Journal of Applied Engineering and Technological Science
Vol 4(2) 2023: 693-710

REGRESSION MODEL-BASED SHORT-TERM LOAD FORECASTING FOR LOAD DISPATCH CENTER

Saikat Gochhait[1]*, Deepak K. Sharma[2]
Symbiosis Institute of Digital and Telecom Management, a constituent of Symbiosis International Deemed University[1 2]
saikat.gochhait@sidtm.edu.in, Deepak.sharma@sidtm.edu.in

Received : 11 February 2023, Revised: 06 April 2023, Accepted : 06 April 2023
**Corresponding Author*

ABSTRACT
In the wake of grid modernization initiatives, such as the integration of renewable energy sources and demand response, as well as the increasing popularity of electric vehicles, a greater degree of uncertainty has been introduced due to the fact that electricity demand has become more active and less predictable, and forecasting load has become increasingly difficult. Since the historical data is irregular, non-linear, non-smooth, and noisy, it is difficult to achieve a satisfactory result. The present study overcomes the challenges with the help of an improved energy demand forecasting model for load dispatch centers as part of an Artificial Intelligence (AI) driven project supported under research grants from the Department of Scientific and Industrial Research. A real-time hourly load consumption dataset was collected from Regional Load dispatch centers from July 1, 2020, to August 22, 2022. In this paper, 24 regression model-based day-ahead load forecasting algorithms are developed and evaluated using the load consumption and meteorological data collected from NASA Power (https://power.larc.nasa.gov/). MATLAB Regression Toolbox offers 24 regression models divided into five families: Linear Regression, Tree Regression, Support Vector Machines (SVM), Gaussian Process Regression (GPR), and Ensemble of Trees. Since GPR models are nonparametric kernel-based probabilistic models, they show the best load forecasting performance. The study recommends two GPR models for load forecasting: Rational Quadratic GPR and Exponential GPR
Keywords: *Short-term load forecasting, regression model, Gaussian process regression, probabilistic models, Subdivision electricity load*

1. Introduction

As the smart grid continues to evolve and demand from industry and academia for more efficient electricity scheduling grows, short-term load forecasting (STLF) has attracted increasing interest. In the context of energy providers, short-term load forecasting (STLF) has a significant impact on generating energy, processing energy distribution efficiently, and optimizing electricity prices (Gochhait et al., 2020; Datar et al, 2021). Load forecasting is a critical component of power grids and is necessary for maintaining the balance between supply and demand (Gochhait et al.,2022). Accurate load forecasts are essential for power grids to operate effectively, and these forecasts must be accurate over a wide range of time periods (Chen et al., 2018). The accuracy of electricity demand models is also critical for power system reliability and operating costs. United Kingdom (UK) utility, reducing load forecasting errors by just 1% resulted in a £10 million per year reduction in operating costs. Therefore, accurate load forecasting is crucial for minimizing operating costs and maintaining reliable power grids (Gilanifar et al., 2020). Electric vehicles, demand response, and renewable have brought about a modernization of the power grid, resulting in increased uncertainty. This transformation has led to more variability in electricity demand, making load forecasting more challenging as predictability decreases (Subhasri & Jeyalakshmi, 2018). A smart distribution grid, which can decentralize demand and generate renewable energy, is the future of the power grid (Patil et al., 2019). However, end-user behavior can significantly affect load, causing sudden variations that make high-resolution and accurate load forecasting crucial for the success of the smart distribution grid. Accurate load forecasting is necessary to minimize the impact of increased demand variability on power grids (Okoye & Madueme, 2016).

Advanced technologies such as smart meters and advanced infrastructure metering have led to the development of load forecasting collection systems (Gochhait et al., 2022). With advances in data analysis and artificial intelligence techniques, academia and industry are

increasingly interested in load forecasting. Load forecasting has three categories: short term, medium term, and long term. Short-term load forecasts are made within a period of a few minutes to a week, while medium and long-term forecasts require at least two weeks of data, and a short-term forecast requires at least three years of data. Accurate load forecasting is essential for minimizing the impact of increased demand variability on power grids and the success of a smart distribution grid, making it a vital area of research and development (Gilanifar et al., 2020; Chang et al., 2019). Short-term load forecasting is critical in various planning scenarios such as demand response, electricity trading, commitment to units, and inverse dispatch. Several techniques, including bidding processes, contingent energy transactions, and clearing processes, can be used to forecast future electricity demand on an hourly basis (Caro et al., 2020). Therefore, accurate short-term load forecasting is crucial for the reliable and efficient operation of the power grid (Feng et al., 2019). Fast restoration methods are essential to minimize the amount of unmet demand in the event of a fault for a robust final restoration. To achieve this, it is necessary to forecast short-term load ahead of time (Feng et al., 2019).

Traditional methods such as expert systems, which mimic the expertise of well-trained experts, have been used for forecasting load. However, these rules for forecasting load can be challenging to translate, and the development of artificial intelligence techniques has led to the creation of adaptive data set training methods (Mele, 2019). The use of artificial intelligence techniques has made it challenging to distinguish between independent variables and dependent variables mathematically (Okoye & & Madueme, 2016). Regression algorithms can be used to predict short-term load forecasting. Regression is a mathematical tool that can establish statistical correlations between variables and provides information about the relationship between parameters that can be explored by examining their magnitude and trend. Regression allows the use of multiple predictors and predicts outcomes even when there are multiple interdependent predictors, making it superior to simplified analyses based on interrelated variables. Furthermore, regression allows the correction of errors resulting from inferences based on previous results, and excellent results can be obtained with relatively modest data sets. Therefore, regression algorithms are a valuable tool for predicting short-term stress. The literature discusses various approaches to load forecasting, including 1) deterministic (point-based) load forecasts (Gilanifar et al., 2020), (Cao et al., 2020). 2) probabilistic load forecasting (Yildiz et al., 2017). 3) hybrid methods combining point forecasts with probabilistic load forecasts (Massana et al., 2015).

Probabilistic load forecasting has been comparatively underexplored compared with traditional forecasting. Electricity demand has become more uncertain and variable in smart grids, which requires probabilistic load forecasting. A number of factors affect the demand for electricity in the residential electricity distribution grid, including demand response programs and feed-in changes. Probability-based load forecasts provide interval load forecasts as well as scenario projections, functional densities, and probability distributions. They have been used to forecast probabilistic prices, plan probability-based lines, or allocate uncertain units (Massana et al., 2016). Based on regression, a load analysis has been developed for the entire country and specific regions, as well as smaller loads such as community microgrids. Variables related to climatic conditions and time can be accounted for in models developed for short, medium, and long-term schedules. Due to its significant impact on power security and load stability (Yildiz et al., 2017), the prediction of time-limited loads is essential.

In regression models, Multiple Linear Regression (MLR), Artificial Neural Networks (ANN) and Support Vector Fig. 1. Monthly Feeder Input (kWh) Regression (SVR) are compared, with Support Vector Machines showing the best performance (Prakash et al., 2018). In this paper, the regression models Autoregressive Integrated Moving Average with exogenous variables (ARMAX) Approach, ANN Approach, MLR Approach, and SVR Approach (Jiang et al., 2016) are examined. The SVR group, Randon Forest (RF), and the Gaussian Process Regression (GPR) group are compared (Deng et al., 2019). The effectiveness of machine learning regression models in predicting short-term loads has been established to be superior to traditional regression models (Khadka et al., 2020). This study aims to compare the effectiveness of different regression models (Cao et al., 2020), namely linear regression, support vector machine (SVM), ensemble trees, and Gaussian process regression (GPR) algorithms in forecasting load. The models are critical in maintaining the integrity and safety of loads, especially short-term loads. To improve the

effectiveness of short-term forecasts, robust regression algorithms should be applied to short-term model forecasts. Regression models were employed to analyze load forecasts for a given region while considering grid microsystem connections to local communities. The models use various parameters related to the local climate and time frames, ranging from short to long-term. While the models are useful for large loads, predicting short-term loads is critical. Robust regression algorithms could be applied to short-term model forecasts to improve the accuracy of empirical short-term forecasts (Liang et al., 2016). The study concluded that machine learning regression models, such as the GPR method, outperformed traditional regression models, including linear regression, SVM, ensemble trees, and neural networks (Hong & Fan, 2019). The weather-sensitive component is forecasted using a Support Vector Regression (SVR) model based on historical load data and meteorological data. Data on the real electricity load is used to test the proposed method (Qiuyu et al., 2017).

The present study proposes the effectiveness of the different models evaluated to identify short-term load forecasting. Therefore, regression models based on machine learning are recommended for developing models that can identify short-term loads (Hammad et al., 2020). The proposed algorithm was adopted to determine the most suitable algorithm for the electricity load profile (Liang et al., 2021). The best regression models are identified by Mean absolute percentage error (MAPE), MAE, and Root Mean Square Error (RMSE). The most efficient regression models were selected from 24 preliminary regression models. This was accomplished by using various actual and predicted plots, response plots, and residual plots. Several iterations of this process were performed to arrive at the recommended final regression model. In addition, this paper presents a viable approach for load forecasting and demonstrates that GPR models are more accurate.

A detailed description of the Load dispatch center (LDC) load is given in Section II, followed by an explanation of the proposed load forecasting approach. The results of the simulations are presented in Section III, followed by the analytical results obtained by using 24 regression models. Fig. 2. Feeder Input (kWh) season-wise in section IV, the simulation results and data analysis are presented and recommendations are derived

2. Literature Review

In this section, we describe the data collection approach and forecasting approach that will be used to gather data from load dispatch centers.

Maharashtra State Electricity Distribution Company Ltd (MSEDCL)

The study considered data from MSEDCL in terms of Region(PUNE REGION), Zone(PUNE ZONE), Circle (PUNE (R) CIRCLE), Division (MANCHAR O&M DIVISION), Subdivision BU(Billing Unit), Substation Name(33/11 KV NETWAD SUBSTATION), Feeder Name (Netwad), Feeder Type Description(Single Phasing Feeders), Month(July 2020 to Sept 2022), Day profile Date(30 Mins Timeframe), Interval(48 Interval), Active Energy (kWh), Reactive Energy (KVarch or reactive energy). If the power factor is less than 90%, will be billed for KVARH. The actual power consumed by the load is called kilowatt power. All the power given to the load is not utilized as useful power. The non-useful power is called reactive power KVARH.), File Source (MRI or AMR).

In this study, data from substation Ale Phata was reviewed and active power (kWh) was calculated according to the daily day profile over a period of 30 minutes on a monthly, yearly, and seasonal basis. We can observe the Feeder input for particular months of 2020 and their seasons respectively in Table 1.

Table 1. Monthly feeder input in kilowatt hours (kWh) from 2020 to 2022.

Months	Feeder Input (kWh)	Season
Jan2020	17097659.18	Winter Season
Feb2020	14409569.88	Winter Season
Mar2020	17453211.60	Spring Season
Apr2020	18040410.80	Spring Season
May2020	17628819.60	Spring Season
Jun2020	11651796.40	Summer Season
Jul2020	11209917.20	Summer Season

Aug2020	10498156.40	Summer Season
Sep2020	10402455.60	Autumn Season
Oct2020	11952921.00	Autumn Season
Nov2020	14611530.00	Autumn Season
Dec2020	15634406.20	Winter Season
Jan2021	10578076.80	Winter Season
Feb2021	11638474.80	Winter Season
Mar2021	14621277.20	Spring Season
Apr2021	12040741.28	Spring Season
May2021	10860208.00	Spring Season
Jun2021	8067561.20	Summer Season
Jul2021	8114217.60	Summer Season
Aug2021	9725654.00	Summer Season
Sep2021	8010404.00	Autumn Season
Oct2021	9678877.00	Autumn Season
Nov2021	10483703.40	Autumn Season
Dec2021	11051949.60	Winter Season
Jan2022	12464362.00	Winter Season
Feb2022	9960312.00	Winter Season
Mar2022	14967082.00	Spring Season
Apr2022	14233508.00	Spring Season
May2022	18968646.00	Spring Season
Jun2022	9619716.00	Summer Season
Jul2022	7517852.00	Summer Season

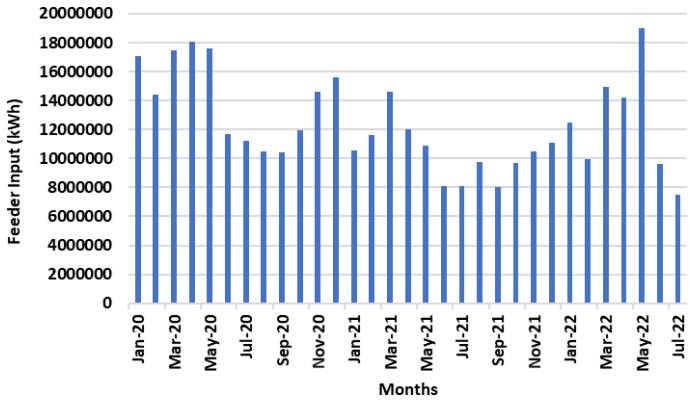

Fig. 1. Monthly Feeder Input (kWh)

Table 2 - Feeder Input (kWh) seasonal change from July 2020 toAugust 2022.

Year	Seasons	Feeder Input (kWh)
2020	Winter Season	47141635.26
2020	Spring Season	53122442.00
2020	Summer Season	33359870.00
2020	Autumn Season	36966906.60
2021	Winter Season	33268501.20
2021	Spring Season	37522226.40
2021	Summer Season	25907432.80
2021	Autumn Season	28172984.40
2022	Winter Season	22424674.00
2022	Spring Season	48169236.00
2022	Summer Season	17137568.00

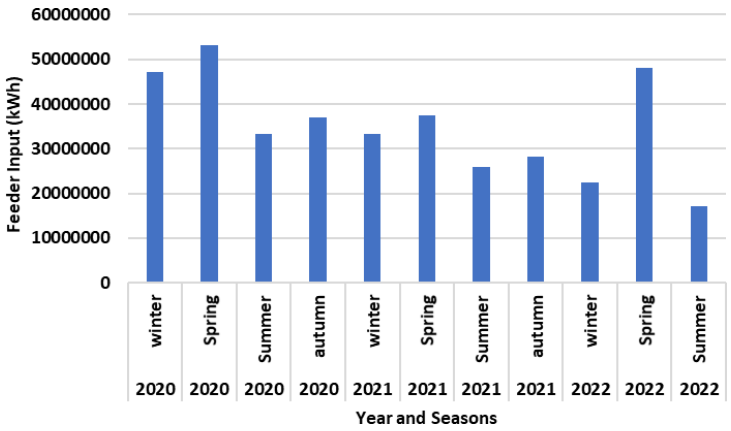

Fig. 2. Feeder Input (kWh) season wise

The seasonal periods are defined as follows:
1) Spring (March to May); 2) Summer (June to August); 3) Fall (September to November); and 4) Winter (December and February). Table 2 and Fig. 2 illustrate the seasonal feeder inputs.

Proposed Load Forecasting Approach and Model Creation

This paper proposes an effective load forecasting method using regression models. Six steps are required to implement the proposed approach

Step 1: Data collection. The two datasets required are historical load demand data and historical meteorological data.

Step 2: In the proposed method, regression models are used to forecast load, and suitable regression models are selected for load forecasting purposes. This paper presents 24 regression models.

Step 3: Input parameters selection. A variety of important input parameters are evaluated and selected, including weather parameters.

Step 4: A regression model was created and a load forecast was conducted based on it. The regression models selected in Step 2 will be trained and tested and then used to forecast load.

Step 5: Regression models are compared in terms of performance. Regression models are evaluated using statistical error matrices in order to compare their performance with actual measured loads.

Step 6: Provide recommendations for regression models that are most accurate. Using the results from the previous steps, the best regression models will be selected.

Data Collection and Preprocessing

As input-output datasets are required to train the model, Step 1 on data collection and pre-processing is crucial. In order for the model to effectively learn the input-output relationship, the raw data must be pre-processed before transforming it. Pre-processing operations include normalization, ranking, and correlation (Olagoke et al., 2016).

Data Collection

Climate data, calendar data, and load demand data are included in the dataset collection process

3. Research Methods

The meteorological dataset was obtained from the NASA Power website (https://power.larc.nasa.gov/data-access-viewer/). The raw data must be pre-processed before being transformed in order for the model to effectively learn the input-output relationship. The pre-processing operations include normalization, ranking, and correlation (Chane et al., 2021). In order to minimize operational costs, load forecasting models use weather forecasts and other elements to predict the future load (Chane et al., 2021).

The weather plays an important role in load forecasting. Weather effects are most prevalent among domestic and agricultural customers, as well as affecting their load profiles.

1) Wet Bulb Temperature at 2 Meters (C)(T2MWET): -The adiabatic saturation temperature which can be measured by a thermometer covered in a water-soaked cloth over which air is passed at 2 meters above the surface of the earth.
2) Dew/Frost Point at 2 Meters (C)(T2MDEW): -The dew/frost point temperature at 2 meters above the surface of the earth.
3) Temperature at 2 Meters (C)(T2M): The average air (dry bulb) temperature at 2 meters above the surface of the earth.
4) Relative Humidity at 2 Meters (%) (RH2M): -The ratio of actual partial pressure of water vapor to the partial pressure at saturation, expressed in percent.
5) 5). Specific Humidity at 2 Meters (g/kg) (QV2M): -The ratio of the mass of water vapor to the total mass of air at 2 meters (kg water/kg total air).
6) Wind Speed at 10 Meters (m/s) (WS10M): -The average wind speed at 10 meters above the surface of the earth.

There is a greater impact of temperature and humidity on load variations than other weather parameters, although all have a direct impact on load requirements.

Time Indicator

In short-term load forecasting, time plays a critical role since it has the greatest impact on the customer's load demand. Time indicators used in this study include the date, weekday, and time (Chane et al., 2021).

Load Parameters

The following load parameters are used for the load demand data in this study: previous Half Hour Active Power in kWh; previous Half Hour Reactive Energy (Rkvah).

Table 3 - Correlation Analysis of Parameters

SN	Parameter	Relation Coefficient
1	T2MWETr	-0.062563795
2	T2MDEW	-0.42138371
3	T2M	0.383244621
4	RH2M	-0.475503209
5	QV2M	-0.555256085
6	WS10M	-0.07992978
7	Reactive Energy (Rkvah)	0.557622633

Correlation analysis has been conducted in order to investigate the relationship between the selected weather parameters and the electrical load. Correlation values close to 1 indicate a strong relationship. Positive signs indicate a proportional relationship, while negative signs indicate an inversely proportional relationship. As a result of the correlation analysis, the following results have been presented in Table 3.

Data Preprocessing: Real-life measurements are susceptible to various degrees of discrepancies including incomplete data, noise, missing values, outliers, redundant data, and inappropriate formatting, which influence the performance of the regressors. Therefore, the data must be pre-processed to ensure data reliability (Garcia et al., 2015; Wahyudi & Arroufu, 2022).

Data Cleaning: It includes filling in the missing values, noise removal, outlier detection, and resolving discrepancies within the dataset. So, the study used the auto-fill feature to fill cells with data that follows a pattern or is based on data in other cells. the partial missing weather data (Han et al., 2016; Wahyudi, T., & Silfia, 2022).

Data transformation: The process of data transformation involves integrating multiple files into a single usable format and scaling the attribute in accordance with specific properties. We created the final predictor's dataset after finding the correlations between data sets and cleaning the data (Han et al., 2016).

Data Reduction: By reducing the number of attributes or by sampling, data reduction attempts to capture most of the properties of the data while removing redundancies

4. Results and Discussions

Initially, a training dataset is used to train a model, and then the results are analysed by varying its parameters until the most effective parameters are identified, resulting in an optimized model for all regression models. The linear Model represents a fitted linear regression model. Regression models describe the relationship between a response and its predictors. Linearity in a linear regression model refers to the linearity of the predictor coefficients. A linear model object can be used to investigate the properties of a fitted linear regression model. A property of an object contains information about coefficient estimates, summary statistics, fitting methods, and input data. Modify, evaluate, and visualize the linear regression model by using the object functions.

Training regression models using Regression Learner includes linear regression models, regression trees, Gaussian process regression models, support vector machines, kernel approximations, ensembles of regression trees, and neural network regression models. In addition to training models, explored data, selected features, specified validation schemes, and evaluated the results. MATLAB code was generated to learn about programmatic regression after exporting a model to the workspace.

In Regression Learner, a model is trained in two stages:

A validated model is one that has been trained using a validation scheme. As a default, the application employs Hold out Validation to prevent overfitting. As an alternative, holdout validation may be chosen. In the application, the validated model can be viewed. The full model is trained on all the data, excluding the test data. This model is trained simultaneously with the validated model by the application. It is, however, not possible to view the model that has been trained on full data in the application. Regression Learner exports the full regression model when you select a regression model to export to the workspace. In the Regression, the results of the validated model are displayed. Diagnostic measures, such as the model accuracy, and response plot or residual plot, reflect the results of the validated model.

A regression model was trained, validation results were compared, and the best model was selected. Regression Learner exported the full model from the chosen model to the workspace. Since Regression Learner created a model object of the full model during training, there was no lag time when we exported the model. The exported model can be used to make predictions based on new data.

Holdout Validation: - This is the simplest form of cross-validation. As opposed to simple or degenerate cross-validation, this method is often classified as a "simple validation". Our data is randomly divided into two sets: Training and Test/Validation, or hold-out data. The model was then trained on the training dataset and evaluated on the test/validation dataset. In order to compute the error on the validation dataset, we use a variety of model evaluation techniques depending on the problem we are solving, such as MSE for regression problems and a number of metrics that indicate the misclassification rate for classification problems in order to find the error. It is a typical method in which the training dataset is larger than the hold-out dataset, so an 80:20 ratio for the training and testing datasets was applied.

Regression Models

In this paper, six families of regression model algorithms provided in the MATLAB Regression are selected to construct the short-term load forecasting model for the Ale Phata Subdivision of Manchar O&M division. They are Linear Regression, Regression Trees, Support Vector Machines (SVM), Gaussian Process Regression (GPR), Ensemble of Trees, and Neural Networks. Table 1V depicts the regression models used in this study.

Linear Regression: Modeling the relationship between the independent variable and the dependent variable, it is one of the simplest regression models used to forecast outcomes. The model attempts to determine the relationship between two or more explanatory variables and a response variable since there are several independent and dependent variables (Jawad et al., 2020).

Regression Trees: All regression models utilize a single output variable (response) and multiple input variables (predictors). There is a numerical output associated with the variable. A combination of continuous and categorical variables may be used as input variables in the general

regression tree structure method. Binary recursive partitioning is the process by which regression trees are constructed. By dividing the data into partitions or branches, and then dividing each section into smaller groups as each upper branch is reached, this method determines the groups for each section of the data. Initially, all training sets are grouped together. A binary split is performed on each field in the first two partitions, or branches, of the algorithm to assign the data. Based on the algorithm, the division that minimizes the sum of the squared variances in the two different partitions from the mean is selected. The new branches are then added to this dividing guideline. The process continues until all nodes exceed the user's minimum node size and become terminal nodes.

Table 4 - Families of Regression Models.

SN	Families of Regression model	Regression Model
1	Linear Regression Models	Linear Interactions Linear Robust Linear Step-wise Linear
2	Regression Trees	Fine Tree Medium TreeCoarse Tree
3	Support Vector Machines	Linear SVM Quadratic SVMCubic SVM Fine Gaussian SVM Medium Gaussian SVMCoarse Gaussian SVM
4	Gaussian Process Regression Models	Squared Exponential GPR Matern 5/2 GPR Exponential GPR Rational Quadratic GPR
5	Ensembles of Trees	Boosted Trees Bagged Trees
6	Neural Networks	Narrow Neural Network Medium Neural Network Wide Neural Network Bilayered Neural Network Trilayered Neural Network

Support Vector Machines: In SVM, kernel functions are used for nonlinear transformations. In this study, standard kernel functions are used, such as linear kernels, polynomial kernels, Gaussian kernels, and radial basis functions. The results of polynomial functions of the lower degree tend to be inadequate because they underfit the model. It is more appropriate for the curve to be fitted as the degree of the polynomial increases.

Gaussian Process Regression: A Gaussian Process Regression (GPR) model is a nonparametric kernel-based probabilistic model with a finite set of random variables with a multivariate distribution. The distribution of any linear combination is equal. As a generalization of multivariate normal distributions, the Gaussian distribution is named after Carl Friedrich Gauss. In statistical modeling, regression to multiple target values, and mapping in higher dimensions, Gaussian processes are used (Semero et al., 2018).

Ensemble of Trees: Ensembles use a variety of algorithms in order to increase their efficiency and predictability egression Model-Based Short-Term Load Forecasting technique that is designed to improve the accuracy and stability of machine learning algorithms. It is used to create a linear combination of model fitting instead of using a single-fit method. Rather than using a single fit method, multiple predictors are constructed and intertwining the best regression models for forecasting short-term load, we performed the following: 1) Training a data set with cross-validation of 5 folds for each regression model; 2) Plotting the behavior of regression models using RMSE, R-Squared Value, MSE, MAE, and 3) analyzing the results to determine whether there is any similarity or difference between the data (Pirbazari et al., 2016).

Neural Networks: Regression Neural Network is a fully connected, feedforward, and trained neural network for regression. There is a connection between the first fully connected layer of the neural network and the network input, and each subsequent layer has a connection from the previous layer.

The proposed method has been shown to be feasible and effective based on application results. The application of neural network prediction shows the efficiency and capability of the proposed techniques for predicting load demand (Badran & Abouelatta, 2012).

In order to identify the best regression model for short-term load forecasting, we performed the following steps:

Training a data set with a holdout validation ratio of 80:20 for all models;

Plotting the behavior of regression models with RMSE, R-Squared Value, MSE, and MAE

Analyzing the data to determine similarities and differences

Datasets Description

The proposed work uses half-hour interval data from July 2020 to July 2022 as the simulation dataset. The columns with completely missing weather data were eliminated.

Validation Data

From the dataset from July 2020 to July 2022, we take out the data from March 2022 to July 2022, forming a new dataset that serves as the validation dataset for the created model. 7,547 rows and 10 columns make up the validation dataset.

Training and Testing Data

After the validation dataset is created, from the rest of the historically recorded dataset, 80% of the data is used for training, and the remaining 20% of data is used for testing. The training data for the regression models are Hold out -validation. The testing data are treated as unseen by the trained model and used to optimize the load forecasting model's control parameters, which helps to optimize and evaluate the performance of the model created.

The validation and training data can be viewed as a matrix with 37733 rows and 10 columns. The rows represent each hour of a day from July 1, 2020, to July 31, 2022, including the validation dataset. The first 9 columns are the predictors or input, and the last column is the training target data, i.e., the load in kWh.

Performance Evaluation Dataset

For performance evaluation after training and testing the model, a random day (one Month data) from the validation dataset which falls in the month of Aug 2022 was selected representing season of the year.

Performance Evaluation Indices

Each regression model is compared with the forecasted load and the actual measured load. Therefore, load forecasting capacity and model accuracy are assessed by calculating three different statistical evaluations, the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE) (Chen et al., 2004; Massanaet al., 2015). There is a difference between RMSE and MAE if RMSE > MAE, it means that there is a variation in errors.

Mean Absolute Error (MAE)

The MAE measures the average magnitude of the errors, which can be calculated by

$$\text{MAE} = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{n}$$

where Yˆt is the prediction, Yt is the true value from field recording, and n is the number of measurement points (Madhukumar et al., 2022).

Mean Absolute Percentage Error (MAPE)

This error percentage is a measure of the prediction accuracy of a forecasting method in statistics, it produces a measure of the relative overall fit, which can be calculated by

$$\text{MAPE} = \frac{\sum_{x=1}^{n} \frac{|y_t - \hat{y}_t|}{y_t}}{n}$$

Where Yˆt is the prediction, Yt is the true value from field recording, and n is the number of measurement points (Ceperic et al., 2013).

Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (y_t - y_t)^2}{n}}$$

where Yˆt is the prediction, Yt is the true value from field recording, and n is the number of measurement points (Ceperic et al., 2013).

Table 5 - Validation Results of the Regression Models

SN	Model/Approach	RMSE of Validation	R^2 of Validation	MSE of Validation	MAE of Validation	Time of training (Sec)
1	Linear	1.8802	0.81	3.5351	1.5055	7.2304
2	Interactions Linear	1.7624	0.83	3.1061	1.3942	7.7953
3	Robust Linear	1.8871	0.81	3.5613	1.4963	4.2474
4	Stepwise Linear	1.7626	0.83	3.1068	1.3944	91.245
5	Fine Tree	1.0073	0.95	1.0147	0.6424	7.6129
6	Medium Tree	0.97901	0.95	0.95846	0.6453	4.7358
7	Coarse Tree	1.0313	0.94	1.0635	0.68931	4.4082
8	Linear SVM	1.9006	0.8	3.6122	1.4903	185.51
9	Quadratic SVM	1.6384	0.85	2.6842	1.2051	346.17
10	Cubic SVM	1.3695	0.9	1.8756	0.99412	1272.6
11	Fine Gaussian SVM	0.82484	0.96	0.68037	0.53716	337.14
12	Medium Gaussian SVM	1.1515	0.93	1.326	0.82051	291.12
13	Coarse Gaussian SVM	1.6497	0.85	2.7214	1.2274	403.47
14	Boosted Trees	1.3153	0.91	1.7299	0.961	345.16
15	Bagged Trees	0.83982	0.96	0.70529	0.56594	358.67
16	Squared Exponential GPR	0.91284	0.95	0.83328	0.64339	1617.2
17	Matern 5/2 GPR	0.7596	0.97	0.57699	0.51619	2126.8
18	Exponential GPR	0.62887	0.98	0.39548	0.38161	2211.7
19	Rational Quadratic GPR	0.64517	0.98	0.41625	0.3913	7383.8
20	Narrow Neural Network	1.1138	0.93	1.2405	0.80984	1643.8
21	Medium Neural Network	1.052	0.94	1.1066	0.74989	1700.3
22	Wide Neural Network	0.91351	0.95	0.8345	0.62238	1918.1
23	Bilayered Neural Network	1.0006	0.95	1.0012	0.69425	1966.3
24	Trilayered Neural Network	0.99506	0.95	0.99014	0.68916	2036.5

Table 6 - Error Parameters for Half Hourly for Validation and Testing.

SN	Model /Approach	RMSE (Validation)	R-Squared (Validation)	MSE (Validation)	MAE (Validation)	Prediction speed (Obs/ sec)	RMSE (Test)	R-Squared (Test)	MSE (Test)	MAE (Test)
1	Exponential GPR	0.6105	0.98	0.373	0. 3541	5600	0.9218	0.82	0.8497	0.692
2	Rational Quadratic GPR	0.62392	0.98	0.3893	0.36345	2600	0.97756	0.8	0.95563	0.728
3	Matern 5 and2 GPR	0.72907	0.97	0.5315	0.47448	4400	0.90906	0.83	0.82639	0.683
4	Bagged Trees	0.83648	0.96	0.6997	0.55738	32000	0.87062	0.84	0.75799	0.64
5	Squared Exponential GPR	0.86667	0.96	0.7511	0.59589	6100	0.93926	0.81	0.88221	0.673
6	Fine Gaussian SVM	0.90178	0.95	0.8132	0.53527	5600	0.91452	0.82	0.83635	0.684

Table 7 - Error Parameters for Hourly Time Duration for Validation and Testing

SN	Model /Approach	RMSE (Validation)	R-Squared (Validation)	MSE (Validation)	MAE (Validation)	Prediction speed (obs/ sec)	RMSE (Test)	R-Squared (Test)	MSE (Test)	MAE (Test)
1	Exponential GPR	3.0048	0.87	9.0288	2.0062	4800	7.9833	-0.9	63.732	0.692
2	Rational Quadratic GPR	3.034	0.87	9.205	2.0355	2800	8.054	-0.94	64.868	6.3881
3	Matern 5/2 GPR	3.1618	0.86	9.997	2.2094	3000	7.7889	-0.81	60.668	6.0174
4	Bagged Trees	3.2381	0.85	10.485	2.2553	32000	7.7823	-0.81	60.564	6.2896
5	Squared Exponential GPR	3.2995	0.84	10.887	2.3724	3700	8.8359	-1.33	78.074	6.7843
6	Fine Gaussian SVM	3.2087	0.85	10.296	2.1074	8300	8.347	-1.08	69.673	7.1238

R-Squared

R-Squared is a statistical measure of how close the fitted regression line is to the results. R-squared lies between 0 and 1. Generally, a higher R-squared value implies that the model matches the data better. The following criteria is used to evaluate load forecasting performance using the error indices:

The RMSE is always positive, and a smaller RMSE value indicates a good model (Madhukumar et al., 2022).

The R-squared lies between 0 and 1. R-Squared indicates a good model near 1.

The MSE is the square of the RMSE, and a smaller MSE value indicates a successful model

The MAE is positive, similar to RMSE, a smaller MAE value suggests a successful model.

An error percentage very close to zero means the predicted values are very relative to actual values.

Predicted versus actual response plots

An actual vs. predicted response plot is a way to compare the actual values of a dependent variable to the predicted values obtained from a regression model. This plot helps to visualize the performance of the model by showing how well the predicted values match the actual values.

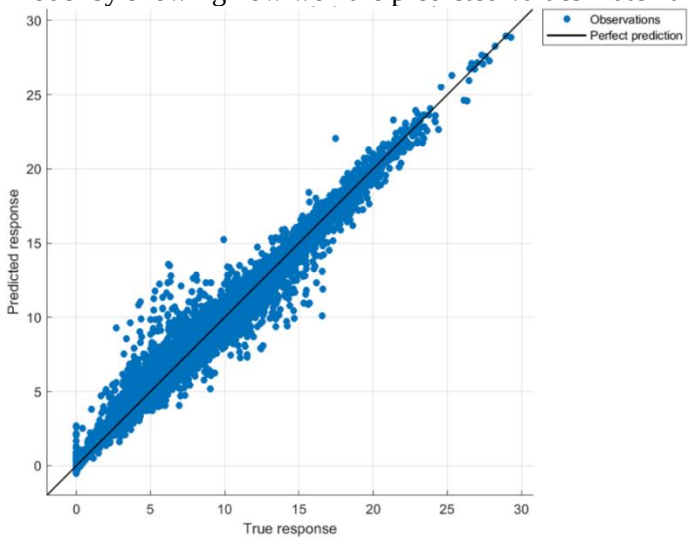

Fig. 3. Predicted Vs Actual Response Plotting of Validation Exponential GPR

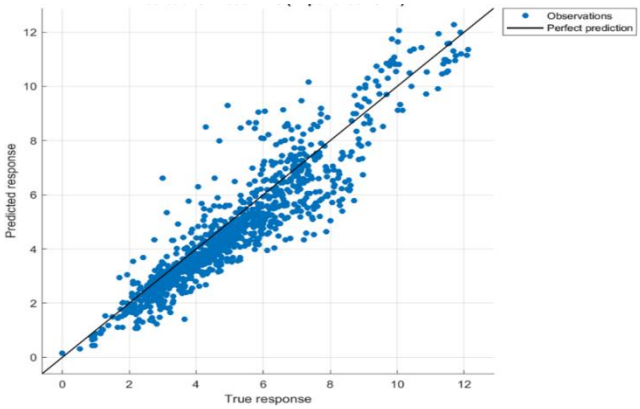

Fig. 4. Predicted Vs Actual Response Plotting of Testing Exponential GPR

Plotting of Residues of Regression

In order for a model to perform well, the following requirements must be met: Residuals are asymmetrically distributed in the vicinity of zero, As seen in the diagram, residuals have a significant impact on size when viewed from the right. As shown in the residual plot, active energy (kWh) is analyzed for the interval (48), along with validation and testing in fig. 5 and 6.

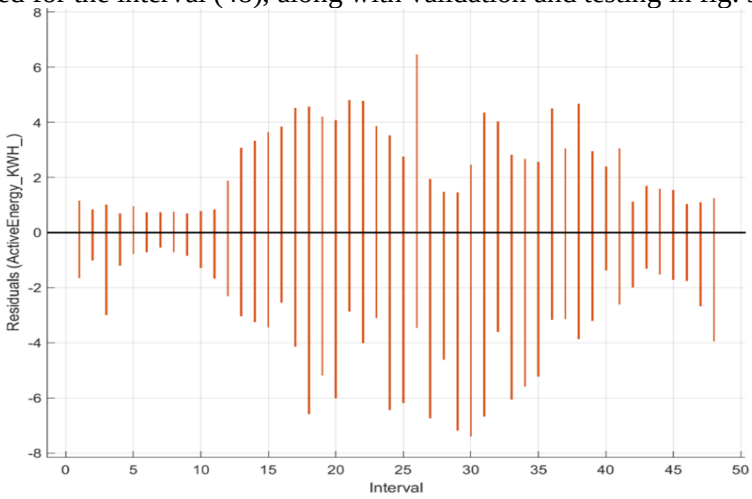

Fig. 5. Residual plot of validation exponential GPR

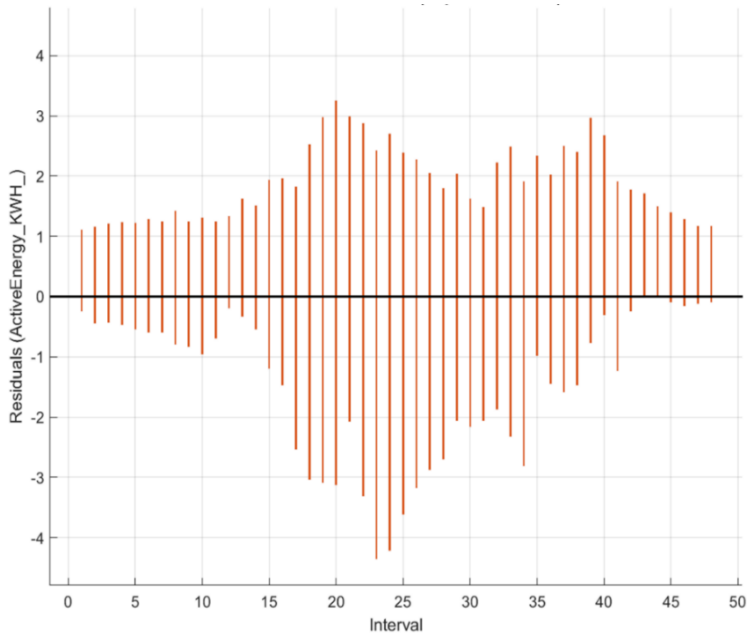

Fig. 6. Residual Plot of Testing Exponential GPR

Response Plots of Regression

Plotting the predicted response against the actual response is shown in the Fig.7 of the response plot. It indicates good performance of a model if the true response and predicted response are identical.

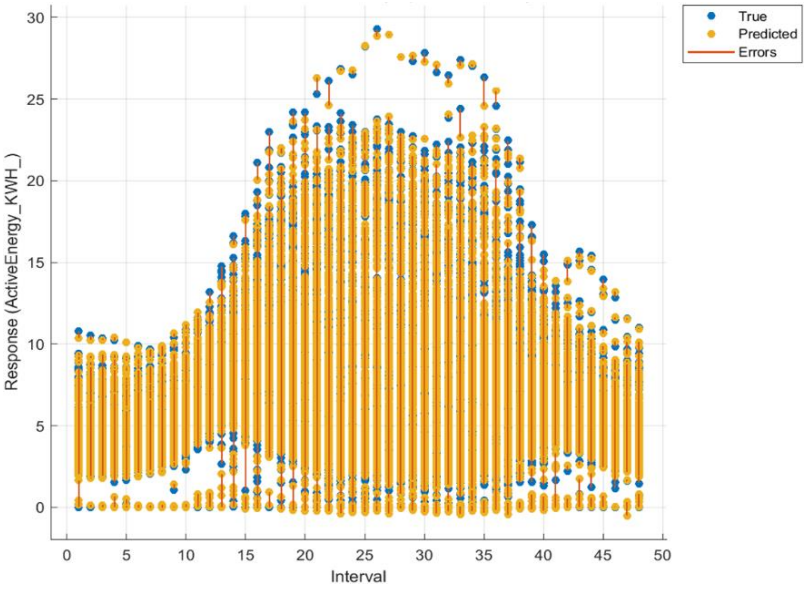

Fig. 7. Response plot of exponential GPR

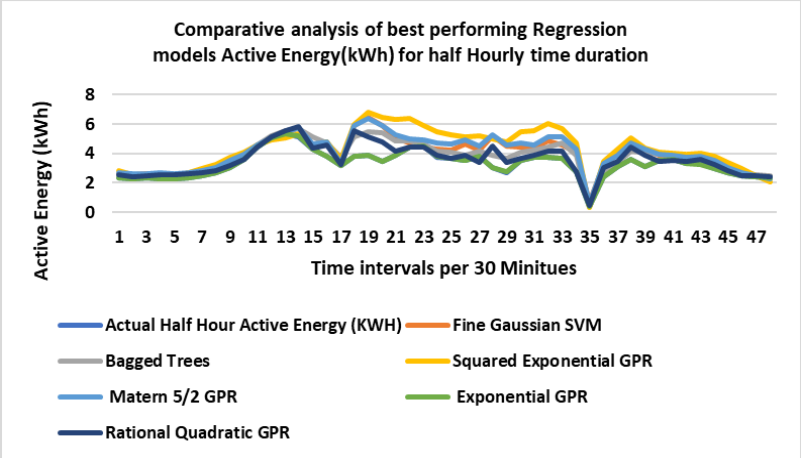

Fig. 8. Comparative analysis Graph of best-performing Regression models Active Energy (kWh) for Hourly time duration

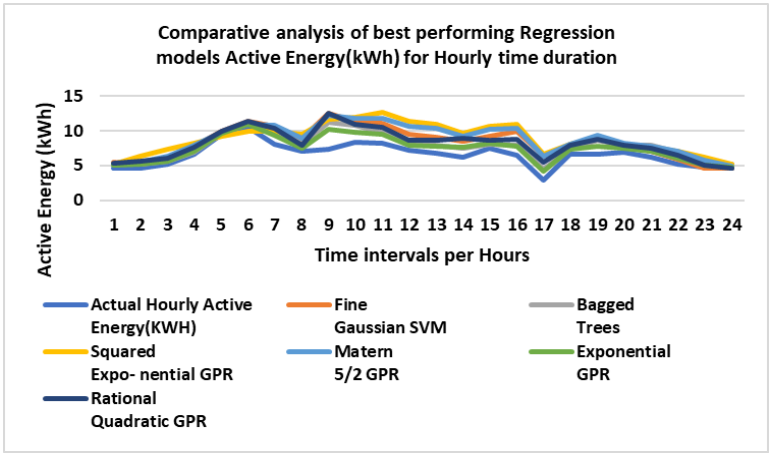

Fig. 9. Comparative analysis Graph of best performing Regression models Active Energy(kWh) for Hourly time duration

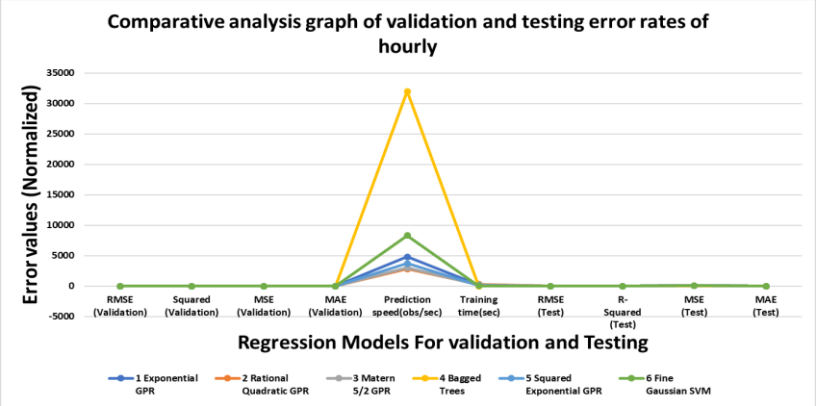

Fig. 10. Comparative analysis graph of validation and testing error rates hourly time interval

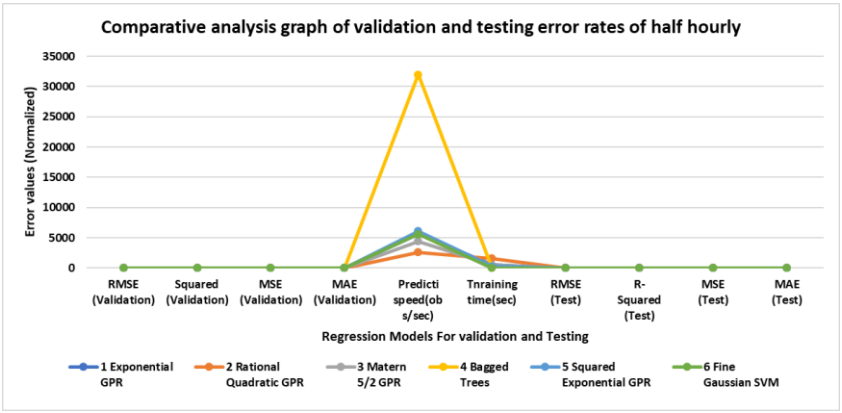

Fig. 11. Comparative analysis graph of validation and testing error rates Half hourly time interval

Simulation Results

The simulations are performed using MATLAB R2021b Regression Toolbox with the Hold out-validation. The outcome of the regression analysis is tabulated in Table 5.

Benchmark Model

Support Vector Machine (SVM) or Support Vector Regressor (SVR) is a widely adopted regressor for developing short-term load forecasting models. Therefore, while proposing improved regressors for short-term load forecasting, SVR is mostly chosen
as the benchmark model (Prakash et al., 2018). Similarly, SVM is chosen as the benchmark model in this paper (Prakash et al., 2018).

Top Six Performance Models

The comparison of the 24 regression models-based forecasted loads with the actual loads for the selected days. Among the regression models, exponential GPR, Rational Quadratic GPR, Maternal 5/2 GPR, Fine Gaussian SVM, Bagged Trees, and Squared Exponential GPR were able to better reproduce the load.
Based on the simulation results and performance evaluation indices among the 24 regression models, the six top-performing models are determined to be:
Exponential GPR

Rational Quadratic GPR

Matern 5/2 GPR

Fine Gaussian SVM

Bagged Trees

Squared Exponential GPR

We evaluated the best models for a particular dataset consisting of 2000 thousand rows and 10 columns (1st July 2020 to 22nd August 2021) data for training 80% (16000 rows from 1st July to 30 May 2121) and validation 20% rows 4000 and 10 columns (30 May to 22nd Aug) and then 1124 rows and 10 columns (23 Aug 2021 to 15 Sep 2021) for testing.

The figures illustrate the performance of the six best-performing models using R-squared plots, residual plots of the predicted model, and response plots of the trained model.

Recommended Final Mode

To be a good model, true and predicted responses should be identical. Comparing the six top-performing models, the analysis confirms that Rational Quadratic GPR and Exponential GPR algorithms are the two recommended final models. They are more accurate and reliable for predicting the load demand throughout every season than other models. The Rational Quadratic model showed excellent results in RMSE, R-Squared, MSE, and MAE values when it came to the validation dataset.

The errors of the predicted model were analyzed for four months out of all the six top-performing models, Exponential GPR can produce more accurate results with less error percentage. Compared to other models, Rational Quadratic GPR and Exponential GPR were able to mimic the actual load pattern more effectively. We have chosen the SVM model as a benchmark in this paper, the two GPR models, Rational Quadratic and Exponential GPR are nonparametric kernel-based probabilistic models and they have outperformed the SVM models.

Hyperparameter setting of the proposed models

Table 8 - Hyperparameter parameters of the proposed models.

SN	Preset	Optimizable GPR
1	Sigma	1.5612
2	Basis functions	Linear
3	Kernel function	Isotropic Matern 3/2
4	Kernel scales	66.3197
5	Optimizer	Bayesian optimization
6	Acquisition function	Expected improvement per second plus
7	Iterations	Iterations 30
8	Signal deviation	2.9663

Adapting a machine learning model to different problems requires tuning its hyperparameters. In machine learning models, selecting the best hyper-parameter configuration has a direct impact on the model's performance. The process often requires a deep understanding of machine learning algorithms and appropriate techniques for optimizing hyperparameters (Yang & Shami, 2020). There are several parameters or hyperparameters that may have a significant impact on the performance of the model. Using an optimization scheme that attempts to minimize the mean squared error (MSE) of a given model type, the approach tries different combinations of hyperparameter values and returns a model with the optimized parameters. For Optimized GPR Hyperparameter is Shown in Table 8.

Using a dataset of 2000 thousand rows and 10 columns (1st July 2020 to 22nd August 2021), we evaluated the best models based on training 80% (16000 rows from 1st July to 30 May 2021) and validation 20% rows 4000 and 10 columns (30 May to 22nd Aug) and testing 1124 rows and 10 columns (23 August 2021 to 15 Sep 2021). Error Parameters are shown in Tables 6 and 7 for validation and testing. There are several parameters or hyperparameters that may have a significant impact on the performance of the model. Using an optimization scheme that attempts to minimize the mean squared error (MSE) of a given model type, the approach tries different combinations of hyperparameter values and returns a model with the optimized parameters. Using a dataset of 2000 thousand rows and 10 columns (1st July 2020 to 22nd August 2021), we evaluated the best models based on training 80% (16000 rows from 1st July to 30 May 2021) and validation.

When load forecasting is considered, there are several components of error, including modeling errors (errors introduced through regression), errors caused by system disturbances like load shedding and irregular events, and errors caused by temperature forecasting. As a result of this model's sensitivity to temperature fluctuations, it requires a highly accurate forecast of temperature. This model needs a very accurate forecast of temperature because even a minor change in temperature will result in a significant change in the prediction of load. This forecasting model uses the next-day temperature forecast as an input which will introduce further errors, as there was not enough data available for temperature forecasting to be included in this regression analysis. Considering the fact that there is no data on temperature forecasting, this was not included in this regression analysis. In addition to the temperature, other weather factors such as humidity, cloud cover, and brightness of the day also affect the load characteristic, so it is very important to include this in future studies with one- dimensional convolutional neural network-long short-term memory (1D CNN-LSTM) model.

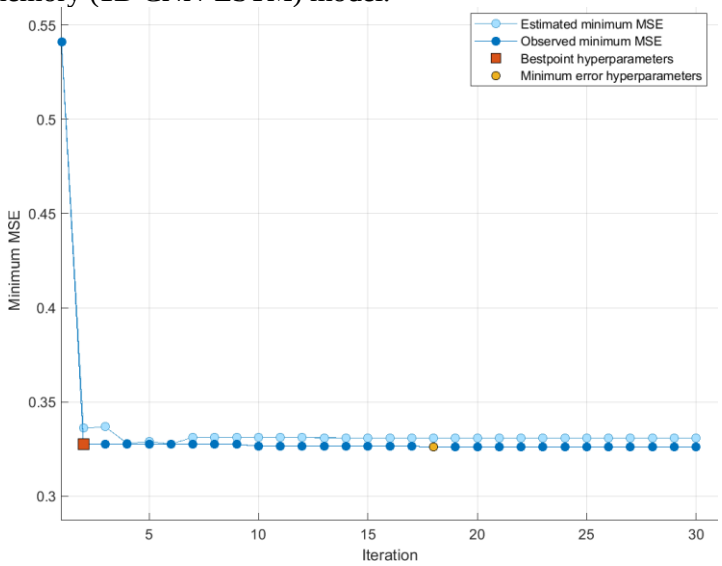

Fig. 12. Optimizable GPR

5. Conclusion

The paper proposes a load forecasting model by combining 24 regression models, with six of the highest-performing models to be evaluated further. According to the study, nonparametric kernel-based probabilistic models such as Gaussian Process Regression (GPR) are viable methods for forecasting load demand. By combining the parameters of all admissible functions, GPR can provide information about consumption trends and do statistical interpolation, which is unlike other models with functional form constraints. The study recommends using exponential GPR algorithms for optimal load forecasting efficiency because GPR is computationally inexpensive, generates a pattern based on the average and standard deviation of the value, and is computationally inexpensive. As part of its evaluation, the paper also uses mean absolute percentage errors (MAPE) and R-squared validation techniques to determine the accuracy of the model.

References

Badran, S., & Abouelatta, O. (2012). Neural network integrated with regression methods to forecast electrical load. *International conference on electrical, electronics and biomedical engineering (ICEEBE2012) Penang (Malaysia)*.

Cao, Z., Wan, C., Zhang, Z., Li, F., & Song, Y. (2019). Hybrid ensemble deep learning for deterministic and probabilistic low-voltage load forecasting. IEEE Transactions on Power Systems, 35(3), 1881-1897. doi:10.1109/TPWRS.2019.2946701

Caro, E., Juan, J., & Cara, J. (2020). Periodically correlated models for short-term electricity load forecasting. *Applied Mathematics and Computation, 364*, 124642. doi:10.1016/j.amc.2019.124642

Ceperic, E., Ceperic, V., & Baric, A. (2013). A strategy for short-term load forecasting by support vector regression machines. *Power Yet IEEE (Trans.), 28*(4), 4356–4364.

Chane, K., Gebru, F. M., & Khan, B. (2021). Short Term Load Forecasting of Distribution Feeder Using Artificial Neural Network Technique. *Journal of Informatics Electrical and Electronics Engineering (JIEEE), 2*(1), 1-22.

Chang, Q., Wang, Y., Lu, X., Shi, D., Li, H., Duan, J., & Wang, Z. (2019, May). Probabilistic load forecasting via point forecast feature integration. In *2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)* (pp. 99-104). IEEE.

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

Chen, B.J., Chang, M.W., & Lin, C. J. (2004). Load forecasting using support Ector machines: A study on EUNITE competition 2001 IEEE (Trans.) *ower Syst, 19*(4), 1821–1830.

Datar, N., Bhoyar, S., Khan, A., Dekapurwar, S., Wankhede, H., Sonone, S.(2021). Solar Power Monitoring system using IoT. *Journal of Emerging Trends in Electrical Engineering, 3*(1).

Deng, Z., Wang, B., Xu, Y., Xu, T., Liu, C., & Zhu, Z. (2019). Multi-scale convolutional neural network with time-cognition for multi-step short-term load forecasting. *IEEE Access, 7*, 88058–88071. doi:10.1109/ACCESS.2019.2926137

Feng, C., Sun, M., & Zhang, J. (2020). Reinforced deterministic and probabilistic load forecasting via Q-learning dynamic model selection IEEE (Trans.) *mart Grid, 11*(2), 1377–1386

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Retrieved from http://link.springer.com/book/10.1007/978-3-319-10247-4#about. Cham, Switzerland: Springer.

Gilanifar, M., Wang, H., Sriram, L. M. K., Oz Guven, E. E., & Arghandeh, R. (June 2020). Multitask Bayesian Spatiotemporal Gaussian Processes for Short-Term Load Forecasting. *IEEE Transactions on Industrial Electronics, 67*(6), 5132–5143. doi:10.1109/TIE.2019.2928275

Gochhait, S., Patil, H., Hasarmani, T., Patin, V., & Maslova, O. (2022, November). Automated Solar Plant using IoT Technology. In *2022 4th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)* (pp. 1-6). IEEE.

Gochhait, S., Asodiya, R., Hasarmani, T., Patin, V., & Maslova, O. (2022, November). Application of IoT: A Study on Automated Solar Panel Cleaning System. In *2022 4th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)* (pp. 1-4). IEEE.

Gochhait, S., Leena, H., Sudheesh, V., Kumar, V., Singh, V., Srinivasan, H., & Badam, D. (2020, June). Smart Lights: How it Enhances Connectivity. In *2020 International Conference for Emerging Technology (INCET)* (pp. 1-4). IEEE.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Amsterdam. The Netherlands: Elsevier.

Hammad, M. A., Jereb, B., Rosi, B., & Dragan, D. (2020). Methods and models for electric load forecasting: a comprehensive review. *Logist. Sustain. Transp, 11*(1), 51-76. doi:10.2478/jlst-2020-0004

Hong, W. C., & Fan, G. F. (2019). Hybrid empirical mode decomposition with support vector regression model for short term load forecasting. *Energies, 12*(6), 1–16. doi:10.3390/en12061093

Jawad, M., Nadeem, M. S. A., Shim, S.O., Khan, I. R., Shaheen, A., Habib, N., Aziz, W. (2020). Machine learning based cost-effective electricity load forecasting model using correlated meteorological parameters. *IEEE Access, 8*, 146847–146864. doi:10.1109/ACCESS.2020.3014086

Jiang, H., Zhang, Y., Muljadi, E., Zhang, J. J., & Gao, D. W. (2016). A short-term and high-resolution distribution system load forecasting approach using support vector regression with hybrid parameters optimization. *IEEE Transactions on Smart Grid, 9*(4), 3341-3350.

Liang, Y., Niu, D., Cao, Y., & Hong, W. C. (2016). Analysis and modeling for China's electricity demand forecasting using a hybrid method based on multiple regression and extreme learning machine: A view from carbon emission. *Energies, 9*(11), 941. doi:10.3390/en9110941

Madhukumar, M., Sebastian, A., Liang, X., Jamil, M., & Shabbir, M. N. S. K. (2022). Regression model-based short-term load forecasting for university campus load. *IEEE Access, 10*, 8891-8905.

Massana, J., Pous, C., Burgas, L., Melendez, J., & Colomer, J. (2015). Short-term load forecasting in a non-residential building contrasting models and attributes. *Energy and Buildings, 92*, 322-330.

Massana, J., Pous, C., Burgas, L., Melendez, J., & Colomer, J. (2016). Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes. *Energy and Buildings, 130*, 519-531.

Mele, E. (2019). A review of machine learning algorithms used for load forecasting at microgrid level. In *Sinteza 2019-International Scientific Conference on Information Technology and Data Related Research* (pp. 452-458). Singidunum University.

Okoye, A. E., & Madueme, T. C. (2016). A theoretical framework for enhanced forecasting of electrical loads, 6(6, June) p. 554, ISSN 2250- 3153.

Olagoke, M. D., Ayeni, A. A., & Hambali, M. A. (2016). Short term electric load forecasting using neural network and genetic algorithm. *Int. J. Appl. Inf. Yst, 10*(4), 22–28.

Patil, V. S., Morey, A. P., Chauhan, G. J., Bhute, S. S., & Borkar, T. S. (2019). A Review Paper on Solar Power Monitoring System using an IoT. *International Journal of Computer Sciences and Engineering, 7*(8). doi:10.26438/ijcse/v7i8.212215

Pirbazari, A. M., Sharma, E., Chakravorty, A., Elmenreich, W., & Rong, C. (2021). An ensemble approach for multi-step ahead energy forecasting of household communities. *IEEE Access, 9*, 36218-36240.

Prakash, A., Xu, S., Rajagopal, R., & Noh, H. (2018). Robust building energy load forecasting using physically based kernel models. *Energies, 11*(4), 862. doi:10.3390/EN11040862

Qiuyu, L., Qiuna, C. A. I., Sijie, L. I. U., Yun, Y. A. N. G., Binjie, Y., Yang, W., & Xinsheng, Z. (2017, November). Short-term load forecasting based on load decomposition and numerical weather forecast. In *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)* (pp. 1-5). IEEE.

Semero, Y. K., Zhang, J., & Zheng, D. (2018). PV power forecasting using n integrated GA-PSO-ANFIS approach and Gaussian process regression used feature selection strategy. *CSEE Journal of Power and Energy Systems, 4*(2), 210–218. doi:10.17775/CSEEJPES.2016.01920

Su, H., & Jung, C. (2018). Perceptual enhancement of low light images based on two-step noise suppression. *IEEE Access, 6*, 7005-7018.

Subhasri, G., & Jeyalakshmi, C. (2018). A study of IoT based solar panel tracking system. *Advances in Computational Sciences and Technology, 11*(7), 537-545.

Wahyudi, T., & Arroufu, D. S. (2022). Implementation of Data Mining Prediction Delivery Time Using Linear Regression Algorithm. *Journal of Applied Engineering and Technological Science (JAETS), 4*(1), 84–92. https://doi.org/10.37385/jaets.v4i1.918

Wahyudi, T., & Silfia, T. (2022). Implementation of Data Mining Using K-Means Clustering Method to Determine Sales Strategy In S&R Baby Store. *Journal of Applied Engineering and Technological Science (JAETS), 4*(1), 93–103. https://doi.org/10.37385/jaets.v4i1.913

Yang, L., & Shami, A. On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. arXiv 2020. *arXiv preprint arXiv:2007.15745*.

Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renewable and Sustainable Energy Reviews, 73, 1104–1122, un. doi:10.1016/j.rser.2017.02.023