**ELECTRICITY CONSUMPTION PREDICTION USING MACHINE LEARNING**

A Project report submitted in partial fulfilment of the requirements for the award of the degree of

**BACHELOR OF TECHNOLOGY**

**IN**

**INFORMATION TECHNOLOGY**

**Submitted by**

**BANDLAMUDI PRASANTHI (319126511005)**

**KAKARA PRIYANKA (319126511018)**

**KOLLI KUSHWANTH SAI (319126511027)**

**KONDAPALLI ANIL KUMAR (319126511030)**

**UNDER THE GUIDANCE OF**

**Dr.P. Padmaja**

**Professor**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES**

(**UGC AUTONOMOUS)**

**(Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with an 'A' Grade)**

**Sangivalasa, Bheemili Mandal, Visakhapatnam dist. (A.P) 2022-2023**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES**

**(UGC AUTONOMOUS)**

**(Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with an 'A' Grade)**

**Sangivalasa, Bheemili Mandal, Visakhapatnam dist. (A.P)**



**CERTIFICATE**

This is to certify that the project report entitled “**ELECTRICITY CONSUMPTION PREDICTION USING MACHINE LEARNING**” submitted by **BANDLAMUDI PRASANTHI (319126511005), KAKARA PRIYANKA (319126511018), KOLLI KUSHWANTH SAI (319126511027), KONDAPALLI ANIL KUMAR (319126511030)** in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Information Technology of Anil Neerukonda Institute of technology and sciences, Visakhapatnam is a record of bonafide work carried out under my guidance and supervision.

**Project Guide: Head of the department:**

**Dr.P. Padmaja Dr.M.Rekha Sundari**

Professor Professor

Department of IT Department of IT

ANITS ANITS

**ACKNOWLEDGEMENT**

We want to express our deep gratitude to our project guide **Dr**.**P. Padmaja** Professor, Department of Information Technology, ANITS, for his/her guidance, unsurpassed knowledge, and immense encouragement. We are grateful to **Dr M. REKHA SUNDAR**I, Head of the Department of Information Technology, for providing us with the required facilities to complete the project work.

We are very much thankful to the **Principal and Management, ANITS**, **Sangivalasa,** for their encouragement and cooperation in carrying out this work.

We thank all **teaching faculty** of the Department of IT, whose suggestions during reviews helped us accomplish our project. We want to thank all **non-teaching staff** of the Department of IT, ANITS, for assisting in accomplishing our project.

We want to thank our parents, friends, and classmates for their encouragement throughout our project period. Finally, we thank everyone for supporting us directly or indirectly in completing this project successfully.

**PROJECT STUDENTS**

**BANDLAMUDI PRASANTHI (319126511005)**

**KAKARA PRIYANKA (319126511018)**

**KOLLI KUSHWANTH SAI (319126511027)**

**KONDAPALLI ANIL KUMAR (319126511030)**

**DECLARATION**

We declare that the project work entitled "**ELECTRICITY CONSUMPTION PREDICTION USING MACHINE LEARNING**” submitted to the Anil Neerukonda Institute of Technology and Sciences is a record of an original work done by **BANDLAMUDI PRASANTHI (319126511005),** **KAKARA PRIYANKA (319126511018), KOLLI KUSHWANTH SAI (319126511027), KONDAPALLI ANIL KUMAR (319126511030)**, under the esteemed guidance of Dr P.Padmaja, Professor of Information Technology, Anil Neerukonda Institute of Technology and Sciences, and this project work is submitted in the partial fulfilment of the requirements for the award of degree Bachelor of Technology in Information Technology. This entire project is done to the best of our knowledge and has not been submitted for the award of any other degree in other universities.

**PROJECT STUDENTS**

**BANDLAMUDI PRASANTHI (319126511005)**

**KAKARA PRIYANKA (319126511018)**

**KOLLI KUSHWANTH SAI (319126511027)**

**KONDAPALLI ANIL KUMAR (319126511030)**

**ABSTRACT**

Accurate electricity demand forecasts that account for the impacts of extreme weather events is needed to inform electric grid operation and utility resource planning and enhance energy security and grid resilience. Three standard data-driven models are used to predict city-scale daily electricity usage: linear regression models, machine learning models for time series data, and machine learning models for tabular data. In this study, we developed and compared seven data-driven models:

1. The five-parameter change-point model
2. The Heating/Cooling Degree Hour model
3. Time series decomposed model implemented by Facebook Prophet
4. The Gradient Boosting Machine implemented by Microsoft lightGBM
5. Three widely used machine learning models (Random Forest, Support Vector Machine, Neural Network

Seven models are applied to the city-scale electricity usage data for three metropolitan areas in India: Hyderabad, Mumbai, and Delhi. Results show that seven models can predict the metropolitan area’s daily electricity use, with a coefficient of variation of the root mean square error (CVRMSE) less than 10%. The lightGBM provides the most accurate results, with CVRMSE on the test dataset of 6.5% for Mumbai, 4.6% for Hyderabad, and 4.1% for the Delhi metropolitan area. Results show that weather-sensitive component accounts for 30%–50% of daily electricity usage. Every degree Celsius ambient temperature increases in summer leads to about 5% (4.7% in Mumbai, 6.2% in Hyderabad, and 5.1% in Delhi) more daily electricity usage compared with the base load in the three metropolitan areas.

**CONTENTS**

**CHAPTER 1 INTRODUCTION page no.**

* 1. Introduction 7
  2. Problem Statement and Motivation 7
  3. Project Objectives 7
  4. Project Scope and Direction 8
  5. Impact, Significance and Contribution 8

**CHAPTER 2 LITERATURE SURVEY 9**

**CHAPTER 3 SYSTEM ANALYSIS**

3.1 Hardware Requirements 11

3.2 Software Requirements 11

**CHAPTER 4 SYSTEM STUDY**

4.1 Feasibility Study 12

4.1.1 Economic Feasibility 12

4.1.2 Technical Feasibility 12

4.1.3 Social Feasibility 12

**CHAPTER 5 METHODOLOGY**

5.1 Linear Model 14

5.2 Machine learning model for time-series data 16

5.3 Machine learning model for tabular data 18

**CHAPTER 6 RESULTS 20**

**CHAPTER 7 CONCLUSION 23**

**CHAPTER 8 REFERENCE 24**

**1. INTRODUCTION**

**1.1 INTRODUCTION:**

The urgency in the demand for electric energy has been accelerated by the recent strong economic development and fast urbanisation. As the cornerstone for making crucial choices in the realm of power system operation and control, power demand forecasting is essential in the electric industry. As comparison to recent years, the global energy consumption increased 0.7% less in 2021 (+0.7% vs. an average of 3% /year from 2000 to 2020).

According to the most recent forecast, the global demand for electricity would rise by 2.4% in 2022 after rising by 6% last year, maintaining up with the average growth rate experienced in the five years before to the Covid-19 outbreak. The prognosis is complicated by the uncertain status of the economy and the effect that fuel prices will have on the mix of energy sources, even if it is projected that the demand for electricity would rise at a constant rate until 2023.

**1.2 PROBLEM STATEMENT AND MOTIVATION:**

Existing machine learning techniques have low prediction accuracy. Three standard data-driven models are used to predict city-scale daily electricity usage: linear regression models, machine learning models for time series data, and machine learning models for tabular data. This project aims to predict ideal electricity usage and recommend changes to the user.

**1.3 PROJECT OBJECTIVES:**

In this project, the proposed model has five objectives:

1. To conduct Exploratory Data Analysis.
2. To pre-process the data.
3. To develop a highly accurate ML model using RF, SVM, NN, GAM, GBM, HCDH.
4. To evaluate the model for improvement.
5. To test the model with new data to check the accuracy.

**1.4 PROJECT SCOPE AND DIRECTION:**

We found many publications on the subject, which has many uses for forecasting electricity at the city scale. Yet after reviewing previous studies, we found a significant study gap: diverse approaches have been provided. Nevertheless, none of the research include publicly accessible or reusable open-source data, models, or programming. As a result, those methodologies are lacking in summary and comparison, especially when it comes to cutting-edge machine learning algorithms and tried-and-true linear regression strategies.

We created a top-down, data-driven system to predict daily electricity use on a city-scale. We investigated three methods to anticipate daily energy consumption at the city-level1: linear regression models, machine learning models for time-series data, and machine learning models for non-time-series data. With the use of power statistics from the Indian metropolises of Mumbai, Hyderabad, and Delhi, we were able to compare and evaluate our methods.

* 1. **IMPACT, SIGNIFICANCE, AND CONTRIBUTION:**

The Electricity consumption prediction have positive impacts and benefits in the power sector:

* + 1. Increasing productivity
    2. Protection against electricity losses and decreased production
    3. Control of electricity usage
    4. Prediction of usage in future

## 2. Literature Survey

## A data-driven technique with known input and output variables was proposed by Corgnati et al. (2013). This information would be used to analyse the system's parameters and create a mathematical model. This data-driven machine-learning technique has been the subject of past studies.

## Fu et al[2015] .'s recommendation to use a Support Vector Machine, ML technique to anticipate load at a building's air conditioning, lighting, electricity, and other systems was based on hourly electrical demand input and weather forecasts. With a mean bias error (MBE) of 7.7% and a root mean square error (RMSE) of 15.2%, the SVM approach accurately forecasted the total electrical load.

## The k-Nearest Neighbor model was used by Valgaev et al. [2016] to estimate the power needed for a smart building as part of the Smart City Demo Aspern project. In order to apply the k-NN forecasting technique, historical data and their successors were obtained. The k-NN algorithm can only locate similar occurrences in a wide feature space, which limits its ability to predict future value. It must therefore be enhanced by temporal information identification across workdays and projections generated for the next 24 hours.

## In order to forecast short-term load, El Khantach et al. [2019] used five machine-learning algorithms. He divided the historical data into time series for each hour of the day, and for each time series, he constructed a 24-time series. The five machine learning techniques are multi-layer perceptrons, SVMs, radial basis function regressors, REPTrees, and the Gaussian process. The experiment was conducted with the aid of data from Moroccan electrical load statistics. The MLP method was discovered to be the most accurate, with a MAPE percentage of 0.96, followed by SVM, which, while less accurate than MLP, was still superior to the others.

## Gonzalez-Brione et al. [2019] noted that although a classification-based machine learning methodology is commonly used to forecast energy usage, the regression method may also be employed. One day earlier, scientists used the use of electricity as a different characteristic in their research criteria. After analysing the data using Linear Regression, Support Vector Regression, Random Forest, Decision Tree, and k-Nearest Neighbor, they created a prediction model. The LR and SVR models outperformed the others, achieving 85.7% accuracy.

## Mean Value Imputation, Last Observation Carried Forward, Maximum Likelihood Estimate, and Multiple Imputation are a few examples of imputation procedures, according to Newgard and Lewis [2015]. The mean value imputation fills in the gaps in data with the aid of the dataset's mean value. For non-strictly random data, this strategy introduces inequality into the data, making it undesirable.The methodologies presented that required the most complexity were Multiple Imputation and Maximum Likelihood Estimation. With each repetition, the Multiple Imputation approach gradually replaces missing data. With the help of statistical analysis based on observable data, this strategy deals with the uncertainty that the missing piece introduces. Multiple Imputation Using Chained Equations is a well-known MI method. By first constructing the parameters and bounds based on the data distribution, the Maximum Likelihood Estimate achieves replacement through assumption. The putative parameters would then be used to perform the imputation. This imputation technique was applied in probabilistic principal component analysis.

**3. SYSTEM ANALYSIS**

The Prediction Model using MLP design is divided into two sections:

1. Hardware

2. Software

**3.1 HARDWARE REQUIREMENTS:**

• Processor: Intel core i5 or above.

• 64-bit, quad-core, 2.5 GHz minimum per core

• Ram: 4 GB or more

• Hard disk: 10 GB of available space or more.

• Display: Dual XGA (1024 x 768) or higher resolution monitors

**3.2 SOFTWARE REQUIREMENTS:**

• Operating system: Windows

* Programming Platform: Microsoft Azure Machine Learning Studio

**4. SYSTEM STUDY**

**4.1 FEASIBILITY STUDY**

This stage involves evaluating the idea's potential and developing a general proposal strategy. System study will include a feasibility assessment of the suggested system. By doing this, the recommended solution is guaranteed to be workable. During a feasibility study, it is essential to comprehend the system's requirements.

Three key considerations involved in the feasibility analysis are

• ECONOMICAL FEASIBILITY

• TECHNICAL FEASIBILITY

• SOCIAL FEASIBILITY

**4.1.1 ECONOMICAL FEASIBILITY:**

To ascertain the project's economic impact, this research is being conducted. The amount of funding for project and system development is constrained. Justification of expenses is crucial. Due to the availability of many of the technologies, the developed system is significantly under budget.

**4.1.2 TECHNICAL FEASIBILITY:**

The purpose of this evaluation is to determine the project's technical viability. The number of technological resources required by the proposed system must be moderate. The designed system must have a low demand because it only needs minor or no alterations to be made.

**4.1.3 SOCIAL FEASIBILITY**:

The purpose of the study is to gauge user acceptance of the technology. Training the user on effective technology use is part of this. The user should embrace the system as a necessity rather than be afraid of it. The techniques used to inform and acquaint users with the system will determine their acceptability. He is the final user of the system; thus he must earn your trust before offering helpful criticism.

**5. METHODOLOGY**

**5. METHODOLOGY:**

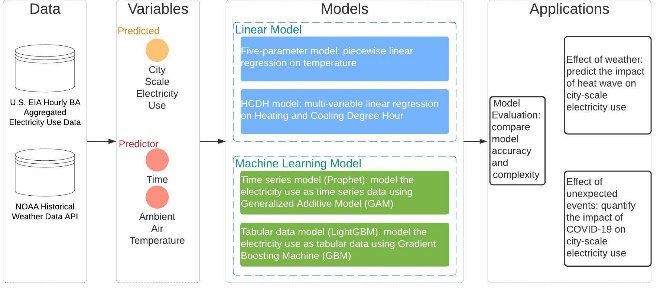


Fig1 Electricity consumption prediction Methodology

Before going into the present city-level electricity usage models, this part introduces the models built for this study. Figure 1 displays the methods of the study in detail. To forecast daily electricity use at the city size, seven data-driven models were created, each of which represented a different modelling approach: linear models, machine learning models for time-series data, and machine learning models for tabular data. Due to its ease of use and compatibility, the ASHRAE Five-Parameter Piecewise Linear Regression Model and the Heating Cooling Degree Hour model were the first two linear models created. Three traditional models were created as the foundation for the machine learning models: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (NN).

Additionally developed and contrasted were two advanced machine learning models, the Generalized Additive Model (GAM) and Gradient Boosting Machine (GBM). GAM and GBM model the data for electricity use in different ways, with GAM modelling the data as a time series and GBM modelling the data as tabular. The main difference between time series modelling and tabular data modelling is the approach taken to represent temporal information. Because energy use has distinct weekly and yearly cycles, using time series modelling to predict electricity demand makes sense. A developing index is used by the time series model to describe temporal information and treats electricity use as time series data.

Data on electricity use can also be displayed in a tabular manner, which necessitates the addition of new elements to capture weekly and yearly cycles, such as the day of the week and the month. To represent temporal information, such as the hour and day of the year, the tabular model also includes other properties. The data sequence contains temporal information in time series models and cannot be changed. This makes it possible to shuffle the order of the data in tabular models. The tabular data format is used to model power usage by the traditional machine learning models (RF, SVM, NN).

The goal of this study is to forecast daily electricity consumption at a city size, including that of buildings, transportation, industry, and public services in the city and neighboring rural areas. The amount of electricity used is significantly influenced by the outside temperature. We contrast the baseline machine learning models, such as linear, with the time series and tabular data models.

**Linear models:**

In city-level energy modelling, linear models are commonly employed because of their ease of use and interoperability. In these models, a linear relationship is used to regress the observed and independent variables. Several studies have demonstrated its effectiveness, including Lindsey et al. (2011)'s development of a linear model to predict the city-level transport energy consumption and greenhouse gas emissions of Chicago [19]. To predict the energy consumption at the neighborhood level, Kuusela et al. (2015) developed a multi-variable linear regression model [20]. Even massive ground source heat pumps, which are intricate systems, have shown to work well using linear models [21].

Because of their interoperability and flexibility to use the coefficients to validate and explain the models, linear models are frequently used. Because of their clarity and simplicity, linear models are a great option as a baseline model for comparison to more complex non-linear models, even if they might not correctly represent the non-linear interactions that are typical in the actual world. Two linear models were created in this work to forecast temperature-sensitive electricity use on a city scale. ASHRAE's change-point model, put forth by ASHRAE in the 1990s, is the first linear model employed in the study. The change point model describes the link between energy consumption and ambient temperature using five factors (base, Th, Tc, h, and c). The base load, which was When the outside temperature is between [Th, Tc], by base, is the lowest energy use. City-level energy consumption rises in reaction to the temperature decrease brought on by the rise in heating demand when the outside temperature drops below the heating change point Th. The change point model is represented in Figure 2 and Equation 1.

On the other hand, when the outside temperature rises above the cooling threshold Tc, there is an increase in the need for cooling, which raises the amount of energy used in the city. The city-level load sensitivity to temperature change is shown by the slopes on the cooling (c) and heating (h) sides. As the emphasis is solely on electricity use, the value of h would be lower than that of c since most air-conditioned buildings use electricity for cooling while many use natural gas for heating. In this work, we forecast city-level energy consumption using the five-parameter change point (5-p) model. The 5-p model is frequently used to gauge the energy performance of buildings and estimate building-level energy consumption.

𝑙𝑜𝑎𝑑(𝑇) = 𝛽𝑏𝑎𝑠𝑒 + 𝛽ℎ × (𝑇ℎ − 𝑇), 𝑖𝑓 𝑇 < 𝑇ℎ

𝛽𝑏𝑎𝑠𝑒 , 𝑖𝑓 𝑇ℎ < 𝑇 <𝑇𝐶

𝛽𝑏𝑎𝑠𝑒 + 𝛽𝑐 × (𝑇 − 𝑇𝑐) , 𝑖𝑓 𝑇𝐶 < 𝑇 (Equation 1)

The second linear model used in this research is the Heating/Cooling Degree Hour (HCDH) model. This method is often used in the HVAC industry to determine the required energy for heating and cooling [27]. The heating and cooling degree day is a crucial proxy variable used to estimate the impact of climate change on power demand [28]. The difference between the ambient temperature and the temperatures at the heating base load (Tbh) and cooling base load (Tbc) bases, which is the cumulative sum, is used to determine the HDH and CDH in Equation 2. The total sum of the difference between the outdoor and base temperatures (Tbh Ti) is an accurate predictor of the required heating. Heating starts when the outside temperature drops below Tbh. HDH and CDH are widely used for calculating building energy demand [29], determining the structure's thermal insulation [30], and other purposes. In this study, the daily city-level energy consumption was regressed as a linear function using the HDH and CDH.

𝐻𝐷𝐻 = ∑max (0, (𝑇𝑏ℎ − 𝑇𝑖)) 24 𝑖=1

𝐶𝐷𝐻 = ∑max (0, (𝑇𝑖 − 𝑇𝑏𝑐)) 24 𝑖=1

𝑙𝑜𝑎𝑑(𝑇) = 𝛽0 + 𝛽1 × 𝐻𝐷𝐻 + 𝛽2 × 𝐶𝐷𝐻 (Equation 2)

According to [31], a similar challenge in the construction of both the HCDH and five-parameter models is the accurate selection of the base temperature (Tbh, Tbc in the HCDH model) and the change temperature (Th, Tc in the five-parameter model). Finding the combinations of change or base temperatures that would produce the most precise linear model was the strategy employed in this study to choose these temperatures. To do this, the best base temperatures for the HCDH model were chosen using the scipy.optimize.curve fit function [32], whereas the best change temperatures for the five-parameter model were chosen using a brute force search.

**Machine learning model for time-series data:**

The first form of time-series modelling method is the Autoregressive Integrated Moving Average (ARIMA). Using its own lagged values (yt1, yt2,...) and prior prediction error, ARIMA predicts a time-series variable, yt (t1, t2,...). The difference between the anticipated value and the actual value, yi, is the prediction error. ARIMA has been used to forecast the demand for electricity and natural gas in Lebanon and Turkey, respectively. The ARIMA model's execution order and the accuracy of the power demand load forecasting models have been determined using Akaike's Information Criterion (AIC) and the Bayesian Information Criterion, respectively (BIC). The wavelet transform has been linked with ARIMA in order to improve the accuracy of its forecasts. Recurrent neural networks (RNN) and its variant short-term long memories are two common machine learning models for time-series data (LSTM). These models use a new approach to define the time dependency and forecast yt using a state from the previous time step. Due to recent advancements in deep learning algorithms and processing power, neural network-based methods for estimating energy consumption are now frequently used. More than 40 research have employed a neural network-based technique for forecasting load over the short, medium, and long durations. In recent publications, RNN and LSTM have been used to predict energy demand at the district level as well as electricity usage for commercial and residential buildings. In research comparing RNN/LSTM and LSTM to ARIMA in the prediction of building load, LSTM outperforms ARIMA because it can capture non-linear correlations between time-series data and exogenous variables.

In this paper, a novel approach to forecasting municipal electricity consumption using time-series modelling is explored. The study included two methodologies: ARIMA and RNN. The three main elements of a deconstructed time-series model used for modelling are exogenous variables, trend, and seasonality. The linear nature of heating and cooling is demonstrated by Equation 3. Degree hours (f(Tempt)) is one of the components. The trend function (g(t)) models non-periodic changes in the time series, while the s(t) function models weekly and yearly seasonality changes. Prophet, an open-source tool from Facebook, was used to put this idea into practice. For further details on the implementation, see the publication [43].

𝑙𝑜𝑎𝑑𝑡 = 𝑓 (𝑇𝑒𝑚𝑝𝑡) + (𝑡) + (𝑡) + 𝜀𝑡 (Equation 3)

The choice of the decomposed time-series model was influenced by two variables. The first is that it consistently projects temperature, financial markets, and daily COVID-19 cases in Bangladesh with high levels of accuracy. Additionally, the model enables the observation of the distinct effects of unexpected public health occurrences on city-level demand as well as the separation of the effects of other factors (such as temperature-dependent load and seasonal time-dependent periodic load). The approach has never been applied to forecast energy consumption at the municipal level, according to the authors. The advantage of the decomposed time series model is that it may separate time series data into discrete components, each with its own implications. Typically, the temperature-dependent load is combined with the temperature-dependent HVAC use (Tempt). The periodic load s(t) model captures variations in load across time, such as seasonal variations in city-level electricity usage. The remaining load variance is represented by the non-periodic load g(t), which may result from recent events like the COVID-19 epidemic or long-term patterns (such as improving building thermal properties and equipment energy efficiency). Results of decomposition can identify the main movers and disclose the size of each component.

It is important to note that the daily heating and cooling degree hour was used while building the time series decomposition model rather than the daily mean temperature. This is so because the f (Tempt) term of the decomposed model is a monotonous linear function. There is a U-shaped link between daily power use in the city and ambient mean temperature, meaning that high electricity consumption occurs when the temperature is either extremely low or extremely high. Since the link between electricity use and degree per hour of heating and cooling is monotonous and a monotonous function cannot capture this U-shape relationship, the daily degree per hour of heating and cooling is used as the regressor in the model.

**Machine learning model for tabular data (non-time-series data):**

It is also possible to model city-level electricity use using tabular data. To record the timing and periodic behaviour of energy usage, new functionalities must be implemented. For example, two new features—the day of the week and the month of the year—must be introduced as input variables to encode the weekly and yearly cycles.

Several research simulate time series energy usage using tabular data. The three main categories of machine learning algorithms for this kind of data are neural network-based algorithms, decision tree-based algorithms, and other approaches. Regression or classification issues are resolved using neural network-based techniques, commonly referred to as artificial neural networks, feedforward neural networks, or multi-layer perception. These methods benefit from being easily parallelizable and versatile. As an illustration, Fernández et al. (2011) forecast building load using a NN-based method. To increase model accuracy and robustness, systems based on decision trees like Classification and Regression Tree, Random Forest, and Gradient Boosting Machine combine numerous decision trees utilising ensemble learning techniques. Yau and Tso (2007) used CART to predict the energy use of buildings in Hong Kong, and Roth et al. (2019) built RF and GBM models to predict the energy use of buildings in New York City. Energy consumption has also been modelled using other tabular data algorithms, including Support Vector Machine, k-Nearest Neighbors, and k-means. For example, Li et al. (2017) forecasted community-level renewable generation using the Support Vector Machine technique, while Al-Qahtani and Crone (2013) predicted UK electricity consumption using k-Nearest Neighbors. Moreover, Fonseca and Schlueter (2015) used k-means to forecast Zurich building energy use at the district level. Not to mention, Kontokosta and Tull (2017) discovered that the geographic resolution may affect whether machine learning technique is better for estimating energy use, with Linear Regression performing best at the building level and SVM working best at the zip code level.

Four tabular data modelling algorithms—RF, SVM, NN, and GBM—are taken into consideration in this work. While the first three algorithms—RF, SVM, and NN—are thoroughly discussed, they are used as the foundational techniques for GBM. GBM has recently gained a lot of interest due to its excellent results in several machine learning projects and competitions. For instance, GBM was a part of the final predictors for all six of the top teams in the ASHRAE Great Energy Predictor III competition. Additionally, earlier studies have demonstrated that, when it comes to predicting building loads, GBM outperforms several other well-known machine learning algorithms, including Ridge regression, Lasso regression, Elastic Net, Support Vector Machine, Random Forest, vanilla Deep Neural Network, and Long Short-Term Memory. As a result, GBM, the most modern technique, is picked and employed in this study as an example of the tabular data modelling strategy.

The initial stage involves utilizing the time index for producing intermediary variables that encode temporal data. Numerous GBM implementation packages are available, and Microsoft's lightGBM [55] was employed in this case due to its user-friendly interface and comprehensive documentation. Proper hyper-parameter tuning is essential to prevent over-fitting or under-fitting while training a GBM.

**6.RESULTS**

Model correctness is one of the most important aspects to take into account when comparing different data-driven models. Data from July 2015 to June 2018 were used to train the model, while data from July 2018 to June 2019 were used to evaluate its performance. This made it possible to compare the models in an accurate manner. The data for 2020 was omitted to avoid any unplanned occurrences affecting the comparison. Three traditional machine learning models (Random Forest, Support Vector Machine, Artificial Neural Network), a time-series decomposed model, a gradient boosting machine model, two linear models (the five-parameter change point model and the Heating and Cooling Degree Hour model), and a time-series decomposed model are among the seven data-driven models that were created. Three metrics were used to assess the models after they had been trained on the training data: mean absolute error (MAE), root mean square error (RMSE), and cross-validation root mean square error (CVRMSE). The CVRMSE for the seven approaches across three metropolitan areas is also shown in Figure 1.

Top-down data-driven algorithms can make precise predictions about the city's electricity needs. All seven of these models can predict the daily electricity consumption of a city with greater than 90% accuracy. For Hyderabad or Delhi, overfitting is acceptable, but in Mumbai, compared to the other two regions on the test dataset, it significantly decreased the model's performance. For the train and test datasets, the models also did well in the two remaining locations. There are two possible explanations for this: either the test dataset's patterns of Mumbai's electricity consumption were adjusted, or the data-driven model discovered and added additional hidden components that significantly affect Mumbai's electricity demand.

Simple linear regression models using piecewise and multivariate techniques can produce precise predictions. The Heating and Cooling Degree Hour model performed worse than the five-parameter change point model in each of the three metropolitan areas. The CVRMSE of the five-parameter model varied from 5.2% to 8.2%, whereas that of the HCDH model ranged from 5.4% to 8.8%. The five-parameter change point approach is also easier to apply since it does not need determining the proper base temperatures for heating and cooling, which might vary between cities with different climates and energy use patterns.

Machine learning models with similar degrees of accuracy include RF, SVM, and NN. Even when using the same hyper-parameters and model architecture, the dataset on which these models are applied affects their efficacy. For instance, SVM can perform poorly in Delhi while producing better results in Mumbai and Hyderabad. As a result, the model's capacity for generalisation is called into question.

The GBM model beats the deconstructed model and other basic machine learning models in all three metropolitan areas (RF, SVM, and NN). In Mumbai, the CVRMSE has decreased from 0.4% to 4.3%. (Hyderabad). This study highlights the idea that, because of the precise encoding of temporal information using two variables, energy usage may not necessarily be viewed as time-series data (month of the year and day of the week). Time-series modelling uses the sequential ordering of input data to describe temporal information and is less robust to missing data than tabular data models. When missing data is imputed in time-series modelling, data preparation becomes more challenging. Nevertheless, tabular data models don't use extra features to encrypt temporal information like the day of the

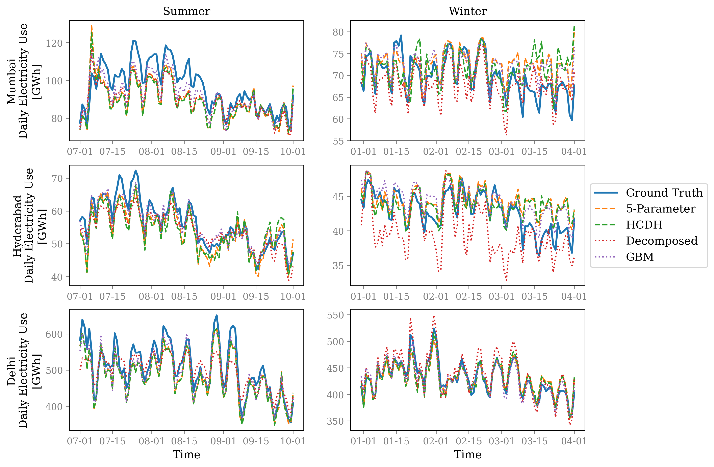
The machine learning model (lightGBM) can increase the model accuracy by a factor of 1.1% to 1.7% when compared to the winner of the linear model. (5-parameter).

Figure 1: The seven algorithms' CVRMSE (Coefficient of Variation of Root Mean Squared Error) values for the three metropolitan areas. Bar chart

Description automatically generated with medium confidence

Figure 2 compares the model prediction to the ground truth of the test dataset. The ground truth was represented by a solid line, whilst the two linear models and the two machine learning models were each represented by a dashed or dotted line. Just the two linear, powerful machine learning models were compared in order to preserve the plot's clarity.

Figure 2: Model prediction of city-scale daily electricity consumption (GWh) on the test dataset



**7.CONCLUSION**

Because heating and cooling take a lot of energy, the quantity of power used locally is dependent on the temperature. As a result of climate change, extreme weather events have become more frequent. To improve the energy security and resilience of the electric system, more precise electricity demand predictions that consider extreme weather events and their effects on load are required (neural network-based, decision tree based, and others). Seven data-driven models were created and compared in this work:

* a five-parameter change-point model
* a Heating Cooling Degree Hour model
* a decomposed time series model implemented by Facebook Prophet
* a Gradient Boosting Trees model implemented by Microsoft lightGBM

Despite the deconstructed model's limited application in this area, LightGBM has been shown to be a top performer in predicting city-scale energy demand. Using information on power use collected at the city level for Hyderabad, Mumbai, and Delhi as well as the surrounding rural areas, we looked at seven models. With a CVRMSE of less than 10%, all models can accurately predict the city-level electricity consumption. In comparison to the HCDH model, the five-parameter model performs better. The Gradient Boosting Machine model is the most precise of the seven. Mumbai's CVRMSE of lightGBM for the test dataset was 6.5%, Hyderabad's was 4.6%, and Delhi's metropolitan areas was only 4.1%. Although less accurate than lightGBM, the decomposed time series model offers us a special chance to separate and

**8.REFERENCES**

[1] C. E. Kontokosta and C. Tull, “A data-driven predictive model of city-scale energy use in buildings,” Appl. Energy, vol. 197, pp. 303–317, Jul. 2017, doi: 10.1016/j.apenergy.2017.04.005.

[2] O. Deschênes and M. Greenstone, “Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US,” Am. Econ. J. Appl. Econ., vol. 3, no. 4, pp. 152–185, Oct. 2011, doi: 10.1257/app.3.4.152.

[3] J. R. Vázquez-Canteli, S. Ulyanin, J. Kämpf, and Z. Nagy, “Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities,” Sustain. Cities Soc., vol. 45, pp. 243–257, Feb. 2019, doi: 10.1016/j.scs.2018.11.021.

[4] S. Davoudi, E. Brooks, and A. Mehmood, “Evolutionary Resilience and Strategies for Climate Adaptation,” Plan. Pract. Res., vol. 28, no. 3, pp. 307–322, Jun. 2013, doi: 10.1080/02697459.2013.787695.

[5] H. Yi-Ling, M. Hai-Zhen, D. Guang-Tao, and S. Jun, “Influences of Urban Temperature on the Electricity Consumption of Shanghai,” Adv. Clim. Change Res., vol. 5, no. 2, pp. 74–80, Jan. 2014, doi: 10.3724/SP.J.1248.2014.074.

[6] J. A. Sathaye, L. L. Dale, P. H. Larsen, and G. A. Fitts, “Estimating impacts of warming temperatures on California’s electricity system,” Glob. Environ. Change, vol. 23, no. 2, pp. 499–511, Apr. 2013, doi: <https://doi.org/10.1016/j.gloenvcha.2012.12.005>.

[7] “Joint Response to Governor Newsom Letter August192020.pdf.” Accessed: Oct. 01, 2020. [Online]. Available: https://www.cpuc.ca.gov/uploadedFiles/CPUCWebsite/Content/News\_Room/NewsUpdates/2020/Join t%20Response%20to%20Governor%20Newsom%20Letter%20August192020.pdf.

[8] M. Carvalho, D. B. de M. Delgado, K. M. de Lima, M. de C. Cancela, C. A. dos Siqueira, and D. L. B. de Souza, “Effects of the COVID-19 pandemic on the Brazilian electricity consumption patterns,” Int. J. Energy Res., vol. 45, no. 2, pp. 3358–3364, 2021, doi: <https://doi.org/10.1002/er.5877>.

[9] A. Bahmanyar, A. Estebsari, and D. Ernst, “The impact of different COVID-19 containment measures on electricity consumption in Europe,” Energy Res. Soc. Sci., vol. 68, p. 101683, Oct. 2020, doi: 10.1016/j.erss.2020.101683.

[10] B. Janzen and D. Radulescu, “Electricity Use as a Real-Time Indicator of the Economic Burden of the COVID-19-Related Lockdown: Evidence from Switzerland,” CESifo Econ. Stud., vol. 66, no. 4, pp. 303–321, Dec. 2020, doi: 10.1093/cesifo/ifaa010.

[11] H. Lim and Z. J. Zhai, “Review on stochastic modeling methods for building stock energy prediction,” Build. Simul., vol. 10, no. 5, pp. 607–624, Oct. 2017, doi: 10.1007/s12273-017-0383-y.

[12] L. G. Swan and V. I. Ugursal, “Modeling of end-use energy consumption in the residential sector: A review of modeling techniques,” Renew. Sustain. Energy Rev., vol. 13, no. 8, pp. 1819–1835, Oct. 2009, doi: 10.1016/j.rser.2008.09.033.

[13] V. Bianco, O. Manca, and S. Nardini, “Electricity consumption forecasting in Italy using linear regression models,” Energy, vol. 34, no. 9, pp. 1413–1421, Sep. 2009, doi: 10.1016/j.energy.2009.06.034.

[14] L. Shorrock and J. Dunster, “The physically-based model BREHOMES and its use in deriving scenarios for the energy use and carbon dioxide emissions of the UK housing stock,” Energy Policy, vol. 25, no. 12, pp. 1027–1037, Oct. 1997, doi: 10.1016/S0301-4215(97)00130-4.

[15] Z. Wang, Z. Zhao, B. Lin, Y. Zhu, and Q. Ouyang, “Residential heating energy consumption modeling through a bottom-up approach for China’s Hot Summer–Cold Winter climatic region,” Energy Build., vol. 109, pp. 65–74, Dec. 2015, doi: 10.1016/j.enbuild.2015.09.057.

[16] C. E. Kontokosta and C. Tull, “A data-driven predictive model of city-scale energy use in buildings,” Appl. Energy, vol. 197, pp. 303–317, Jul. 2017, doi: 10.1016/j.apenergy.2017.04.005.

[17] Y. Chen, T. Hong, and M. A. Piette, “Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis,” Appl. Energy, vol. 205, pp. 323– 335, Nov. 2017, doi: 10.1016/j.apenergy.2017.07.128.

[18] N. Abbasabadi and M. Ashayeri, “Urban energy use modeling methods and tools: A review and an outlook,” Build. Environ., vol. 161, p. 106270, Aug. 2019, doi: 10.1016/j.buildenv.2019.106270.

[19] M. Lindsey, J. L. Schofer, P. Durango-Cohen, and K. A. Gray, “The effect of residential location on vehicle miles of travel, energy consumption and greenhouse gas emissions: Chicago case study,” Transp. Res. Part Transp. Environ., vol. 16, no. 1, pp. 1–9, Jan. 2011, doi: 10.1016/j.trd.2010.08.004.

[20] P. Kuusela, I. Norros, R. Weiss, and T. Sorasalmi, “Practical lognormal framework for household energy consumption modeling,” Energy Build., vol. 108, pp. 223–235, Dec. 2015, doi: 10.1016/j.enbuild.2015.09.008.

[21] S. K. Park, H. J. Moon, K. C. Min, C. Hwang, and S. Kim, “Application of a multiple linear regression and an artificial neural network model for the heating performance analysis and hourly prediction of a large-scale ground source heat pump system,” Energy Build., vol. 165, pp. 206–215, Apr. 2018, doi: 10.1016/j.enbuild.2018.01.029.

[22] J. D. Olden, M. K. Joy, and R. G. Death, “An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data,” Ecol. Model., vol. 178, no. 3, pp. 389– 397, Nov. 2004, doi: 10.1016/j.ecolmodel.2004.03.013.

[23] J. K. Kissock, J. S. Haberl, and D. E. Claridge, “Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models, ASHRAE Research Project 1050-RP, Final Report,” Energy Systems Laboratory, Texas A&M University, Technical Report, Nov. 2002. Accessed: Jan. 19, 2021. [Online]. Available: <https://oaktrust.library.tamu.edu/handle/1969.1/2847>.

[24] Y. Zhang, Z. O’Neill, B. Dong, and G. Augenbroe, “Comparisons of inverse modeling approaches for predicting building energy performance,” Build. Environ., vol. 86, pp. 177–190, Apr. 2015, doi: 10.1016/j.buildenv.2014.12.023.

[25] Y. Geng, W. Ji, B. Lin, J. Hong, and Y. Zhu, “Building energy performance diagnosis using energy bills and weather data,” Energy Build., vol. 172, pp. 181–191, Aug. 2018, doi: 10.1016/j.enbuild.2018.04.047.

[26] H. Li, A. Bekhit, C. Szum, C. Nesler, S. Lisauskas, and S. C. Snyder, “Targeting Building Energy Efficiency Opportunities: An Open-source Analytical & Benchmarking Tool,” ASHRAE Trans., vol. 125, pp. 470–478, 2019.

[27] O. Büyükalaca, H. Bulut, and T. Yılmaz, “Analysis of variable-base heating and cooling degree-days for Turkey,” Appl. Energy, vol. 69, no. 4, pp. 269–283, Aug. 2001, doi: 10.1016/S0306-2619(01)00017- 4.

[28] S. N. Chandramowli and F. A. Felder, “Impact of climate change on electricity systems and markets – A review of models and forecasts,” Sustain. Energy Technol. Assess., vol. 5, pp. 62–74, Mar. 2014, doi: 10.1016/j.seta.2013.11.003.

[29] A. D’Amico, G. Ciulla, D. Panno, and S. Ferrari, “Building energy demand assessment through heating degree days: The importance of a climatic dataset,” Appl. Energy, vol. 242, pp. 1285–1306, May 2019, doi: 10.1016/j.apenergy.2019.03.167.

[30] A. Bolattürk, “Optimum insulation thicknesses for building walls with respect to cooling and heating degree-hours in the warmest zone of Turkey,” Build. Environ., vol. 43, no. 6, pp. 1055–1064, Jun. 2008, doi: 10.1016/j.buildenv.2007.02.014.

[31] Gh. R. Roshan, A. A. Ghanghermeh, and S. Attia, “Determining new threshold temperatures for cooling and heating degree day index of different climatic zones of Iran,” Renew. Energy, vol. 101, pp. 156– 167, Feb. 2017, doi: 10.1016/j.renene.2016.08.053.

[32] “scipy.optimize.curve\_fit — SciPy v1.5.2 ReferenceGuide.” https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve\_fit.html (accessed Oct. 01, 2020).

[33] J. G. De Gooijer and R. J. Hyndman, “25 years of time series forecasting,” Int. J. Forecast., vol. 22, no. 3, pp. 443–473, Jan. 2006, doi: 10.1016/j.ijforecast.2006.01.001.

[34] E. Erdogdu, “Natural gas demand in Turkey,” Appl. Energy, vol. 87, no. 1, pp. 211–219, Jan. 2010, doi: 10.1016/j.apenergy.2009.07.006.

[35] S. Saab, E. Badr, and G. Nasr, “Univariate modeling and forecasting of energy consumption: the case of electricity in Lebanon,” Energy, vol. 26, no. 1, pp. 1–14, Jan. 2001, doi: 10.1016/S0360- 5442(00)00049-9.

[36] S. Sp. Pappas, L. Ekonomou, D. Ch. Karamousantas, G. E. Chatzarakis, S. K. Katsikas, and P. Liatsis, “Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models,” Energy, vol. 33, no. 9, pp. 1353–1360, Sep. 2008, doi: 10.1016/j.energy.2008.05.008.

[37] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, “Day-ahead electricity price forecasting using the wavelet transform and ARIMA models,” IEEE Trans. Power Syst., vol. 20, no. 2, pp. 1035– 1042, May 2005, doi: 10.1109/TPWRS.2005.846054.

[38] L. Suganthi and A. A. Samuel, “Energy models for demand forecasting—A review,” Renew. Sustain. Energy Rev., vol. 16, no. 2, pp. 1223–1240, Feb. 2012, doi: 10.1016/j.rser.2011.08.014.

[39] A. Rahman, V. Srikumar, and A. D. Smith, “Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks,” Appl. Energy, vol. 212, pp. 372–385, Feb. 2018, doi: 10.1016/j.apenergy.2017.12.051.

[40] W. Wang, T. Hong, X. Xu, J. Chen, Z. Liu, and N. Xu, “Forecasting district-scale energy dynamics through integrating building network and long shortterm memory learning algorithm,” Appl. Energy, vol. 248, pp. 217–230, Aug.2019, doi: 10.1016/j.apenergy.2019.04.085.

[41] Z. Wang, T. Hong, and M. A. Piette, “Predicting plug loads with occupant count data through a deep learning approach,” Energy, vol. 181, pp. 29–42, Aug. 2019, doi: 10.1016/j.energy.2019.05.138.

[42] “Prophet,” Prophet. http://facebook.github.io/prophet/ (accessed Oct. 02, 2020).

[43] S. J. Taylor and B. Letham, “Forecasting at Scale,” Am. Stat., vol. 72, no. 1, pp. 37–45, Jan. 2018, doi: 10.1080/00031305.2017.1380080.

[44] J. Asha, S. Rishidas, S. SanthoshKumar, and P. Reena, “Analysis of Temperature Prediction Using Random Forest and Facebook Prophet Algorithms,” in Innovative Data Communication Technologies and Application, Cham, 2020, pp. 432–439, doi: 10.1007/978-3-030-38040-3\_49.

[45] W.-X. Fang, P.-C. Lan, W.-R. Lin, H.-C. Chang, H.-Y. Chang, and Y.-H. Wang, “Combine Facebook Prophet and LSTM with BPNN Forecasting financial markets: the Morgan Taiwan Index,” in 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Dec. 2019, pp. 1–2, doi: 10.1109/ISPACS48206.2019.8986377.

[46] S. Mahmud, “Bangladesh COVID-19 Daily Cases Time Series Analysis using Facebook Prophet Model,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3660368, Jul. 2020. doi: 10.2139/ssrn.3660368.

[47] I. Fernández, C. E. Borges, and Y. K. Penya, “Efficient building load forecasting,” in ETFA2011, Sep. 2011, pp. 1–8, doi: 10.1109/ETFA.2011.6059103.

[48] G. K. F. Tso and K. K. W. Yau, “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks,” Energy, vol. 32, no. 9, pp. 1761–1768, Sep. 2007, doi: 10.1016/j.energy.2006.11.010.

[49] J. Roth, A. Bailey, S. Choudhary, and R. K. Jain, “Spatial and Temporal Modeling of Urban Building Energy Consumption Using Machine Learning and Open Data,” pp. 459–467, Jun. 2019, doi: 10.1061/9780784482445.059.

[50] Y. Li, Z. Wen, Y. Cao, Y. Tan, D. Sidorov, and D. Panasetsky, “A combined forecasting approach with model self-adjustment for renewable generations and energy loads in smart community,” Energy, vol. 129, pp. 216–227, Jun. 2017, doi: 10.1016/j.energy.2017.04.032.

[51] F. H. Al-Qahtani and S. F. Crone, “Multivariate k-nearest neighbour regression for time series data — A novel algorithm for forecasting UK electricity demand,” in the 2013 International Joint Conference on Neural Networks (IJCNN), Aug. 2013, pp. 1–8, doi: 10.1109/IJCNN.2013.6706742.

[52] J. A. Fonseca and A. Schlueter, “Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts,” Appl. Energy, vol. 142, pp. 247–265, Mar. 2015, doi: 10.1016/j.apenergy.2014.12.068.

Q1

[53] C. Miller et al., “The ASHRAE Great Energy Predictor III competition: Overview and results,” Sci. Technol. Built Environ., vol. 0, no. 0, pp. 1–21, Aug. 2020, doi: 10.1080/23744731.2020.1795514.