Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9966395, 13 pages
https://doi.org/10.1155/2021/9966395

Research Article

Power Prediction of Combined Cycle Power Plant (CCPP) Using Machine Learning Algorithm-Based Paradigm

Raheel Siddiqui,[1] Hafeez Anwar,[1] Farman Ullah,[1] Rehmat Ullah,[2] Muhammad Abdul Rehman,[1] Naveed Jan,[3] and Fawad Zaman[4]

[1]*Department of Electrical & Computer Engineering, COMSATS University Islamabad-Attock Campus, Pakistan*
[2]*Department of Computer Systems Engineering, University of Engineering and Technology Peshawar, Peshawar, Pakistan*
[3]*Department of Information Engineering Technology, University of Technology Nowshera, KPK, Pakistan*
[4]*Department of Electrical & Computer Engineering, COMSATS University Islamabad, Pakistan*

Correspondence should be addressed to Hafeez Anwar; hafeez.anwar@cuiatk.edu.pk
and Farman Ullah; farmankttk@ciit-attock.edu.pk

Received 29 March 2021; Revised 17 July 2021; Accepted 3 November 2021; Published 23 December 2021

Academic Editor: Javier Prieto

Power prediction is important not only for the smooth and economic operation of a combined cycle power plant (CCPP) but also to avoid technical issues such as power outages. In this work, we propose to utilize machine learning algorithms to predict the hourly-based electrical power generated by a CCPP. For this, the generated power is considered a function of four fundamental parameters which are relative humidity, atmospheric pressure, ambient temperature, and exhaust vacuum. The measurements of these parameters and their yielded output power are used to train and test the machine learning models. The dataset for the proposed research is gathered over a period of six years and taken from a standard and publicly available machine learning repository. The utilized machine algorithms are K-nearest neighbors (KNN), gradient-boosted regression tree (GBRT), linear regression (LR), artificial neural network (ANN), and deep neural network (DNN). We report state-of-the-art performance where GBRT outperforms not only the utilized algorithms but also all the previous methods on the given CCPP dataset. It achieves the minimum values of root mean square error (RMSE) of 2.58 and absolute error (AE) of 1.85.

1. Introduction

The accurate prediction of power generated by a plant helps in reducing various related issues such as power outages, economic, and technical difficulties [1, 2]. In particular, an inaccurate prediction results in the rise of per unit cost of electric power [3] due to the high fuel consumption. Hence, in this paper, we aim at achieving a precise prediction of electric power of a base load CCPP on full load conditions thus ensuring decreased cost of per unit of electric power [4].

The power of thermodynamic power stations can be calculated using complex mathematical models [5]. These models involve a vast range of assumptions and parameters to reflect the actual uncertainty of the system. However, these mathematical models are time consuming and are based on a deterministic approach [5]. On the other hand, supervised machine learning (ML) algorithms incorporate probabilistic approaches for power prediction instead of mathematical modeling [6]. With the availability of data, the prediction done with the ML approach is far more convenient, scalable, and flexible thus preventing to model the whole the system. It can also be observed in other similar approaches, for instance, predicting groundwater hardness vulnerability [7], estimating soil erosion susceptibility [8], groundwater level prediction [9], and groundwater potential prediction [10]. Our proposed ML algorithms assess historical data of a power plant, operating under a variety of environmental conditions, in order to provide optimized power forecasts in less time [11]. However, being probabilistic in nature, the predictions made by ML algorithms involve errors. Due to this reason, we propose to evaluate several algorithms for the task of

power prediction of a CCPP. Furthermore, we also search for the parameters of these algorithms where they give the least error on the current dataset.

The generated electric power of a CCPP unexpectedly fluctuates throughout the whole year due to several parameters, such as ambient temperature, atmospheric pressure, humidity, and exhaust vacuum [12]. Consequently, these parameters directly and indirectly influence the output power [13] of a CCPP. Therefore, power production can be improved and fuel consumption can be reduced by optimally controlling these parameters [14]. The primary focus of this research is to analyze the influence of ambient parameters on output power prediction rather than controlling the parameters. For this purpose, these environmental parameters are used to predict the electric power through various machine learning algorithms [4].

Nonetheless, various probabilistic approaches are previously used for CCPP power prediction including bagging and regression ANN [4, 15]. However, their prediction error is slightly high [16], due to which, this paper proposes the machine learning algorithms for the prediction of electric power of CCPP operated on full load using the previously mentioned four parameters. Gradient-boosted regression tree (GBRT), linear regression (LR), artificial neural network (ANN), and K-nearest neighbor (KNN) are used to improve the power prediction. The individual and cumulative effects of each parameter are evaluated on output power prediction using these four machine learning algorithms. These algorithms are compared with the previous research work to find improved results. The best performance among these four machine learning models is analyzed by choosing the least RMSE and AE values.

The rest of the paper is organized as follow: Section 2 outlines the related work. Section 3 introduces the proposed methodology, and the results and discussion are briefly discussed in Section 4. Finally, we conclude the paper in Section 5.

2. Related Work

Pourbeik et al. [17] proposed the basic working of a combined cycle power plant (CCPP). The main parts of the power plant are gas turbine, steam turbine, heat recovery steam generator (HRGS), and electric generator. In the gas turbine, compressed high-pressure air combined with fuel inside the combustion chamber, which makes hot pressurized air to strike the turbine blades and shaft coupled with electric generator for power generation. Due to the usage of gas turbine, the temperature of residual exhaust air is still high, and thus, it is used as an input of the heat recovery steam generator (HRSG). After that, hot air converts water into steam which rotates the steam turbine and power is generated by using both gas and steam turbines. The overall efficiency of CCPP is about 55 percent. Tfekci [4] presented to forecast the hourly based electrical power of a combined cycle power plant (CCPP) operated on base load for full load conditions. The benefit of operating in full load is to escalate the turnover of available hours. Several machine learning regression algorithms used for increasing the performance of predicting models and compared with their output results in the form of RMSE to find the best results. Firstly, it fetched the best subset that contained all input features. Secondly, it identified the best regression algorithm among the other fifteen algorithms. As a result, the bagging algorithm with the REPtree [4] algorithm to observed the best method applied to the best subset with a minimum RMSE of 3.787 and absolute error (AE) of 2.818. Yeom and Kwak [18] presented the Takai-Sugeno-Kang- (TSK-) based extreme learning machine (ELM) for power estimation. This algorithm is designed by an efficient approach to generate automatic fuzzy if then rules. It has mainly two steps. Firstly, the initial matrix of random partitions generated, and cluster centers are calculated for random clustering. These centers are used to decide the nearer part of fuzzy rules. Secondly, least square estimate (LSE) used to estimating the linear parameters of the TSK fuzzy type. The results depict that the TSK ELM method showed less RMSE of 3.93 as compared to ordinary ELM as shown in Table 1.

Lorencin et al. [16] proposed a genetic algorithm (GA) approach to a multilayer perceptron (MLP) design in order to predict of the CCPP electrical power output. A heuristic algorithm to increase the regression performance of MLP compared to those available in the literature. The GA was applied by using crossover procedures and processes based on mutation. These methods are implemented in 50 different generations for the design of 20 different chromosomes. Using Bland-Altman (BA) analysis, MLP configurations, which are devised with GA implementation, are validated. Five hidden layers of MLP, 25, 80, 65, 75, and 80 nodes using GA, respectively, are built. K-fold cross-validation is performed to assess average performance of the abovementioned MLP. The RMSE value obtained with the abovementioned MLP is 4.305, which is considerably less than the MLP provided in the existing literature, but still greater than numerous complicated algorithms, such as K star and tree-based algorithms. Bandić et al. [22] described the random forest algorithm for estimating the output power of a CCPP at full load. The analysis is conducted between twofold where in the first fold all the features are utilized, whereas in the second fold only three features are used. Random forest, random tree, and adaptive neurofussy inference system (ANFIS) algorithms are used for power prediction. Performance evaluated by taking RMSE, absolute error (AE), and correlation of predicted model. After applying all algorithms, the best result is obtained by the random forest. Results acquired on all features showed RMSE of 3.0271 MW, where 90 percent of the data used for training while the remaining 10 percent used for testing. Elfaki and Ahmed [12] presented the regression artificial neural network (ANN) to estimate the electric power of the CCPP. A total of seven experiments are performed on the ANN model. Each time the performance of the ANN model is different due to random initial weights and bias. It perceived that the performance improved by increasing the number of data records of available features, and performance oscillates when new features are added into the dataset. Consequently, higher number of training records led to increased performance. After that, correlations among features and then between features and output parameters are calculated. Two training functions are applied to the same datasets and performance is observed. In

Table 1: Achieved results in the literature on the current dataset.

Sr. no	Algorithm	RMSE	AE
1	KNN [19](2012)	N/A	3.51
2	Bagging with REPTree [4] (2014)	4.23	3.22
3	C-CRF [20] (2016)	3.97	2.97
4	TSK-ELM [18] (2018)	3.93	N/A
5	Regression ANN [12] (2018)	4.32	N/A
6	MLP with GA [16] (2019)	4.30	N/A
7	TOB Matching [21] (2020)	2.89	N/A

this study, Bayesian regularization showed better performance than the Levenberg-Marquardt method. Moreover, the output result is compared with original result, and it is almost the same with a minimal value of standard deviation. The error between the original and output results also the least. Şen et al. [13] described the vital role of input features to estimate the output power of CCPP. There are four input features used in CCPP, namely, ambient temperature, pressure, humidity, and vacuum. One main feature among all others is the ambient temperature, which causes significant variation in the performance of CCPP. Depending on the environmental conditions, the temperature also deviates between 8°C to 23°C. The efficiency of CCPP observed at 8°C is 42.7% and generated power is 227.7 MW. On the other hand, at 23°C temperature, the observed efficiency of 43.3% and generated power of 197.3 MW are recorded. This happens when the inlet air temperature of the gas turbine increases, which in turn decreases the amount of oxygen in the air per unit volume. Due to less amount of oxygen, the burning rate becomes lower in the combustion chamber. It has a negative influence on the output power of the gas turbine. Thus, combustion will be higher when the amount of oxygen is high. Therefore, for maximum power generation, a proper cooling system should be installed on the inlet side, which could reduce the temperature of air. Rashid et al. [23] proposed a novel approach of swarm optimization-based feed-forward neural network that used to design the predicting model. All input variables of the plant are used as an input of the feed-forward neural network. Particle swarm optimization (PSO) is a learning algorithm. Performance is analyzed by using mean square error (MSE). PSO trained feed-forward neural network which depicted favorable results for power estimation. MSE of the training dataset is observed to be 1.019*e*-04; on the other hand, for the testing dataset, MSE is 0.0055. Burkov [24] presented an accurate and reliable way of estimating the hourly electrical power of a combined cycle power plant. For designing, the local and global predictive models, many algorithms are used in this paper, such as additive regression, *K*-means clustering, feed-forward ANN, KNN, and conventional multivariate regression. The model's performance is analyzed by using a mean relative error (MRE%) and mean absolute error (MAE). Among all the mentioned algorithms, KNN results are found more efficient and reliable, with a relative error of less than 1. Han [25] proposed the computational intelligence algorithm like tree architecture of fuzzy neural networks that are used for power prediction. It has a benefit of selecting the minimum rules by opting neurons as nodes and significant inputs as leaves. There are two primary optimization method genetic algorithm (GA) and random signal-based learning (RSA) for prediction. GA optimized the binary structure of the networks by opting for the leaves and nodes as binary. RSA further refines the binary connection in the interval. 70% of data is used for training and 30% used for testing. Performance is accessed by using RMSE. The result of GA for testing data depicts RMSE of 3.31. Most recently, the so-called transparent open-box machine learning algorithm is used by [21] to achieve an RMSE of 2.89% on the dataset. However, some of the records from the UCI CCPP dataset are removed to achieve this reduced RMSE.

3. Proposed Methodology for Power Prediction of Combined Cycle Power Plants Using Machine Learning (ML) Algorithms

This section explains the methodology of power prediction using machine learning algorithms. The main steps in this research are feature extraction of the collected data from sensors, performance evaluation of the ML algorithms, and performance parameter calculation. The prime goal is to find the algorithm that predicts the power of a CCPP with least error. For this purpose, four ML algorithms are evaluated which are gradient-boosted regression tree (GBRT), *K* -nearest neighbor (KNN), linear regression (LR), and artificial neural networks (ANN). The prediction is done by taking four parameters related to the CCPP, which are ambient inlet air temperature, atmospheric pressure, relative humidity, and vacuum (gas turbine exhaust pressure) are used. Results are predicted using Rapid Miner [26] which is a machine learning and data mining software suite. Figure 1 depicts the proposed architecture for the prediction of CCPP power using ML algorithms.

3.1. Dataset. Table 2 depicts a part of the whole dataset where the complete dataset comprises of 9568 records gathered from an operational CCPP over a period of 6 years. This dataset is taken from the UCI machine learning repository [4, 19].

3.2. Feature Extraction. The features acquired from the sensors data are measurements per unit time, mean, variance, standard deviation cross-correlation, autocorrelation, maximum value, and minimum value.

3.2.1. Measurements per Unit Time. The hourly based data (averaged) is obtained from various sensors installed outside the plant which record ambient variables every second. The values of these variables are used without being normalized.

3.2.2. Mean Value. There are four features and the mean value of each feature is given in the dataset of CCPP. Total number of samples is 9568.

$$\mu = \left(\frac{1}{N}\right)\sum_{i=0}^{N} x_{i,} \tag{1}$$

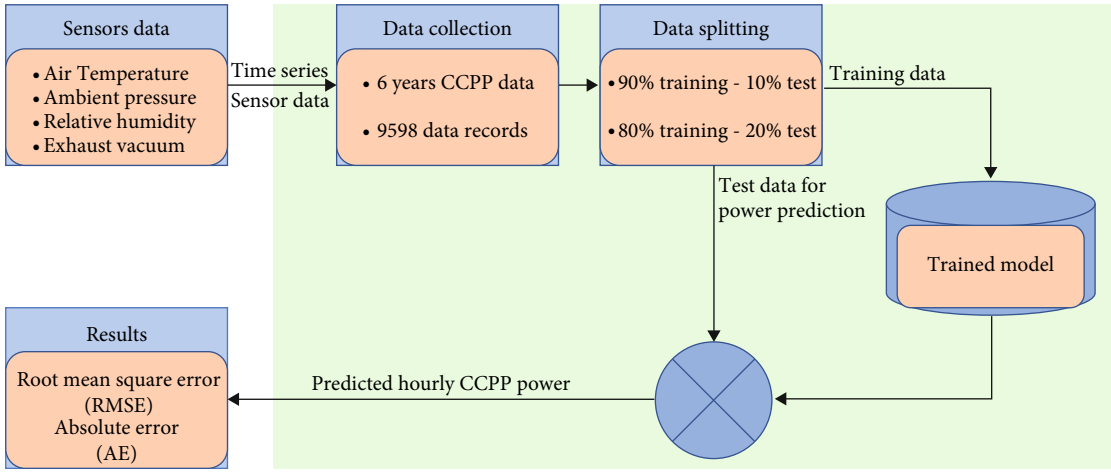

Figure 1: Proposed architecture of the prediction of CCPP power using ML algorithm.

Table 2: Input and output attributes of CCPP, where the temperature is measured in celcius, vaccum is measure in cm-Hg, pressure is measured in millibars, humidity is in percentage, and the unit of predicted power is megawatts.

Temperature	Vacuum	Pressure	Humidity	Predicted power
14.96	41.76	1024.07	73.17	463.26
25.18	62.96	1020.04	59.08	444.37
5.11	39.4	1012.16	92.14	488.56
20.86	57.32	1010.24	76.64	446.48
10.82	37.5	1009.23	96.62	473.9
26.27	59.44	1012.23	58.77	443.67
15.89	43.96	1014.02	75.24	467.35
9.48	44.71	1019.12	66.43	478.42
14.64	45	1021.78	41.25	475.98
11.74	43.56	1015.14	70.72	477.5

where N represents the total number of samples of one feature and x is the sensor output at i^{th} sample [27].

3.2.3. Variance. It is used to measure the spread of data points among themselves and from the mean. Consequently, it is calculated as the average of the squared distances from each point to the mean.

3.2.4. Standard Deviation. It finds the spread in the sensors data around the mean value.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{n}(x_i - \mu)^2}, \tag{2}$$

where N gives the total number of records, x represents the current sensor output, and μ shows the mean value of a particular feature [27]. Table 3 shows the extracted feature details of the dataset.

3.2.5. Correlation. Correlation is a relationship of one feature with another feature. Table 4 shows that increase or decrease in the values of one feature tends to be paired with relative increment or decrement in the values of another.

$$\text{Cor}(x, y) = \left(\frac{(\text{Cov}(x, y))}{\sqrt{\text{Var}(x)\text{Var}(y)}}\right). \tag{3}$$

3.3. Algorithms. In this paper, four ML algorithms are evaluated for power prediction, which are gradient-boosted regression tree, K-nearest neighbor, artificial neural network, and linear regression. A brief explanation of each algorithm is given in the following.

3.3.1. Linear Regression (LR). Linear regression [28, 29] is used to predict the dependent variable (CCPP's power in this case) (y) based on the four independent variable (x_1, x_2, x_3, and x_4). Figure 2 shows the flow chart of the LR algorithm while its mathematical model [24] is given below.

$\theta_0, \theta_1 \cdots \theta_4$ are the five initial weights where they are assigned values between 0.5 and 2.5:

$$h_\theta(x) = \sum_{i=0}^{m}(\theta_i x_i), \tag{4}$$

where $h_\Theta(x)$ is the predicted value.

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=0}^{m}\left(h_\theta\left(x^i\right) - y^i\right)^2, \tag{5}$$

where $J(\theta)$ is the cost function. It should be minimized in order to achieve maximum prediction accuracy. y^i is the actual value.

$$\text{temp}\left(\theta_j\right) = \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta_0, \theta_1), \tag{6}$$

Table 3: Feature characteristics of CCPP dataset with 9568 data records.

Features	Units	Type	Min	Max	Mean	Variance	Std
Temperature	Celcius	Input	1.81	37.11	19.65	55.54	7.45
Pressure	Millibars	Input	992.89	1033.30	1013.40	35.27	5.93
Humidity	Percent	Input	25.56	100	73.30	213.17	14.6
Vacuum	cm-Hg	Input	25.36	81.56	54.30	161.49	12.70
Power	MW	Output	420.26	495.76	454.36	291.28	17.06

Table 4: Correlation matrix among input and output features.

	Temperature	Vacuum	Pressure	Humidity	Output power
Temperature	1	0.84	-0.50	-0.54	-0.98
Vacuum	0.84	1	-0.41	-0.31	-0.87
Pressure	-0.50	-0.41	1	-0.10	0.51
Humidity	-0.54	-0.31	-0.10	1	0.39
Output power	-0.98	-0.87	0.51	0.39	1

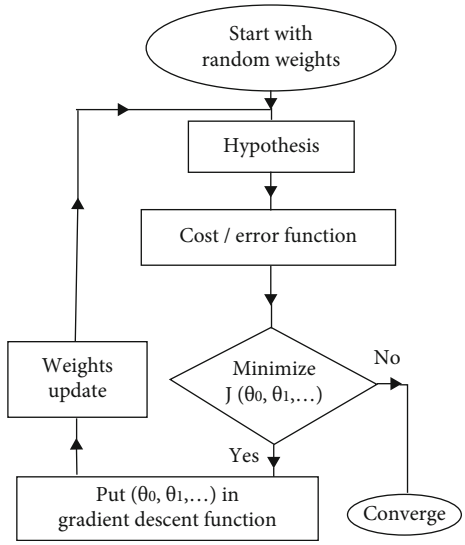

Figure 2: Flow chart of linear regression.

$$\text{temp}(\theta_j) = \theta_j - \frac{\alpha}{2m}\sum_{i=0}^{m}\left(h_\theta\left(x^i\right) - y^i\right)^2 x_j^i. \tag{7}$$

Equation (6) shows the gradient descent function where m gives the total number of sample in the dataset, α denotes the learning rate whose value ranges between 0 and 0.5. For every value of J repeat, the abovementioned equation until convergence is obtained [30].

3.3.2. Gradient-Boosted Regression Tree. In gradient-boosted regression [31–33] $(x_1, x_2, x_3, \text{and}\, x_4)$ are used as input parameters of CCPP and (y) as the output power. Figure 3 represents the tree structure of gradient-boosted regression. Learning rate α value is set as 0.1, and numbers of trees are varied for power prediction. In each tree, the depth is 20. For power prediction of CCPP, the "quantile function" is used as a distribution function. α is learning rate used to scale the tree and gradually improve the tree performance; its value is between 0 and 1. This whole process is continued until *ri* approaches to a minimum or stable value [30].

3.3.3. Artificial Neural Network. The ANN [34–36] is found to be a very useful Algorithm in machine learning [36]. ANN is an information management model that is similar to the function of biological nerves of the human brain. In Figure 4, there are 4 input parameters and one output which is the produced electric power by the CCPP. In this scenario, the ANN has two hidden layers where each layer comprises of 100 neurons. The learning rate α is 0.01. The Momentum value is set as 0.9 where in ANN, the momentum simply adds a fraction of the previous weight update to the current one. This prevents local maxima and provides smoothly optimized results. Following are the steps required to train the ANN.

(1) Initialize the weights arbitrarily:

$$w = \frac{1}{\text{total hidden layers}} \tag{8}$$

(2) Calculate the sum of inputs and weights products:

$$V = wx_i + b \tag{9}$$

(3) Analyze the activation (Sigmoid) function response:

$$h_\theta = \frac{1}{1 + e^{-V}} \tag{10}$$

(4) Take the input from the training data of (input, actual output) and enter it to the neural network.

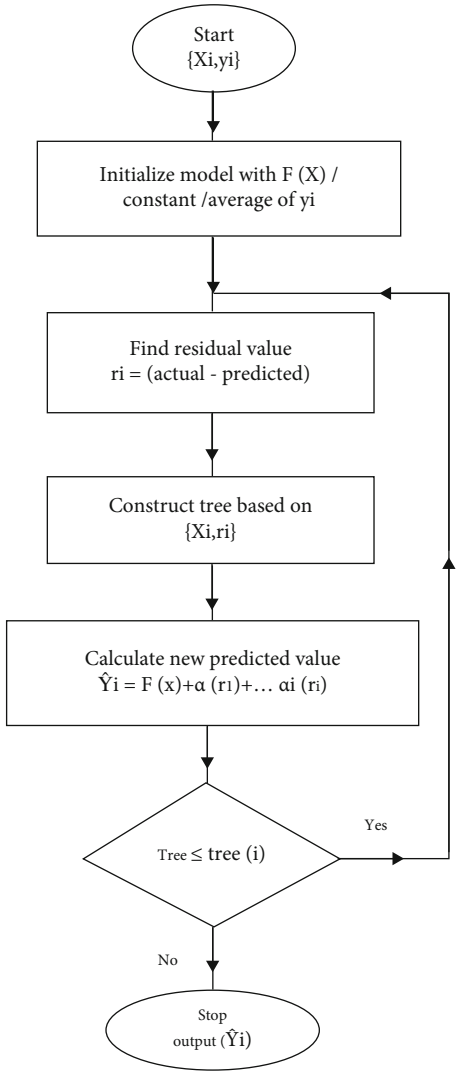

Figure 3: Flow chart of GBRT.

After that, calculate the error:

$$e_i = d_i - y_i \tag{11}$$

(5) Calculate the weight updates according to the following delta rule:

$$\Delta w_i j = \alpha e_i x_i \tag{12}$$

(6) Adjust the weights as

$$w_i j \longleftarrow w_i j(\text{old}) + \Delta w_i j \tag{13}$$

(7) Repeat steps 2–5 until the error is in an acceptable range [34]

3.3.4. K-Nearest Neighbor. KNN [36–39] stores all available scenarios and predict the numerical value based on a similarity measure such as the Euclidean distance [39] which is calculated between two points p and q as

$$\text{Dist} = \sqrt{\sum_{i}^{n}(q_i - p)2}. \tag{14}$$

For KNN, in the first step, find the K neighbor values which are near to the new feature whose value is anonymous. This nearness is found in terms of Euclidean distance between new data point and every data point present in the training set. For optimal results, K must be selected as an odd number [39]. Figure 5 depicts an example of 3-nearest neighbors where a new test sample will be labeled according to the label of the 3 nearest samples having shortest distance from the new sample [40].

3.4. Performance Metrics. For performance evaluation of the models, root mean square error (RMSE) and absolute error (AE) are used.

$$\begin{aligned} \text{RMSE} &= \sqrt{\left(\frac{\sum_1^N \left(Y - \bar{Y}\right)^2}{N}\right)}, \\ AE &= \frac{1}{N}\sum_1^N \left|Y - \bar{Y}\right|, \end{aligned} \tag{15}$$

where Y and $\bar{Y}$ are the actual and predicted output values of the CCPP plant, N is the total number of records [41].

4. Result Analysis

This section outlines the detailed results by discussing the influence of various parameters of the utilized algorithms on the predicted power. The experiments are performed in a systematic manner as outlined in the following:

(1) As a first step, we show the effect of feature selection on the predicted power achieved by each algorithm. For this, the default parameters of each algorithm are used as set in the Rapid Miner software [26]. This is helpful to find the features that have a significant effect of the generated power. In addition to that, a reduced features set will also make the rest of the analysis convenient and efficient by reducing the time taken in training a machine learning model

(2) The effect of the most crucial parameters of the respective algorithms on the predicted power is presented in terms of both RMSE and AE. Furthermore, for the training and testing of each algorithm, the dataset is split randomly into 90-10 and 80-20 where the larger number represents the percentage of dataset used for training the algorithm while the smaller is the percentage of the dataset used to test the trained algorithm. In the subsequent discussion we use 9 : 1 and 8 : 2 for the respective splits

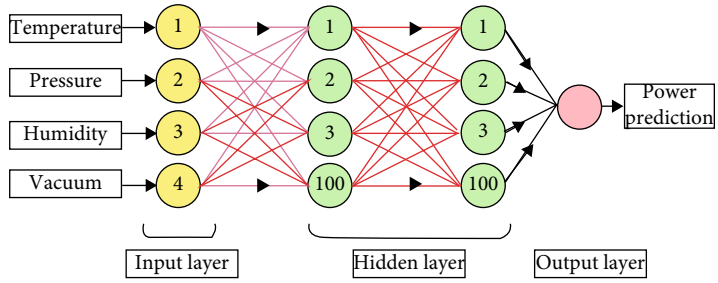

Figure 4: Structure of artificial neural network (ANN) [34].

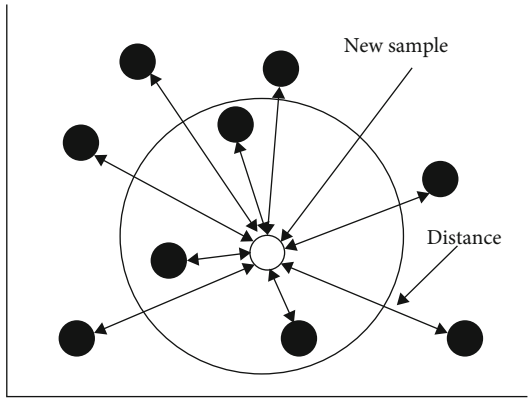

Figure 5: An example of 3-nearest neighbors [40].

(3) With the best values of the parameters for each algorithm, the actual and predicted powers for randomly selected samples from the dataset are shown

(4) Finally, the best results achieved by all the algorithms in terms of RMSE and AE are discussed along with the best results reported by the previous works on the same dataset

4.1. Effect of Features Selection on Predicted Power. In order to examine the effect of each of the input features, i.e., temperature (TEMP), vacuum (VAC), pressure (PRE), and humidity (HUM), on the predicted power, the experiments are performed by making their 15 combinations. Each of these combinations are then used to train and test each of the five algorithms for power prediction. Since, the dataset is randomly split between the training and test sets, the experiments for each of the combinations are performed 10 times. The average RMSE and AE over all the 10 runs achieved by each of the algorithms are presented in Table 5.

Columns 12 and 13 of the table are used to rank the feature combinations from worst to best. Column 12 is the average of all RMSE values achieved by all the five algorithms for a given feature combination. For instance, the mean RMSE for HUM (16.53, row 1, column 12) is calculated as the average of RMSEs achieved by LR (15.66), GBRT (17.31), KNN (18.19), ANN (15.86), and DNN (15.64). The average AE for each of the combinations is calculated in a similar manner. It should be noted that the features combinations are ranked in a descending manner with respect to their average RMSE and AE (columns 12 and 13). For instance, the highest RMSE and AE averaged over all the five algorithms is for humidity (HUM). It is noticeable that the presence of TEMP has a significant effect on the performance of all the algorithms.

Similarly, using the rest of the three features, i.e., HUM, VAC, and PRE without TEMP give high RMSE and AE values regardless of the utilized algorithm. The least RMSE of 2.63 is achieved by GBRT on the combination of TEMP-VAC-PRE whose mean RMSE and AE values are second best (column 12 and 13, second to last row). Hence, it can safely be concluded that the usage of only these three features is enough to perform the rest of the analysis. However, the complete set of the input features achieves the least RMSE and AE (column 12 and 13, last row) due to which dropping just a single feature in the name of achieving efficiency is not technically convincing. Therefore, in the subsequent experiments, all the four features are used.

4.2. Performance Evaluation of ML Algorithms Based on their Important Parameters

4.2.1. K-Nearest Neighbor (KNN). The most important parameter of the K-nearest neighbor algorithm is the value of K as explained in Section 3.3.4. We empirically select the values of K from the set of odd number ranging from 3 to 17. For training and testing, the algorithm uses two phases where in the first phase, we use 90% of the dataset as training data and 10% as test data. Similarly, in phase two, 80% is used for training and 20% of the data is used as the test set. Figure 6 shows the effect of various values of K for both the settings. The minimum RMSE value of 3.28 is achieved at $K = 5$ for 9 : 1 split, while at the same of K, the least RMSE of 3.51 is achieved when the algorithm evaluated on 8 : 2 split.

Similarly, Figure 6 shows the absolute error (AE) at various values of K for both the settings of data split. The minimum value of AE is found to be 2.374 at $K = 3$ for 9 : 1, while at the same value of K, the least AE of 2.49 is achieved at 8 : 2.

4.2.2. Gradient-Boosted Regression Tree (GBRT). The GBRT algorithm is used to predict the output power where the RMSE values are analyzed by varying the number of trees from 50 to 400, as shown in Figure 7. For 9 : 1 split, the minimum RMSE value of 2.581 is achieved for 550 trees. However, there is no significant drop of RMSE from 2.61 achieved at 200 trees to 2.581 achieved at 550 trees. Hence,

Table 5: Effect of feature selection on RMSE and AE for all the algorithms. The table is horizontally partitioned to emphasize on the combinations with and without the temperature (TEMP).

Features combinations	LR		GBRT		KNN		ANN		DNN		Mean RMSE and AE for combinations	
	RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE	RMSE-FEATURES	AE-FEATURES
HUM	15.66	13.15	17.31	13.95	18.19	14.72	15.86	13.69	15.64	13.15	16.53	13.73
PRE	14.39	11.67	15.41	11.62	16.5	12.93	15.55	13.24	14.15	11.42	15.2	12.18
PRE-HUM	13.12	10.63	13.72	10.4	14.19	1.01	13.92	11.7	12.82	10.21	13.55	8.79
VAC	8.43	6.56	5.72	3.99	6.16	4.41	7.78	6.05	7.64	6.02	7.15	5.41
VAC-HUM	8.15	6.39	4.63	3.39	7.48	5.5	7.55	5.93	7.36	5.76	7.03	5.39
VAC-PRE	7.85	6.14	4.45	3.11	6.05	4.24	7.1	5.49	6.93	5.29	6.48	4.85
VAC-PRE-HUM	7.52	5.85	4.15	2.86	5.75	4.08	6.72	5.21	6.35	4.79	6.1	4.56
TEMP	5.07	4.11	5.1	4.03	5.51	4.41	4.89	3.91	4.7	3.75	5.05	4.04
TEMP-PRE	5.01	4.09	4.9	3.71	5.15	3.93	4.67	3.77	4.53	3.63	4.85	3.83
TEMP-HUM	4.43	3.55	4.45	3.5	4.69	3.74	4.39	3.43	4.12	3.25	4.42	3.49
TEMP-PRE-HUM	4.43	3.55	4.01	3.01	4.2	3.12	4.31	3.39	4.06	3.16	4.2	3.25
TEMP-VAC	4.7	3.77	2.9	2.22	4.12	3.09	4.41	3.52	4.2	3.35	4.07	3.19
TEMP-VAC-HUM	4.27	3.4	2.82	2.11	3.79	2.86	4.19	3.33	3.84	3.01	3.78	2.94
TEMP-VAC-PRE	4.63	3.7	2.63	1.96	3.57	2.53	4.16	3.32	3.88	2.96	3.77	2.89
TEMP-VAC-PRE-HUM	4.26	3.4	2.64	1.93	3.32	2.37	4.16	3.34	3.61	2.82	3.6	2.77

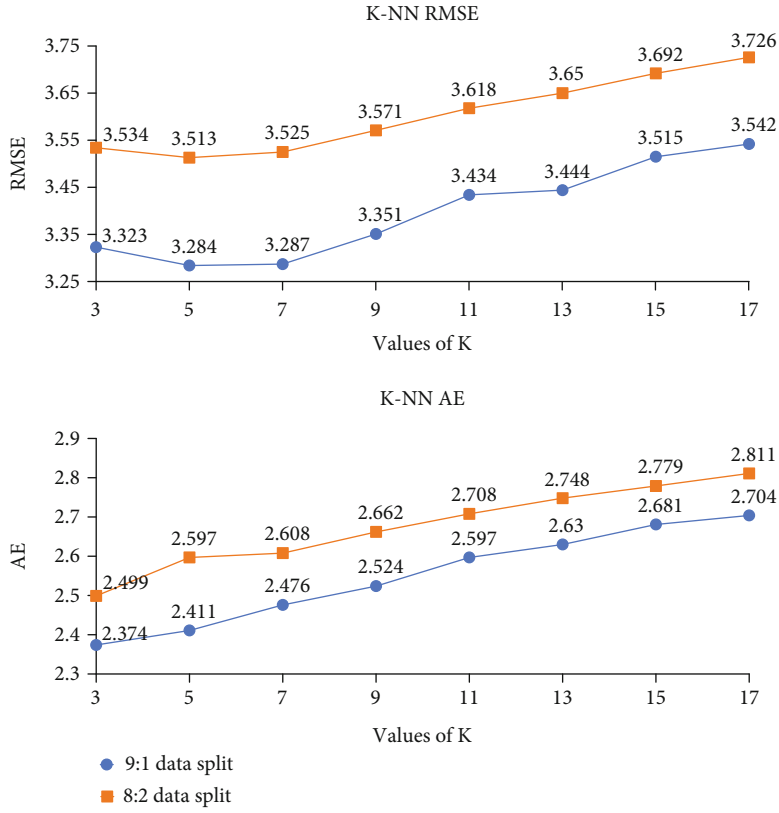

Figure 6: RMSE and AE values achieved by KNN for various values of K on both 9 : 1 and 8 : 2 splits

on the current dataset, for 9 : 1 split, 200 trees are advised for GBRT in order to optimize the usage of computational resources. Furthermore, for 8 : 2 split, the RMSE values are higher than those of 9 : 1. Figure 7 shows a similar trend in the values of AE for various values of trees and for both the settings of dataset split. The least AE for 9 : 1 split is

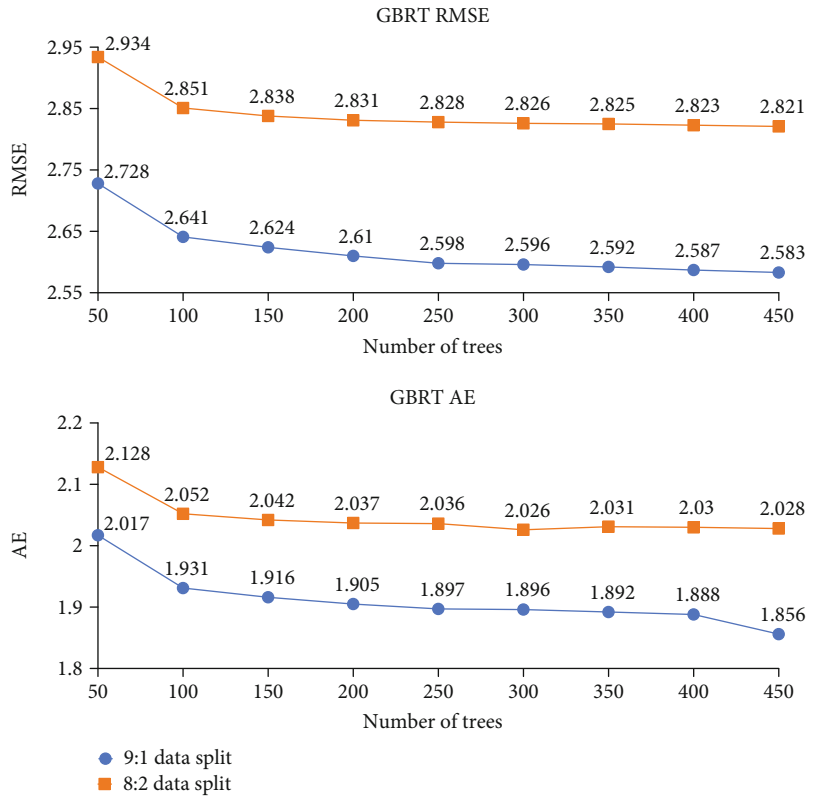

Figure 7: RMSE and AE values achieved by GBRT for various number of trees on both 9 : 1 and 8 : 2 splits.

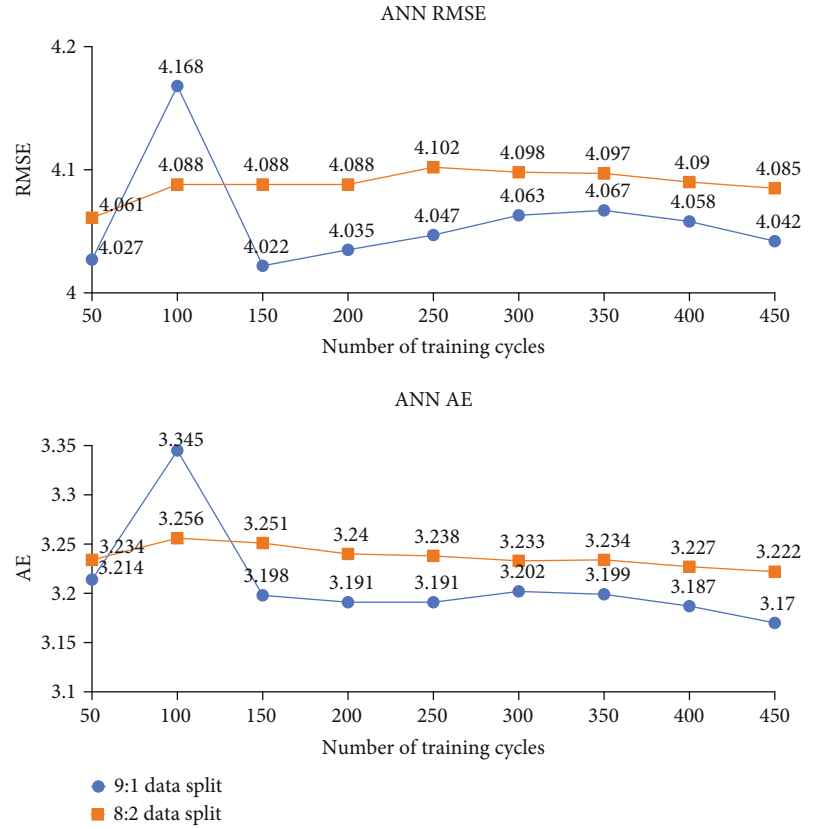

Figure 8: RMSE and AE values achieved by ANN on various number of training cycles for both 9 : 1 and 8 : 2 splits.

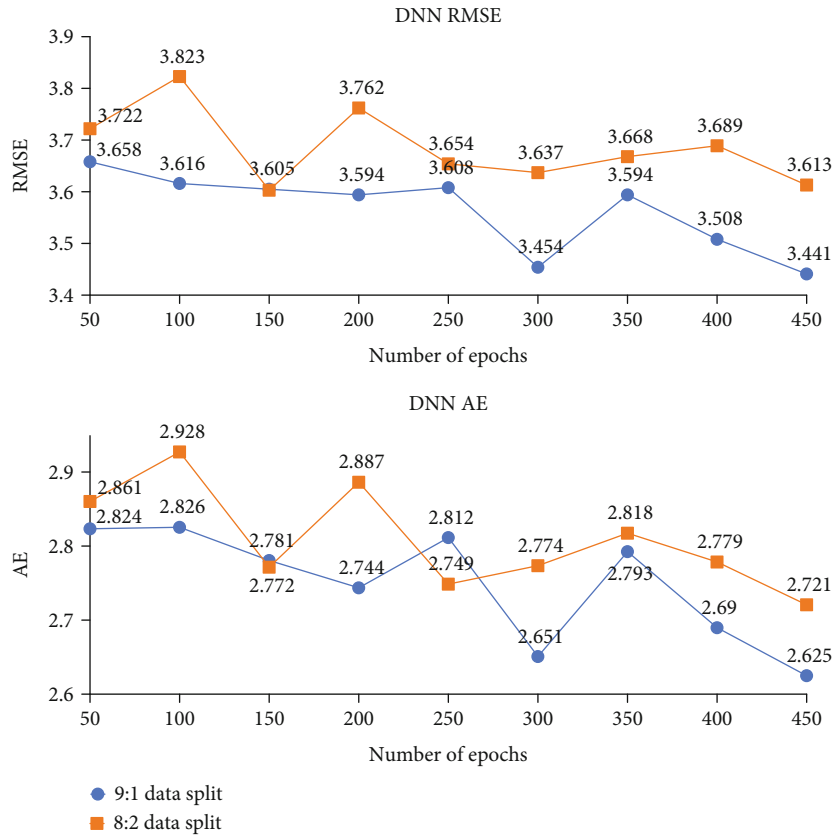

Figure 9: RMSE and AE values achieved by DNN on various training Epochs for both 9 : 1 and 8 : 2 splits.

achieved at 550 trees; however, after 250 trees, the drop is not significant and hence recommended for efficient usage of computational resources on the current dataset. Similarly, the AE values for 8 : 2 split are higher than those of 9 : 1 split.

4.2.3. Artificial Neural Network (ANN). The ANN algorithm of Rapid Miner is used for power prediction on the current dataset. We evaluate the number of training cycles while keeping all other parameters to the default values such as those of the hidden layers and the number of activation functions per layer. Figure 8 shows the achieved RMSE values for various numbers of training cycles for both 9 : 1 and 8 : 2 splits. The least RMSE value of 4.022 is achieved when the training is performed for 150 cycles with 9 : 1 dataset split. For the 8 : 2 split, the minimum RMSE of 4.085 is achieved at 550 training cycles; however the second best RMSE of 4.088 is achieved on 100 training cycles which is comparably acceptable than 4.085 achieved at 550cycles.

In Figure 8, the value of AE are depicted for various training cycles for power prediction. The minimum value of absolute error (AE) at 9 : 1 is observed as 3.15 at 500 training cycles. Using 8 : 2 data split, the minimum AE is 3.222 at 500 training cycles.

4.2.4. Deep Neural Network (DNN). We also use the deep neural network algorithm provided in the Rapid Miner suite which is a variant of the neural network. The default parameters are used while the number of training epochs are varied to observe the effect on the achieved RMSE and AE. Consequently, for 9 : 1 data split, the least RMSE of 3.441 is achieved on 450 epochs. However, the second best RMSE of 3.445 is obtained on 300 epochs and hence recommended for on the current dataset. For the data split of 8 : 2, all the RMSE values are higher than 9 : 1 split. Nontheless, the least achieved RMSE of 3.592 for 8 : 2 split is achieved on 500 epochs. The values of RMSE for both the data splits on various epochs are shown in Figure 9.

Table 6: Comparison of results achieved by each of the algorithms on both the dataset splits. GBRT outperforms the rest of the algorithms by achieving the least RMSE and AE such that the number of trees is 100.

(a)

	RMSE				
Data splits	KNN	GBRT	NN	DL	LR
Training data 90%	3.323	**2.641**	4.168	3.616	4.263
Training data 80%	3.534	**2.851**	4.088	3.823	4.347

(b)

	AE				
	KNN	GBRT	NN	DL	LR
Training data 90%	2.374	**1.931**	3.345	2.826	3.402
Training data 80%	2.499	**2.052**	3.256	2.928	3.469

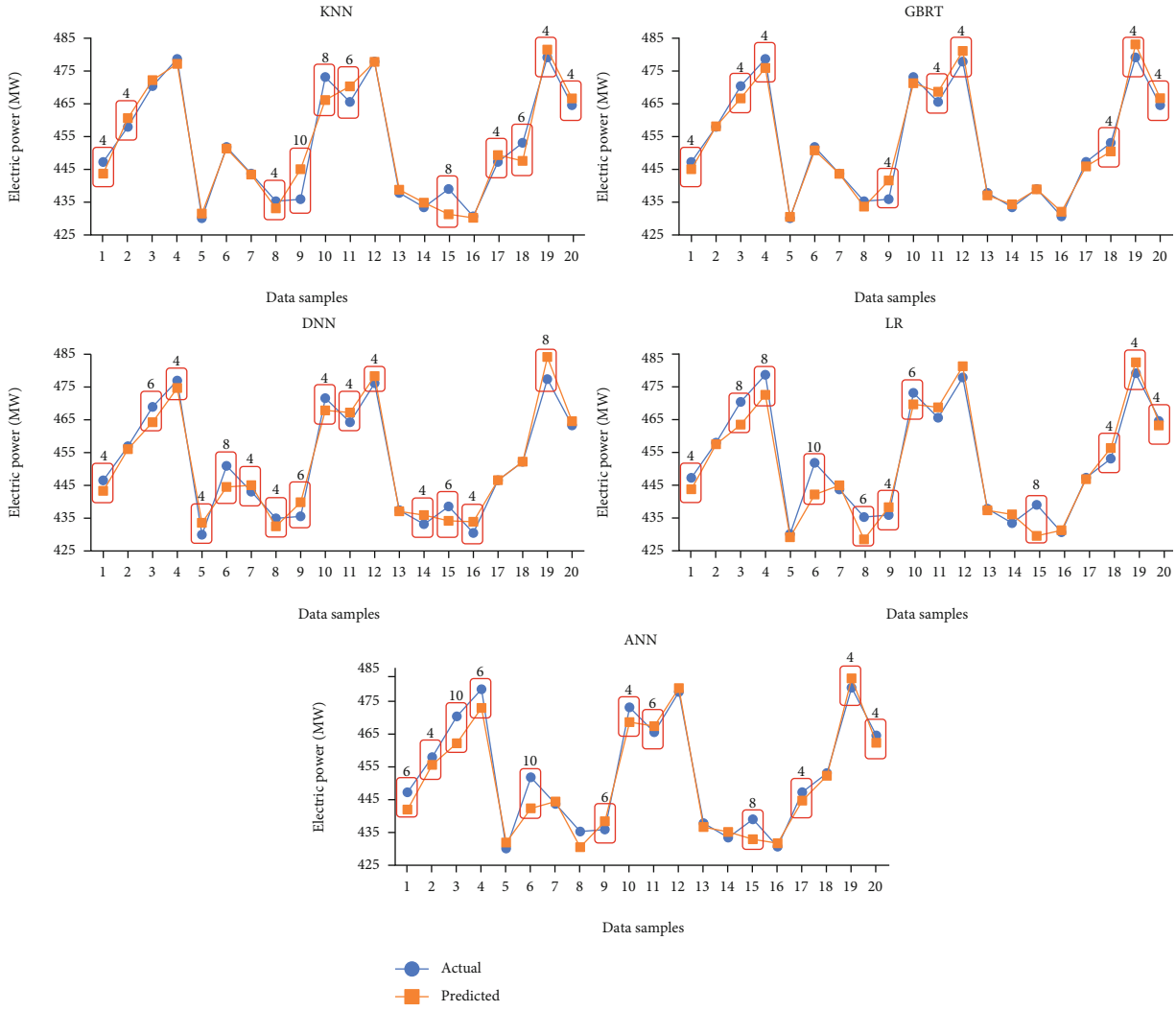

Figure 10: Actual vs. predicted power for 20 random samples of the dataset using five ML algorithms.

Similarly, in Figure 9, AE is shown for both the settings of data split and all the empirically selected values of epochs. The minimum value of AE, using 9 : 1 split, is found to be 2.625 on 450 training epochs. At 8 : 2 split, the minimum AE value is 2.694 achieved on 500 training epochs. To summarize, we show a comparison of all the five algorithms in Table 6 in terms of RMSE and AE for both the settings of data split. The results achieved by each of the algorithms on the current dataset are shown where the value of K in KNN is 3, the number of trees for GBRT is 100, number of training cycles for ANN is 100, and the number of training epochs for DNN is also 100. It can clearly be observed that GBRT outperforms the rest of the algorithms by achieving the least RMSE and AE on the current dataset.

4.2.5. Comparison of Actual and Predicted Powers for each ML Algorithm. In order to demonstrate the prediction results achieved by each algorithm, we randomly select 20 samples from the test dataset such that the actual power value is known for each of them. We then provide the input features of these 20 samples to the trained models of each of the algorithms for power prediction. It should be noted that the models are trained with the best parameters of the respective algorithms as found in the previous section. The actual power values along with the predicted values achieved by each algorithm are shown in Figure 10. The samples where the absolute difference of the actual and predicted power values are greater than 2 MW are highlighted and shown for each algorithm. It can be observed that GBRT achieves power predictions closer to the actual values and the samples where the difference is greater than 2 MW are comparatively less than other algorithms. Hence, it can be concluded that GBRT gives better power predictions than other algorithms.

4.2.6. Performance Comparison with the Literature. Table 7 shows our results in comparison on the previous methods proposed for CCPP power prediction on the same dataset. We clearly achieve superior performance from all those methods using GBRT where the number of tree is 450 and the size of the training set is 90% of the whole dataset. The achieved RMSE and AE values are 2.58 and 3.51, respectively. Similarly, our achieved RMSE and AE values on

Table 7: Results comparison with the previous methods on the CCPP dataset.

	RMSE	AE
Previous literature		
KNN (2012) [19]	N/A	3.51
Bagging with REPTree [4]	4.23	3.22
C-CRF (2016) [20]	3.97	2.97
TSK-ELM (2018) [18]	3.93	N/A
Regression ANN (2018) [12]	4.32	N/A
TOB matching (2020) [21]	2.89	N/A
Ours		
GBRT (450 trees)	**2.58**	**1.85**
KNN ($K = 3$)	3.32	2.37
DNN (450 epochs)	3.44	2.62
ANN (450 cycles)	4.04	3.17
LR	4.26	3.40

KNN and DNN are also better than those methods. Hence, at the moment, we can convincingly say that on the current dataset our results are the best.

5. Conclusion and Future Work

We perform power prediction of a CCPP on hourly basis using machine learning paradigm. In this regard, we use a publicly available dataset that is collected over a period of 6 years such that the power plant is operating on full load. The dataset measures output power as a function of four input parameters which are temperature, humidity, pressure, and vacuum. We evaluate five machine learning algorithms, namely, K-Nearest Neighbors, Linear Regression, Gradient-boosted Regression Tree, Artificial Neural Network, and Deep Neural Network of the Rapid Miner software suite. Keeping the default parameters, we evaluate the most crucial parameters of each algorithm to find the best of them that achieves minimum RMSE and AE. We also evaluate the effect of training set size and number of features on the achieved results. Consequently, GBRT outperforms the rest of the algorithm by achieving the least RMSE and AE with 450 trees while training on 90% of the dataset. Interestingly, it also exceeds in performance from all the previously proposed methods on the same datasetby achieving the least RMSE and AE.

In future, the output power can be controlled by changing the value of the parameters. Moreover, by incorporating these parameters as well as increasing the number of input parameters, the power prediction of different types of power plants can be done by using more advance machine learning algorithms.

Data Availability

The data that support the findings of this study are all briefly introduced, and all information is available in the manuscript.

Disclosure

The authors carried out the research as a part of employment at COMSATS University Islamabad, Attock Campus, UET Peshawar, UoT Nowshera, KPK.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The second author (Hafeez Anwar) is supported by the Ernst Mach Follow-up grant awarded by the Austrian Agency for International Cooperation in Education and Research (OeAD).

References

[1] J. L. Lobo, I. Ballesteros, I. Oregi, J. del Ser, and S. Salcedo-Sanz, "Stream learning in energy IoT systems: a case study in combined cycle power plants," *Energies*, vol. 13, no. 3, 2020.

[2] A. Castillo, "Risk analysis and management in power outage and restoration: a literature survey," *Electric Power Systems Research*, vol. 107, pp. 9–15, 2014.

[3] B. Çetin, S. H. S. Erdem, S. H. Sevilgen, and A. V. Akkaya, "Electricity production cost analysis of a combined cycle power plant," *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 3, no. 3, pp. 224–232, 2008.

[4] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *International Journal of Electrical Power & Energy Systems*, vol. 60, pp. 126–140, 2014.

[5] U. Kesgin and H. Heperkan, "Simulation of thermodynamic systems using soft computing techniques," *International Journal of Energy Research*, vol. 29, no. 7, pp. 581–611, 2005.

[6] E. Rich and K. S. Knight, "Chapter 3," in *Artificial Intelligence*, McGraw Hill, New York, USA, 2nd Ed. edition, 1991.

[7] A. Mosavi, F. S. Hosseini, B. Choubin et al., "Susceptibility prediction of groundwater hardness using ensemble machine learning models," *Water*, vol. 12, no. 10, 2020.

[8] A. Mosavi, F. Sajedi-Hosseini, B. Choubin, F. Taromideh, G. Rahi, and A. Dineva, "Susceptibility mapping of soil water erosion using machine learning models," *Water*, vol. 12, no. 7, p. 1995, 2020.

[9] B. Choubin, F. S. Hosseini, Z. Fried, and A. Mosavi, "Application of bayesian regularized neural networks for groundwater level modeling," in *2020 IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)*, pp. 000209–000212, Budapest, Hungary, 2020.

[10] A. Mosavi, F. Sajedi Hosseini, B. Choubin, M. Goodarzi, A. A. Dineva, and E. Rafiei Sardooi, "Ensemble boosting and bagging based machine learning models for groundwater potential prediction," *Water Resources Management*, vol. 35, no. 1, pp. 23–37, 2021.

[11] F. Jiménez-Espadafor Aguilar, M. T. García, E. C. Trujillo, J. A. Becerra Villanueva, and F. J. Florencio Ojeda, "Prediction of performance, energy savings and increase in profitability of

two gas turbine steam generator cogeneration plant, based on experimental data," *Energy*, vol. 36, no. 2, pp. 742–754, 2011.

[12] E. A. Elfaki and A. H. Ahmed, "Prediction of electrical output power of combined cycle power plant using regression ANN model," *Engineering*, vol. 6, no. 12, pp. 17–38, 2018.

[13] G. Şen, M. Nil, H. Mamur et al., "The effect of ambient temperature on electric power generation in natural gas combined cycle power plant-a case study," *Energy Reports*, vol. 4, pp. 682–690, 2018.

[14] A. González-Díaz, A. M. Alcaráz-Calderón, M. O. González-Díaz, Á. Méndez-Aranda, M. Lucquiaud, and J. M. González-Santaló, "Effect of the ambient conditions on gas turbine combined cycle power plants with post-combustion CO_2 capture," *Energy*, vol. 134, pp. 221–233, 2017.

[15] F. Chu, J. Wang, L. Nannan, T. Tan, and F. Wang, Eds., "Prediction of ccpp output based on improved fuzzy analytical hierarchy process," in *2017 29th Chinese Control And Decision Conference (CCDC)*, pp. 3636–3641, Chongqing, China, 2017.

[16] I. Lorencin, N. Anđelić, 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, 2019.

[17] P. Pourbeik, "Modeling of combined-cycle power plants for power system studies," in *2003 IEEE Power Engineering Society General Meeting (IEEE Cat. No.03CH37491)*, Toronto, ON, Canada, 2003.

[18] C.-U. Yeom and K.-C. Kwak, "A design of TSK-based elm for prediction of electrical power in combined cycle power plant," in *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Bangkok, Thailand, 2018.

[19] H. Kaya, P. Tüfekci, and F. S. Gürgen, "Local and global learning methods for predicting power of a combined gas & steam turbine," in *Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE*, pp. 13–18, Dubai, UAE, 2012.

[20] G. Ahn and S. Hur, "Continuous conditional random field model for predicting the electrical load of a combined cycle power plant," *Industrial Engineering and Management Systems*, vol. 15, no. 2, pp. 148–155, 2016.

[21] D. A. Wood, "Combined cycle gas turbine power output prediction and data mining with optimized data matching algorithm," *Applied Sciences*, vol. 2, no. 3, pp. 1–21, 2020.

[22] L. Bandić, M. Hasičić, and J. Kevrić, "Prediction of power output for combined cycle power plant using random decision tree algorithms and ANFIS," in *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies*, pp. 406–416, Sarajevo, Bosnia and Herzegovina, 2019.

[23] M. Rashid, K. Kamal, T. Zafar, Z. Sheikh, A. Shah, and S. Mathavan, "Energy prediction of a combined cycle power plant using a particle swarm optimization trained feedforward neural network," in *2015 International Conference on Mechanical Engineering, Automation and Control Systems (MEACS)*, pp. 1–5, Tomsk, Russia, 2015.

[24] A. Burkov, *The Hundred-Page Machine Learning Book, Volume 1*, Andriy Burkov Canada, 2019.

[25] C.-W. Han, *Output Power Prediction of Combined Cycle Power Plant Using Logic-Based Tree Structured Fuzzy Neural Networks*, AIP Publishing LLC, 2019.

[26] I. Mierswa and R. Klinkenberg, "Rapid miner studio," 2018.

[27] X. Wan, W. Wang, J. Liu, and T. Tong, "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range," *BMC Medical Research Methodology*, vol. 14, no. 1, 2014.

[28] S. Rong and Z. Bao-wen, "The research of regression model in machine learning field," *MATEC Web of Conferences*, vol. 176, article 01033, 2018.

[29] A. Schneider, G. Hommel, and M. Blettner, "Linear regression analysis: part 14 of a series on evaluation of scientific publications," *Deutsches A¨rzteblatt International*, vol. 107, no. 44, p. 776, 2010.

[30] J. H. Friedman, "Gradient boosting (regression)," 2021, https://statweb.stanford.edu/jhf/ftp/stobst.pdf.

[31] X. Li and R. Bai, "Freight vehicle travel time prediction using gradient boosting regression tree," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1010–1015, Anaheim, CA, USA, 2016.

[32] K. Gao, H. Chen, X. Zhang, X. K. Ren, J. Chen, and X. Chen, "A novel material removal prediction method based on acoustic sensing and ensemble XGBoost learning algorithm for robotic belt grinding of Inconel 718," *The International Journal of Advanced Manufacturing Technology*, vol. 105, no. 1-4, pp. 217–232, 2019.

[33] C. Persson, P. Bacher, T. Shiga, and H. Madsen, "Multi-site solar power forecasting using gradient boosted regression trees," *Solar Energy*, vol. 150, pp. 423–436, 2017.

[34] R. Siddiqui, S. Umer, A. Iqbal, F. Ullah, A. Khan, and K. S. Kwak, "Estimation of solar panel output based on weather parameters using machine learning algorithms," in *Proceedings of the Korea Telecommunications Society Conference*, pp. 459–462, Korea, 2020.

[35] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: a review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[36] X. Wenchao, Y. Pang, Y. Yang, and Y. Liu, "Human activity recognition based on convolutional neural network," in *In 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 165–170, Beijing, China, 2018.

[37] R. Goyal, P. Chandra, and Y. Singh, "Suitability of KNN regression in the development of interaction based software fault prediction models," *Ieri Procedia*, vol. 6, pp. 15–21, 2014.

[38] T. Lee, T. B. M. J. Ouarda, and S. Yoon, "KNN-based local linear regression for the analysis and simulation of low flow extremes under climatic influence," *Climate Dynamics*, vol. 49, no. 9-10, pp. 3493–3511, 2017.

[39] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: theoretical background," *International Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605–610, 2013.

[40] Z. Yao and W. L. Ruzzo, "A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data," *BMC Bioinformatics*, vol. 7, Supplement1, 2006.

[41] A. Qazi, A. W. Fayaz, A. Wadi, R. G. Raj, N. A. Rahim, and W. A. Khan, "The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review," *Journal of Cleaner Production*, vol. 104, pp. 1–12, 2015.