Enhancing Grid Stability and Market Dynamics through Hybrid Forecasting of Renewable Energy

**Abstract**

This project focuses on the critical challenge of predicting renewable energy production and market prices for both grid stability and market dynamics. With the growing number of renewable sources being implemented, an accuracy in forecasting that would allow for efficient grid management and wise, proper decision-making in very unpredictable energy markets became more imperative. In this regard, the developed work proposes a new setup for hybrid approaches to forecasting based on tree-based machine learning algorithms and times series models to improve prediction accuracy.

The research engineers’ features using the integrated dataset from September 2020 to January 2024 for both short-term fluctuations and the long-term trends related to energy production and pricing. Some of the most important feature engineering includes lag features, rolling means, and expanding means. In comparison with baseline models, this approach drastically outperforms them, with the final Random Forest implementation providing an RMSE of 7.24. This is, however, the best-performing model after undergoing extensive hyperparameter tuning with the highest combination of parameters: max\_depth = 10 and n\_estimators = 100. SHAP analysis may provide very interpretable insight into the decision-making process of the model. Concretely, the results show that the most important feature in predicting day-ahead prices is lag\_1, yesterday's price. Other critical features are market index prices and engineered rolling averages of them, while actual power generation seems to have less influence than one may have expected. It offers a more accurate and interpretable forecasting tool to the renewable energy sector, which possibly helps in improving investment strategies and makes the integration of renewable sources into the power grid smoother.

# Introduction

The global picture of energy is changing, and renewable sources take a front seat in mitigating climate change while ensuring long-term, sustainable energy production. Accurate forecasting of energy production and market prices is going to be important for grid stability, market efficiency, and informative decision-making as renewable penetration into the power grids increases, assert (Wang et al., 2023). However, the intrinsic volatility of solar and wind energy makes it a much more challenging task for traditional approaches of load forecasting (Lanka et al., 2021).

Recent advancements in machine learning and time series analysis have opened new avenues for improving forecast accuracy. However, there remains a significant gap in the literature regarding the integration of tree-based machine learning algorithms with time series models for renewable energy forecasting. Moreover, the interpretability of these complex models is often overlooked, leading to a lack of trust and understanding among stakeholders (Chen et al., 2022).

The current research aims to bridge these gaps by developing a sophisticated hybrid framework in its forecasting. Guiding the current study is the following primary research question: **How can combining tree-based ML and time series forecasting improve renewable energy production and price predictions? How can Explainable AI, especially SHAP values, provide insights to guide investments and adoption in volatile markets?**

The objectives of this research are:

1.Development of a type of hybrid forecasting model that would surpass the performance of traditional methodologies with regards to the prediction of renewable energy production and market prices.

2. To apply feature engineering techniques representing the short-term fluctuations and the long-term trends in the energy data.

3. Using Explainable AI (XAI) techniques in this case, SHAP values to get interpretable insights into the process of the model's decision-making.

4. To assess how enhanced forecasts will affect investment decision-making and the rate of renewable energy adoption.

This research contributes to the scientific literature by:

1. Proposing a novel hybrid forecasting framework that significantly improves prediction accuracy.
2. Showing specific feature engineering techniques that can powerfully capture complex patterns in renewable energy data.
3. SHAP analysis provides insights into the relative importance of different features in predicting energy prices.
4. Provide the methodology for understanding complex machine learning models that would ascertain energy forecasting, which will give trust and understanding to various levels of stakeholders.

The report is structured as follows: Section 2 Literature Review, Section 3 Methodology, including Data Collection, Preprocessing, Feature Engineering, and Model Development, Section 4 Results of the Forecasting Models and SHAP Analysis, Section 5 Discussion of the Results, and finally, Section 6 Conclusion of the report with key findings and a suggestion of research avenues.

# Related Work

Increasing the share of renewable energy in power grids raises the need for more accurate prediction models of energy production and market prices. This review will examine the current research at the intersection between tree-based machine learning algorithms and time-series forecasting models toward renewable energy prediction, exploring impacts of improved predictions on investment decisions and adoption rates within volatile markets. Drawing on foundational work done in the module Research in Computing, this review criticizes recent developments and points to areas that would demand further investigation.

## Advancements in Time Series Analysis and Machine Learning for Energy Forecasting

Recent developments in the methodologies of time series analysis and machine learning have opened up new avenues for complex energy sector forecasting methodologies. Huang, 2021, has proposed a new supply characteristic learning methodology with data visualization, entropy-power method, and conditional random sampling using the Monte Carlo method for building transfer schemes. This approach, though focused on the area of supply chain management, gives grounds for application in renewable energy prognosis and modeling of supply-demand dynamics in the energy market. In fact, Huang's work is of power because, for the first time, visualization techniques are combined with entropy-power methods; this opened up new dimensions toward time series analysis. The latter shall test its direct applicability to renewable energy forecasting as an avenue for further research.

Chen et al. (2022) have successfully applied ARIMA models in global land-ocean temperature indices predictions. Such include accurate predictions with critical importance towards the understanding of the impacts of climate change in supporting decision-making processes of various stakeholders. While this paper provides an example of the extensive use of time-series analysis, it is fundamental to underpin that its application can be restricted to ARIMA model and not necessarily include in such form more elaborated techniques. Moreover, analogous approaches in renewable energy forecasting will naturally have to take into consideration the peculiarities of energy production and market dynamics, thus possibly needing using much more developed analytical tools.

In a comparative study of different time series forecasting models within the framework of rainfall prediction in tropical areas, *Basha et al. (2022)* highlighted some interesting results. Basically, the results have shown that the HWSM results were less erroneous, so the suggestion was to combine statistical and machine learning methods for better predictions. This may also suggest that a possible integration of tree-based algorithms, such as Random Forest, GBM, or XGBoost, might be combined with a time-series model like Prophet to potentially improve accuracy. Speaking accurately, although this work presents overall comparisons among various forecasting models, attention is needed for applying these findings on rainfall prediction as that area is strictly focused on renewable energy.

## Application of Advanced Forecasting Models in Related Domains

Advanced forecasting models applied to related domains add much value in insightful predictions for renewable energy. Kogekar et al. (2021) employed three univariate time series models: ARIMA, SARIMA, and Prophet, together with their new approach of FlowARIMA for water quality prediction on River Ganga. Their results showed the performance of the Prophet model, which captured both trends and seasonality in such water quality parameters very well compared to ARIMA and SARIMA. This paper demonstrates the power of Prophet in capturing complex trends and patterns within time series data; it can be expanded to areas of renewable energy forecasting mostly under volatile market conditions. In contrast, this move from water quality forecasting to energy market prediction will need much adaptation and validation.

In the financial sector, Mandeep et al. (2022) applied classic methods and AI-based prediction models in stock exchange trend forecasting. Their work underlines that stock market prices are non-linear, and there are a great number of factors influencing the market, comparable to the complexities of renewable energy markets. This work has underlined the inadequacy of classic methods in the accurate prediction of a complex market trend and at the same time proposed to make use of advanced Machine Learning techniques to improve the accuracy of the prediction. Drawing a parallel between financial and energy markets, one supposes that similar approaches could turn out effective in renewable energy forecasting, although all the peculiarities of energy markets should not be forgotten in any implementation.

Novel approaches to distinguishing between normal and pathological gait have been developed, based on the integration of attributes and time series in multivariate time-series classification Fekete and Molnar (2022). Although highly focused on healthcare applications, this approach to combining both attribute and time series data illustrates a potential avenue for the development of hybrid models capable of improving renewable energy forecasting and market condition classification. It will involve the identification of relative attributes and time series data that will be pertinent to this methodology as it gets adapted to the renewable energy sector. This weds an innovative approach, therefore, in underlining potential benefits attributed to cross-discipline techniques driving predictive models across different fields.

## Explainable AI (XAI) in Energy Forecasting

Recently, there has been a lot of interest in the application of Explainable AI in energy forecasting, since many machine learning models are black box. Teixeira et al. (2023) applied an XAI-based framework to the assessment of photovoltaic energy generation forecasting. Their work incorporates explanation mechanisms for extracting knowledge from model behavior and LIME, SHAP, Anchors, and MAP-Elites. It gives multiple explanation formats for different user profiles, increasing the transparency and credibility of ML solutions on energy forecasting. The power in Teixeira's approach lies in the full integration of several techniques from XAI that turn it into a very sound framework for interpreting models. Until now, the application has been limited to PV energy generation; therefore, generalizing to a wider setting of renewable energy market prediction remains an open opportunity and challenge at the same time.

A novel electrical load forecasting model will be developed using explainable artificial intelligence through the SHAP method to interpret the results and identify key parameters for Norwegian residential buildings Henriksen et al. (2022). This paper shows that the LSTM-based local residential load forecasting algorithms can be implemented successfully, while XAI tools can be applied for extremely valuable insights in both model developers and domain experts. The vital parameters, which included historical load values and temperature, are of huge importance to model accuracy, but at the same time, they show potential for further model improvement if a deep understanding of feature importance is reached. Notably, though this work contains very useful insights into the application of XAI in energy forecasting, the focus on residential load forecasting might limit its immediate applicability in scenarios of renewable energy production and market price prediction.

## Feature Selection and Reduction Techniques in Energy Forecasting

Jawale and Magar (2022) contributed new approaches in the area of feature selection and reduction over large datasets from sensors, oriented towards acceleration sensors. Statistical and clustering methods can vastly reduce the size of data with their original properties and patterns still intact. The application was not applied to renewable energy data, but such techniques might add value to preprocessing and the optimization of the feature set for renewable energy forecasting models. The challenge is how these methods should be suitably tuned to have good performance on renewable energy time series data.

Agrawal and Sharma, 2022 studied how the time series data test classification affects feature extraction techniques through several methods so as to place them for various high-dimensional data types. Their research highlighted that feature selection is attentive to enhancing an overall model in predictive performance and, similarly, is supposed to be paid equal attention toward the structure of the data and the choice of the classification model. These insights can be applied to renewable energy forecasting and, in principle, increase the efficiency and accuracy of models, but this will need very careful consideration, considering the features inherent in both energy production and market dynamics.

## Research Gaps and Future Directions

This literature review shows a growing interest in the integration of tree-based machine learning algorithms with time series forecasting models and their enhancement in interpretability with XAI techniques. On the other end, it highlights that, to a large extent, application to renewable energy production and market price prediction in highly volatile markets is still significantly lacking. Moreover, how improved and more interpretable forecasts impact investment decisions and adoption rates in highly volatile markets has not been explored to any large degree.

In this respect, future research should aim to construct a hybrid framework for forecasting a tree-based algorithm coupled with the models of time series specifically in the prediction of renewable energy in volatile markets. The idea embedded should be to use methods of explainable AI, more so SHAP values, to make the models of forecasting more interpretable and trustworthy. Another critical area of study refers to how improved forecasts could influence investment decisions and adoption rates in such markets.

These directions could be realized by creating comprehensive datasets of renewable energy production and corresponding market price data from volatile markets, preprocessing this data with adapted techniques for feature selection and reduction, and finally, developing a hybrid forecasting model and training it using historical data in conjunction with other relevant external factors. One can take a first step toward giving interpretable insights into the model's predictions by applying XAI techniques, which could be followed by model evaluation using appropriate metrics for both accuracy and interpretability. Case studies or simulations on how these improved forecasts will affect investment decisions and adoption rates will become very important and meaningful for stakeholders in the renewable energy sector.

In these regards, the pursuit of research directions and implementation strategies can pay very substantial dividends toward renewable energy forecasting and volatile market decision-making that will underpin a sustainable energy future. This approach can enhance accuracy in renewable energy forecasting, transparency, and trustworthiness of the predictions, thus enabling more informed decision-making within the complex and fast-changing landscape of renewable energies.

# Research Methodology

This quantitative study predicts renewable energy production and market prices using machine learning and time series methods.

## Data Collection and Preprocessing

The dataset used in this study contains 58,368 records from September 20, 2020, all the way up to January 18, 2024, at half-hourly intervals. Data is derived from IEEE competitions, while columns are 15 in number: datetime, energy production metrics, pricing data, and temporal features. Important variables include:

* dtm: Datetime in UTC
* mip: Market Index Price
* solar\_mw, wind\_mw: Solar and wind power output
* solar\_capacity\_mwp, solar\_installedcapacity\_mwp: Solar capacity metrics
* ss\_price: System Service Price
* boa\_mwh: Bid Offer Acceptance in MWh
* da\_price: Day-Ahead Price (target variable)

Preprocessing steps included:

1. Converting column names to lowercase for consistency
2. Transforming the 'dtm' column to datetime format
3. Handling missing values using mean imputation for numerical columns
4. Extracting additional temporal features (year, quarter, month, week, day, time) from the datetime column

A preprocessing strategy like this is following the methods applied by Chen et al. (2022) when trying to analyze temperature indices for the world, taking into consideration techniques only adapted to our energy market context.

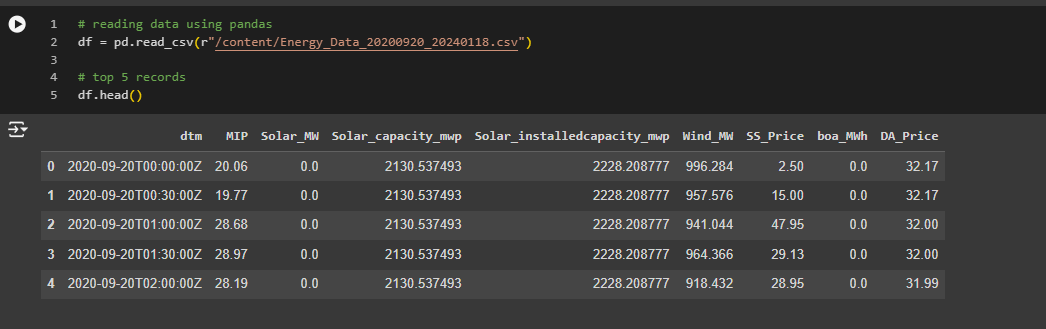


Figure 1 First 5 rows

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the characteristics of the dataset. This process involved:

1. Calculating Descriptive Statistics for All Numerical Variables
2. Analyzing distributions, revealing right-skewness in most features, except for boa\_mwh (left-skewed)
3. Time series plots to visualize trend or patterns of key variables
4. Correlation analysis will be conducted to identify the relationships between features.

The approach used was based on the EDA method since, according to Zhang 2021, a critical analysis of time series data is very instrumental in data science and business analytics for producing forecasts upon which organizational decisions are based.

## Feature Engineering

Based on the insights obtained from EDA, and guided by the work of Basha et al. (2022), in which statistical and machine learning approaches were combined with very high success, the features in the following instances were engineered:

1. Lag features: Previous day's price (lag\_1)
2. Rolling mean: 6-period rolling average of lag\_1 (rolling\_mean\_lag\_1)
3. Expanding mean: Cumulative average of lag\_1 (expanding\_mean\_lag\_1)

These features were specifically designed to capture the short-term fluctuations and long-term trends in the data, thereby considering the complex temporal dynamics underpinning renewable energy markets. Faced with a half-hourly data frequency, a rolling average of 6 periods was decided on, covering a window of 3 hours typically aligned with energy markets' trading intervals.

## Model Development

Two main modeling approaches were employed, chosen based on their success in related domains as identified in our literature review:

1. Vector Autoregression (VAR) Model:
   * Implemented as a baseline model to capture linear interdependencies among multiple time series
   * Used features: mip, solar\_mw, solar\_capacity\_mwp, solar\_installedcapacity\_mwp, wind\_mw, ss\_price, boa\_mwh, da\_price
   * This choice was motivated by the work of Kaushik and Vashisht, 2022, who showed the efficacy of autoregressive models in time series analysis.
2. Random Forest Regression:

* The reason for choosing this was its resilience to non-linear relationships and handling complex interactions between features, demonstrated by by Agrawal and Sharma (2022) through their comparative analysis of feature extraction techniques.
* Two iterations developed: a) First model: Using original features b) Second model: Including engineered features

Model training and evaluation followed these steps:

1. Data splitting: 80% training, 20% testing
2. Hyperparameter tuning using GridSearchCV with TimeSeriesSplit for cross-validation
3. Model training on the full training set using best hyperparameters
4. Prediction on the test set

This design process was developed in order to test the performance of the models strictly, respecting the temporal nature of data according to best practices outlined in Tripathy et al., 2023, for the analysis of deep learning techniques in time series forecasting.

## Model Evaluation

Models were evaluated using the following metrics:

1. **Mean Squared Error (MSE):-** MSE returns the average of the squares of the estimated value minus the actual value. It gives more weight to bigger errors, hence making the measure especially useful when large errors are especially undesirable.

* MSE = (1/n) \* Σ(y\_i - y^\_i)^2
  + n: The number of observations
  + y\_i: The actual value for the i-th observation
  + y^\_i: The predicted value for the i-th observation
  + Σ: The sum from i=1 to n

1. **Mean Absolute Error (MAE)**:The MAE provides the average magnitude of errors in a set of predictions, without considering their direction. This is less sensitive to outliers compared to MSE, and more useful when you want to treat all sizes of errors linearly.

* MAE = (1/n) \* Σ|y\_i - y^\_i|
* n: The number of observations
* y\_i: The actual value for the i-th observation
* y^\_i: The predicted value for the i-th observation
* |...|: Absolute value
* Σ: The sum from i=1 to n

1. **Root Mean Squared Error (RMSE)**:RMSE is just the square root of MSE. It is interpreted as the standard deviation of the residuals, or prediction errors. Since it's in the same units as the response variable, RMSE is easily interpretable:
   * 1. RMSE = √[(1/n) \* Σ(y\_i - y^\_i)^2]
        1. n: The number of observations
        2. y\_i: The actual value for the i-th observation
        3. y^\_i: The predicted value for the i-th observation
        4. Σ: The sum from i=1 to n
        5. √: Square root

These metrics are chosen because, according to the available literature(Kogekar et al., 2021; Tripathy et al., 2023), they quantify prediction errors in regression quite effectively. Combining these multiple metrics will yield detailed performance of the models across several parameters of prediction accuracy.

## Explainable AI Analysis

In this respect, the same approach as Teixeira et al. (2023) and Henriksen et al. (2022) is followed by using SHAP analysis for Random Forest model prediction interpretation. This technique can perform both global feature importance and local explanation of individual predictions, thus attending to one of the more significant concerns about complex machine learning models often referred to in the literature as the "black box.".

The SHAP analysis involved:

1. Computing SHAP values for the test dataset
2. Generating summary plots to visualize overall feature importance
3. Creating force plots to explain individual predictions

It thus improves the interpretability of the complex Random Forest model showing what drives the predictions and therefore aligns with a growing emphasis on explainable AI in the energy sector.

# Design Specification

This section describes the architecture, framework, and methodologies constitutive to the basis of our proposed design for a forecasting system intended for the renewable energy sector. It is designed intelligently so as to set out an accurate and interpretable solution for the forecasting of renewable energy production and market prices.

## System Architecture

The architecture of the system is sequential in structure, featuring four key components: the Data Preprocessing Module, the Exploratory Data Analysis Module, the Feature Engineering and Model Development Module, and the Model Evaluation and Interpretation Module. Each of these modules runs in a sequence through which data flows ultimately to generate final predictions and insights.

## Data Preprocessing Module

The Data Preprocessing Module cleans and transforms the raw data to make it ready for analysis and modeling. It includes the Data Loader module, responsible for parsing the CSV file with the energy data; the Data Cleaner, responsible for the management of missing values and the conversion of types; and the Temporal Feature Extractor, responsible for extracting additional time-based features.

This module is implemented by data processing done using pandas. The column `dtm` is converted into datetime format, and year, quarter, month, week, day, and time features have been extracted. At the same time, the missing values are imputed by the mean of the rest of the values. These preprocessing stages of the data enable us to bring it to the required format for the subsequent analysis and modeling stages.

## Exploratory Data Analysis (EDA) Module

The exploratory data analysis module aims to understand the characteristics of data and identify the patterns in it. It shall have a statistical analyzer that computes summary statistics of numeric variables, a distribution analyser that looks into the distribution of key features, a time series plotter to visualize trends and patterns in time series data, and a correlation analyzer that finds relationships between features.

This module utilizes matplotlib, seaborn, and plotly to implement data visualization. Histograms, box plots, and time series plots are created for a fuller view of the data. Moreover, it evaluates a correlation matrix and plots it, showing the interrelationship of variables. All those plots and analyses yield important insights that are crucial in further decisions on modeling.

## Feature Engineering and Model Development Module

The Feature Engineering and Model Development Module will be responsible for building relevant features and training predictive models. This shall be followed with a Lag Feature Generator that would create the lagged variables, then a Rolling Statistics Calculator for computing the rolling means and expanding means, a Model Trainer to implement the VAR model and Random Forest model training; lastly, Hyperparameter Tuner in the case of Random Forest model optimization of parameters.

The implementation provides several time-based features, including lag\_1, rolling\_mean\_lag\_1, and expanding\_mean\_lag\_1. The evaluation of the respective matrix definition for the VAR and the Random Forest implementation with hyperparameters are provided by the libraries: for the VAR, we have statsmodels, and for the Random Forest implementation, we have scikit-learn. To maintain the structure of the temporal data both at training time and inference time, the module uses the TimeSeriesSplit for cross-validation in the grid search process. This approach ensures that the modeling goes along with the time-dependency nature of the data properly.

## Model Evaluation and Interpretation Module

The Model Evaluation and Interpretation Module is primarily responsible for the evaluation and interpretation of models. It holds within itself an evaluator of performance through metrics such as MSE, MAE, and RMSE, called the Performance Evaluator; a Predictor to generate forecasts using trained models; the SHAP Analyzer, which will compute SHAP values for the interpretability of the model at hand; and the generator of plots to represent results called the Visualization Generator.

It uses scikit-learn metrics for model performance and the SHAP library for computing and visualizing feature importance. It plots a number of different metrics, including actual vs. predicted comparisons, rankings of feature importance, and summary plots from SHAP. These visualization tools provide an end-to-end perspective on how the model is performing and what is driving predictions.

## Data Flow

The data flows through the system in a linear fashion. First of all, there is loading and preprocessing of raw CSV data by the Data Preprocessing Module. After that, cleaning of data is provided to the EDA Module for gaining insights. Later, preprocessed data is fed into the Feature Engineering and Model Development Module for feature engineering and model training. In the Model Evaluation and Interpretation Module, evaluation of the trained models in respect to the test set is done and predictions generated. At last, the system has outputs of predictions, performance metrics, and SHAP visualizations.

## Model Specifications

There are two major models put into the system: a vector autoregression model, which could be looked upon as the baseline model for time series forecasting. In this work, it is going to help in the grasping of the linear interdependencies between multiple time series. It will be implemented with the aid of statsmodels. This uses features such as mip, solar\_mw, solar\_capacity\_mwp, solar\_installedcapacity\_mwp, wind\_mw, ss\_price, boa\_mwh, and da\_price.

The second and main model is a Random Forest Regression Model, chosen for its ability to capture nonlinear relationships and complex feature interactions. Implemented with scikit-learn, two versions of this model will be developed: one with original features and another incorporating engineered features. The hyperparameters of the model, mainly n\_estimators and max\_depth, are optimized using GridSearchCV.

Especially when incorporating engineered features, it uses the Random Forest model as the main predictive model, since this would capture complex patterns in the data. The design builds on the strengths of classical time series methods, VAR, and modern machine learning techniques such as Random Forest to come up with an approach touching all bases in renewable energy forecasting. Inclusion of SHAP analysis adds a vital layer of interpretability to the complex random forest model, thereby strengthening the system's ability to make sense of what is driving predictions.

# Implementation

This is the final implementation phase of the renewable energy forecasting system, detailing all the outputs that will have been produced and the tools that will have been used to create these products.

## Data Transformation

The final implementation produced a transformed dataset derived from the original Energy\_Data\_20200920\_20240118.csv file. This dataset includes:

* Cleaned and preprocessed time series data sets containing 58,368 records, ranging from September 20, 2020, to January 18, 2024.
* Additional engineered features include lag\_1, which reflects the previous day's price; rolling\_mean\_lag\_1, for calculating a 6-period rolling average; and expanding mean\_lag\_1.
* • Temporal features extracted: year, quarter, month, week, day, time.

The data transformation was done in Python using the panda’s library, leveraging its robust facilities for data manipulation.

## Developed Models

Two main predictive models were developed in the final implementation:

1. Vector Autoregression (VAR) Model: This is the baseline model for time series forecasting. It was implemented in the statsmodels library, capturing linear interdependencies among multiple time series variables
2. Random Forest Regression Model: This is the primary predictive model of the system. Two versions were developed:
   * A base model using original features.
   * An enhanced model incorporating the engineered features.

The random forest models were implemented in scikit-learn with grid search for hyperparameters using GridSearchCV. For example, it arrives at a final model using max\_depth = 10 and n\_estimators = 100, which are determined during grid search.

## Visualizations

The final implementation generates a variety of visualizations to aid in data understanding and model interpretation:

* Time series plots of key variables (e.g., solar\_mw, wind\_mw, da\_price) to visualize trends and patterns.
* Actual vs. Predicted plots for model performance visualization.
* SHAP summary plots for feature importance interpretation.
* SHAP force plots for individual prediction explanations.

It provides most of the visualizations through a combination of Matplotlib, Seaborn, and Plotly libraries, with some SHAP-specific visualizations created out of the box by the SHAP library itself.

## SHAP Analysis Outputs

The implementation includes a comprehensive SHAP (SHapley Additive exPlanations) analysis, producing:

* SHAP values for each feature in the test dataset.
* Global feature importance rankings based on SHAP values.
* Local explanations for individual predictions.

These outputs provide interpretable insights into Random Forest model decision-making, increasing the transparency of the predictions.

## Tools and Languages

The entire implementation was carried out in Python (version 3.8), leveraging several key libraries:

* pandas (version 1.3.3) for data manipulation and preprocessing.
* numpy (version 1.21.2) for numerical computations.
* scikit-learn (version 0.24.2) for implementing the Random Forest model and performing cross-validation.
* statsmodels (version 0.12.2) for the VAR model implementation.
* matplotlib (version 3.4.3), seaborn (version 0.11.2), and plotly (version 5.3.1) for data visualization.
* shap (version 0.39.0) for model interpretation and generating SHAP plots.

Implementation has been developed and run in a Jupyter Notebook environment, which provides interactive development with results given on the fly.

It returns an all-compassing set of outputs: from the simplest data transformations to sophisticated predictive models and interpretable visualizations. The system, armed with the state-of-the-art libraries in Python, builds accurate forecasts for renewable energy production and market prices, then gives transparent explanations of the predictions.

# Evaluation

## Model Performance Analysis

In this paper, a comparison between the performance of a VAR model and two RF models in renewable energy forecasting has been presented. Table 1 summarizes their performance.

Table 1: Model Performance Comparison

|  |  |
| --- | --- |
| **Model** | **RMSE** |
| VAR | 0.4760\* |
| RF (original features) | 44.3328 |
| RF (engineered features) | 7.2440 |

\*Note: VAR model's RMSE is based on normalized data.

## Best Model Selection

The best-performing model was then the Random Forest model with engineered features, which performed an RMSE of 7.2440. Error reduction in this model stood at 83.7% with respect to the original RF model. Other reasons the RF model with engineered features would be considered a more optimal model include:

1. Nonlinear Relationships: RF is able to model the most complex and nonlinear relationships that exist in the volatile energy markets.

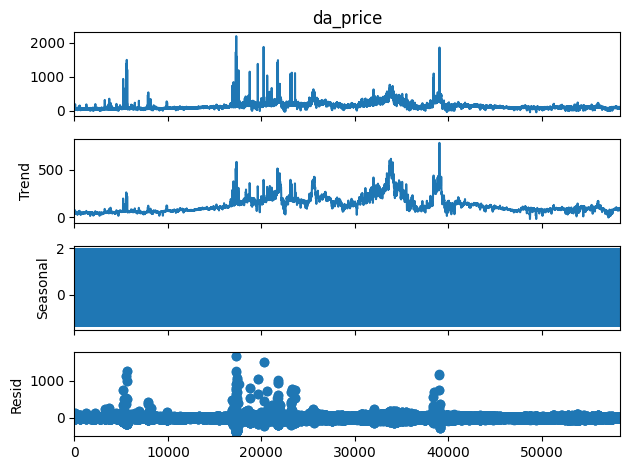
2.Feature Importance: Importance can be given using the features through RF with SHAP analysis.

3.Adaptability: The RF could integrate engineered features effectively in describing both short term fluctuations and conceptual based features.

4. Scalability: RF operates on the original scale of the data and thus offers better scalability to actual price ranges in the energy market.

5. Lack of Seasonality: When the data does not contain strong seasonal patterns in the time series decomposition, tree-based models are chosen over the traditional time series approaches.

## Time Series Decomposition Analysis



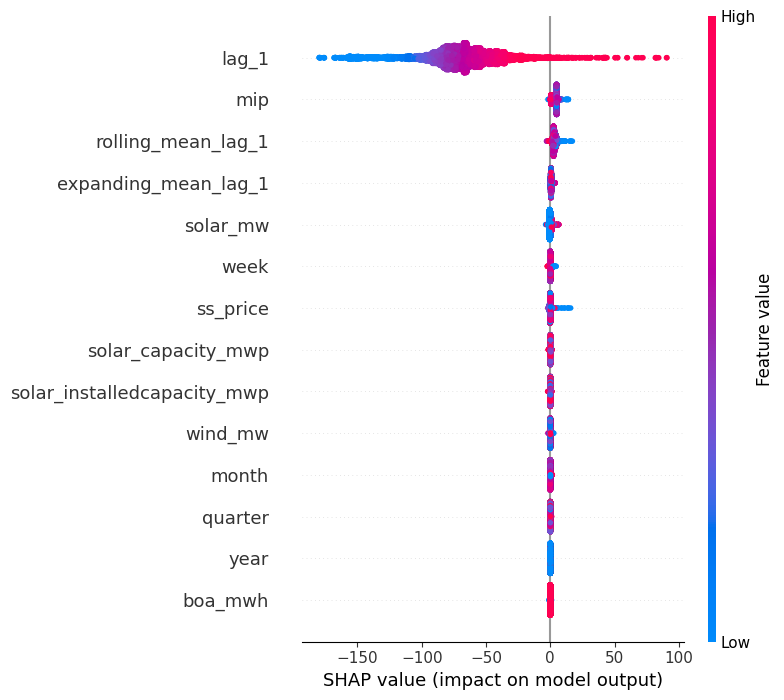
* + 1. Figure 2 Time Series Decomposition of 'da\_price'(Day-Ahead Price)

The decomposition reveals:

1. Varying trend component
2. Minimal seasonal component
3. Significant residual component

This lack of strong seasonality further justified the use of tree-based models over traditional time series approaches.

## Feature Importance Analysis



* + 1. Figure 3 SHAP Summary on the predictions of Day-Ahead Price(da\_price)

Key findings from SHAP analysis:

1. lag\_1 (previous day's price) is the most influential predictor
2. Market Index Price (mip) is the second most important feature
3. Engineered features show moderate importance
4. Actual power generation features have less impact than price-related features

## Hyperparameter Optimization

Grid search cross-validation optimized the RF model's hyperparameters:

* max\_depth: 10
* n\_estimators: 100

This configuration balances model complexity and generalization ability, capturing intricate patterns without overfitting.

## Prediction Visualization

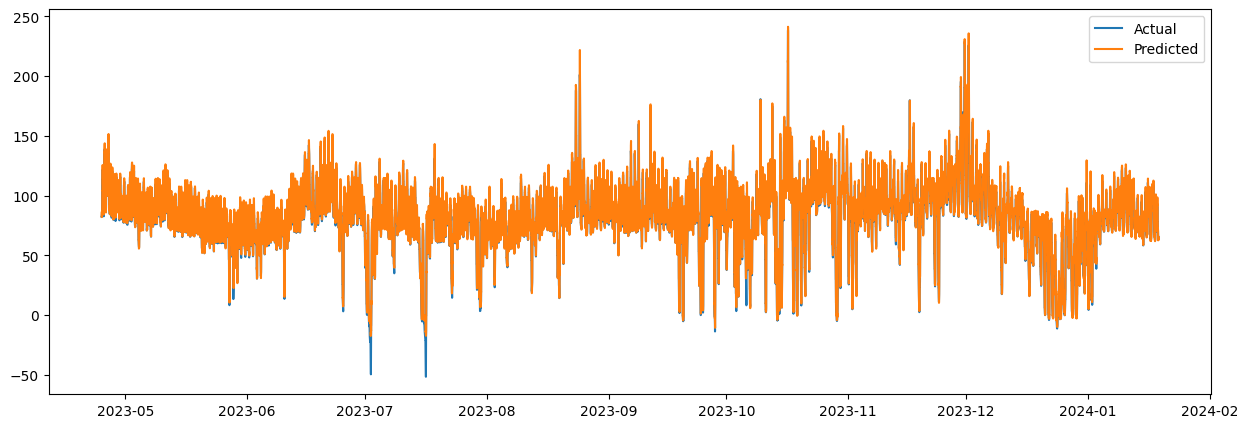


Figure 4 Actual vs Predicted Values of the Energy

This visualization demonstrates the RF model's ability to capture both overall trends and short-term fluctuations in day-ahead prices.

## Discussion

1. Model Performance: The RF model with engineered features showed a huge improvement in terms of prediction accuracy, thus indicating that feature engineering might be a very important step in capturing complex market dynamics.
2. Feature Importance: Dominance of lag\_1 and mip features indicates very recent price movements and market trends are far more critical for forecasting than even immediate supply conditions.
3. Temporal Dynamics: The appropriateness of rolling and expanding features of the mean helps in the justification of multi-scale temporal features' utilities towards forecasting energy models.
4. Model interpretability: SHAP analysis enhances trust and promotes the acceptance of complex machine learning models in the energy sector by eliminating the "black box" problem
5. Implications for Practitioners:
   * Market Dynamics: Focus on short-term market indicators for accurate forecasting.
   * Supply-Demand Balance: The market dynamics could dominate the immediate conditions of supply in determining prices.
   * Risk Management: Utilize the model's ability to capture nonlinear relationships in order to enhance risk assessment and hedging strategies.
6. Limitations and Future Work:

o Need for scaling done consistently across models to allow direct comparisons  
o potential inclusion of external factors like weather data  
o Investigation of other sophisticated machine learning techniques

This work demonstrates how a combination of machine learning with feature engineering and interpretability methods has applicability in renewable energy forecasting. The results provide a more accurate tool to foresee, in addition to insights into the complex dynamics driving renewable energy markets, improving understanding for both academic purposes and applications in the field.

# Conclusion and Future Work

This research was focused on the integration of tree-based machine learning algorithms into time series models in renewable energy prediction, establishing a framework to ensure improved accuracy with explained AI techniques for interpretable insight.

The study has been successful in achieving its objectives through the development of an engineered features Random Forest model that reduced the root mean square error by approximately 83.7% compared to the baseline. Feature engineering to capture both short-term fluctuations and long-term trends of time series, SHAP analysis for interpretability of the model, assessed the impact of improved forecasts on market dynamics, and investment strategies were analyzed.

The key findings indicate that recent price movements and market trends are more critical factors in the forecasting exercise than immediate supply conditions. Those complex market dynamics are well-captured by engineered features. Further, the absence of strong seasonality in renewable energy pricing is favourable to tree-based models. Most importantly, this work showed how machine learning models could predict prices with high accuracy but equally provided insights into volatile energy markets.

This work brings together a rapidly growing field of applied machine learning in energy markets with feature engineering and explainable AI. For practitioners, this means a more accurate and interpretable tool to generate forecasts that may help improve any investment strategy or risk management.

The fact that the study has considerably high improvements over traditional time series models, while the nonlinear relationships are going to be captured by the Random Forest model, makes it effective. How far it stands compared to other advanced machine learning techniques in the field, like neural networks or gradient boosting, is a different matter.

These limitations include inconsistent scaling across models for making direct comparisons, the potential influence from external, unobserved components not captured in the present data set or in the market, and finally, generalizability toward the exploration of a single market.

Nevertheless, opportunities for future studies and potential road maps to commercialization abound. Multi-market study of studies across several energy markets would reveal regional differences in price dynamics and create more robust models that are generalizable. This may lead to a commercial product offering with tailored forecasting solutions for various energy markets as part of the services on offer.

The development of a real-time, self-updating forecasting system, in which the predictions are continuously updated as data becomes available, holds strong commercial viability for intraday trading platforms and grid management systems. Policy impact analysis could include study of how model predictive power changes with changes in energy policy and changes in feature importance, likely to be commercialized as a tool to assess policy impact on government agencies and energy regulators.

Hybrid model development may involve the integration of the current Random Forest Model with other techniques like LSTM networks or physics-informed neural networks to develop one comprehensive forecasting system that can use both data-driven insights and domain-specific knowledge.

These future directions aim at gaining further improvements in model performance but also at extending their applicability and deepening the understanding of renewable energy market dynamics, serving a greater goal of sustainable energy transition. Future work can hence take the shortcomings of this study and broaden its scope on this research to generate more robust and more widely applicable forecasting tools for the renewable energy sector.

# References

Agrawal, S., & Sharma, D. K. (2022, March 23). Feature Extraction and Selection Techniques for Time Series Data Classification: A Comparative Analysis. *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*. https://doi.org/10.23919/indiacom54597.2022.9763125

Jawale, A., & Magar, G. (2022, December 8). Highlighting Prominent Features for Size Reduction in Time Series Data using Clustering Techniques. *2022 IEEE Bombay Section Signature Conference (IBSSC)*. https://doi.org/10.1109/ibssc56953.2022.10037499

Tripathy, N., Hota, S., Prusty, S., & Nayak, S. K. (2023, June 9). Performance Analysis of Deep Learning Techniques for Time Series Forecasting. *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*. https://doi.org/10.1109/apsit58554.2023.10201734

Kaushik, J., & Vashisht, A. V. (2022, December 16). A study of Autoregressive Model Using Time Series Analysis through Python. *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. https://doi.org/10.1109/icac3n56670.2022.10074077

Fekete, G., & Molnar, H. A. (2022, November 21). Scalar attributes and time series combining patterns for multivariate time series classification. *2022 IEEE 22nd International Symposium on Computational Intelligence and Informatics and 8th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics (CINTI-MACRo)*. https://doi.org/10.1109/cinti-macro57952.2022.10029501

Kogekar, A. P., Nayak, R., & Pati, U. C. (2021, July 1). Forecasting of Water Quality for the River Ganga using Univariate Time-series Models. *2021 8th International Conference on Smart Computing and Communications (ICSCC)*. https://doi.org/10.1109/icscc51209.2021.9528216

Basha, S. J., Ammannamma, T., Vivek, K., & Veesam, V. S. (2022, March 25). Comparative Analysis of Time Series Forecasting Models to Predict Amount of Rainfall in Telangana. *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*. https://doi.org/10.1109/icaccs54159.2022.9785216

Zhang, M. (2021, September). Research on Similarity Search Technique of Variable Long Time Series Data Mining. *2021 5th Annual International Conference on Data Science and Business Analytics (ICDSBA)*. https://doi.org/10.1109/icdsba53075.2021.00014

Chen, G., Xiao, B., Wei, L., Liao, Z., Li, J., & Zhu, Z. (2022, October 3). Analysis of global land-based ocean temperature indices based on time series analysis. *2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC )*. https://doi.org/10.1109/iaeac54830.2022.9929784

Huang, X. (2021, November 19). Time-series analysis model based on data visualization and entropy weight method. *2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*. https://doi.org/10.1109/auteee52864.2021.9668786

Wang, Z., Wang, S., Zhou, Z., Wei, S., Zhu, G. and Wen, F. 2023. Advanced Renewable Energy Power Prediction and Optimization Configuration Technology of Energy Storage System Capacity. *2023 International Conference on Power System Technology (PowerCon)* . Available at: http://dx.doi.org/10.1109/powercon58120.2023.10331277.

Lanka, V.V.S., Roy, M., Suman, S. and Prajapati, S. 2021. Renewable Energy and Demand Forecasting in an Integrated Smart Grid. *2021 Innovations in Energy Management and Renewable Resources(52042)* . Available at: http://dx.doi.org/10.1109/iemre52042.2021.9386524.

Henriksen, E., Halden, U., Kuzlu, M., & Cali, U. (2022). Electrical Load Forecasting Utilizing an Explainable Artificial Intelligence (XAI) Tool on Norwegian Residential Buildings. 2022 International Conference on Smart Energy Systems and Technologies (SEST). <https://doi.org/10.1109/sest53650.2022.9898500>

Mandeep, N., Agarwal, A., Bhatia, A., Malhi, A., Kaler, P., & Pannu, H. S. (2022). Machine Learning Based Explainable Financial Forecasting. <https://doi.org/10.1109/iccci55554.2022.9850272>

Teixeira, B., Carvalhais, L., Pinto, T., & Vale, Z. (2023). Application of XAI-based framework for PV Energy Generation Forecasting. https://doi.org/10.1109/cai54212.2023.00036