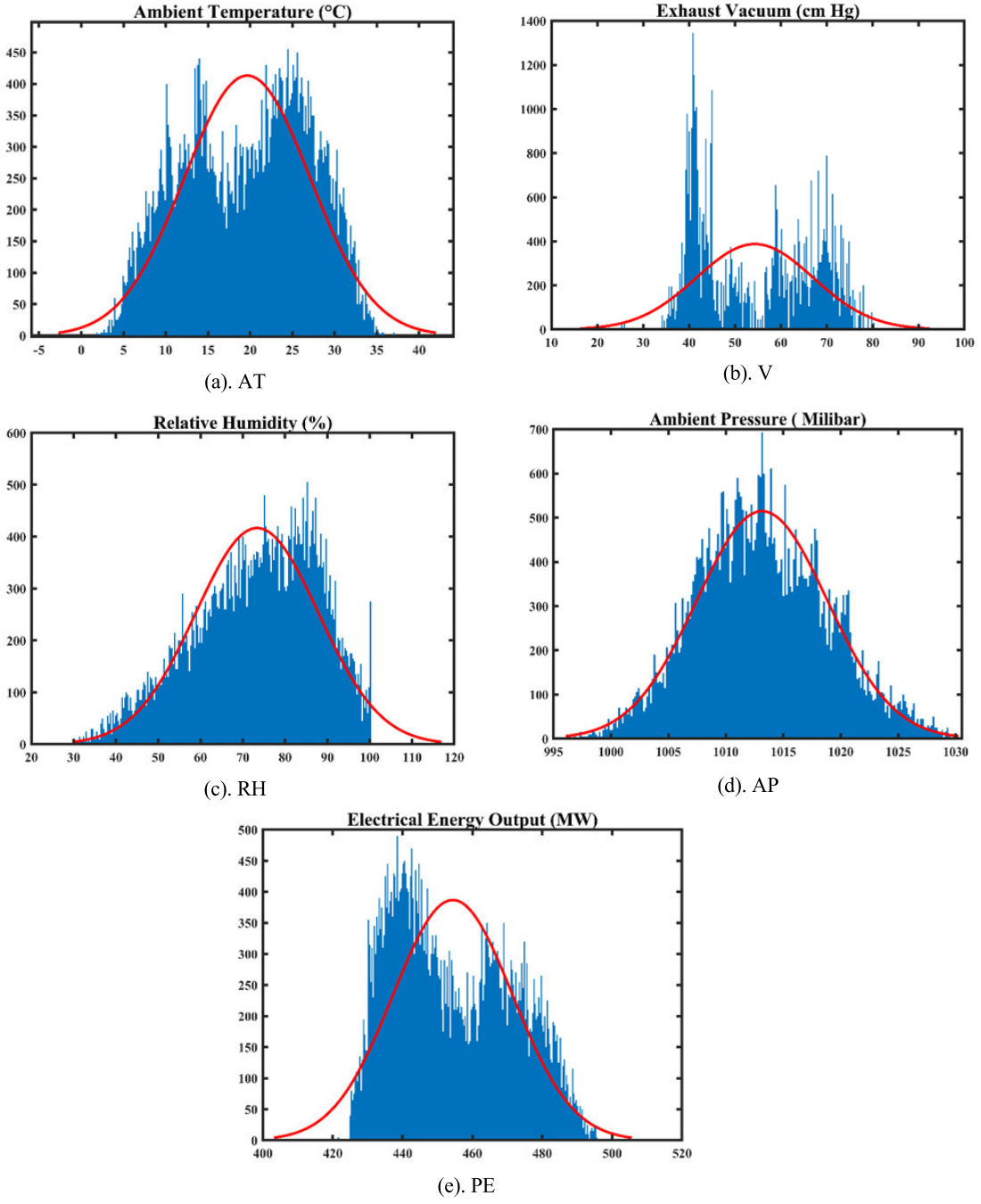


FIGURE 2. Representation of Input and output attributes of CCPP.

and rejection (O_R) method is applied, which is given by equation (1)

$$O_R(l) = \begin{cases} l & if\ Q_1 - 1.5 \times IQ_R \leq l \leq Q_3 + 1.5 \times IQ_R \\ reject & otherwise \end{cases} \quad (1)$$

where, l is the input or output attribute that lies in m-dimensional space ($l \in R_m$). Q_1, Q_3 and, IQ_R represents the 1st, 3rd and interquartile range of an input or output attribute, respectively. Further, a median by target method (equation (2)) is employed to fill the missing (M_V) terms in the attributes as follows.

$$M_V(l) = \begin{cases} median(l) & if\ l = missed/Nan \\ l & otherwise \end{cases} \quad (2)$$

After preprocessing, the next step is sampling of the dataset into training (70%), validation (15%), and testing

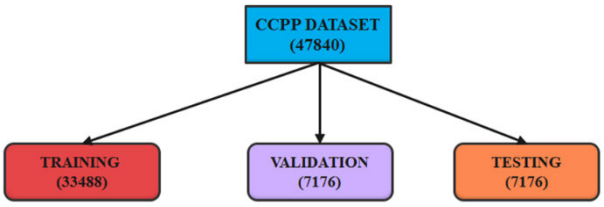


FIGURE 3. Sampling of the CCPP dataset in training, validation and testing subset.

(15%) subsets. Figure 3 shows the sampling of the CCPP dataset into three subsets as follows.

III. GENERALIZED ADDITIVE MODEL (GAM)

As per literature, various machine learning algorithms like boosted tree, RF, MLP, SVR, and DNN are utilized for the energy prediction of CCPP. These algorithms provide accurate and precise predictive regression models for low and high-dimensional predictive problems. Furthermore, in several applications, whatever is learnt is just as essential as predictive accuracy. As a result, the profound precision of complicated models comes at the cost of interpretability, i.e., the influence of a specific input on the predictive output of a complex model is cumbersome to interpret. Generalized additive model (GAM) can easily address the interpretability issue of complex models [24]. GAM is the extended version of LR models. A conventional LR models, gives the linear correlation between input and output attributes. let Y is the output attribute with normal distribution mean γ and variance η^2. The linear relationship between Y and input attributes X_j is given as follows

$$\gamma = \lambda_0 + \sum_{j=1}^{N} \lambda_j X_j \tag{3}$$

where,

γ = estimated value of Y
λ_0 = intercept
λ_j = j^{th} predictor attribute coefficient
X_j = j^{th} predictor attribute value
N = No. of predictor attributes

Furthermore, equation (3) can be rewritten by considering link function h, which relates γ to X_j, as follows

$$h(\gamma) = \lambda_0 + \sum_{j} \lambda_j X_j \tag{4}$$

The equation (4) is a functional form of generalized linear models (GLMs). GAM is the extended version of GLMs, which introduces the non-linear form of predictor attributes. Such non-linear predictors are linked to the predicted value of the dependent variables using an appropriate link function and are therefore expressed as:

$$h(\gamma) = \lambda_0 + \sum_{j} \psi_j f_j(X_j) \tag{5}$$

where,

$f_j(i)$ = j^{th} basis function
Ψ_j = parameter of j^{th} basis function

In order to improve the accuracy of the conventional GAM, a pairwise interactions are added to it, then equation (5) can be modified as [25].

$$h(\gamma) = \lambda_0 + \sum_{j} \psi_j f_j\left(X_j\right) + \sum_{j \neq i} f_{ji}\left(X_j, X_i\right) \tag{6}$$

However, training the GAM is dependent on two critical factors: (1) the shape function selection and (2) the learning algorithm used to train the GAM. In this work, boosted trees and gradient boosting are used as shape function and learning method to train the GAM.

A. GRADIENT BOOSTING (GB)

GB is a repetitive process that starts by estimating the function while considering a constant offset, which does not fit the data adequately. After each iteration, fit is improved by fitting the base learner to the negative gradient of a pre-specified error function. GB enhances the predictive performance of the model along with attribute selection and model identification. It has significant benefits over other approaches. If GB stops suddenly before getting convergent, then it improves predictive accuracy by decreasing regression coefficients to 0, a strategy similar to lasso regression, ridge regression, and shrinkage smoothing. GB is used to achieve attribute selection by setting certain components to 0. Another advantage of GB over other regression is its ability to integrate nonlinear correlations and spatial impact [26], [27]. As per the GB method, the estimation of the optimal prediction function $f*$ to realize the output attribute Y from the input attribute X is as follows.

$$f^* = \arg\min_{g} E_{Y,X}\left[\rho\left(Y, f(X)\right)\right] \tag{7}$$

where, $f*$ minimizes the cost function ρ over the all possible values of input attribute X. $f*$ can be any function that minimizes $E_{Y,X}[\rho\ (Y, f(X))]$. The correct distribution of X and Y is not known, so GB reduces the following empirical risk (ER).

$$ER = \frac{1}{n} \sum_{j=1}^{n} \rho\left(Y_j \cdot f(X_j)\right) \tag{8}$$

where ER is the approximation of $E_{Y,X}[\rho\ (Y, f(X))]$, it relates the mean to $-\rho(Y_j, f(X_j))$, $j = 1, 2, \ldots n$ of the sample site. Steps for the implementation of GB technique [28]–[30].

I. Set the initial value for the n-dimensional shape function vector $\hat{f}^{[0]}$ with a preliminary estimate $\hat{f}_1, \ldots\ldots, \hat{f}_n$ to 0.
II. Select the base-learner for GB. Set the current value of the boosting iteration to $m = 0$.
III. After incrementing the boosting iteration m by 1, evaluate the negative gradient $-\left(\frac{\partial ER}{\partial f_1}, \ldots.., \frac{\partial ER}{\partial f_n}\right)$, and $\hat{f}^{[m-1]} = \left[\hat{f}_1^{[m-1]}, \ldots\ldots, \hat{f}_n^{[m-1]}\right]$ respectively, in order to calculate $U^{[m-1]} = -\left(\frac{\partial}{\partial f_1} \rho\left(Y_j, \hat{f}_j^{[m-1]}\right)\right)_{j=1,\ldots\ldots,n}$

IV. Fit the negative gradient to each base learner to acquire the predicted output values. whereas predicted values of vector $\hat{U}^{[m-1]}$ from the excellent-fit base learner is depends on the predictor values.
V. Update the shape function by adding the fraction of $\hat{U}^{[m-1]}$ to it as $\hat{f}^{[m]} = \hat{f}^{[m-1]} + \beta\hat{U}^{[m-1]}$, where β is the step-length.
VI. Repeat steps III-V till the stopping criteria is not achieved (m_{Stop}). After that, we have the final predictive value of $\hat{f}^{[m_{Stop}]}$, which evaluates the optimum forecasting function.

Figure 4 shows the schematic representation of the overall methodology considered in this work. Furthermore, three performance indices, i.e., root mean square error (RMSE), mean absolute error (MAE), and R-squared (R^2) are considered for the performance investigation of the proposed ML algorithm towards energy prediction.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\hat{Y}_k - Y_k)^2} \tag{9}$$

$$MAE = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{Y}_k - Y_k\right| \tag{10}$$

$$R^2 = 1 - \frac{\sum_{k=1}^{N}\left(\hat{Y}_k - Y_k\right)^2}{\sum_{k=1}^{N}(Y_k - Y_{mean})^2} \tag{11}$$

where, $\hat{Y}_k$ and Y_k are predicted and actual values under the k^{th} independent variable, N is the total number of samples in the CCPP dataset.

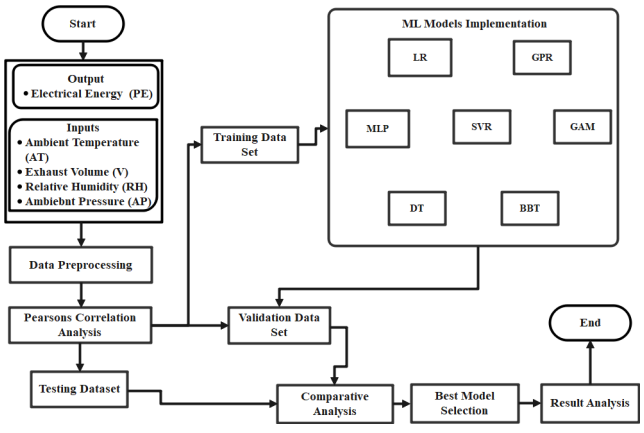


FIGURE 4. Layout of overall methodology of the work.

IV. RESULT AND DISCUSSION

As an illustration, CCPP uses gas and steam turbines to produce 50% more power in comparison to a single-cycle plant. Further, the development of a mathematical model for CCPP under full load is a tedious job due to its dependencies on various factors. Hence, a predictive model is required for the improvement of the plant's efficiency and financial operations. Therefore, in this work gradient boosting based GAM is proposed for the prediction of energy generation by CCPP. Preprocessing of the data is required before designing the predictive models in order to remove outliers. They introduce skewness and kurtosis, which make the algorithm overfit or underfit to the predicted output values.

	AT	V	AP	RH	PE
AT	1	0.8441	-0.5030	-0.54030	-0.9481
V	0.8441	1	-0.4119	-0.3109	-0.8697
AP	-0.5030	-0.4119	1	0.1055	0.5159
RH	-0.5403	-0.3109	0.1055	1	0.3883
PE	-0.94812	-0.8697	0.5159	0.3883	1

FIGURE 5. Pearson correlation matrix for CCPP.

Figure 5 shows the Pearson correlation matrix for CCPP input and output attributes after preprocessing. The Pearson correlation coefficient provides an indication of the level of gradual shift of independent parameters in order to accurately examine the influential aspects of the data. The negative value of coefficients shows the inverse correlation between the variables, whereas the positive value suggests a positive correlation. If the value of the coefficient is 0, it means both the variables are uncorrelated. It can be observed from Figure 5 that input attributes AT and V are negatively correlated to the output PE. Whereas AP and RH have a positive correlation with PE. There is a strong positive correlation between AT and V, and both the input attributes have a negative correlation with AP and RH, respectively. After preprocessing, the dataset is divided into three subsets: training (70%), validation (15%), and testing (15%). Thus, 33488, 7176, and 7176 samples have been chosen randomly as the train, validate, and test subset. It is a well-known fact that, the predictive accuracy of any ML algorithm greatly depends on the values of its hyper-parameters. A trial and error method is performed in order to evaluate the optimal values of the GAM hyper-parameters (ILRP = Initial Learn Rate for Predictors; IN = Interactions; MNSP = Maximum Number of Splits Per Predictor; ILRI = Initial Learn Rate for Interaction; NTP = Number of Trees Per Predictor; MNSI = Maximum Number of Splits Per Interaction). Firstly, GAM is trained for the randomly selected hyper-parameter values. Secondly, its performance has been evaluated on a validation dataset and, simultaneously, the values of RMSE, MAE, and R^2have been recorded. Finally, the hyper-parameter values for which RMSE and MAE are lower with a higher value of R^2 is selected. Table 3 shows the quantitative analysis of estimating the optimal values of GAM hyper-parameters. It can be observed from the table that the values of RMSE and MAE on the validation sets are least for the ILRP = 0.9987; IN = 3; MNSP = 65; ILRI = 0.9999; NTP = 47; MNSI = 30 with a higher value of correlation coefficient, respectively. So, for further investigation, these values are considered. Furthermore, predictive models for CCPP using

TABLE 3. Quantitative analysis of estimating the optimal values of GAM hyper-parameters.

S.No.	ILRP	IN	MNSP	ILRI	NTP	MNSI	RMSE (Validation)	MAE (Validation)	R^2 (%)
1	0.1254	1	60	0.2210	38	10	3.0938	2.3370	96.73
2	0.2265	2	50	0.4500	39	15	2.3269	1.7367	98.07
3	0.3825	2	55	0.4800	40	18	2.0832	1.5367	98.53
4	0.4056	3	58	0.5480	45	20	1.6362	1.2295	99.08
5	0.5698	3	60	0.6209	47	21	1.5429	1.1575	99.19
6	0.3698	2	39	0.7210	54	23	1.7849	1.3256	98.91
7	0.6675	3	67	0.6651	47	23	1.4624	1.0925	99.27
8	**0.9987**	**3**	**65**	**0.9999**	**47**	**30**	**1.1614**	**0.8716**	**99.54**
9	0.9987	4	78	0.8765	89	20	1.2267	0.9164	99.49
10	0.9880	3	70	0.9423	42	27	0.9550	1.2759	99.44

Note: Bold values denotes the best hyper-parameter values for GAM.

linear regression (LR), support vector regression (SVR), Gaussian process regression (GPR), multilayer perceptron neural network (MLP), decision tree (DT), and bootstrap-aggregated tree (BBT) ML algorithms are also designed for comparison purposes.

As discussed previously, 7176 data points are considered for validation and testing purposes. Figure 6 shows the regression plots for all the algorithms in the validation data set. According to the regression plots, GAM had the highest R^2 value of 99.54%, followed by BBT (99.32%), DT (99.21%), SVR (98.11%), GPR (95.82%), and MLP (95. 30%). However, LR has the lowest value of R^2 (94%) in comparison to other models. Table 4 demonstrates the performance comparison for all the designed algorithms in terms of RMSE and MAE. From Table 4, it can be observed that GAM attains the lowest RMSE and MAE when compared to other techniques.

TABLE 4. Quantitative comparison amongst all the ML techniques for validation data set.

Methods	RMSE (MW)	MAE (MW)
LR	4.1891	3.3660
GPR	3.4970	2.6094
MLP	3.6849	2.8201
SVR	2.3502	2.0289
DT	1.5178	0.9566
BBT	1.4099	0.9151
GAM	**1.1614**	**0.8716**

Note: Bold values denote the best performance measures among the models

The next step is to investigate the performance of the predictive models for the testing data subset. Figure 7 shows the individual tracking plots of all the designed ML algorithms for the testing dataset. It can be observed from Figure 7, that GAM is able to track the testing dataset with the highest R^2 value of 99.58%, followed by BBT (99.28%), DT (99.12%), SVR (97.93%), GPR (95.65%), and MLP (95.20%). In this case also, LR attains the lowest value of R^2 (93.70%) compared to other algorithms. Finally, Figure 8 shows the tracking performance of ML algorithms on 50 data points (6000–6050) for a better understanding of the comparison. It is observed that the predicted values by GAM are nearer the testing data points compared to the other ML algorithms.

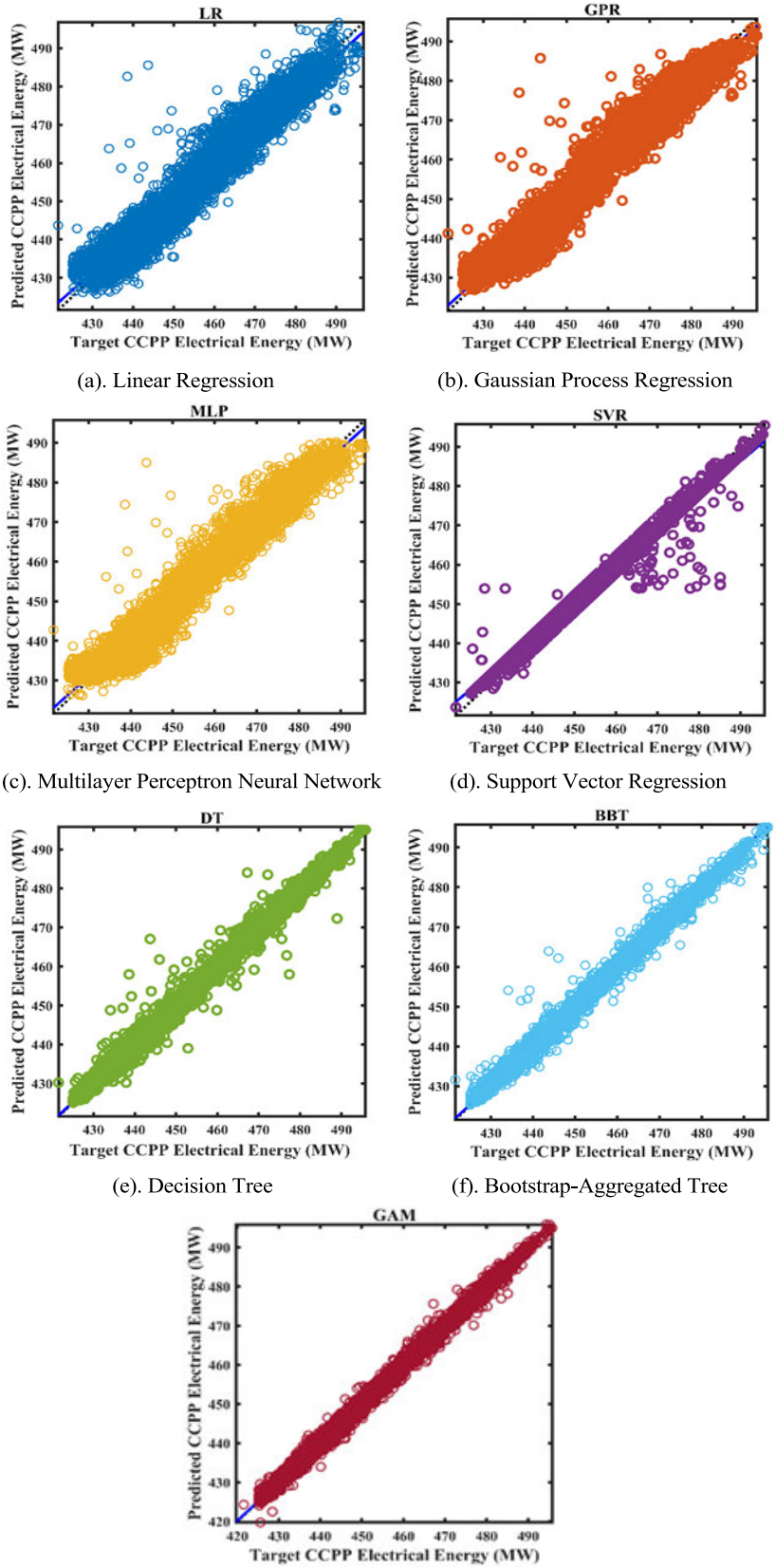
Figure 9 shows the error distribution plot of predictive models for CCPP electrical energy prediction. Table 5 displays the maximum and minimum error deviations for all the predictive models. The maximum and minimum deviations attain by GAM are 11.2470 and −10.9319, respectively. Hence, from table 5 it can be revealed that the deviation of the error is less in the case of GAM in comparison to other ML algorithms. It can also be further concluded by the visual inspection of Figure 9. In addition to this, a non-parametric 'Mann–Whitney U' test [31] is also performed to investigate the normality and probability distribution of actual and predicted values for all the developed models.

The M-W test compared the actual and predictive outputs to investigate whether both the outputs are derived from the same distribution or whether there is a difference in their median values. Table 6 shows the outcomes of the M-W test for all the predictive models. After comprehensive analysis, it can be observed that the Z value is highest for SVR (1.8437) and the smallest for GAM (0.0294), respectively. There is a homogeneity in the 1t-P and 2t-P values, meaning no large deviations are observed. Further, GAM has the larger values of 1t-P (0.4882) and 2t-P (0.9764) compared to other techniques, which shows the effectiveness of the proposed model.

Furthermore, the performance of GAM with other designed models and models existing in the literature are estimated on the basis of performance indexed (PI) and rank analysis values. For this analysis, all the relationships evaluated using RMSE and MAE presented in Table 7 are considered. The PI can be defined by following equation [32].

$$PI_a = \frac{1}{2}\left(\frac{RMSE_a}{RMSE_{\max}} + \frac{MAE_a}{MAE_{\max}}\right) \tag{12}$$

where, a is related to every predictive model. From table 6, it can be seen that predictive models based on GAM (0.2514), BBT (0.2830), DT (0.2987) and SVR (0.5925) have better accuracy in comparison to existing models. However, the predictive model developed in [11] has a PI value equal



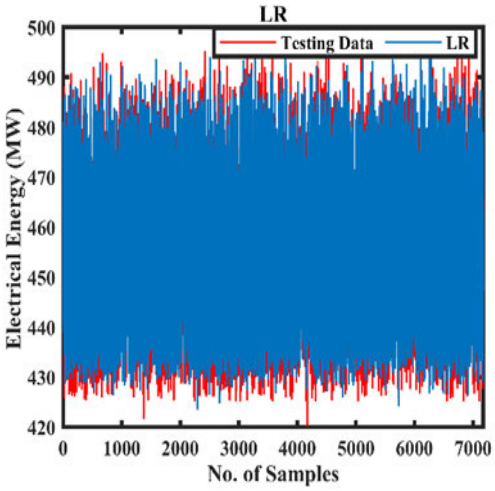
(a). Linear Regression (b). Gaussian Process Regression

(c). Multilayer Perceptron Neural Network (d). Support Vector Regression

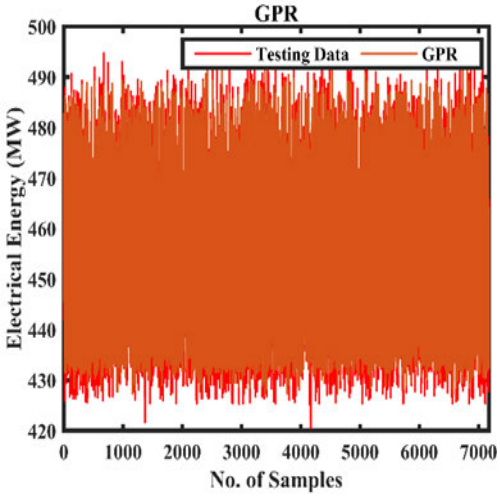
(e). Decision Tree (f). Bootstrap-Aggregated Tree

(g). Generalized Additive Model

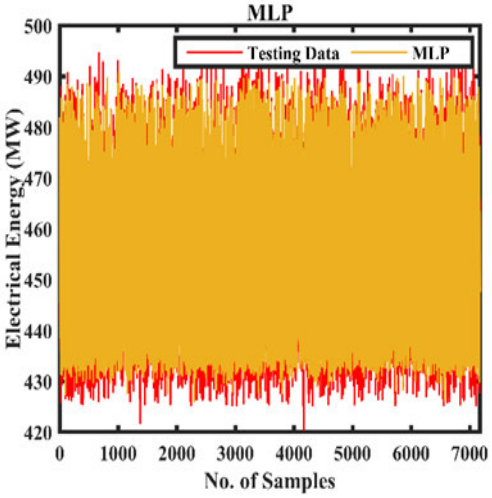
FIGURE 6. The regression plots of all the designed ML techniques predictive models for validation data set.



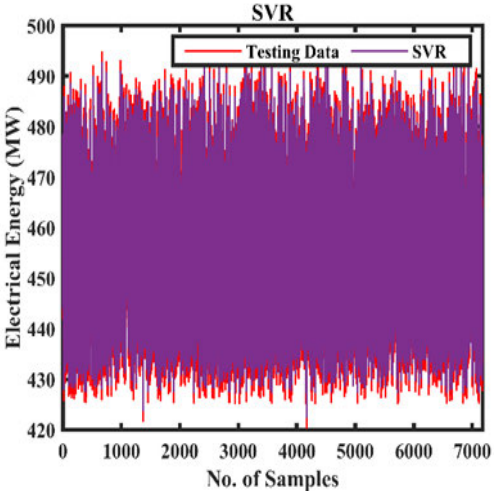
(a). Linear Regression



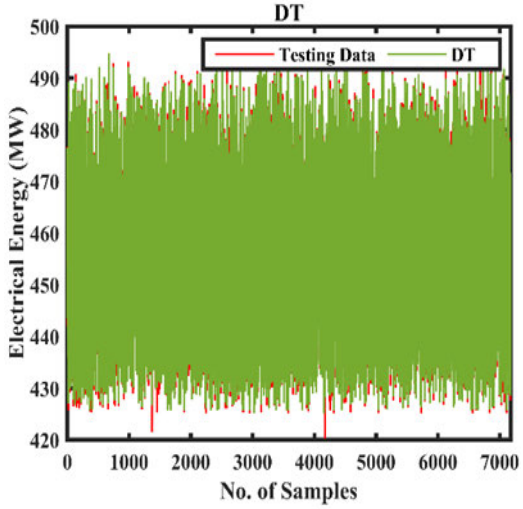
(b). Gaussian Process Regression



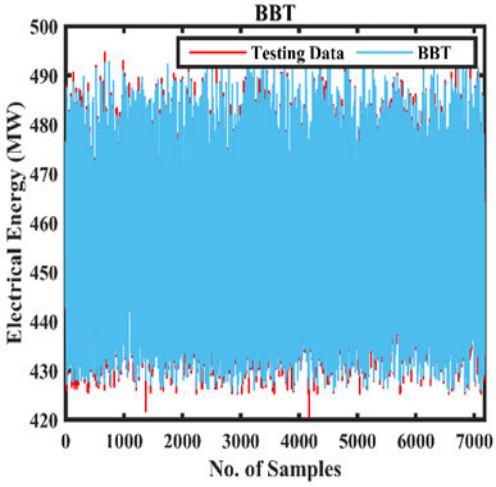
(c). Multilayer Perceptron Neural Network



(d). Support Vector Regression

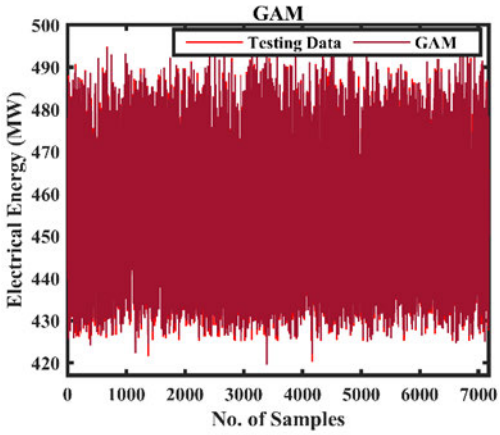


(e). Decision Tree



(f). Bootstrap-Aggregated Tree

FIGURE 7. **Prediction performance of all the designed ML techniques for testing data.**



(g). Generalized Additive Model

FIGURE 7. *(Continued.)* **Prediction performance of all the designed ML techniques for testing data.**

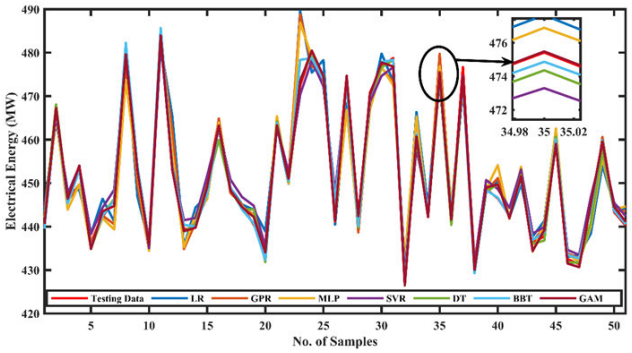


FIGURE 8. **Prediction performance of all the designed ML techniques for 50 testing samples.**

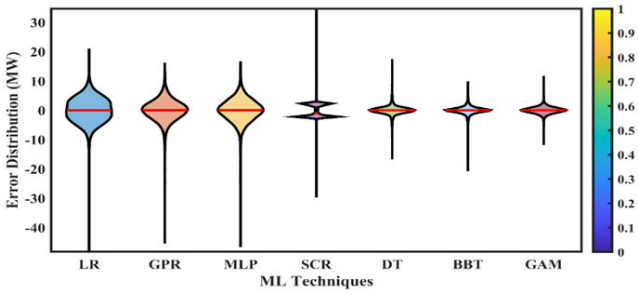


FIGURE 9. **The error distribution of all the ML models in violin plot for CCPP electrical energy prediction.**

to 0.7677, which is better than LR (1), MLP (0.8581) and GPR (0.8091), respectively

A. UNCERTAINTY ANALYSIS

The confidence ranges of forecast errors ($CL^{\pm}$) are calculated by the given equation (13) to measure the uncertainty related with all the predictive models [33].

$$\text{uncertainity band} = CL^{+} - CL^{-}$$
$$\text{where, } CL^{\pm} = \xi \pm D_{\lambda}\omega. \tag{13}$$

TABLE 5. **Quantitative analysis of error distribution for all designed ML techniques.**

ML models	Maximum Error Deviation (MW)	Minimum Error Deviation (MW)
LR	18.6760	-45.8796
GPR	14.5893	-43.4150
MLP	14.8498	-44.4359
SVR	33.7076	-28.4276
DT	16.9827	-15.8400
BBT	9.4219	-19.9590
GAM	**11.2470**	**-10.9319**

Note: Bold values denote the best performance measures among the models

TABLE 6. **Mann–Whitney U test.**

ML models	Z value	1t-P value	2t-P value [0-1]
LR	0.4083	0.3415	0.6830
GPR	0.0814	0.4675	0.9351
MLP	0.0529	0.4788	0.9577
SVR	1.8437	0.0326	0.0652
DT	0.0488	0.4805	0.9610
BBT	0.0527	0.4789	0.9579
GAM	**0.0294**	**0.4882**	**0.9764**

Note: Bold values denote the best performance measures among the models.

where, ξ and ω are the mean and standard deviation of the forecasted error, respectively. D_{λ} is the standard variable with λ % of significance level. Figure 10 shows the uncertainty band bar graph of all the models for CCPP electricity generation prediction. It can be observed that GAM possesses the lowest forecasted uncertainty value of 4.3773 for 5 % of significance level. On the other hand, BBT, DT, SVR, GPR, MLP, and LR have uncertainty band values equal to 5.2978, 5.5546, 9.6780, 14.0466, 14.7453, and 16.8523 respectively. LR has the highest level of forecasted uncertainty. Finally, it can be concluded from the above findings that GAM is superior and robust in comparison to other designed models developed in this study.

B. SENSITIVITY ANALYSIS

One of the most important components of a forecasting model is sensitivity assessment, which evaluates the significance of every input attribute to the forecasting of the target attribute. The level of dependency (S_A) between target attribute (O_j) and predictive variables (I_i) is given as follows [34], [35]

$$S_{A(i,j)} = \frac{\sum_{k=1}^{N} I_{i,k} O_{j,k}}{\sqrt{\sum_{k=1}^{N} I_{i,k}^2 \sum_{k=1}^{N} O_{j,k}^2}} \tag{14}$$

TABLE 7. Performance assessment of GAM in comparison to other designed models and models existing in the literature.

Methods	RMSE (MW)	MAE (MW)	PI	Rank
LR	4.2553	3.3660	1	10
GPR	3.5469	2.6417	0.8091	7
MLP	3.7233	2.8317	0.8581	8
SVR	2.4457	2.0547	0.5925	4
DT	1.4026	0.9016	0.2987	3
BBT	1.3378	0.8473	0.2830	2
GAM	**1.1053**	**0.8187**	**0.2514**	**1**
Hundi & Shahsavari [11]	3.5000	2.4000	0.7677	5
Tüfekci [3]	3.8870	2.9770	0.8989	9
Qu et.al. [12]	3.4320	2.5840	0.7871	6

Note: Bold values denote the best performance measures among the models

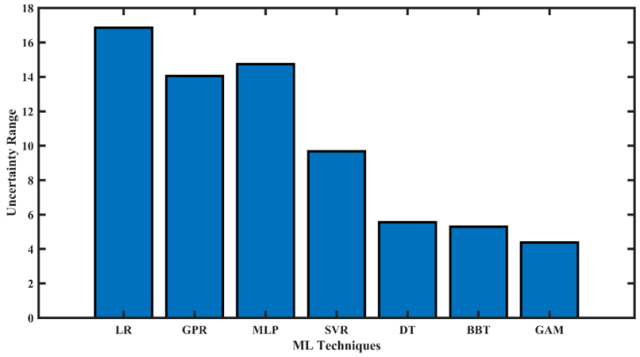


FIGURE 10. Uncertainty band of all the models for CCPP electrical energy prediction.

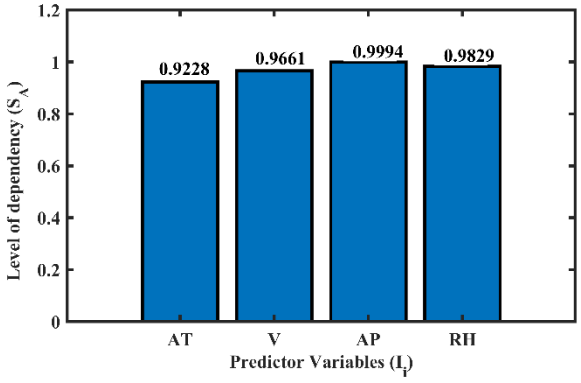


FIGURE 11. Sensitivity Analysis of predictor variables for CCPP electrical energy prediction.

whereas, for every I_i the higher value of $S_{A\,(i,j)}$, shows the greater dependency of that predictive variable on the target attribute (O_j). Figure 11 shows that the S_A values for the input attributes AP (0.9994) and RH (0.9829) are higher than AT (0.9228) and V (0.9661), respectively. This suggests that AP and RH are the most important elements in estimating PE.

V. CONCLUSION

In this article, a gradient boosted generalized additive model (GAM) ML algorithm is proposed for the development of a predictive model for combined cycle power plant (CCPP). Initially, preprocessing of a CCPP dataset is completed by removing the outliers using a quartile-based method and replacing the missing values using the median method. The next step is to split the preprocessed dataset into training, validation, and test subsets. Furthermore, optimal values of GAM hyper-parameters are estimated using the trial and error method. In addition to this, predictive models based on LR, GPR, MLP, SVR, DT, and BBT are also designed for the performance comparison of GAM. The performance of the presented models has been analyzed with different statistical measures like RMSE, MAE, and R^2 respectively. A detailed comparison has been carried out among all the predictive models on the basis of violin plots and the nonparametric M-W test. Results also suggested that GAM shows the best performance amongst the seven models, with RMSE = 1.1053, MAE = 0.8187, and PI = 0.2514. Finally, an uncertainty analysis was also conducted for all the models. GAM shows the least uncertainty in predicting the electrical energy of CCPP. As a result of this study and the overall review, it can be said that the proposed model has a better ability to improve plant reliability and financial performance than other predictive models.

REFERENCES

[1] U. Kesgin and H. Heperkan, "Simulation of thermodynamic systems using soft computing techniques," *Int. J. Energy Res.*, vol. 29, no. 7, pp. 581–611, 2005.
[2] A. Ticà, H. Guéguen, D. Dumur, D. Faille, and F. Davelaar, "Design of a combined cycle power plant model for optimization," *Appl. Energy*, vol. 98, pp. 256–265, Oct. 2012.
[3] P. Táfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *Int. J. Electr. Power Energy Syst.*, vol. 60, pp. 126–140, Sep. 2014.
[4] J. L. Calvo-Rolle and E. Corchado, "A bio-inspired knowledge system for improving combined cycle plant control tuning," *Neurocomputing*, vol. 126, pp. 95–105, Feb. 2014.
[5] B. Akdemir, "Prediction of hourly generated electric power using artificial neural network for combined cycle power plant," *Int. J. Electr. Energy*, vol. 4, no. 2, pp. 91–95, 2016.
[6] G. Ahn and S. Hur, "Continuous conditional random field model for predicting the electrical load of a combined cycle power plant," *Ind. Eng. Manage. Syst.*, vol. 15, no. 2, pp. 148–155, Jun. 2016.
[7] J. Janouáek, P. Gajdoá, P. Dohnálek, and M. Radecký, "Towards power plant output modelling and optimization using parallel regression random forest," *Swarm Evol. Comput.*, vol. 26, pp. 50–55, Feb. 2016.
[8] E. A. Elfaki and A. H. Ahmed, "Prediction of electrical output power of combined cycle power plant using regression ANN model," *J. Power Energy Eng.*, vol. 6, no. 12, pp. 17–38, 2018.
[9] I. Lorencin, V. Mrzljak, and Z. Car, "Genetic algorithm approach to design of multi-layer perceptron for combined cycle power plant electrical power output estimation," *Energies*, vol. 12, no. 22, p. 4352, Nov. 2019.
[10] D. A. Wood, "Combined cycle gas turbine power output prediction and data mining with optimized data matching algorithm," *Social Netw. Appl. Sci.*, vol. 2, no. 3, Mar. 2020.
[11] P. Hundi and R. Shahsavari, "Comparative studies among machine learning models for performance estimation and health monitoring of thermal power plants," *Appl. Energy*, vol. 265, May 2020, Art. no. 114775.
[12] Z. Qu, J. Xu, Z. Wang, R. Chi, and H. Liu, "Prediction of electricity generation from a combined cycle power plant based on a stacking ensemble and its hyperparameter optimization with a grid-search method," *Energy*, vol. 227, Jul. 2021, Art. no. 120309.

[13] M. Karaçor, A. Uysal, H. Mamur, G. Åžen, M. Nil, M. Z. Bilgin, H. Doáan, and C. Åahin, "Life performance prediction of natural gas combined cycle power plant with intelligent algorithms," *Sustain. Energy Technol. Assessments*, vol. 47, Oct. 2021, Art. no. 101398.
[14] H. Moayedi and A. Mosavi, "Electrical power prediction through a combination of multilayer perceptron with water cycle ant lion and satin bowerbird searching optimizers," *Sustainability*, vol. 13, no. 4, p. 2336, Feb. 2021.
[15] Y. D. Arferiandi, W. Caesarendra, and H. Nugraha, "Heat rate prediction of combined cycle power plant using an artificial neural network (ANN) method," *Sensors*, vol. 21, no. 4, p. 1022, Feb. 2021.
[16] H. Preeti, R. Bala, A. Dagar, and R. P. Singh, "A novel online sequential extreme learning machine with $L_{2,1}$-norm regularization for prediction problems," *Int. J. Speech Technol.*, vol. 51, no. 3, pp. 1669–1689, Mar. 2021.
[17] A. Afzal, S. Alshahrani, A. Alrobaian, A. Buradi, and S. A. Khan, "Power plant energy predictions based on thermal factors using ridge and support vector regressor algorithms," *Energies*, vol. 14, no. 21, p. 7254, Nov. 2021.
[18] G. Revathy, S. Z. Affan, M. Suriya, P. S. Kumar, and V. Rajendran, "Optimization study on competence of power plant using gas/steam fluid material parameters by machine learning techniques," *Mater. Today, Proc.*, vol. 37, pp. 1713–1720, Dec. 2021.
[19] C. A. Saleel, "Forecasting the energy output from a combined cycle thermal power plant using deep learning models," *Case Stud. Thermal Eng.*, vol. 28, Dec. 2021, Art. no. 101693.
[20] L. X. Niu and X. J. Liu, "Multivariable generalized predictive scheme for gas turbine control in combined cycle power plant," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, Sep. 2008, pp. 791–796.
[21] V. Ramireddy. *An Overview of Combined Cycle Power Plant*. Accessed: Feb. 3, 2013. [Online]. Available: http://electricalengineering-portal.com/an-overview-of-combined-cycle-power-plant
[22] D. C. Dua, *UCI Machine Learning Repository* (School of Information and Computer Science). Irvine, CA, USA: Univ. California, 2019.
[23] H. Gupta, H. Varshney, T. K. Sharma, N. Pachauri, and O. P. Verma, "Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction," *Complex Intell. Syst.*, vol. 2021, pp. 1–15, May 2021.
[24] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 150–158.
[25] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 623–631.
[26] S. Dhulipala and G. R. Patil, "Freight production of agricultural commodities in India using multiple linear regression and generalized additive modelling," *Transp. Policy*, vol. 97, pp. 245–258, Oct. 2020.
[27] K. O. Maloney, M. Schmid, and D. E. Weller, "Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages," *Methods Ecology Evol.*, vol. 3, no. 1, pp. 116–128, Feb. 2012.
[28] T. Hastie and R. Tibshirani, *Generalized Additive Models*. Evanston, IL, USA: Routledge, 2017.
[29] L. Xue, A. Qu, and J. Zhou, "Consistent model selection for marginal generalized additive model for correlated data," *J. Amer. Stat. Assoc.*, vol. 105, no. 492, pp. 1518–1530, Dec. 2010.
[30] S. N. Wood, *Generalized Additive Models: An Introduction with R*. Boca Raton, FL, USA: CRC Press, 2006.
[31] N. Kardani, A. Bardhan, D. Kim, P. Samui, and A. Zhou, "Modelling the energy performance of residential buildings using advanced computational frameworks based on RVM, GMDH, ANFIS-BBO and ANFIS-IPSO," *Jour. Buil. Eng.*, vol. 35, Dec. 2021, Art. no. 102105.
[32] M. Gholizadeh, M. Jamei, I. Ahmadianfar, and R. Pourrajab, "Prediction of nanofluids viscosity using random forest (RF) approach," *Chemometric Intell. Lab. Syst.*, vol. 201, Jun. 2020, Art. no. 104010.
[33] I. Ahmadianfar, M. Jamei, M. Karbasi, A. Sharafati, and B. Gharabaghi, "A novel boosting ensemble committee-based model for local scour depth around non-uniformly spaced pile groups," *Eng. Comput.*, vol. 2021, pp. 1–23, Mar. 2021.
[34] A. Bahrami, M. Monjezi, K. Goshtasbi, and A. Ghazvinian, "Prediction of rock fragmentation due to blasting using artificial neural network," *Eng. Comput.*, vol. 27, no. 2, pp. 177–181, Apr. 2011.
[35] Y.-H. Jong and C.-I. Lee, "Influence of geological conditions on the powder factor for tunnel blasting," *Int. J. Rock Mech. Mining Sci.*, vol. 41, no. 3, p. 461, Apr. 2004.

NIKHIL PACHAURI (Member, IEEE) received the bachelor's degree in electronics and instrumentation from the Institute of Engineering and Technology, M. J. P. Rohilkhand University, Bareilly, India, in 2009, the M.Tech. degree in control and instrumentation from the National Institute of Technology Jalandhar, India, in 2012, and the Ph.D. degree in instrumentation and control engineering from the University of Delhi, India, in 2018. He is currently working as a Postdoctoral Research Fellow with the Gwangju Institute of Science and Technology, Gwangju, South Korea. He has published several research articles in international journals and conferences. His research interests include process control, biomedical signal processing and control, artificial intelligence, and energy management.

CHANG WOOK AHN (Member, IEEE) received the Ph.D. degree from the Department of Information and Communications, Gwangju Institute of Science and Technology (GIST), South Korea, in 2005. From 2005 to 2007, he worked with the Samsung Advanced Institute of Technology, South Korea. From 2007 to 2008, he was a Research Professor with the GIST. From 2008 to 2016, he was an Assistant/Associate Professor with the Department of Computer Engineering, Sungkyunkwan University (SKKU), South Korea. He is currently working as a Professor with the School of Electrical Engineering and Computer Science, GIST. His research interests include genetic algorithms/programming, multi-objective optimization, neural networks, and quantum machine learning.