Bootstrap aggregating approach to short-term load forecasting using meteorological parameters for demand side management in the North-Eastern Region of India

Dipu Sarkar1 · Taliakum Ao1 · Sravan Kumar Gunturi2

Received: 10 June 2021 / Accepted: 9 January 2022 / Published online: 1 March 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2022

Abstract
Electricity is an essential commodity that must be generated in response to demand. Hydroelectric power plants, fossil fuels, nuclear energy, and wind energy are just a few examples of energy sources that significantly impact production costs. Accurate load forecasting for a specific region would allow for more efficient management, planning, and scheduling of low-cost generation units and ensuring on-time energy delivery for full monetary benefit. Machine learning methods are becoming more effective on power grids as data availability increases. Ensemble learning models are hybrid algorithms that combine various machine learning methods and intelligently incorporate them into a single predictive model to reduce uncertainty and bias. In this study, several ensemble methods were implemented and tested for short-term electric load forecasting. The suggested method is trained using the influential meteorological variables obtained through correlation analysis and the past load. We used real-time load data from Nagaland’s load dispatch centre in India and meteorological parameters of the Nagaland region for data analysis. The synthetic minority over-sampling technique for regression (SMOTE-R) is also employed to avoid data imbalance issues. The experimental results show that the Bagging methods outperform other models with respect to mean squared error and mean absolute percentage error.

Keywords Demand side management · Ensemble Learning · Random Forest · Extra Trees · Short term load forecasting · SMOTE

1 Introduction
Load forecasting is a pivotal step for electric grids’ efficient operation and management in the day-ahead electricity market (Amato et al. 2021). Owing to the rapid growth of renewable energy and the implementation of electric vehicles and other emerging technologies, the electric power grid has undergone significant changes in the last few years, both on the supply and demand lines (Andriopoulos et al. 2021). To sustain an equilibrium between supply and demand in this setting, electricity suppliers must take advantage of the next-generation power grid’s opportunities. Furthermore, in the event of unusually high peak energy demand, electric utility companies must be prepared in advance with contingency plans. As a result, it is essential to forecast future load demand to reduce the cost of generating electricity (Li and Jin 2019). Electricity load forecasting is the process of predicting future load based on various features, including weather conditions, time details such as the month and hour, economic conditions, energy tariffs, regional conditions are examples. Since accurate electricity load forecasting is critical in the power system, even minor improvements in load forecasting accuracy can result in substantial cost savings and environmental benefits (Chen et al. 2018). Besides that, any forecast error results in a substantial cost increase. As a result, effective load forecasting is critical for power-grid construction, investment, and transactions in order to ensure reliable and cost-effective power system operation (Moradzadeh et al. 2021).

In recent times, researchers have suggested multiple models for forecasting electric load for different periods. Short-term load forecasting (SLF) models forecast electricity demand every half-hour or hour for the next few hours.
to two weeks. Medium- and long-term models are established when long-term load forecasting is needed (Fan and Hyndman 2011). Regression models are good at estimating stationary time series (Miraftabzadeh et al. 2019). Since the time series of load demand is nonlinear and non-stationary, the Autoregressive Moving Average (ARMA) method (Li and Zhen-gang 2009; Din et al. 2018) and Support Vector Regression (SVR) (Velasco et al. 2018; Yang et al. 2019) methods were proposed. A single model is incapable of accurately capturing the inherent characteristics of electricity load demand. Hence, Li and Zhen-gang (2009) integrated the ARMA approach with the other machine learning methods to boost the efficiency of combined load forecasting. Nepal et al. (2020) developed a hybrid model that combines clustering with the integrated ARMA model.

The SLF problem has been also studied using a variety of sophisticated data-driven models. Fuzzy logic (Sadaei et al. 2014), artificial neural networks (ANNs) (Cecati et al. 2015; Yi et al. 2021), Extreme Learning (Li et al. 2016; Park et al. 2020), and exponential smoothing (Sudheer and Suseelatha 2019) are some of the models that have been used. Numerous hybrid models have been developed by combining multiple models in order to increase forecasting accuracy. Yang et al. suggested an advanced empirical mode decomposition approach to solve the net outcome and the envelope fitting limitations inherent in traditional empirical mode decomposition-based models (Yang et al. 2015). Liang et al. proposed a framework that combines EMD, minimal consistency, maximal relevance, and a neural network with an optimization method (Yi et al. 2019). Talaat et al. suggested a novel approach of a multi-layer front-propagated neural net and the grasshopper optimization technique to achieve accurate load forecasting (Talaat et al. 2020). The majority of the methods reported previously have the following limitations:

- The mean absolute percentage error is high for conventional methods.
- Only time-series load data is considered.
- Most only focus on boosting strategies, not bagging approaches like Random Forest and Extra-Trees.
- The majority of methods, such as SVR and artificial neural networks, need additional training and testing time.
- Several techniques have failed to solve the issue of data imbalance caused by oversampling.

The current research makes four significant contributions: (1) For specific-region SLF implementation, the proposed framework makes use of real-time load and meteorological data. (2) The implementation of the synthetic minority oversampling technique for regression (SMOTE-R) addresses data imbalance. (3) The most advanced load forecasting models are checked and compared. (4) The proposed method achieves the least mean absolute percentage error by using Bootstrap aggregating (Bagging) models.

The remainder of the article is organized as follows: Section 2 deals with the influence of climatic variables on load consumption of Nagaland State. Section 3 addresses the theoretical basis of ensemble learning models. Section 4 discusses the proposed methodology. Section 5 gives the conclusion and future directions.

2 The influence of climatic variables on load consumption

2.1 Geography of Nagaland

Nagaland is located in India's north-eastern section; the north-eastern part of India is situated between 21.57 and 29.30 ° North latitude and 88 to 97.30 ° East longitude and covers an area of approximately 262,185 km² (Borah 2018). Dimapur is the capital city located in the north-western part of Nagaland, and the district is surrounded by the state of Manipur to the north, and Assam to the west, Arunachal Pradesh and Assam to the north, and Manipur to the south. Moreover, peak power shortages of 18.2% and 17.24% during 2013–2014 and 2014–2015, respectively, and energy shortages of 13.7% and 5.04% (Aier et al. 2021). According to 2011 census data, Nagaland has approximately four lakh households, about 2.85 lakh were in rural areas, and 1.15 lakh were in urban areas. Rural regions were electrified in 75.09% of cases, whereas urban areas were in 97.40%. The Nagaland Department of Power is in charge of power generation, transmission, and distribution. The state has enormous hydroelectric potential but lacks power due to the construction of the power plant. It is determined mainly by the distribution of power from central public sector organizations. This reliance on power from other sources necessitates additional financing and inadequacy in power delivery during peak seasons. The load consumption during the year 2018 is illustrated in Fig. 1.

The load profile shows that the load consumption slowly rises from day one to the middle of the year and then gradually drops at the end of the year. It is because Dimapur’s winter months are at the beginning and end of the year, and it has a scorching summer because it is a warm and humid place. Winter energy consumption is low because the chilly weather in Dimapur is tolerable. Nonetheless, the summer months are energy demanding due to Dimapur’s warm and humid atmosphere, resulting in a rise in energy usage during the summer months. In Nagaland, the rainy
season lasts around nine months and has approximately 85–95% relative humidity (Singh et al. 1998). The yearly average temperature in Nagaland is forecast to climb by roughly 1.6 to 1.8 °C. Annual rainfall is also expected to rise between 2020 and 2050 (Mishra et al. 2019).

2.2 Humidity

Humidity is considered since a high temperature does not always indicate a hot day. Humans experience heat differently depending not only on the day’s temperature but also on the humidity, as human skin utilizes the surrounding air to wick away moisture (in the form of sweat). A high humidity level indicates that the air is completely saturated with water vapor. As a result, sweat evaporates more slowly, making a person feel hotter than the actual temperature number. On the other hand, a low humidity level indicates that the air is dry, which aids in the evaporation of sweat, making a person feel cooler than the actual temperature number. Thus, increased humidity means that more cooling loads (cooling equipment) will be utilized. The same is valid to a certain extent when the humidity level is low enough to support heating loads.

2.3 Dew point

The dew point is defined as the temperature at which the partial vapor pressure of water in wet air is sufficient to saturate the air thoroughly (Wood 1970). In other words, the partial vapor pressure at any temperature equals the partial saturation vapor pressure at the dew point. In practice, relative humidity can be perplexing at times. For example, if the outside temperature is 30°C and the dew point temperature is also 30°C, the relative humidity will be 100%. If the temperature is 80°C and the dew point is 60°C, a relative humidity of 50% is obtained, a person will feel more uncomfortable in the second scenario. Thus, we may see the significance of the dew point. The higher the dew point temperature, the higher the electricity consumption due to increased cooling loads.

2.4 Day and time of the year

Electric energy usage varies according to the time of day or range of times of the day. Assume increased electrical energy consumption during the day because most workplaces and schools operate during this period and may require more electricity. Additionally, it is necessary to consider the year’s day, as the seasons fluctuate. Thus the conditions for power consumption alter according to the day or range of days of the year. It is also worth noting that weekdays may have different load requirements than weekends. Special events that occur on the same day each year will also increase load demand compared to regular days.

2.5 Cloud coverage

With a clear sky, there is an increase in the consumption of cooling loads due to the higher solar radiation (Yuk 1998; Jhajharia et al. 2009). Cloud cover’s effect on electricity demand is time-dependent and is dependent on the cloud’s height and thickness. For instance, on a hot summer day, if the cloud cover is dense, the heat from the sun is absorbed by the cloud. This has the effect of reducing cooling loads to a certain extent. Whereas, at night, the cooling impact of a cloud is minor due to the absence of sunshine.

2.6 Precipitation

Precipitation is a broad phrase that refers to the amount of rainfall, snowfall, hail, and other types of liquid water that fall on a particular location in millimetres (Trenberth 2011; Ray et al. 2021). When precipitation occurs in the form of heavy rain or snow during the winter, the demand for electric energy may increase due to drying loads or, if there...
is snow, heating loads. On the other side, if it rains or snows during the summer, the demand for electrical energy may decrease.

2.7 Wind chill factor

The wind chill factor is the difference in air temperature felt on the skin due to the wind. When the wind blows against a hot body, it replaces the hot air with colder air. If the wind is blowing quickly, the rate of hot air replacement by colder air will be faster. This is referred to as wind chill, and it results in a drop in the temperature of a hot body. This impact results in increased heating load usage throughout the winter season. The wind chill is given by an equation as follows:

\[ WC = 33 + \frac{(10 \times \sqrt{W_S} + 10.45 - W_S) \times (T_A - 33)}{22} \]  

(1)

where WC is wind chill in °C, \( W_S \) is the wind speed in m/s, \( T_A \) is the air temperature in °C (Oliveira et al. 2011).

3 Ensemble learning

Ensemble machine learning is a series of models that train a diverse group of potential learners and incorporate their hypotheses to teach the target’s particular viewpoint (Wan and Yang 2013). Ensemble machine learning regressors were evaluated in this analysis, including Ada-Boost, Cat-Boost, Extreme-Gradient-Boost, Light-Gradient-Boost, Random Forest, and Extra-Trees (Pedregosa et al. 2011).

3.1 Boosting techniques:

Boosting is an ensemble technique in which every successive model tries to eliminate the previous model subsets’ errors, as shown in Fig. 2.

Boosting templates such as Ada-Boost, Cat-Boost, Extreme-Gradient-Boost, and Light-Gradient-Boost were introduced. Ada-Boost (Zhao et al. 2019) is used in aggregation with other algorithms to reduce the prediction error. Training a sequence of weak regressor models from the observed data package and combining them into a robust regressor is one way to build the most potent forecast model. Extreme-Gradient-Boost (Georganos et al. 2018) expands Classification And Regression Tree (CART)–compliant classical boosting strategies. The fundamental role in the ensemble tree-boosting method predicts the group’s new membership for each iteration. This is achieved in the most efficient way possible, ensuring that forecasts are made with soft regressors continuously constructed over the preceding regressor’s fault. Following that, incorrectly classified models are assigned advanced weights, letting the regressor focus on subsequent recurrences’ performance. The final categorization is critical because it requires the combined growth of all past decision trees. The evaluation of such regressors depends on the specific function, which defines the training failure and regularization.

Cat-Boost (Jhaveri et al. 2019; Kumar et al. 2019) is an application for gradient boosting that excels at handling categorical results. It makes no attempt to substitute discrete categorical attributes. It determines the average mark value by permuting a random dataset, such as adding the change’s identical class value before the specified one. When confronted with definite category features, Cat-Boost avoids over-fitting. Light-Gradient-Boost (Li et al. 2018) is a decision tree–focused ensemble model. The advancing propagation process is used for this system. In each epoch, the residual is close-fitting with a negative slope for studying the decision tree. Light-gradient-Boost (introduced by Microsoft in 2016) is a significant Extreme-Gradient-Boost update that can fix issues by improving the sample size while the node is isolated. Light-Gradient-Boost employs a leaf-based method to increase forecast accuracy (Ke et al. 2017).

3.2 Bagging techniques:

The objective of bagging is to achieve a standardized outcome by combining the outcomes of several decision trees. As seen in Fig. 3, the bootstrap aggregating approach employs subsets to maintain a correct representation of the whole set. It is a technique of selecting several random samples from a training feature dataset. Random Forest (Zhang et al. 2018) is a learning method that provides planning to construct multiple decision trees, resulting in a class that each tree averages or votes on. By adding a randomness layer to the bagging operation, Genuer et al. suggested a Random-Forest method (Genuer et al. 2010). Random-Forest is helpful for regression, grouping, and feature collection, among other things. Random-Forest has three distinct advantages. To begin, it reduces noise by selecting variables and data at random to create multiple regression trees. Second, Random-Forest can handle high-dimensional data and adapts to various datasets. It will work for both continuous and discrete data without the need for normalization. Third, the parallel arithmetic’s rapid learning speed reduces computational costs.

Finally, an Extra-Trees regressor uses a meta-learning algorithm to train several randomized decision trees. The averaging method is used to increase prediction accuracy and reduce over-fitting. Each decision tree is evaluated at each intersection by executing the following steps: (1) Select K attributes randomly, (2) calculate the split for each feature, and (3) select the feature that boosts the score using a normalized method of feature selection.
4 Proposed method

This section addresses the proposed load forecasting model, which uses real-time load data and meteorological parameters as training data. The proposed scheme made use of data analysis collected from the Nagaland Load Dispatch Centre. The proposed system’s block diagram is depicted in Fig. 4.

As shown in the block diagram, the proposed system can be explained in three steps: (1) feature engineering, (2) features extraction, (3) use ensemble learning for load forecasting.
4.1 Feature engineering

Feature engineering is a method of converting operational variables into functions that best reflect the primary problem in analytical models and improve system performance for data that has not been used before. We used SMOTE-R to solve the data imbalance problem, which is discussed in the next section.

4.1.1 SMOTE for regression (SMOTE-R)

In classification problems, SMOTE is a tool for sampling imbalanced class distributions. This approach has the advantage of combining normal class sub-sampling with minority class over-sampling. Bowyer et al. demonstrate the benefits of this method over other alternate sampling models on several real-world difficulties using several classification
systems (Chawla et al. 2002). SMOTE-R, a version of SMOTE for regression tasks that correctly estimate rare extreme values, is proposed by Luis Torgo et al., who make a significant contribution (Torgo et al. 2013). The original SMOTE algorithm constructs fake cases with a rare target value using an over-sampling technique. To build these artificial cases, Fernandez et al. implemented an interpolation strategy (Fernández et al. 2018). The procedure involves selecting one of the k close neighbours for each event from a list of observations of uncommon values. These two results result in a new scenario that combines attribute values with the two original examples’ values. Both cases and synthetic cases belong to the same class as SMOTE is used to address classification issues using an unique class of concern in terms of the target variable. To build a general structure for predicting rare maximum values in a continuously varying target, the authors used a SMOTE-R technique combined with any other regression algorithm.

4.2 Important feature extraction

The feature pre-processing technique removes unimportant features from the CSV data before feeding it to the ensemble model. Feature importance metrics can aid in data interpretation, but they are often used to rank and pick features most relevant to a predictive model. The relative scores will indicate which features are most important to the target and, conversely, which features are least relevant. Using significance scores, feature importance can be used to construct a statistical model. Features with low scores will be removed, while features with high scores will be retained. The feature selection can help to simplify the problem being modelled, speed up the modelling process (deleting features is referred to as dimensionality reduction), and, in some cases, increase the model’s efficiency.

To effectively deal with the system’s unforeseen operating circumstances, the qualified model must be updated in actual implementations. As a result, a model upgrade mechanism is introduced to boost the generalization ability of the stability assessment model and achieve seamless online evaluation. Variable reduction is a technique for reducing the number of input variables. The more input features there are, the more complex the predictive modelling task becomes. Though dimensional reduction techniques are widely used for data visualization, they may also be used for classifiers or regression models with higher-dimensional data.

4.3 Ensemble load forecasting

As discussed in Section 3, we use different ensemble learning models to forecast the load and we found that the Random Forest model is giving the best precision over the remaining ensemble models. The proposed ensemble load forecasting is illustrated using flow chart as seen in Fig. 5.

4.4 Blended model

The term blending was used to refer to models that were constructed by stacking hundreds of forecasting models. Blending ensembles are a subclass of stacking in which the meta-model is fitted with predictions from a holdout validation dataset rather than out-of-fold forecasts. A stacking model’s architecture consists of two or more base models, which are frequently referred to as zero level models, and a meta-model that aggregates the predictions of the base models, which is referred to as a first level model. The meta-model is trained using out-of-sample predictions generated by basic models.

5 Experimental results

5.1 The dataset and feature extraction

Ensemble SLF model is validated using real-time data from the Nagaland Load Dispatch Center in India’s North-
Fig. 5 Flow diagram of Proposed Ensemble SLF

Eastern area (Abida Choudhury et al. 2019; Jhajharia et al. 2012). The dataset includes meteorological parameters, categorical time data, and load demand for a defined sampling time. Table 1 provides additional information about the meteorological parameters used in the dataset. The proposed load forecasting model was trained using data from Nagaland’s electric load demand between January 1st and December 31st, 2018. Hourly load data and 16 meteorological parameters were obtained over 2 years. Thus a total of 17,520 samples (730 days × 24 h = 17,520 samples) are being used as a database, as seen in Fig. 6. As a result, the size of the data is 17,520 × 17. This research used feature selection to increase the model’s accuracy and effectiveness. Without feature selection, researchers would be forced to conduct additional unnecessary data training with and without specific features in order to assess whether or not those attributes have an impact on the prediction’s accuracy.

The current work divides the total data set into training and testing data sets with 80% and 20% split, respectively as seen in Fig. 7. Figure 8 illustrates the histogram plot of the dataset used in training.

We selected features using the feature selection approach based on correlation coefficient and eliminated redundant features by evaluating the feature correlation matrix. The correlation matrix is an important data analysis metric used to analyze data to understand the relationship between different variables and make sound decisions. Each row and column in this matrix represents a variable with a value representing the coefficient of correlation between the variables defined by that row and column.

The Pearson correlation coefficient is the most basic method for determining the association between features and the target variable. The results’ interval is [−1, 1], where −1 represents a complete weak association (one variable rises while the other decreases), +1 represents a complete strong association, and 0 represents no linear association. Figure 9 shows the correlation map for the proposed system’s training data.

Each cell in the grid represents the correlation coefficient between two variables. It is a square matrix with each row containing a vector and all columns corresponding to the rows. The total number of columns is the same as the total number of rows. A high positive value (near 1.0) implies

Table 1 Meteorological parameters and their units

| Meteorological parameter | Units                  |
|--------------------------|------------------------|
| Temperature              | Degree Centigrade      |
| Wind Direction           | Degrees                |
| Humidity (Jhajharia et al. 2012) | Grams per cubic metre |
| Precipitation            | Millimetres            |
| Wind speed               | Kilometer per hour     |
| Pressure                 | Inches                 |
| Visibility               | Kilo meter             |
| Cloud Cover              | Okta                   |
| Heat Index               | Degree Centigrade      |
| Dew point                | Degree Centigrade      |
| UV index                 | Index number           |
| Wind Chill               | Degree Centigrade      |
| Feels Like               | Degree Centigrade      |
| Wind Gust                | Kilometer per hour     |
Bootstrap aggregating approach to short-term load forecasting using meteorological parameters for demand...
an excellent positive correlation, which means that when one variable's value rises, so does the other's value. The relationship between a and b is the same as the relationship between b and a. Diagonal cells reflect each variable's correlation with itself. Hence, all diagonal elements are equal to one.

Along with correlation matrix, we have also used another method for selecting important features, i.e., Recursive Feature Elimination (RFE). The RFE looks for a subset of the test-set attributes that begin with all of them and then successfully eliminates them until the correct list remains in place. This is accomplished by re-fitting the learning model, ranking features by score, discarding the least important features, and re-fitting it. This procedure is replicated until only a few features remain. By using RFE process, obtained the plot between number of features selected and its cross validation score as shown in Fig. 10. From this analysis, three features, i.e., day, time, and dew point, are extracted as important features. The time series feature importance also calculated as seen in Fig. 11.

5.2 Tuning of hyperparameters

Hyperparameter tuning is needed to obtain the ideal ensemble weights, which are then necessary to increase the accuracy of the ensemble load forecasting model. Table 2 lists the optimal tuning values for various hyperparameters for various models.

5.3 Evaluation of ensemble learning models

In the present study, we have used six ensemble regressor models including Ada-Boost, Extreme-Gradient-Boost, Cat-Boost, Light-Gradient-Boost, Random Forest, and
Extra-Trees Regressor models. Long Short-Term Memory (LSTM) and ARMA networks were also implemented and compared to demonstrate the proposed system’s feasibility. First, we used cross-validation to determine which ensemble learning model will be the best fit for the proposed SLF. Figure 12 shows the comparison of different machine learning models cross-validation performance with respect to Negative mean square error. The mean absolute error (MAE), mean absolute percentage error (MAPE), and mean square error (MSE) metrics have been used to assess the performance of ensemble learning algorithms. If $\hat{y}_i$ indicates the expected value of $i$th sample and $y_i$ indicates the corresponding actual value. Then the MSE over $n_{samples}$ may be given as follows:

$$\text{MSE} = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2.$$  (2)
Table 2 Optimum Hyperparameter values

| Model                        | Hyperparameters         | Optimum values               |
|------------------------------|-------------------------|------------------------------|
| SVR                          | Kernel, gamma           | ‘rbf’, 0.5                   |
| ARMA                         | lag values (p), residual error lag values (q) | p=4, q=1                     |
| LSTM                         | Epochs, batch-size      | Epochs=1000, batch size=4    |
| AdaBoost                     | n-estimators            | 350                          |
| Extreme-Gradient-Boost       | n-estimators            | 350                          |
| Light-Gradient-Boost         | n-estimators            | 100                          |
| Cat-Boost                    | n-estimators            | 300                          |
| Random Forest                | Forest-size             | 50                           |
| Extra-Trees                  | Forest-size             | 100                          |

Fig. 12 Comparison of machine learning models

Table 3 Evaluation of machine learning models on test set without feature importance

| Model                        | MSE    | MAE    | MAPE   |
|------------------------------|--------|--------|--------|
| SVR                          | 109.441| 8.145  | 19.412 |
| ARMA                         | 42.316 | 4.194  | 10.584 |
| LSTM                         | 90.120 | 6.239  | 14.412 |
| Ada-Boost                    | 113.619| 8.619  | 24.663 |
| Extreme-Gradient-Boost       | 49.737 | 4.704  | 11.560 |
| Light-Gradient-Boost         | 39.282 | 4.0767 | 10.335 |
| Cat-Boost                    | 37.535 | 4.016  | 10.031 |
| Random Forest                | 33.895 | 3.758  | 9.675  |
| Extra-Trees                  | 39.543 | 4.077  | 10.408 |
| Blended Model                | 42.434 | 3.889  | 10.561 |
### Table 4: Evaluation of machine learning models on test set with feature importance

| Model                    | MSE    | MAE    | MAPE  |
|--------------------------|--------|--------|-------|
| SVR                      | 91.941 | 7.678  | 18.211|
| ARMA                     | 30.359 | 3.818  | 9.415 |
| LSTM                     | 84.978 | 7.114  | 16.254|
| AdaBoost                 | 65.124 | 6.415  | 16.670|
| Extreme-Gradient-Boost   | 34.007 | 4.167  | 10.104|
| Light-Gradient-Boost     | 26.009 | 3.649  | 8.612 |
| Cat-Boost                | 28.968 | 3.649  | 8.612 |
| Random Forest            | 25.463 | 3.462  | 8.482 |
| Extra-Trees              | 28.461 | 3.722  | 9.057 |
| Blended Model            | 36.434 | 5.344  | 8.033 |

The MSE calculates accuracy by comparing the amount of squared errors between forecasting and real values. The MSE value will evaluate the error variation; the smaller the MSE value, the more stable the prediction model will be. The MAE is a measure of the average amount of combined absolute error between predictions and actual values, which can help avoid error mutual cancellation. As a consequence, MAE will correctly represent the absolute prediction error.

\[
\text{MAE} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |(y_i - \hat{y}_i)|. \tag{3}
\]

The relative difference between exact errors and real values is denoted by MAPE. The MAPE is a common metric for evaluating regression tasks.

\[
\text{MAPE} = \frac{100}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \tag{4}
\]

The Table 3 summarizes machine learning models’ evaluation using the MSE, MAE, and MAPE metrics without feature importance. As seen in the table, Random forest outperformed all other models, including LSTM, ARMA, and machine learning models. All state-of-the-art models were re-evaluated merely based on the important attributes obtained in the previous Section 5.1. The Random forest algorithm, as seen in Table 4, has the lowest MSE (25.463), MAE (3.462), and MAPE (8.482) values of the other algorithms.

### 6 Conclusion

In this analysis, an ensemble regressor was used to demonstrate a process of short term load forecasting in North-Eastern part of India. The suggested scheme is dependent on meteorological parameters and historical patterns of load use. At the pre-processing data stage, SMOTE-R analysis was used to solve the data imbalance issue. Numerous ensemble algorithms based on boosting and bagging techniques were investigated, including Ada-Boost, Extreme-Gradient-Boost, Cat-Boost, Light-Gradient-Boost, Random Forest, and Extra-Trees. The proposed design was tested and checked using historical load data collected from Nagaland’s Load Dispatch Unit in real-time. For the Random Forest and Extra-Tree Regressors, bagging methods yielded the lowest MAE values of 3.75 and 4.07, respectively. Boosting methods generated higher MAE and MAPE values than bagging methods and were thus unsuitable for real-time scenarios. Therefore, real-time use of bagging techniques such as Random forest and Extra-Trees is facilitated. Additionally, the mean absolute error and mean absolute percentage error study revealed that bagging methods outperform the other methods. The authors plan to continue their work by developing a generative adversarial network-based energy forecasting scheme based on domain information.

### Author contribution
All the authors participated and contributed equally in the analysis and interpretation of the results and data, drafting the article or revising it critically and preparing the final version.

### Availability of data and material
The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

### Code availability
The custom code used in the current study is available from the corresponding author on reasonable request.

### Declarations

**Conflict of interest** The authors declare no competing interests.

### References

Abida Choudhury B, Saha SK, Konwar M, Sujith K, Deshamukhya A (2019) Rapid drying of northeast India in the last three decades: climate change or natural variability. Journal of Geophysical Research: Atmospheres 124(1):227–237
Aier I, Sakhya AK, Kaushal P, Vijay VK (2021) Scope of renewable energy intervention for energy sufficiency in Nagaland. Integrated Approaches Towards Solid Waste Management. Integrated Approaches Towards Solid Waste Management, pp 317–328

Amato U, Antoniadis A, Feis ID, Goude Y, Lagache A (2021) Forecasting high resolution electricity demand data with additive models including smooth and jagged components. Int J Forecast 37(1):171–185

Andriopoulos N, Magklaras A, Birbas A, Papalexopoulos A, Valouxis C, Daskalaki S, Birbas M, Housos E, Papaioannou GP (2021) Short term electric load forecasting based on data transformation and statistical machine learning. Appl Sci 11(1):158

Borah P, Singh MK, Mahapatra S (2015) Estimation of degree-days for different climatic zones of North-East India. Sustainable Cities and Society 14:70–81

Cecati C, Kolbusz J, Różycki P, Siano P, Wilamowski BM (2015) A novel rbf training algorithm for short-term electric load forecasting and comparative studies. IEEE Trans Ind Electron 62(10):6519–6529

Chawla NV, Bowyer KW, Hall LO, Philip Kegelmeyer W (2002) Smote: Synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Chen K, Chen K, Wang Q, He Z, Hu J, He J (2018) Short-term load forecasting with deep residual networks. IEEE Trans Smart Grid 10(4):3943–3952

Din GMU, Mauthe AU, Marnerides AK (2018) Appliance-level short-term load forecasting using deep neural networks. In: 2018 International conference on computing, networking and communications (ICNC), IEEE, pp 53–57

Fan S, Hyndman RJ (2011) Short-term load forecasting based on a semi-parametric additive model. IEEE Trans Power Syst 27(1):134–141

Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets. Springer

Genuer R, Poggi Jean-Michel, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recogn Lett 31(14):2225–2236

Georganos S, Grippa T, Vanhuyse S, Lennert M, Shimoni M, Wolff E (2018) Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. IEEE Geosci Remote Sens Lett 15(4):607–611

Jhajharia D, Dinpashoh Y, Kahya E, Singh VP, Fakheri-Fard A (2012) Trends in reference evapotranspiration in the humid region of northeast India. Hydrology 26(3):421–435

Jhajharia D, Shrivastava SK, Sarkar DAS, Sarkar S (2009) Temporal characteristics of pan evaporation trends under the humid conditions of northeast India. Agricultural and Forest Meteorology 149(5):763–770

Jhajharia D, Yadav BK, Maske S, Chattopadhyay S, Kar AK (2012) Identification of trends in rainfall, rainy days and 24 h maximum rainfall over subtropical Assam in Northeast India. Compt Rendus Geosci 344(1):1–13

Jhaeveri S, Khedkar I, Kantharaya I, Jaswal S (2019) Success prediction using random forest, catboost, xgboost and adaboost for kickstarter campaigns. In: 2019 3rd international conference on computing methodologies and communication (ICCMC), IEEE, pp 1170–1173

Ke G, Qi M, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu Tie-Yan (2017) Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems, pp 3146–3154

Kumar S, Singh K, Kumar S, Kawaiyta O, Cao Y, Zhou H (2019) Delimited anti jammer scheme for internet of vehicle: machine learning based security approach. IEEE Access 7:113311–113323

Li S, Goel L, Wang P (2016) An ensemble approach for short-term load forecasting by extreme learning machine. Appl Energy 170:22–29

Li F, Jin G (2019) Research on power energy load forecasting method based on knn. International Journal of Ambient Energy, pp 1–6

Li F, Li Z, Chen B, Gao D, Cheng Y, Zhang X, Yang Y, Gao K, Huang Z, Peng J (2018) A light gradient boosting machine for remaining useful life estimation of aircraft engines. In: 2018 21st international conference on intelligent transportation systems (ITSC), IEEE, pp 3562–3567

Li W, Zhen-gang Z (2009) Based on time sequence of arima model in the application of short-term electricity load forecasting. In: 2009 International conference on research challenges in computer science, IEEE, pp 11–14

Miraftabzadeh SM, Foniadelli F, Longo M, Pasetti M (2019) A survey of machine learning applications for power system analytics. In: 2019 IEEE International conference on environment and electrical engineering and 2019 IEEE industrial and commercial power systems europe (EIEEIC/i&CPS europe), IEEE, pp 1–5

Mishra G, Jangir A, Francaviglia R (2019) Modeling soil organic carbon dynamics under shifting cultivation and forests using Rothc model. Ecol Model 396:33–41

Moradzadeh A, Moayyed H, Zakeri S, Mohammadi-Ivatloo B, Pedro Aguinar A (2021) Deep learning-assisted short-term load forecasting for sustainable management of energy in microgrid. Inventions 6(1):15

Nepal B, Yamaha M, Yokoe A, Yamaji T (2020) Electricity load forecasting using clustering and arima model for energy management in buildings. Japan Architectural Review 3(1):62–76

Oliveira MO, Marzec DP, Bordin G, Bretas AS, Bernardon D (2011) Climate change effect on very short-term electric load forecasting. In: 2011 IEEE Trondheim powertech, IEEE, pp 1–7

Pal S, Dutta S, Nasrin T, Surajit C (2020) Hurst exponent approach through rescaled range analysis to study the time series of summer monsoon rainfall over northeast India. Theor Appl Climatol 142(1):581–587

Park R-J, Song K-B, Kwon B-S (2020) Short-term load forecasting algorithm using a similar day selection method based on reinforcement learning. Energies 13(10):2640

Pedregosa F, Varoquax G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Ray SN, Bose S, Chattopadhyay S (2021) A Markov chain approach to the predictability of surface temperature over the northeastern part of India. Theor Appl Climatol 143(1):861–868

Sadac HJ, Enayatifar R, Abdullah AH, Gani A (2014) Short-term load forecasting using a hybrid model with a refined exponentially weighted fuzzy time series and an improved harmony search. International Journal of Electrical Power & Energy Systems 62:118–129

Singh AK, Bijimol G, Singh VB (1998) Performance of gladiolus in the application of short-term electricity load forecasting. In: 2009 International conference on research challenges in computer science, IEEE, pp 11–14

Sudheer G, Suseelatha A (2015) Short term load forecasting using clustering and arima model for energy management in buildings. Japan Architectural Review 3(1):62–76

Torgo L, Ribeiro RP, Pfahringer B, Branco P (2013) Smote for delimitated anti jammer scheme for internet of vehicle: machine learning based security approach. IEEE Access 7:113311–113323

D. Sarkar et al.
Trenberth KE (2011) Changes in precipitation with climate change. Clim Res 47(1-2):123–138
Velasco LCP, Polestico DLL, Abella DMM, Alegata GT, Luna GC (2018) Day-ahead load forecasting using support vector regression machines. Int J Adv Comput Appl (IJACSA) 9(3):22–27
Wan S, Yang H (2013) Comparison among methods of ensemble learning. 2013 International Symposium on Biometrics and Security Technologies, pp 286–290
Wood LA (1970) The use of dew-point temperature in humidity calculations. Journal of Research of the National Bureau of Standards–C. Engineering and Instrumentation C 74:117–122
Yang Y, Che J, Deng C, Li Li (2019) Sequential grid approach based support vector regression for short-term electric load forecasting. Appl Energy 238:1010–1021
Yang X, Cheng G, Liu H (2015) Improved empirical mode decomposition algorithm of processing complex signal for IOT application. International Journal of Distributed Sensor Networks 11(10):862807
Yi D, Dong Z, Zhao T, Li Z, Ding Z (2021) Short term load forecasting with markovian switching distributed deep belief networks. International Journal of Electrical Power & Energy Systems 130:106942
Yi L, Niu D, Hong W-C (2019) Short term load forecasting based on feature extraction and improved general regression neural network model. Energy 166:653–663
Yuk YY (1998) Climate and residential electricity consumption in Hong Kong. Energy 23(1):17–20
Zhang D, Qian L, Mao B, Huang C, Huang B, Si Y (2018) A data-driven design for fault detection of wind turbines using random forests and xgboost. IEEE Access 6:21020–21031
Zhao H, Yu H, Li D, Mao T, Zhu H (2019) Vehicle accident risk prediction based on adaboost-so in vanets. IEEE Access 7:14549–14557

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.