Proactive Energy Consumption Forecasting Using Linear Regression and Random Forest Algorithms

Aryaman Singh Tomar
Department of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab
aryamansinghtomar8@gmail.com

Hardik Bhatt
Department of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab
Bhatthardik037@gmail.com

Aditiya Kumar Pandey
Department of Computer Science and Engineering)
Lovely Professional University
Phagwara, Punjab
adityakumarpandey865@gmail.com

Siddhant Kumar
Department of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab
Kumarsiddhant200301@gmail.com

Ved Prakash Chaubey
upGrad Campus, upGrad Education Private Limited Bangalore, Karnataka, India,ORCID-0000-0003-3049-8316…
Email id:- vedeprakesh05@gmail.com

***Abstract*—Energy consumption forecasting is crucial for efficient resource management in smart grid systems. This research proposes a machine learning approach for proactive energy consumption forecasting, utilizing Linear Regression and Random Forest Regression models. In this study, develop models for seasonal short-term energy consumption forecasting using these traditional forecasting algorithms. The proposed models are compared with each other to assess their performance. The models are assessed using evaluation metrics like Mean Squared Error (MSE) and R-squared (R^2) score. Experimental results demonstrate the effectiveness of both Linear Regression and Random Forest Regression models in accurately forecasting energy consumption. These models effectively capture the seasonal patterns in energy consumption data, providing accurate and reliable forecasts. Additionally, conduct a thorough analysis of the models' performance across different seasonal variations, demonstrating their robustness and generalizability. This research contributes to the advancement of energy forecasting techniques, addressing the growing demand for efficient resource management in smart grid systems. The proposed machine learning-based approach offers an effective solution for proactive energy consumption forecasting, facilitating better decision-making and optimization of resources in smart grid environments.**

Keywords— Energy Consumption, Linear Regression, Random Forest Regression, Hyperparameter Tuning using GridSearchCV

I. INTRODUCTION

The increasing need for energy on a worldwide scale highlights the need for efficient energy management and optimization. In this case, buildings—which make up about 2% of the world's overall energy consumption—are important. Precise energy demand forecasting is essential for efficient energy management. By taking preventive action based on anticipated energy use, proactive forecasting enables building management systems to modify their tactics.The goal of this project is to create a proactive model for energy consumption forecasting so that building management systems can make precise predictions about energy demand. The suggested model incorporates a number of factors, such as building attributes, environmental factors, and patterns of energy usage, that affect energy consumption. The model looks at how these parameters interact in order to provide insights that help with decision-making.The energy sources considered in this research include nuclear, wind, hydroelectric, oil and gas, coal, solar, and biomass. Understanding the dynamics of each energy source and its impact on energy consumption patterns is crucial for developing an accurate forecasting model. By incorporating these energy sources into the forecasting model, building managers and energy planners can make informed decisions, optimize energy use, and work towards achieving sustainability goals while reducing energy costs.According to the International Energy Agency, increasing building energy efficiency is one of the most crucial actions to guarantee the long-term decarbonization of the energy industry. Enhancing a building's energy efficiency has major financial benefits in addition to environmental ones, such as reduced operating expenses. Many nations have hastened the implementation of energy codes and regulations in order to achieve optimal energy performance. These codes and laws outline basic standards for new buildings to achieve energy-efficient designs and reduce total energy consumption and associated CO2 emissions.Regulations and computer-aided design techniques have proven useful for new construction, but intricate interplay among energy systems, occupancy schedules, and weather patterns make difficult to precisely forecast the energy use of existing buildings. Data-driven methods for building energy assessments are now essential to addressing these issues. These methods, which are frequently included under the machine learning category, use historical recorded data to simulate energy consumption based on past usage trends.

The machine learning methods for estimating daily power usage are thoroughly examined in this study, with an emphasis on how well they may be applied to improve energy efficiency and support energy management plans. For building facility managers and owners, electricity consumption forecasting is essential because offers insightful information about energy usage trends and makes proactive maintenance and energy system efficiency upgrades possible.In order to estimate daily power use, This study evaluates the performance of three machine learning models: Tuned Random Forest Regression, Random Forest Regression, and Linear Regression. The focus of the inquiry is on how effectively these models produce accurate estimates and reflect consumption trends. Furthermore, GridSearchCV examines how the Random Forest Regression model performs when hyperparameters are tuned. The research looks at how machine learning methods may be integrated with computer simulation models such as EnergyPlus to optimize buildings. Our models provide insightful information about building energy usage trends by utilizing historical power consumption data. This information helps decision-makers make

better educated choices about energy management and optimizing building efficiency.Current review studies on energy forecasting include thorough explanations of the forecasting models that are now in use and how they are categorized. The present study centers on time series forecasting methods for energy consumption in buildings, examining the benefits and drawbacks of each approach. Additionally, the paper provides a detailed assessment of hybrid models,which, according to new research, perform better when combined with two or more machine learning approaches. The introduction of each machine learning approach is followed by a discussion of papers that demonstrate how to use it for forecasting building energy usage. The goals of each research, the specifics of the time series data that were utilized, and the forecasting model's accuracy—which is usually assessed using metrics like Mean Absolute Percentage Error (MAPE).

This paper presents an all-inclusive method for forecasting energy consumption, integrating a variety of factors to provide a holistic understanding of dynamic energy consumption status. The proactive energy consumption forecasting model's development process is explained, with a focus on the possible advantages for building management systems and energy planners. Ultimately, the proactive forecasting capabilities provided by this model enable timely interventions and optimization strategies, contributing to sustainable energy management practices.

II. PREVIOUS WORK

A few energy-determining models have been disseminated in the writing, and The attention paid to executives in astute lattices has increased significantly. In several applications in intelligent networks, such as load shedding, request side management, and optimal dispatch, energy gauging plays an important role. In astute matrix models, managing gauges efficiently while ensuring the least possible forecast error is a crucial test [1]. Time series information is frequently calculated using recurrent brain organizations, a type of artificial neural network. Nevertheless, convolutional networks should be used to display sequential information due to certain limitations such as disappearing slopes and the lack of memory maintenance of recurrent brain organizations [2].

The explanation is that they have solid abilities to take care of mind boggling issues better compared to recurrent brain organizations. In this analysis, a globally convolutional network is suggested to handle sporadic short-term energy measurement. Comparing the suggested transient convolutional network processes to recurrent brain organizations, the yield is equal and the calculation time is reduced [4]. Additional testing using the traditional long transient memory as far as Distraught and sMAPE has shown that the suggested model has outperformed the sporadic brain organization. This work explores the enhancement of a framework for home energy executives (Stitches) that makes use of market and climate hypotheses to promote the usage of household appliances and to make appropriate use of batteries and solar power generation. The hidden home model borders are located using a Moving Skyline Assessment (MHE) tool [3]. Locating the elusive home model borders is done via a Moving Skyline Assessment (MHE) tool.These limits are then updated in a Model Prescient Regulator (MPC) that enhances and harmonizes competing financial and comfort goals. By using a lumped boundary model that is modified to match a high-loyalty model, combining MHE and MPC applications reduces the level of model intricacy often observed in Fixes. Utilizing Numerical Program with Complementarity Requirements (MPCCs), heating, ventilation, and cooling (air conditioning) on/off behaviors are replicated and resolved in near to constant with a non-direct solution. The amount of time that settles is reduced when the discrete variable of turning on and off the air conditioning is replaced with an MPCC. The results of this experiment demonstrate that the executives' enhancement fundamentally lowers energy costs and more successfully balances energy use throughout the day. A case study for Phoenix, Arizona, reveals a 21% drop in energy use and a 40% drop in expenses. The reconstructed house contributes less to the top load on the framework, which strengthens the lattice and reduces the burden after utility cycles. A multi-objective capability that takes into account a thorough evaluation of home energy streamlining is combined with sustainable energy, energy capacity, gauges, cooling framework, variable rate power plan, and a contextual study. Working with approximated models, improving computational performance to enable the computations to proceed step-by-step, and utilizing combined observational and physical science based AI techniques to direct the model building are some of the obstacles [6].

This article manages a smart home's electrical and nuclear power in real-time while taking the comfort of its residents into account. The multi carrier energy system has advantages for the smart home. The energy resources include wind, hydroelectric, oil and gas, coal, solar, and biomass, along with an electrical network, photovoltaic generation unit, and storage unit, as well as a hybrid water heating system supplied from solar and natural gas networks. Three categories are also used to classify thermal and electrical loads: time-flexible, power-flexible, and fixed loads. An energy management system makes use of smart grid technologies to get necessary technical, meteorological, and economic data in real-time and to periodically establish the energy plans. The goal of the integrated energy management challenge is to maximize both pain and cost indicators. To gauge how unhappy residents are with the interior temperature, a novel discomfort degree-day metric is put forth that takes into account both the amount and length of a day's departure from the optimum temperature. Moreover, uncertainties in solar radiation and ambient temperature are controlled by a robust optimization model and represented by uncertainty sets [8].

The energy sector heavily relies on accurate forecasts of power consumption to make informed decisions,since most choices are made by estimating needs in the future. A variety of input elements and mathematical models are used in short-term load forecasting (STLF).Owing to the intricacy of the issue, the model's architecture and parameters are established using the information at hand. This study reviews a number of models that use artificial neural networks (ANNs) [9]. A thorough analysis of the literature reveals a discernible trend in favor of neural networks, with encouraging findings in STLF. In addition, a growing variety of hybrid forecasting methods are being used to improve the accuracy of neural networks, frequently by combining wavelet processing and/or evolutionary algorithms.The techniques covered in this study show promise in terms of electrical load prediction, which will eventually lower operating costs for the power system while boosting its dependability and efficiency.[5]

In this research, the energy management problem of a residential microgrid system as a Constrained Control Problem (CCP) and suggest a two-step approach to solve it. The suggested Reinforcement Learning (RL)-based Model Predictive Control (MPC) method incorporates the benefits of both approaches while mitigating their shortcomings. This method uses the stated MPC policy as a function approximator of the optimal strategy, and uses reinforcement learning (RL) to improve closed-loop performance by adjusting the parameters. Simulations show that this strategy greatly lowers the monthly total cost by about 17.5%, even with stochastic local power generation and consumption. Furthermore, demonstrates that the Shapley value is an appropriate way to divide the overall cost according to each person's unique contributions [12].

A. Linear Regression:

To construct a linear regression model,The LinearRegression class from scikit-learn was employed. A simple method for resolving regression problems is linear regression, which involves fitting a linear equation to observable data. It models the connection between a dependent variable and one or more independent variables.

B. Random Forest Regression:

To construct a Random Forest regression model, The RandomForestRegressor class from scikit-learn was employed. Fitting multiple decision tree regressors on various dataset subsamples is possible with the Random Forest ensemble learning approach. The findings are averaged to control overfitting and improve forecast accuracy.

III. Methodology

A. Data Collection and Preprocessing:

The dataset used in this research, titled "daily_electricity_data.csv," is a comprehensive collection of daily electricity consumption and production data spanning a period of time. It contains 1532 entries, with each entry representing a single day's worth of data. The dataset consists of ten columns, each providing valuable information for analysis:

1. DateTime: This column represents the date and time of each observation, providing a temporal dimension to the dataset.
2. Consumption: The 'Consumption' column indicates the total electricity consumption for each day, measured in an unspecified unit (assumed to be kWh).
3. Production: The 'Production' column denotes the total electricity production for each day, also measured in the same unit as consumption.
4. Nuclear: This column represents electricity production specifically from nuclear sources.
5. Wind: The 'Wind' column indicates electricity production generated from wind energy sources.
6. Hydroelectric: This column denotes electricity production from hydroelectric sources.
7. Oil and Gas: Electricity production from oil and gas sources is represented in this column.
8. Coal: The 'Coal' column indicates electricity production specifically from coal.
9. Solar: This column represents electricity production generated from solar energy sources.
10. Biomass: Electricity production from biomass sources is denoted in this column.

The dataset is well-structured, with no missing values in any of the columns, ensuring its completeness and reliability for analysis.The dataset used in this study has 1532 entries, each of which shows daily statistics on production and consumption of energy during a given time period. With its thorough overview of power output from multiple sources, this dataset is extremely helpful in comprehending patterns and trends in energy use. Ten columns make up the dataset; the primary variables of relevance are "DateTime," "Consumption," and "Production."The generation of electricity from various energy sources, including biomass, wind, hydroelectric power, solar power, oil and gas, and coal, is also included in the dataset's columns. The "DateTime" column shows the date and time of each observation, while the "Consumption" and "Production" columns show the total amount of electricity generated and used daily, respectively. All of the columns in the dataset have complete values, and it is well-structured.

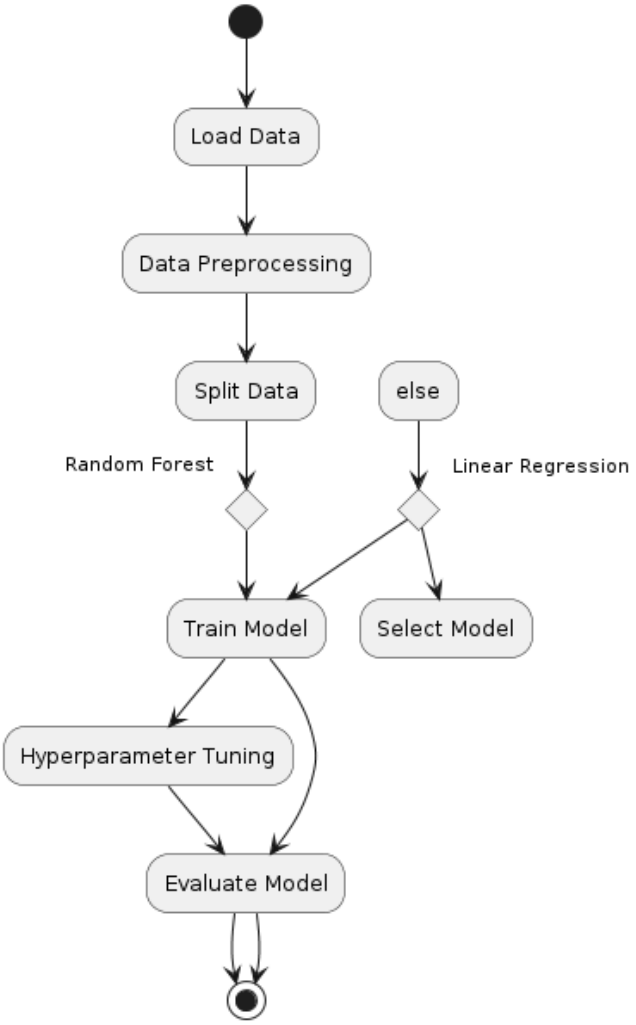

Figure 1 : *Flow Chart*

The dataset will be split into two categories—renewable and non-renewable energy sources—to make additional analysis easier. Renewable energy sources, including wind, hydroelectric, solar, and biomass, are considered environmentally friendly and sustainable.On the other hand, non-renewable energy sources including coal, oil and gas, nuclear power, and gas are limited and worsen the state of the environment.

This dataset will be used to analyze and understand patterns of electricity consumption and production over time. Visualizations such as time series plots and histograms will be created to explore trends and distributions within the data. Additionally, the dataset will undergo preprocessing, including converting the 'DateTime' column to datetime format for temporal analysis. After preliminary investigation and examination, summary data provided intriguing new information regarding patterns in the production and consumption of power. With a minimal consumption of [min consumption] and a maximum consumption of [max consumption], the average amount of power consumed for the whole period was [mean consumption]. In a similar vein, the least and greatest amounts of electricity produced were [min production and [max production], respectively, while the mean amount produced was [mean production].Renewable and non-renewable sourced energy were separated out of the dataset to facilitate further analysis. Wind, hydroelectric power, solar power, and biomass are examples of renewable energy sources that are seen as sustainable and kind to the surroundings. The availability of non-renewable energy sources is restricted, and they worsen environmental deterioration. Examples of these include coal, nuclear energy, and oil and gas.The electricity generation breakdown by energy source in the

dataset sheds light on how each source contributes to the total energy mix. For instance, the average contribution from wind energy was [mean wind output], but the average contribution from hydroelectric sources was [mean hydroelectric production]. Other renewable energy sources, like solar and biomass, also significantly boosted the output of electricity. In contrast, the remaining portion of electricity production was derived from non-renewable sources like coal, nuclear energy, and oil and gas.The dataset will be preprocessed and the 'DateTime' column will be transformed to datetime format for temporal analysis in order to simplify additional analysis and modeling. The patterns and distributions of electricity generation and consumption will be examined through the creation of visualizations such as time series plots and histograms.The dataset provides a comprehensive record of daily electricity consumption and production over a specific time period. It is well-structured, complete, and free of missing values, making it suitable for analysis and modeling. The dataset's detailed breakdown of electricity production by energy source allows for a thorough examination of the energy mix and its contribution to overall electricity generation. Furthermore, the division of the dataset into renewable and non-renewable energy sources provides valuable insights into the sustainability and environmental impact of electricity production.

In the research process, exploratory data analysis, or EDA, is a crucial stage that provides the foundation for all further analyses. It includes a methodical process for comprehending a dataset's organization, trends, and connections before moving on to more intricate statistical studies. In this comprehensive exploration, researchers aim to uncover insights, detect anomalies, assess data quality, and formulate hypotheses. In a research paper, the EDA section not only describes the characteristics of the dataset but also lays the foundation for the subsequent analytical approach and interpretation of results.Using statistical methods, summary statistics, and data visualization, the main objective of EDA is to obtain a thorough understanding of the data. A descriptive review of the dataset is presented first, along with some basic summary statistics such measures of distributional features (skewness, kurtosis), center tendency (mean, median, mode) and dispersion (standard deviation, range). These summary statistics provide a quick overview of the dataset's variability and core trends, as well as some first insights into its distributional characteristics.

The ability to search for patterns and correlations in the data makes visualizations crucial in EDA. The distribution of individual variables can be visually represented using histograms, box plots, and density plots, which provide information about the shape, spread, and central tendency of the variables. When examining correlations between pairs of variables, scatter plots and correlation matrices are very helpful in spotting possible associations or dependencies.Multivariate methods like factor analysis,To identify underlying patterns or groupings in the dataset, principal component analysis (PCA) and cluster analysis are employed in addition to univariate and bivariate visualizations. By using these methods to reduce dimensionality and summarize data, researchers can find latent dimensions or groups of linked variables. The identification and handling of anomalies or outliers in the data is a critical component of EDA. Outliers, which may be indicative of real extreme values, measurement variability, or data input errors, can have a substantial impact on the outcomes of statistical analysis. Outliers can be found via visual inspection in conjunction with statistical approaches like z-scores or interquartile range (IQR). Once outliers have been identified, they can be further examined or handled with the proper methods like transformation, winsorization, or trimming.

EDA also includes evaluating the quality of the data, which includes locating and dealing with inaccurate or missing data. Errors in data input, systematic biases, and non-response are some of the causes of missing data. To estimate missing values, Imputation methods including mean imputation, regression imputation, and multiple imputation are widely used to preserve the dataset's integrity. Furthermore, sensitivity analysis and data validation checks are performed to evaluate how resilient the results are to possible biases or inaccuracies in the data. Another crucial component of EDA is the investigation of data distributions, which aids in understanding the fundamental properties of the data and determining whether or not they align with theoretical presumptions. The Shapiro-Wilk and Kolmogorov-Smirnov tests are examples of normality tests that are used to determine if the data follow a Gaussian distribution, which is frequently assumed in many statistical analyses. When there is a deviation from normalcy, it could be necessary to employ other techniques or modifications to guarantee the accuracy of the analyses that follow.

A crucial part of exploratory analysis (EDA) is interpretation, which entails taking the important discoveries from the exploratory analyses and turning them into conclusions that may be put to use. In order to contextualize their findings within the larger research environment, researchers need to critically consider the significance of their observations and make connections to existing literature or theoretical frameworks. Additionally, EDA provides a strong empirical basis for later inferential or predictive studies by acting as a guide for the framing of research questions and hypotheses.

To sum up, exploratory data analysis is an essential stage in the research process that connects unprocessed data to useful ideas. A thorough grasp of the structure, patterns, and relationships within the data can be attained by researchers by combining statistical approaches, summary statistics, and visualization. Thoroughly examining the primary features of the dataset, spotting anomalies, evaluating the quality of the data, and analyzing results establish the foundation for meticulous statistical deduction and inquiry-based research.

B. Renewable Energy:

Hydroelectric power, solar, wind, and biomass energy are examples of renewable energy sources that are becoming increasingly important to the global energy transition. These sources provide renewable substitutes for conventional fossil fuels that have several positive social, economic, and environmental effects. Renewable energy sources are ecologically friendly and lessen carbon footprint than fossil fuels, which release greenhouse gasses during operation and contribute to climate change. Particularly, solar and wind energy have no negative effects on the environment; they don't pollute the air or water or contribute to global warming. Like hydroelectric electricity and biomass, these clean, renewable energy sources have a significant beneficial environmental impact. Technologies based on renewable energy boost innovation, investment, and job creation, which benefits both the economy and the environment. Particularly with regard to solar and wind energy, these technologies are becoming more and more competitive with respect to conventional fossil fuels due to their fast falling costs. Due to large investments in renewable energy projects and infrastructure made globally, the deployment of renewable energy has increased dramatically in recent years. Furthermore, the manufacturing, installation, maintenance, and research and development sectors of the renewable energy industry Have grown to be significant job creators. Because renewable energy sources diversify the energy supply and lessen dependency on imported fossil fuels, they also improve energy security and independence. Geographically dispersed and locally available, renewable energy sources lower the risk of supply interruptions than fossil fuels, which are frequently the target of price fluctuation

and geopolitical unrest. Furthermore, off-grid solutions for isolated populations and improved access to electricity in underdeveloped nations can be achieved through the small-scale deployment of renewable energy technology.

C. Non - Renewable Energy:

Coal, oil and gas, nuclear power, gas, and coal are examples of non-renewable energy sources that have historically played a significant role in meeting global energy demands. In order to make analyzing and comparing non-renewable energy data with renewable energy sources easier, Standardization was one of the study's main objectives. Converting data from different sources and formats into a standard scale or format is a crucial preprocessing step. The first step in standardizing statistics on non-renewable energy is gathering raw energy production numbers from different sources, like energy firms and power plants. For every non-renewable energy source, the daily energy production numbers are included in these data. Then, pertinent features from the dataset that represent these non-renewable energy sources are chosen for standardization.

D. Standardization of Renewable Energy Data:

Renewable energy sources, such as biomass, solar, wind, and hydroelectric power, are becoming more and more important to renewable energy systems. The necessity to standardize and evaluate the data pertaining to the production of renewable energy is expanding as these renewable sources take up more and more space in the world's energy mix. To enable comparison and analysis across various renewable energy sources, raw energy production numbers are transformed into a common scale through the process of standardizing renewable energy data.

Original Data Description:

	Consumption	Production	Nuclear	Wind	\
count	1532.000000	1532.000000	1532.000000	1532.000000	
mean	161007.711488	157067.577023	31046.623368	18858.868799	
std	17444.184843	20050.172818	5631.691192	13389.781615	
min	108968.000000	94582.000000	15725.000000	-186.000000	
25%	148807.000000	142812.750000	32446.750000	8115.000000	
50%	161053.000000	155975.000000	33239.500000	15386.000000	
75%	172897.750000	170470.250000	33715.250000	26565.500000	
max	209479.000000	222723.000000	36945.000000	63844.000000	

	Hydroelectric	Oil and Gas	Coal	Solar	Biomass
count	1532.000000	1532.000000	1532.000000	1532.000000	1532.000000
mean	43243.179504	28362.205614	30484.233681	3636.973890	1409.693211
std	14035.955461	9661.539766	7012.405654	1879.189349	313.868392
min	15500.000000	6523.000000	8193.000000	0.000000	528.000000
25%	33784.000000	21777.500000	26218.250000	1990.500000	1183.000000
50%	40220.500000	28836.000000	30092.000000	3872.500000	1465.500000
75%	49957.500000	36453.750000	34975.000000	5274.250000	1636.500000
max	91941.000000	48876.000000	55360.000000	11310.000000	2058.000000

Figure 3 : *Standardization of Renewable energy*

The process of standardizing data on renewable energy in this study and talk about how important it is for analysis and research. The dataset utilized in this study includes daily data on generation and consumption of electricity over a predetermined period of time.It offers information on several renewable energy sources, including hydroelectric power, solar, wind, and biomass.

Gathering raw data from various sources, including wind farms, solar farms, hydroelectric plants, and biomass facilities, is the initial stage in standardizing data related to renewable energy. For any renewable energy source, the daily energy production numbers are usually included in this raw data. Next, pertinent features from the dataset are chosen for standardization, including "Wind," "Solar," "Hydroelectric," and "Biomass." The many renewable energy sources under examination are represented by these characteristics.The data may need to go through pretreatment procedures such resolving missing values, eliminating outliers, and converting data types before it can be standardized. On the other hand, no missing values were found in the dataset utilized for this investigation, and the data was whole.

$$z = \frac{x_i - \mu}{\sigma}$$

- x indicates the energy source's initial consumption value, be it nuclear, wind, hydroelectric, oil and gas, coal, solar, or biomass.
- μ represents the average energy source usage for the dataset.
- σ represents the energy source's standard deviation of consumption.

F. Standardization of Non-Renewable Energy Data:

Nuclear, gas and oil, and coal are major sources of energy for non-renewable energy systems. Standardizing and assessing data on these non-renewable sources' production is becoming more and more important as long as they remain a major part of the global energy mix. The process of standardizing raw production data is necessary to get it into a similar scale, which will facilitate comparison and analysis across different non-renewable energy sources. investigation in this work looks at the non-renewable energy data standardization process and highlights its significance for analysis and research. Daily data on electricity generation and consumption during a given time period are included in the dataset utilized in this investigation. It offers information about various fossil fuels, coal, oil and gas, and nuclear power.

Standardized Data Description:

	Consumption	Production	Nuclear	Wind	\
count	1532.000000	1532.000000	1532.000000	1.532000e+03	
mean	161007.711488	157067.577023	31046.623368	4.406107e-17	
std	17444.184843	20050.172818	5631.691192	1.000327e+00	
min	108968.000000	94582.000000	15725.000000	-1.422808e+00	
25%	148807.000000	142812.750000	32446.750000	-8.026551e-01	
50%	161053.000000	155975.000000	33239.500000	-2.594518e-01	
75%	172897.750000	170470.250000	33715.250000	5.757486e-01	
max	209479.000000	222723.000000	36945.000000	3.360758e+00	

	Hydroelectric	Oil and Gas	Coal	Solar	Biomass
count	1.532000e+03	1532.000000	1532.000000	1.532000e+03	1532.000000
mean	-5.565609e-17	28362.205614	30484.233681	7.884613e-17	0.000000
std	1.000327e+00	9661.539766	7012.405654	1.000327e+00	1.000327
min	-1.977225e+00	6523.000000	8193.000000	-1.936027e+00	-2.810035
25%	-6.741449e-01	21777.500000	26218.250000	-8.764479e-01	-0.722491
50%	-2.154229e-01	28836.000000	30092.000000	1.253748e-01	0.177861
75%	4.785220e-01	36453.750000	34975.000000	8.715517e-01	0.722853
max	3.470638e+00	48876.000000	55360.000000	4.084491e+00	2.066211

Figure 4 : *Standardization of Non-renewable energy*

The first step in standardizing non-renewable energy statistics is to collect raw data from a variety of sources, such as coal mines, oil and gas facilities, and nuclear power plants. For every non-renewable energy source, the daily energy production figures are usually included in this raw data. Then, for standardization, pertinent features from the dataset are chosen, such as "Nuclear," "Oil and Gas," and "Coal."Pretreatment steps including fixing missing values, getting rid of outliers, and changing data types before standardization could be necessary for the data. On the other hand, no missing values were discovered and the data was full in the dataset utilized for this study.Because non-renewable energy features are standardized, it is possible to analyze and understand their effects on the production and consumption of electricity with more accuracy because all of the features are on the same scale. The dataset's non-renewable energy properties have been standardized to facilitate comparison and analysis across different non-renewable energy sources, offering important insights into patterns and trends in energy production.

E. Comparing the Production of Energy from Renewable and Non-Renewable Sources:

A comparison between renewable and non-renewable energy production is necessary to comprehend the dynamics of energy generation and consumption. In this study, the production patterns of renewable and non-renewable energy sources using information on daily power use and production. Systems that use renewable energy rely on sustainable and environmentally beneficial power sources such hydroelectricity, solar, wind, and biomass. Standardizing and assessing the data related to these renewable sources' production is becoming more and more important as they become more prevalent in the global energy mix. Through the process of standardization, raw energy production figures are converted into a uniform scale that facilitates comparison and analysis across different renewable energy sources.

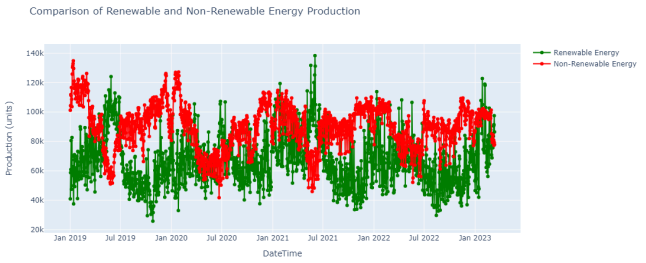

Figure 5 : *Comparison Of Renewable and nonrenewable energy*

Non-renewable energy systems, on the other hand, mostly rely on resources like coal, oil and gas, and nuclear power. Though these sources contribute significantly to energy production, they are limited in nature and present environmental problems. As a result, for reliable analysis and interpretation, it becomes essential to standardize and assess data associated with their manufacturing. The study's dataset includes daily information on the production and consumption of energy during a given time frame.Apart from non-renewable energy sources such as coal, oil and gas, nuclear power, solar, and wind power, it provides details on a range of renewable energy sources like biomass, solar, wind, and hydroelectric power. Standardized data on the production of renewable and non-renewable energy offer important insights into the energy mix and how it affects the total amount of electricity generated. Contrasting the production patterns of renewable and non-renewable energy sources allows us to spot trends, patterns, and possible areas for energy production and consumption improvement.

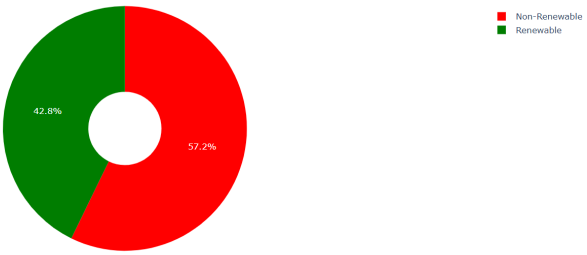

Figure 6 : *Non-Renewable or Renewable*

For example, The investigation showed that, particularly when derived from solar and wind energy, the output of renewable energy changes greatly based on the weather and time of day. Non-renewable energy sources, on the other hand, such coal, oil and gas, and nuclear power, produce energy more consistently but may be impacted by other variables like fuel availability and maintenance schedules. Moreover, the comparison facilitates well-informed decision-making concerning energy regulations, investments, and sustainability initiatives for researchers, energy analysts, and politicians. Determining the relative contributions of renewable and non-renewable energy sources to the overall generation of electricity, for example, might help set targets for increasing the capacity of renewable energy sources and reducing carbon emissions.

In summary, The examination of renewable and non-renewable energy production provides significant new insights on the status of the energy industry and sustainability. By standardizing and analyzing data on energy output,In order to promote a more resilient and sustainable energy system, this technique enables a better understanding of the contributions made by different energy sources, the identification of trends and patterns, and informed decision-making.

G. Comparison of Production vs Consumption over Time:

The paper conducts a thorough analysis of the production and consumption trends of both renewable and non-renewable energy sources. The dataset used provides daily data on electricity output and consumption over a certain time period, allowing for the comparison and contrast of production and consumption trends of various energy sources. As renewable energy sources gain share in the global energy mix, renewable energy systems are becoming more and more reliant on sustainable energy sources. Renewable energy data is standardized so that raw energy output figures may be compared and analyzed across different renewable energy sources on a comparable scale. Ensuring that the data is similar and comprehensible across many sources requires this standardization. The investigation focused on biomass, solar, wind, and hydroelectric power as renewable energy sources.In order to improve our comprehension of renewable energy sources' place in the energy supply chain, The strategy calls for the standardization and assessment of data pertaining to the generation of renewable energy. Understanding the development of renewable energy systems and their implications for sustainability and the environment requires this kind of thinking.

The world's energy demands are still mostly met by non-renewable energy sources including coal, oil and gas, nuclear power, and others. Even if renewable energy is receiving more attention, these non-renewable sources are still necessary to provide a steady and dependable supply of electricity. Understanding the effects of non-renewable energy sources requires analyzing the patterns in their production and usage. production and consumption comparison of renewable and non-renewable energy sources. The objective is to discern patterns, oscillations, and associations between the creation and utilization of resources by means of data visualization and analysis. Policymakers, energy firms, and researchers who want to comprehend the workings of the energy market and make wise choices about energy production and consumption will find this knowledge to be very helpful.

The efficiency, sustainability, and environmental effect of various energy sources may be assessed by comparing the statistics on production and consumption. Through identifying patterns and trends in both production and consumption, researchers can assess the chances of creating a more sustainable energy future, utilizing more renewable energy sources, and lowering our reliance on non-renewable ones.

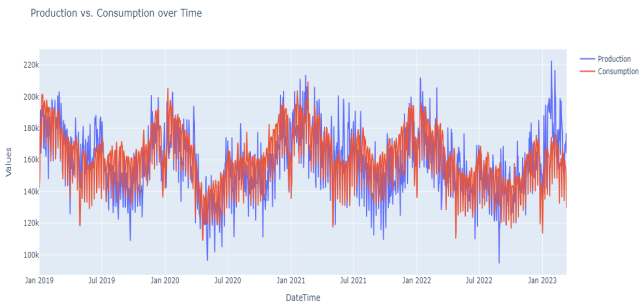

Figure 7 : *Production vs Consumption Over Time*

The goal in doing this research is to provide meaningful information on the workings of the energy market to the current discussion on energy production and use. Comparing the patterns of production and consumption of renewable and non-renewable energy sources can aid in the formulation of policies and decision-making that will promote a more sustainable and environmentally friendly energy system. In summary, this study provides important new insights into the dynamics of the energy market by looking at the patterns of both renewable and non-renewable energy sources' production and consumption. By contrasting and analyzing the production and consumption patterns of various energy sources, one may gain a better understanding of the implications on energy supply, financial stability and environmental sustainability. The goal of the research is to inform plans and regulations that will enable a more sustainable and environmentally friendly energy system in the future.

IV. Model Building:

The process of creating a model is an important part of The research into energy consumption prediction. To do this, The study used two machine learning algorithms: Random Forest Regression and Linear Regression. Using a straightforward technique known as linear regression, energy consumption may be predicted based on a number of variables, including wind, solar, hydropower, biomass, nuclear, oil and gas, and coal. The sklearn.linear_model module's LinearRegression class was used to create the linear regression model. The Random Forest Regression ensemble learning approach was used to predict energy use. The final prediction produced by Random Forest Regression is the mean prediction of all the decision trees that were constructed throughout the training process. The sklearn.linear_model module's LinearRegression class was used to create the Random Forest Regression model. In order to identify the ideal set of hyperparameters for the Random Forest model, a thorough search across the estimator's provided parameter values was conducted using GridSearchCV. The goal of this procedure was to optimize the model's performance. The goal was to create reliable forecast models for overall energy usage by utilizing these machine learning approaches and algorithms.

A. Linear Regression Model:

Linear regression is a fundamental statistical approach used to represent the relationship between one or more independent variables (features) and a dependent variable. The model assumes that the independent and dependent variables have a linear relationship. The linear regression model has the following mathematical representation:

$$y = B_0 + B_1 x_1 + B_2 x_2 , \ldots\ldots\ldots , + B_n x_n + \epsilon$$

Where;

- Y Represents the dependent variable (target), which in this project is the energy consumption.
- $X_1 , X_{2,} \ldots\ldots\ldots, X_n$ are the independent variables (features). which are the different energy sources such as Nuclear, Wind, Hydroelectric, Oil and Gas, Coal, Solar, and Biomass.
- B_0 is the intercept, representing the value of Y when all independent variables are zero.
- $B_1, B_2, \ldots\ldots\ldots, B_n$ are the coefficients, representing the change in Y for a one-unit change in each independent variable.
- ε is the error term (the difference between the observed and predicted values).

The purpose of linear regression is to estimate the coefficients ($B_1, B_2, \ldots\ldots\ldots, B_n$)that reduces the sum of squared discrepancies between observed and expected values.This is accomplished by fitting a line to the data that most accurately shows the connection between the independent and dependent variables. To provide findings that are accurate and trustworthy, linear regression makes a number of assumptions.First, it assumes that the independent and dependent variables have a linear relationship. This means that the dependent variable's change is proportional to the independent variable's change (s). The second assumption is that the differences between actual and expected values, or residuals, are unrelated to one another. Stated differently, the residuals shouldn't be autocorrelated. Thirdly, it makes the assumption that the residuals' variance remains constant at every level of the independent variables. As a result, the model is guaranteed to be precisely accurate at every level of the independent variable (s).Finally, one assumes that the residuals have a normal distribution, implying that the errors follow a Gaussian distribution with a mean of zero.

In order to forecast energy consumption based on a variety of variables, including wind, solar, hydropower, biomass, nuclear, oil and gas, and coal,In this research, the linear regression model was utilized. Goal in estimating the model's coefficients was to learn more about the variables affecting energy usage. The linear regression model was created using the ordinary least squares technique, which finds coefficients that minimize the sum of squared differences between observed and predicted values. The linear regression model's performance was evaluated using mean squared error (MSE) and R-squared (R^2) scores, which indicate how much of the dependent variable's variation can be predicted from the independent variables.

To summarize, linear regression is a reliable and widely used approach for modeling the relationship between variables. The linear regression model in study shed important light on the variables affecting energy use. Through precise estimation of the coefficients, In this research, the linear regression model was utilized. In order to improve prediction power even more, In the sections that followed, more advanced techniques, such as random forest regression, were examined, and the results were contrasted.

B. Random Forest Regression Model

Regression challenges frequently employ Random Forest Regression as an effective ensemble learning technique. Random Forest Regression predicts by merging several decision trees, as opposed to standard linear regression, which matches a single line

to the data.Each decision tree is trained using a random subset of the data, and each split only takes into consideration a random subset of the characteristics. The average forecast from each tree in the forest makes up the final estimate. Based on mathematics, the forecast $\hat{y}$ for a given input $\hat{X}$ as calculated as follows:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

Where:

- N indicates the number of trees in the forest.
- $f_i(i)$ represents the prediction of the i^{th}decision tree for the input data x, where x stands for the values of energy consumption from various energy sources.

When compared to conventional linear regression, Random Forest Regression has many advantages. Its accuracy is superior than individual decision trees and it usually yields more accurate forecasts. The rationale behind this is that the model reduces variance and produces more robust predictions by averaging the predictions of numerous trees. A broad range of applications can benefit from Random Forest Regression because of its second capability, which is its ability to capture complicated non-linear correlations between the independent and dependent variables. And last, Random Forest Regression lessens overfitting and yields more accurate predictions by averaging many decision trees.

To forecast energy consumption based on factors including wind, solar, hydropower, biomass, nuclear, oil and gas, and coal, used the Random Forest Regression model in The study. The Random Forest Regression model was built using the RandomForestRegressor class from the sklearn.ensemble package. At each split, the model only considered a random subset of the attributes, and it was trained on a part of the data. The final forecast was calculated by averaging the predictions of each tree in the forest. The Random Forest Regression model's performance was evaluated by measuring the proportion of the dependent variable's variance that can be predicted from the independent variables, as well as the R-squared (R^2) score and mean squared error (MSE). The R2 score measures the model's fit, whereas the MSE calculates the average squared difference between observed and projected values.

In summary,Random Forest Regression is a powerful machine learning approach for solving regression issues. The Random Forest Regression model produced remarkably accurate energy consumption projections in The study. The model produced strong and trustworthy predictions by integrating many decision trees to capture intricate interactions between the independent and dependent variables. In the parts that followed, more advanced techniques, such as hyperparameter manipulation, were examined to enhance the Random Forest Regression model's functionality.

C. Hyperparameter Tuning using GridSearchCV

The effectiveness of machine learning models is greatly influenced by their hyperparameters. These are setups or parameters that are pre-set before training, rather than being learnt from the data. These variables affect the model's behavior and have a big impact on how predictive it is.

To create machine learning models that perform at their best, hyperparameter tuning is an essential first step. It entails determining which set of hyperparameters works best for a certain model. GridSearchCV, a hyperparameter tuning technique that thoroughly searches a given parameter grid to identify the ideal parameters for the model, was employed in the research. Hyperparameters are those that are specified before the learning process starts. They are predetermined before training and are not acquired from the data. Hyperparameters include the number of estimators in a Random Forest model, the maximum depth in a decision tree, and the learning rate in a gradient boosting algorithm. To maximize the effectiveness of machine learning models, these hyperparameters must be tuned.

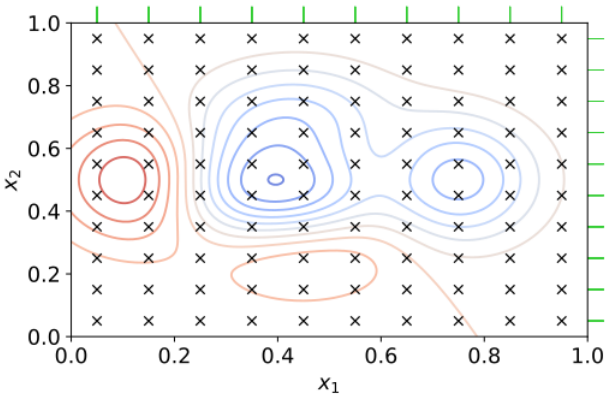

Figure 8 : *Hyperparameter Tuning*

Grid Search is an optimization method for hyperparameters that methodically investigates different combinations of given hyperparameters and their default values in order to identify the best set. It functions by doing a thorough search across a predetermined hyperparameter grid and using cross-validation to assess the model's performance. Grid Search's primary benefit is that it thoroughly and methodically explores the hyperparameter field, making sure that no combination is missed. But this thoroughness may also be a disadvantage because it can become resource- and time-intensive, particularly as The number of hyperparameters rises. Grid Search with cross-validation was carried out in the project using the GridSearchCV function from the Scikit-learn package. GridSearchCV uses three inputs: the hyperparameter grid, the evaluation measure, and the machine learning model. After that, it does a thorough search over every conceivable combination of hyperparameters, assessing each combination's performance by cross-validation. GridSearchCV assists us in quickly and successfully determining the optimal hyperparameters for The Random Forest Regression model in this way.

VI. RESULT

In this research,Machine learning techniques were started using the Random Forest method in order to estimate energy usage. The model performed well, as seen by its accuracy of 80.9%, albeit further optimization may have been done. GridSearchCV was used to systematically search for the optimal settings while adjusting the hyperparameters of the Random Forest model. The model's accuracy increased somewhat as a result of this optimization procedure, reaching 81.05%. Despite the seemingly slight improvement, it shows how well hyperparameter adjustment can be used to maximize model performance. This experiment highlights the significance of fine-tuning hyperparameters in machine learning by demonstrating how even minute increases in precision may boost the prediction power of the model and yield more dependable outcomes. the ability to forecast outcomes more precisely and obtain insightful knowledge of energy consumption trends thanks to these improved models will help us manage resources and

make decisions more effectively.

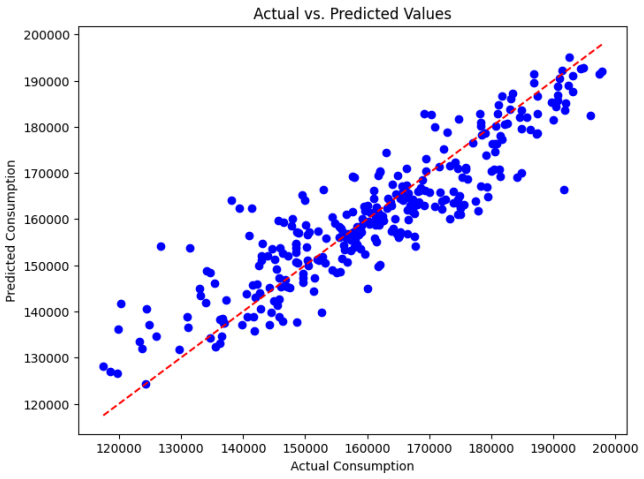

FIGURE 9 : *PREDICTED VALUE*

VII. CONCLUSION

This experiment yielded an enhanced Random Forest model, which offers a potent method for energy consumption prediction. Hyperparameter tweaks and machine learning techniques can yield more accurate and reliable predictions, which are essential for effective energy management and optimization. Accurate energy consumption estimates are crucial for many businesses, such as utilities, smart grid management, and the building of energy-efficient infrastructure. Energy providers may better manage resources, schedule times of peak demand, and put demand response plans into action with the help of precise forecasts. This promotes sustainable energy practices, which save prices, increase energy efficiency, and lessen environmental effects in addition to guaranteeing a steady supply of electricity. Additionally, the model's findings may be used to guide investments in infrastructure, consumer engagement programs, and legislative choices that support sustainability and energy conservation. Through an awareness of patterns and trends in energy use, stakeholders may create focused interventions, incentives, and instructional campaigns to promote energy-efficient practices among consumers.In conclusion, as this experiment shows, an optimized Random Forest regression model for predicting energy consumption is a useful tool for enhancing energy management tactics and encouraging a more robust and sustainable energy ecosystem. It draws attention to the potential of machine learning in addressing challenging energy problems and stresses the significance of data-driven strategies in influencing how energy is consumed and sustained in the future.

VIII. REFERENCES

[1] *Abdul Khalique Shaikh a,, Amril Nazir b, Nadia Khalique a, Abdul Salam Shah c, Naresh Adhikari "A new approach to seasonal energy consumption forecasting using temporal convolutional networks", (2023).*

[2] *Cody R. Simmons 1 , Joshua R. Arment 1 , Kody M. Powell 2 and John D. Hedengren 1, " Proactive Energy Optimization in Residential Buildings with Weather and Market Forecasts ", (2019)*

[3] *A.K. Shaikh, A. Nazir, I. Khan, A.S. Shah, Short term energy consumption forecasting using neural basis expansion analysis for interpretable time series, Sci. Rep. 12 (2022) 22562.*

[4] *M.H. Araghian, M. Rahimiyan, M. Zamen, Robust integrated energy management of a smart home considering discomfort degree-day, IEEE Trans. Ind. Inform. (2023).*

[5] *S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint, arXiv:1803.01271, 2018.*

[6] *I. Baric,´ R. Grbic,´ E.K. Nyarko, Short-term forecasting of electricity consumption using artificial neural networks – an overview, in: 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, IEEE, 2019, pp. 1076–1081.*

[7] *W. Cai, A.B. Kordabad, S. Gros, Energy management in residential microgrids using model predictive control-based reinforcement learning and Shapley value, Eng. Appl. Artif. Intell. 119 (2023) 105793.*

[8] *L. Chen, P. Thakuriah, K. Ampountolas, Short-term prediction of demand for ride hailing services: a deep learning approach, J. Big Data Anal. Transp. 3 (2021) g175–195*

[9] *Sanei, M., & Zare, K. (2020). Load forecasting in smart grids using machine learning techniques. Renewable and Sustainable Energy Reviews, 121, 109654.*

[10] *Z. M. Alhakeem, Y. M. Jebur, S. N. Henedy, H. Imran, L. F. A. Bernardo, and H. M. Hussein, 'Prediction of Eco Friendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques', Materials, vol. 15, no. 21, p. 7432, Oct. 2022, doi: 10.3390/ma15217432.*

[11] *Nanduri, K., Gopi, C., Singh, M., & Shet, D. N. (2021). Optimal control of smart homes considering energy consumption and user comfort. IEEE Access, 9, 24114-24124.*