BEYKOZ ÜNİVERSİTESİ

GRADUATE PROGRAMS INSTITUTE

Optimized Random Forest Model For Prediction Solar Irradiance and Photovoltaic Total Power Energy by Using JAYA Algorithm

MARIAM KENAWY

MASTER'S THESIS

Istanbul, July 2025

Optimized Random Forest Model For Prediction Solar Irradiance and Photovoltaic Total Power Energy by Using JAYA Algorithm

MARIAM KENAWY

MASTER'S THESIS

ARTIFICIAL INTELLIGENCE DEPARTMENT
COMPUTER ENGINEERING MASTER'S PROGRAM WITH THESIS

MASTER'S THESIS ADVISOR
Assist.Prof. Dr. İnal Begüm TURNA DEMIREL

MASTER'S THESIS JURY MEMBERS
Assist.Prof. Dr.Gizem TEMELCAN RGENECOŞAR

Assist.Prof. Dr.Engin HENGİRMEN

BEYKOZ ÜNİVERSİTESİ

T.C.
BEYKOZ ÜNİVERSİTESİ
LİSANSÜSTÜ PROGRAMLAR ENSTİTÜSÜ MÜDÜRLÜĞÜ

01/08/2025

Yüksek Lisans Tez Onay Belgesi

Enstitümüz Bilgisayar Mühendisliği Anabilim Dalı Yapay Zeka Tezli Yüksek Lisans Programı 2330200003 numaralı öğrencisi Mariam Ibraheem Elsayed Hassaan Kenawy'nin "Optimized Random Forest model for prediction solar Irradiance and photovoltaic total power energy by using JAYA Algorithm" adlı tez çalışması Enstitümüz Yönetim Kurulunun ../../…. tarih ve 20../.. sayılı kararıyla oluşturulan jüri tarafından oybirliği ile Tezli Yüksek Lisans tezi olarak …kabul…edilmiştir.

Öğretim Üyesi Adı Soyadı	**İmzası**

Tez Savunma Tarihi :01/08/2025

1)Tez Danışmanı: Dr. Öğr. Üyesi İnal Begüm TURNA DEMİREL

2) Jüri Üyesi : Dr. Öğr. Üyesi Gizem TEMELCAN ERGENECOŞAR

3) Jüri Üyesi : Dr. Öğr. Üyesi Engin HENGİRMEN

Not: Öğrencinin Tez savunmasında **Başarılı** olması halinde bu form **imzalanacaktır**. Aksi halde geçersizdir.

BEYKOZ ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

JAYA ALGORITMASI KULLANILARAK GÜNEŞ IŞINIMI VE FOTOVOLTAIK TOPLAM GÜÇ ENERJISININ TAHMINI IÇIN OPTIMIZE EDILMIŞ RANDOM FOREST MODELI

YÜKSEK LİSANS TEZİ

MARIAM KENAWY
2330200003

Yapay Zeka Bölümü

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Assist.Prof. Dr.İnal Begüm TURNA DEMIREL

TEMMUZ 2025

Mariam Kenawy, a M.Sc. student of Beykoz University, student ID **2330200003**, successfully defended the thesis entitled "Optimized Random Forest model for prediction of solar Irradiance and Photovoltaic total power energy by using JAYA Algorithm
", which she prepared after fulfilling the requirements specified in the associated legislation, before the jury whose signatures are below.

Thesis Advisor :Assist.Prof. Dr.İnal Begüm TURNA DEMIREL .....
Beykoz University

Jury Members :Assist.Prof. Dr.Gizem Temelcan Ergenecoşar .....

Beykoz University

Jury Members :Assist.Prof. Dr.Engin HENGİRMEN……..

Istinye University

Date of Submission: 04 July 2024
Date of Defense: 1 Aug 2025

FOREWORD

I would like to express my sincere gratitude to my unique advisor, DR/ INAL BEGÜM for supporting and making all difficulties easy. My thanks also for the jury members for giving time and instructions for enhancing my work.

This thesis is dedicated to the memory of my beloved father, whose values continue inspires me.

I want to extend my thanks to my family for supporting and believing in me always until I believed I can do it, and I did.

MARIAM KENAWY

July 2025

TABLE OF CONTENTS

Page

ABBREVIATIONS

AI : Artificial Intelligence

h : Hour angle

PV : Photovoltaic

RF : Random Forest

GHI : Global Horizontal Irradiance

BHI : Beam Horizontal Irradiance

SYMBOLS

θ theta : Zenith Angle

ϕ phi : Latitude

δ delta : Solar Declination Angle

← : Assignment

LIST OF TABLES

Page

LIST OF FIGURES

Page

Optimized Random Forest model for prediction solar Irradiance and Photovoltaic total power energy by using JAYA Algorithm

SUMMARY

This research deals with the solar photovoltaic power generation prediction through a machine learning–based architecture combining the Random Forest algorithm with the JAYA metaheuristic optimization. As the demand for renewable energy has been increasing, the establishment of correct and expeditive forecasting models is vital for the system operators, along with the energy managers, for effective planning, efficacy, and assured cost-effectiveness.

The data, obtained for Chapel Lane substation in the United Kingdom (2019–2022) at 15-minute integration, went through the pre-processing, such as feature creation and normalizations. Random Forest was utilized for the extraction of the most pertinent parameters for the requirement of prediction, whereas the JAYA algorithm has been used for hyperparameter optimization for the better performance of the designed models.

The performance accuracy of the models has been observed through established performance indicators, i.e., the coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE). Through the results, the intention is demonstrated by projecting the predicted respective outputs with fewer faults, indicating its potential for solar energy prognostication along with the power management activities.

Optimized Random Forest model for prediction solar Irradiance and Photovoltaic total power energy by using JAYA Algorithm

ÖZET

Bu çalışma, Random Forest Algoritması ve JAYA metasezgisel optimizasyon algoritmasını birleştiren makine öğrenimi tabanlı bir çerçeve uygulayarak güneş enerjisi fotovoltaik veri kümelerini tahmin etmeye odaklanmaktadır. Yenilenebilir enerjinin artan önemi nedeniyle, üretilen güçleri en iyi sonuçlarla tahmin edebilecek etkili bir model çıkarmaya çalışacağız. Güneş ve fotovoltaik enerjinin doğru tahmini, sistem sahipleri ve enerji yöneticileri için planlama, verimlilik ve maliyet etkinliğini iyileştirmek açısından önemlidir. Verilerimizin makine öğrenimi ve analiz çalışmaları için hazır olduğundan emin olmak için bazı ön işleme adımları uygulayacağız. Ardından, Random Forest makine öğrenimi algoritmasını kullanarak tahmin modeli için önemli parametreleri belirleyeceğiz. Ardından, ulaşabileceğimiz en iyi çözümlere ulaşmak için JAYA metasezgisel algoritmasını kullanarak optimizasyon modeli için uygun parametreleri belirleme adımını tekrarlayacağız.

Bu çalışma ayrıca, bu sürecin kalite oranı için ölçüt olarak R2, MSE ve MAE gibi belirli değerlendirme metriklerini alarak daha az hata içeren tahminlerin pratik etkisini vurgulamaktadır. Modeli iyileştirmeye çalışırken, bu metrikler bize modelimizin ne kadar iyi çalıştığına dair bir gösterge sunmaktadır.

1. INTRODUCTION

In this study, the Random Forest machine learning algorithm is applied for the forecast of the key solar energy parameters, Beam Horizontal Irradiance (BHI), and overall power yield by the photovoltaics. Random Forest provides an optimal baseline for forecasting due to its ability to model non-linear relations found in the data. Aside from the improvement in the predictive capability, the JAYA metaheuristic search library was applied for parameter optimization, for the extraction of optimal solutions, to further enhance the reduction in the prediction errors.

The proposed methodology has valuable practical applications for the stakeholders within the renewable power industry, with the accurate forecast enabling better operational scheduling, durability for the systems, as well as the successful utilization of solar energy sources.

the data sets were recorded from a substation located in the United Kingdom called Chapel Lane (2019-2022), with intervals of 15 minutes. Preprocessing steps were applied, such as feature engineering, normalization, and performance measures of the model (MAE, MSE, R^2) were utilized to compare performance. Random Forest models provided good predictions with satisfactory accuracy. To further improve the performance, the metaheuristic JAYA algorithm hyperparameter optimization was used to justify that aim. Some preprocessing techniques were carried out to improve the data quality and to ensure the model's performance reliability.

Model accuracy is measured depending on some statistical indicators such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2).

Integration of JAYA optimization algorithm and Random Forest not just to enhance the accuracy of the model, but also demonstrates the value of hybrid approaches in the renewable Energy forecasting sector.

1.1 Literature Review

The increased global interest in clean energies has propelled solar photovoltaics (PV)systems to be among the most promising prospecting sources of energy. Coupled with the rapid decrease in costs associated with solar energy, the IEA has also identified solar PV as one of the fastest growing technologies in the world, due to the slope PV systems' versatility, spanning from single rooftop arrangements to expansive solar farms spanning several megawatts. Nonetheless, these promising prospects of solar energy are tempered by several concerns, including the intermittent solar radiation, the dependence of strong solar radiation on favorable weather conditions, and the need for accurate predictions of the grid to determine its stable operations and reliable energy planning.

Research pertaining to the prediction of solar and PV power has traditionally relied on the use of statistical models. While valuable insights were derived from these types of models, they, unfortunately, are not useful in capturing the negative, dynamic behavior of solar radiation under varying conditions. Recent studies in this field have turned to neoteric methods, including advanced machine learning and deep learning techniques, among which were the earlier studies by Küçüktopçu et al. (2024), who applied the hybrid ML-WT approach and found the models to be inordinately sensitive to the mean squared error (MSE) of their parameters.

This study addresses those gaps with the implementation of a supervised machine learning approach where the Random Forest (RF) model's regressive component is paired with the JAYA metaheuristic optimizing algorithm. JAYA is less computationally intensive. RF is a strong baseline for non-linear datasets. JAYA is less computationally intensive than other optimization methods, while RF is a strong baseline for non-linear datasets. JAYA is less complex. This dual-application strategy solves the main issues previous works have had, dramatically lessening computational strain while also minimizing the risk of overfitting, while still ensuring reasonable predictions for both components of solar radiation, Beam Horizontal Irradiance (BHI), and the total power output of the PV system. Here is a table showing a comparison between some previous studies Vs this study.

Table 1.1 : Comparison of the Last Studies and This Study

Author & Study	**Algorithm(s):**	**Accuracy Metrics**	**Key Findings**
E. Küçüktopçu et al., 2024	Hybrid ML-WT techniques	The MSE (estimated from RMSE ≈ 2.17–2.436 MJ/m²) is approximately 4.7 to 5.94. After ML-WT MSE≈3.28–4.28	Performance heavily relies on hyperparameter tuning, which is not recommended for computational
Z. Hou et al., (2024)	LSTM, WOA-LSTM, VMD-LSTM, WOA-VMD-LSTM	Before (LSTM): MSE ≈ 22,124.4 (kW)² After MSE ≈ 390.2 (kW)²	High computational complexity, models like LSTM and WOA can be complex and time-consuming
This Thesis	Random Forest + JAYA Algorithm	RF solar MSE=6.44 JAYA solar MSE=0.28 RF pv MSE=1.43 JAYA pv MSE=0.93	Find optimal solutions without complex parameter tuning

The results in this case study also validate the proposed framework. For instance, although the base RF model's results yielded a solar MSE of 6.44 and a PV MSE of 1.43, JAYA optimization brought the solar MSE down to 0.28 while the PV MSE dropped to 0.93. This RF–JAYA framework achieves optimal results with the lowest complexity in comparison to the other methods available, making it a less computationally intensive option for solar and PV forecasting.

1.2 Hypothesis

The random forest prediction algorithm is designed to yield accurate prediction results, making it a valuable forecasting Machine learning algorithm.

The random forest demonstrates interpretability with respect to historical data for solar and photovoltaic data provided in that study.

The results of the random forest are good, but the study aims to further increase accuracy by optimizing the outputs of the random forest, as it will never be totally accurate.

It's hypothesized that optimization with JAYA will further improve the predictive efficiency of the Random Forest.

Choosing the JAYA algorithm was a good choice, as it can easily capture the behavior of both solar and photovoltaic data sets.

The structure of this study aims to find the optimal solution for the predicted elements by using MSE as an indicator for the optimization process, as the target is

to reduce the errors of the predictions and optimize the predictions, according to the principle of JAYA is to move faster forward the optimal solution and keep being away from the worst solution JAYA also gives us the right to detect the termination point in case we reached the targeted point we can stop the algorithm from continuing the process to save the time and prevent the over iterations.

1.3 Objectives

The main objective of this study is to develop a machine learning framework that can forecast solar radiation and photovoltaic (PV) power output accurately.

This framework will provide practical support for Farmers, system designers, and energy stakeholders. The particular goals are: To enhance model learning through pre-processing and feature engineering from electrical (voltage, current, harmonics) and meteorological (irradiance) datasets.

To train a Random Forest regression model to predict total PV power, beam horizontal irradiance (BHI), and global horizontal irradiance (GHI). For a more trusted evaluation, model performance will be evaluated by using metrics like the coefficient of determination (R2), mean squared error (MSE), and mean absolute error (MAE).To use the JAYA optimization algorithm to automatically adjust important Random Forest hyperparameters, improving the accuracy level and avoiding errors.

To guide stakeholders in sensor placement and system monitoring design by using importance analysis to identify key predictive features.

To provide stakeholders and energy managers with actionable insights that will allow: to improve solar energy availability forecasting, more intelligent energy storage and dispatch, and better off-grid and grid-tied photovoltaic system planning.

1.4 Scope and Limitations

The main focus of this study is to develop a predictive model for solar radiation and Photovoltaic power generation depending on machine learning mainly the Random Forest algorithm the study further combining JAYA optimization algorithm to enhance model performance by tuning hyperparameters, to achieve more accurate and reliable predictions, the used data sets contains historical metreological data including solar

radiations and weather variables such as (temperatures, wind speeds), are used for training and evaluating the model, The scope also embodies exploring the practical applications of these predictions for managing energy, in agricultural systems and also for hybrid solar energy systems by providing insights for the planning and the optimization of energy resources.

While the study emphasizes model development and performance evaluation, it is chiefly grounded in standard error metrics such as Mean Squared Error (MSE), Coefficient of determination(R2), and Root Mean Squared Error (RMSE) to evaluate model accuracy, without employing enhanced statistical validation techniques at this stage.
this study has range of limitations should be mentioned the model is evaluated by using single train test split but advanced statistical validation methods such as k-fold cross-validation or bootstrapping were not included in this study, integrating these techniques in the future work will add more systematic assessments and external validity for the model, some missing influential features for the Photovoltaic data set such as shading and panel degradation.
Finally, the Random Forest-JAYA hybrid model is a good step towards improving model performance and accuracy within the domain of Solar and PV energy; still, statistical validation methods are important as they give more validity to the model.

1.5 Traditional Statistical Forecasting Methods

Traditional statistical forecasting methods, such as Simple Exponential Smoothing (SES), Holt, Theta, and Error-Trend-Seasonal (ETS) models, as mentioned in the study by Grebovic et al. (2022), focus on linear relationships, limited or specific predictors, and fixed forms of data (e.g., trends or seasonality). Key pros of traditional statistical forecasting methods are: they are simple, are intuitively interpretable, have a long-established mathematical theory behind them, work well for linear and well-behaved data, and have low computability; while their fundamental limitations are: have limited ability to correctly model non-linear relationships, limited to linearity along with specific patterns in the data, have no or negligible ability to work well with complex and/or higher dimension to the data, and have normative behaviours in the presence of fixed patterns, and/or types of data structures. Neural networks (NNs), as Multilayer Perceptron (MLP), Bayesian Neural Networks (BNN), Generalized Regression Neural Networks (GRNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM), are unique.

The challenges can be effectively resolved by integrating Random Forest and introducing JAYA as the optimization algorithm. Random Forest is an ensemble learning technique that can capture complex relationships containing non-linear relationships between solar radiation conditions and PV power output, allowing for a stronger ability to handle high-dimensional datasets containing many predictors versus standard statistical models. The application of randomly drawn bootstrap samples and aggregates reduces the occurrences of overfitting while achieving stable and robust levels of prediction under various weather conditions. Coupled with hyperparameter optimization, the JAYA metaheuristic algorithm automatically changes parameters while leading to the best parameters corresponding to improved predictive accuracy without any trial-and-error. The adaptation offered by this originating work routines a flexible approach and captures patterns from data without strict statistical assumptions; maintains a balance of minimal realism with regards to robustness, computational cost, and accuracy, offering a reliable alternative to off-the-shelf statistical models such as regression and more empirically complicated neural network design alternatives for predicting solar and PV output.

2. Data Set

2.1 Solar Energy Data Set

The dataset used for the purpose of this study is a time-series measurement of solar irradiance and important environmental variables, which can be used to simulate and infer solar energy generation. This dataset was recorded at the Chapel Lane Substation in the UK(2019-2022) with latitude=53.323, and longitude=-1.912) at 15-minute intervals. The reference clock reading for this dataset is True Solar Time (TST).

Figure(2) Location map of the used data set

The dataset contains variables Global Horizontal Irradiance (GHI), which is the total solar radiation incident upon a horizontal surface (W/m²), Beam Horizontal Irradiance (BHI), which is the direct component of solar radiation, ambient temperature, which directly affects the efficiency of PV panels, and wind speed, which affects thermal and operational output of the system.

This study considers the approach of estimating GHI (which is the most important indicator of available solar energy) from predicted BHI estimates using statistical and mathematical modelling techniques. This approach allows for a more accurate understanding of solar potential over various environmental variable ranges.

The Estimation of GHI is built on equation (2.1)

$$GHI=DHI+BHI\cdot Cos(\theta) \tag{2.1}$$

As shown in equation (2.1), calculating GHI (Global Horizontal Irradiance)is about adding DHI (Diffuse Horizontal Irradiance), which is a part of GHI coming from scattering of sunlight by molecules measured on a horizontal surface.

to the BHI (Beam Horizontal Irradiance) multiplied by the Cos(zenith angle).

The zenith angle is defined as the angle known between the sun ray and the vertical line at a determined area. It plays a vital role in detecting the solar panel efficiency and the energy output, as shown in Figure 2.1, created using Canva.

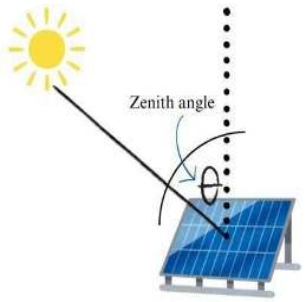


Figure 2.1. Zenith angle illustration.

in this study the Random Forest machine learning algorithm is employed to forecast the Beam Horizontal irradiance (BHI) and use the predicted values of (BHI) to calculate the Global horizontal Irradiance (GHI) as shown in the given equation, in case the GHI values are not provided directly by following this process, the GHI can be reached in another practical way, which provides the stakeholders with alternative solutions to help them, by having the zenith angle θ we can estimate the GHI.

2.1.1 Calculating Zenith angle

Estimating equation (2.1) requires some essential previous steps, one of which is calculating the zenith angle itself as follows: equation (2.1.1)

$$\text{Cos}(\theta) = \text{Sin}(\delta) \text{ x } \text{Sin}(\phi) + \text{Cos}(\phi) \text{ x } \text{Cos}(\delta) \text{ x } \text{Cos}(h) \qquad (2.1.1)$$

As shown in equation (2.1.1), calculating zenith angle(θ) depends on calculated values for other angles, such as Solar Declination angle and hour angle.

2.1.2 Calculating Solar Declination Angle

The solar declination angle (δ) is the angle of the line between the center of the Earth and the center of the Sun to the equator of the Earth. It is an important factor in describing how the Sun moves over the Earth during the year. The declination angle is positive when the Sun is above the equator (Northern Hemisphere) and negative when the Sun is below the equator (Southern Hemisphere) as shown in figure 2.2.

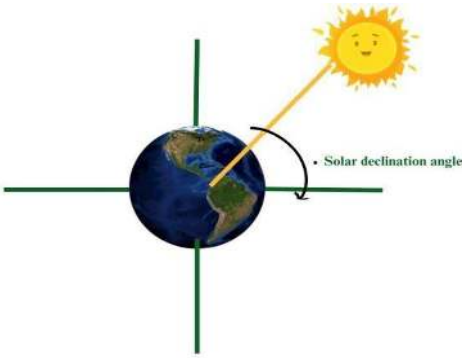


Figure 2.2 Solar Declination Angle illustration

Figure 2.2 created using Canva.

The declination angle has a value of approximately +23.5 degrees, and during the winter solstice, it is marked as -23.5 degrees

Accordingly, the solar declination angle varies from -23.5 < d < 23.5 degrees,as shown in equation(2.1.2).

$$\delta = 23.45° \times \sin\left(\left(\frac{360°}{365}\right) \times (n-81)\right) \qquad (2.1.2)$$

This equation includes more than one variable, such as Variable 1: 23.45°, which is the Maximum point of the sun`s Declination

Variable 2: 360°where 360° is the degree of a full circle of the sun, and 365 is the total number of days 365 in a year.

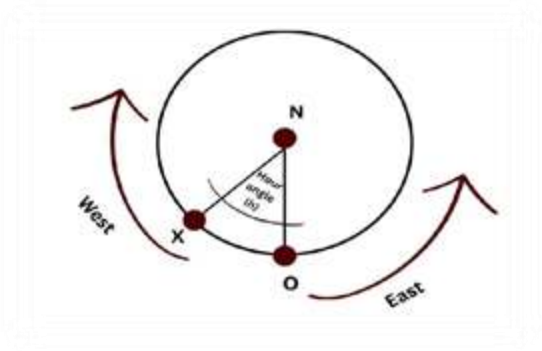
Variable 3: 81 is the estimated day of the year when spring occurs on March 21, where declination approximately =0°.

Variable 4: n, which is the day number in the year.

Collecting those values facilitates calculating the value of the declination angle.

2.1.3 Calculating the Hour Angle (h)

The hour angle (h) indicates the angular displacement between the observer's local meridian and the Meridian containing the Sun. Hour angle is defined as zero at solar noon, and thereafter increases by 15 degrees each hour away from solar noon. The hour angle is primarily used in solar positioning calculations, especially in direct calculations of the solar zenith angle and solar altitude, as shown in Figure 2.1.3.



Figure(2.3) The hour angle (h) illustration

The figure 2.3) created by using canva after searching and understanding the hour angle details.

issue is even more pronounced on mobile devices, where navigating multi-level filtering menus like category → brand → price → color is tedious. According to Nielsen Norman Group (2022), over 70% of mobile users abandon a website if the search experience does not provide relevant results quickly.

In addition, the hour angle, along with other parameters, can also help in estimating the irradiance at a given location and time, thereby also playing a vital role in estimating the actual output of operational PV (Photovoltaic) panels and reliability of solar energy estimates, as equation 2.1.3 shows

$$h = 15° \times (\text{Local solar time} - 12) \qquad (2.1.3)$$

Key limitations of traditional systems include: Vague keyword matching, Static filter

This equation includes: Variable 1: 15° is calculated from 360°/24=15°, which is the rotation of Earth during 24 hours, Variable 2: 12.00 is the solar noon time when the sun in highest point in the sky for that location.

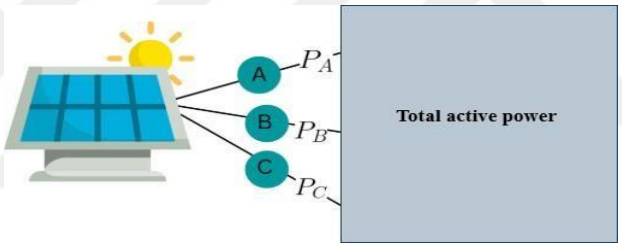
It's necessary to use latitude (φ) in decimals; in case the Northern Hemisphere, able to use positive values; if you are in the Southern Hemisphere, use negative values.

Correct latitude is crucial because it has a direct impact on how precisely it also can determine the Sun's position. This is particularly important when estimating solar radiation or forecasting how well a solar energy system will perform.

Additionally, use local solar time rather than the time displayed on a conventional clock when applying the hour angle formula. Both longitude and the equation of time. Even minor time errors can cause solar angle calculations to be off, but these adjustments can have a significant effect.

2.2 Photovoltaic Data Set

The PV data set obtained from Chapel Lane substation in the United Kingdom and contains several key features relevant to photovoltaic (PV) system performance, contains the PV total power in (KW) representing the total generated power also contains current and voltage across three phases(A, B and C) these metrics are necessary for overall evaluating the performance and efficiency of the PV system, making it suitable for Forecasting and machine learning models well.figure2.2 illustrates the power flow of the PV



Figure(2.4) The total photovoltaic power flow

Figure 2.4) created by using Canva after researching and understanding the flow Pv total power through a panel illustrates the flow of power from PV (photovoltaic) generation across the three electrical phases: Phase A, Phase B, and Phase C. Each line seems to denote a phase where the active power for the phase is recorded and then separately identified. Specifically, active power recorded on Line A denotes active power from Phase A, active power recorded on Line B denotes active power from Phase B, and active power recorded on Line C denotes active power from Phase C. The system adds these separate contributions together and, in doing so, creates an algebraic sum of the total active power output of the PV system.

A three-phase system is very important for providing consistent usable energy production since spreading the load over three lines stabilizes loads while minimizing losses. By measuring the output of each phase separately, issues can be identified, such as

imbalances in the overall system, different levels of inefficiencies, or distortions related to a phase that would distinguish the energy output of the PV system. Yet once individual measurements are summed, the total active power is a very reliable indication of the generating capacity of the PV system at any one instant. This same information is also particularly useful when forecasting, modeling, and machine learning utilize individual behavior of each phase and the aggregate system dynamics.

According to the equation(2.2) of the PV power total :

$$P_{total} = P_{A+}P_{B+}P_{C} \tag{2.2}$$

As the deduction of the shown equation is that the PV total power is the summation of the power of each phase along the three phases (A, B, and C), that is what the prediction algorithm will follow.

Moreover, the three-phase representation of PV power increases system monitoring and evaluation. In real PV installations, the analysis for any issues was phase-specific, such as partial shading with the PV or failure of the inverter, or, in some cases, load imbalance, which affects the total power. From an advancement perspective, performance forecasting could utilize phase-based power data because the representation can better model variation and nonlinear functions. This is critical to strengthening machine learning models and thus can adjust to solar irradiation, with corresponding system operating conditions.

Also, the three-phase aggregation representation aligns with the grid compliance policy, i.e., the majority of utility-scale PV is three-phase distribution. Comparing contributions from each phase will reflect the value of generation leadership, as well as contribute to grid reliability, power quality, and harmonics assessment. Notably, the equation for total PV power summation, once recognized in the context of data acquisition, is an effective connection between predictive modeling and systems optimization.

3. Methodology

This chapter outlines the methodology approach which taken in this study to address the core idea of the JAYA-Random Forest Hybrid Model.

3.1 Prediction Model

In this study, we chose the Random Forest model as our predictive model due to its suitability for the more complex nonlinear data to be used in photovoltaic (PV) power forecasting. Random Forest (RF) is an ensemble-learning algorithm that constructs multiple decision trees and takes the mean of them for a more plausible and accurate output. It is a bootstrapped aggregating method where each tree in the forest is fitted to a random subset of the data sampled with replacement. With this approach, you would expect variation from the algorithm to control variance, and it leverages multiple trees, meaning it doesn't just rely on a single decision tree for output.

A major advantage of Random Forest in comparison to previous ensemble models, such as Gradient Boosting and XGBoost, is its ability to generate interpretable measures of feature importance. Feature Importance enables us to isolate the most significant variables in any prediction, solar irradiance, temperature, relative humidity, or cloud cover, to name a few. The interpretability of results, in energy forecasting, can provide knowledge and understanding of the driving factors that result in variability of PV performance.

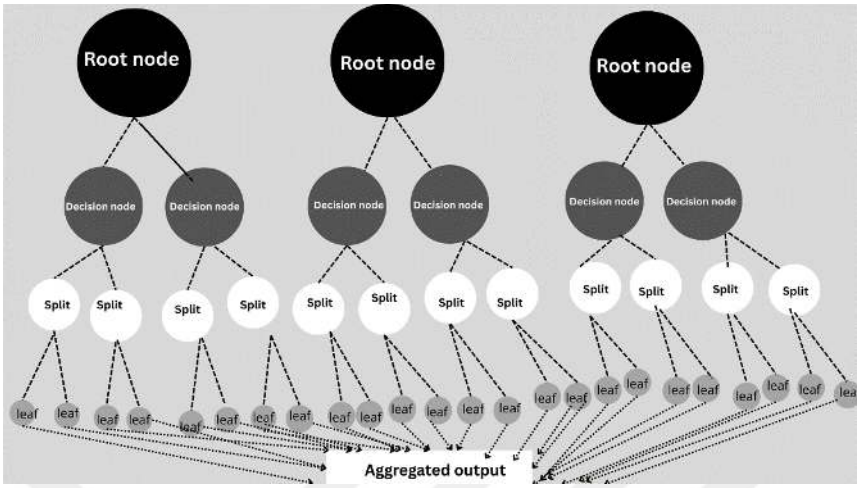
Additionally, Random Forest requires much less hyperparameter tuning than other ensemble methods and can also mitigate issues of overfitting, which is an issue for most machine learning models. These factors, in conjunction with the robustness of Random Forest and its being computationally efficient, validated its use as the predictive model for this analysis.

3.1.1 Random Forest

Random Forest Algorithm is an ensemble machine learning algorithm. The ensemble learning method used by the Random Forest allows for resilience and generalization compared to individual decision trees that would tend to overfit and be sensitive to noise in observations.

It consists of decision trees; each tree of the Random Forest collection includes a root node, internal decision nodes, intermediate splits, and terminal leaf nodes that provide

individual predictions. In the regression approach, we take the outputs of all trees by averaging, which tends to reduce variance and reduce overfitting, as Figure 3 shows



Figure(3) Structure Of Random Forest

The power of a Random Forest is in the ability to increase the overall performance of weak learners on the composite model. Each decision tree is constructed independently from random samples of the data and features, and the independence of the trees allows there to be variation between the trees. This is important because variation among the trees creates diversity, which reduces correlation between component trees, increasing both accuracy and stability of the aggregated output of Random Forest. Random Forest provides better bias-variance tradeoffs than individual decision trees and other models.

For the context of this study, Random Forest is appropriate for photovoltaic (PV) and solar forecasting applications. Solar energy state data is very susceptible to inherent variability operating from purely environmental impacts (irradiance, cloud cover, temperature, and wind speed). Variability, in this case, also indicates solar energy lack of a clear empirical relation, which effectively allows Random Forest to discover non-linear correlations and interactions uses many different variables and do so to reflect the complexity that exists for the interactions in solar power generation as much as possible; moreover Random Forest provides interpretable information on feature importance, which could demonstrate to researchers which environmental variables best explain energy output.

In addition to this, Random Forest is unique in being able to aggregate many predictions using multiple decision trees to maximize the predictive accuracy when using an accurate and timely computational model. The strength of the prediction outcomes gathered collectively does not purely provide predictions, but broader benefits of guarantees to

expectancy for when forecasting predictions are used in real-life terms. With all of these reasons, Random Forest is popular in the renewable energy literature and was an appropriate method to apply to the predictive tasks of this thesis.

3.1.2 Random Forest Pseudocode

The presented Pseudocode below explains the Random Forest algorithm workflow, which depends on building multiple decision trees and then combining their results to improve prediction accuracy, at the same time reducing possible overfitting.

Pseudocode of Random Forest

Precondition:
A training set S $(x1, y1)$, (xn, yn),
X_i : feature vector of the ith sample
Y_i : corresponding label
F: the set of features
T: number of trees to be built in the forest
Function of Random Forest (S, F)
(S, F) is responsible for building the entire ensemble of decision trees.
H ← ɵ
An empty set H was created to store separately all the decision trees
For i ϵ 1, T do
This step repeats the incoming steps T times to create T trees
$S^{(i)}$ ← A bootstrap sample from S
A bootstrap sample is a random sample with replacement taken from the S data set, which means part of data may appear more than time, and others may not appear at all
h_i ←Randomized Tree Learn ($S^{(i)}$,F)
By using randomized tree learn, the decision tree h_i is trained on bootstrap $S^{(i)}$
H ← H U {hi }
The trained h_i decision tree is saved in set **H**
end for
After looping through all T iterations, the T decision trees are already built
return H
Set H, which contains all decision trees, is returned while this set forms the Model of Random Forest
Function Randomized Tree Learn (S, F)
This function gives back a description of how each decision tree is created
At each node :
The following procedures will be applied to the following nodes
f ← **f** very small subset of **F**
feature bagging step which means instead of carrying out all features, a random subset **f** ⊂ **F** is selected
Split on the best feature on F
From the subset f choose the feature that gives the best split for the data
return the learned tree
after splitting nodes recursively by using best features the full decision tree will be returned .
Function is end .

This pseudocode captures two fundamental concepts driving Random Forest, one of them is Bootstrap Aggregation (Bagging): Each individual tree is built using a bootstrap sample, or

random resample of the data set, which allows trees to add variability to a model. Bagging reduces variance (and thus overfitting, which is often a problem with single decision trees), and the second is Random subset of features: At each split, a random subset of features is considered, instead of the full set of features. This adds diversity to each tree and also reduces the correlation between trees. Lower correlation increases the robustness of an ensemble prediction.

As a result, Random Forest is able to balance accuracy and generalization through two mechanisms: Bagging is used to reduce variance by averaging a set of diverse trees, and randomly selecting features evaluates the fit of the full set of features. When many predictors are strongly correlated, the individual tree does not dictate the final model. The resulting combination of properties builds a model that is not just stable, but also trustworthy and interpretable - especially when building models for high-dimensional or noisy data sets.

This research is focused on predicting solar irradiance and photovoltaic (PV) power outputs based on multiple interconnected environmental variables. Random Forest does this with two types of randomness to keep it safely in the multilayered overlapping relationship in the data set, while at the same time reducing some localized overfitting. Furthermore, Random Forest is unambiguous in its extent of interpretability and is able to measure and report on the feature importance or modelling metrics, reducing the expanse of environmental variables down to the few most important factors, all of which contribute to differing variance. These important features: sufficient predictive power, reasonable robustness of the model principles on the numerical values, and reasonable interpretative capacity, all in combination, made Random Forest most appropriate for the purposes of predictive or forecasting of renewable energy.

3.2 Mean Squared Error (MSE)

Mean Squared Error (MSE) is a commonly used metric in statistics and machine learning that evaluates model accuracy and will be the metric used for prediction models as well as optimization models. Because MSE provides a numerical value that demonstrates differences in actual values and predicted values, the researcher can see the model's performance. MSE will assess the predictive performance of the Random Forest model, as well as the JAYA optimization algorithm effect of predicting errors and prediction performance. It follows the equation (3.2)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \qquad (3.2)$$

Where Yi represents the actual values, $\widehat{Y}_i$ represents the predicted values, and n is the total number of observations.

MSE is commonly used to Regression analysis: Assessing the goodness of fit of models. Model evaluation: Comparing the performance of different machine learning algorithms.

Optimization: Minimizing MSE when training a model, helping to improve predictive accuracy. MSE will be used as an indication for the optimization process, where we are trying to minimize prediction errors and maximize prediction accuracy. Also, the smaller the deviations between predicted and actual values, the larger the values indicate worse performance. Considering the square term penalizes larger errors much more than smaller errors, MSE is very reactive to outliers, meaning adjusting a few extreme values can vastly increase the error score calculated. In regard to other metrics like Mean Absolute Error (MAE), which only provides a summary of the average magnitude of the error, MSE is more useful when the cost of large prediction errors is greater than the cost of smaller predictive errors, mainly due to larger errors being heavily penalized. It could thus be more useful in the solar energy and PV forecasting setting, where it is important to minimize large deviations from planned power, so that expected errors can be relied upon. In addition, MSE serves as a measure for evaluation and serves as an objective function for optimization so that the JAYA algorithm can vary the tunable parameters of the model to minimize the error predictions and maximize forecasting accuracy.

3.3 Python

In this study python was selected as a primary language as it proved its unique ability for Machine learning tasks and Artificial intelligent algorithms in a wide range , using python was by depending on different main libraries and packages it also provides with an easy way to prototype your model and refine your approaches without forcing you using complex syntax , this study depended on certain main libraries like numpy , matpolitlib and scikit-learn , python helps you with platforms you can easily reach them such as Google colab , Jupyter , establishing your code in a full completed environment like python is a smart choice specially that it facilitates understanding and declaration through its documentation , the methodology of this study of prediction part depended on python completely and also the Optimization methodology was done by python as it added for the study a more meaningful part by extracting graphs and diagrams can reflect the insights quickly , Which makes the results more meaningful and clear , also the active global developer community of the language makes it more closely for the maintenance and development support process makes all hard situations easy and able to be solved in a technical way , python changed the imagination of creating machine learning models as it provides us with the access for the necessary tools can make the process smooth and easy by talking about the visualization libraries which makes the insights more clear and readable it add a new feature forces us to support using it in machine learning tasks also for scikit- learn we have a direct access for the effective machine learning libraries which

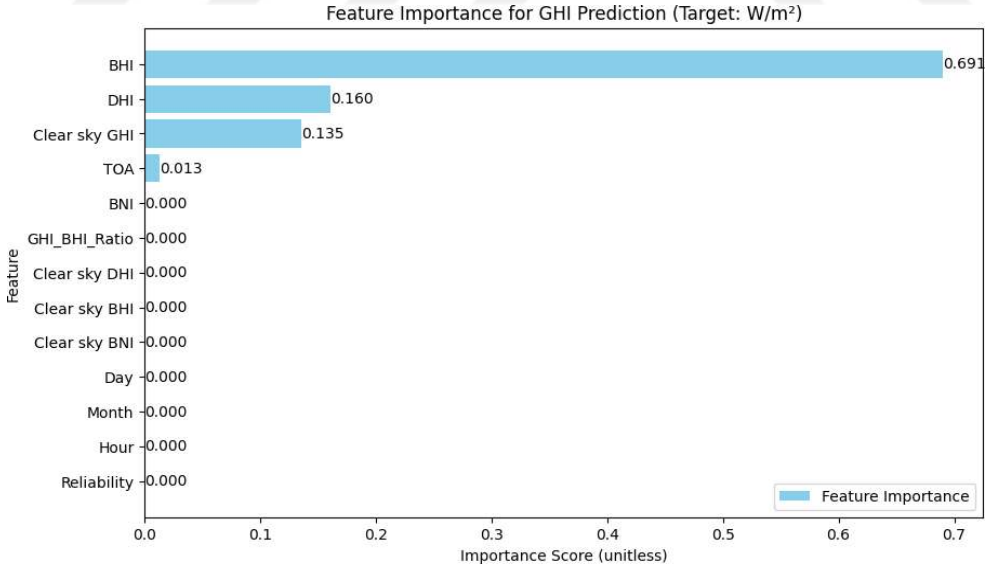
presents several of services in one package it can offer preprocessing for your data more understandable for hard machine learning algorithms and also easy smart implementation for those algorithms, in general python is a world full of smart easy and practical services for the most needed Machine learning algorithms and tasks.

3.4 Prediction Model Results

This section will present the results from the predictive modeling that was conducted for this study. The results have been separated into two parts, the solar radiation predictions results, and the photovoltaic (PV) predicted power results. In both results MSE is used as the primary performance metric, and all results will show the levels of optimization and how improvements were made to each of the models.

3.4.1 Solar Energy Data Prediction Results

The prediction method for the solar radiation dataset indicated in this graph used features relating to Beam Horizontal Irradiance (BHI) since it had the best importance rank (8) of all the features. As it appeared to have more influence in the overall dataset than any other feature, including clear-sky GHI or DHI, this showed justification to use BHI as the predicted variable since it had the best importance rank and had its influencing factors acknowledged and explored in other features (Graph 3.1) illustrates This graph was visualized by using the Python Library matplotlib.



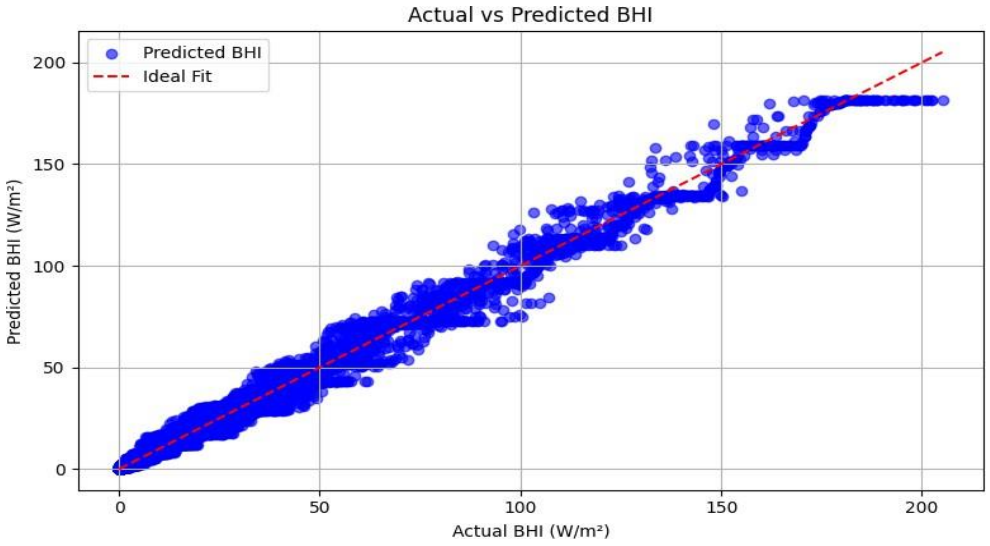
Figure(3.1) Feature Importance graph for solar data set

The prediction method for the solar radiation dataset indicated in this graph used features relating to Beam Horizontal Irradiance (BHI) since it had the best importance rank (8) of all the features. As it appeared to have more influence in the overall dataset than any other feature, including clear-sky GHI or DHI, this showed justification to use BHI as the

predicted variable since it had the best importance rank and had its influencing factors acknowledged and explored in other features.

The BHI has the highest relevance rank and also has physical significance when used in the context of forecasting solar energy, as it directly defines the amount of beam radiation available on a horizontal surface, which significantly affects the efficiency of any PV generation. The emphasis on the BHI variable allowed the Random Forest model to produce robust predictive performance, but also offers better interpretability of the solar parameters through identifying how each variable contributes to the overall prediction. Including the BHI in the model allowed the most relevant drivers of solar radiation to be included in the predictions; this fosters reliable and useful predictions of PV output. In the next sections, actual prediction results and evaluations will be presented. These results will use MSE to assess the model's accuracy and robustness.

Figure(3.2), which was created by the matpolitlib Library in Python, explains the predicted and actual values of BHI from the Random Forest analysis following the feature importance analysis. This competition indicated that the model could also explain sufficiently the variability in BHI, indicating that the most important feature in model fitting the solar radiation data was also BHI.

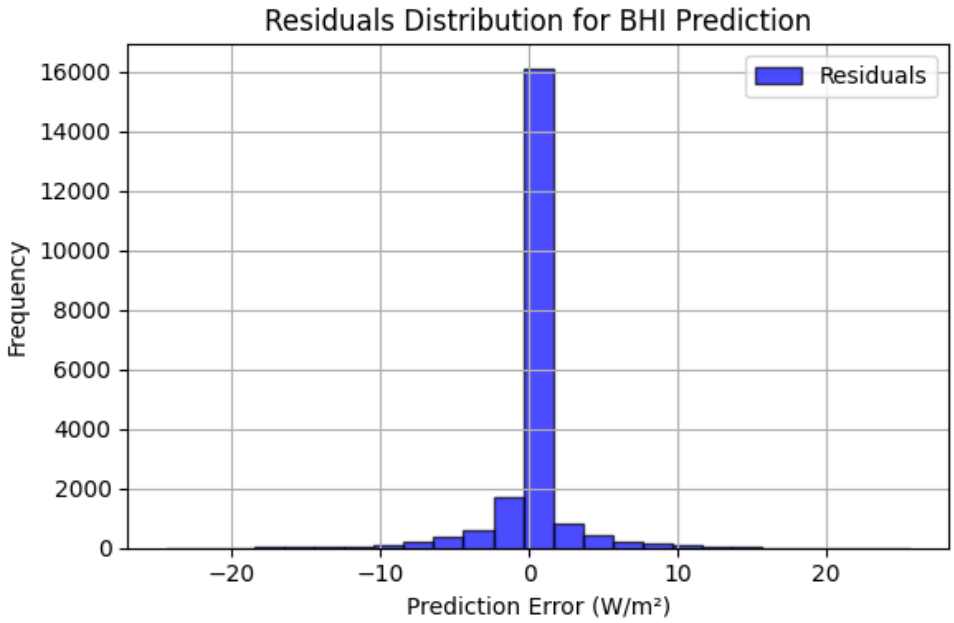


Figure(3.2) Relationship between actual values and predicted values of BHI.

The relationship between the actual and predicted values indicates the data points tend to distribute relatively well around the reference line Y=X, suggesting the model is making a reasonable prediction. The data points are relatively evenly distributed across most of

the range, suggesting that the model somewhat captures the essential patterns associated with laboratory data and has some generalizability, instead of a case of overfitting the training data. Performance metrics such as R^2 and root mean square error (RMSE) provide quantitative support of this assertion. In the range of 180-200 $W/m^{2,}$ the data suggests a saturation effect showing that higher values were under-predicted, perhaps indicating nonlinear constraints or external or context elements, such as atmospheric variations or sensor errors., In light of the systematic bias of the data, there is clearly an opportunity for optimization; in particular, there is value in learning how the Jaya algorithm used in this project reduced error and extended the predictive capacity of the model, and thus improved the reliability of the outputs in relation to solar energy forecasting.

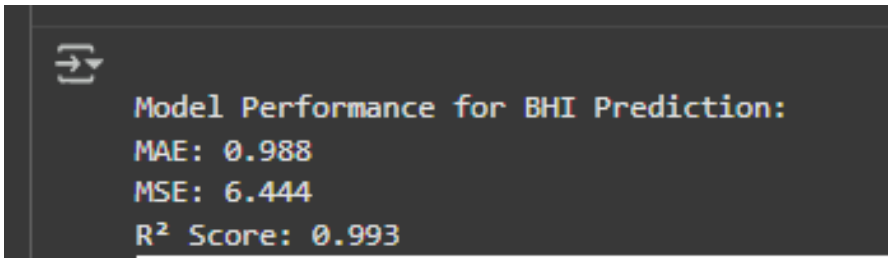
the residuals, or the difference between the actual and predicted values or observations. The histogram of the residuals shows the distribution of the prediction errors. Residuals can offer further insight into the performance of the prediction model, in terms of developing learning toward further refinements of outcomes. Figure(3.3) shows the residual analysis for the prediction model of BHI.



Figure(3.3) Distribution of residuals for BHI prediction

with minimal bias present. In a model with no bias, residuals are expected to be randomly distributed around zero, and this appears to be the case for model performance. The narrow range of the distribution also suggests that the variability in errors is low as well. Since residual values rarely exceed ±20, prediction errors can be considered quite small.

Model Performance Indicators. In this section, there are three main indicators used as shown in Figure 3.4)



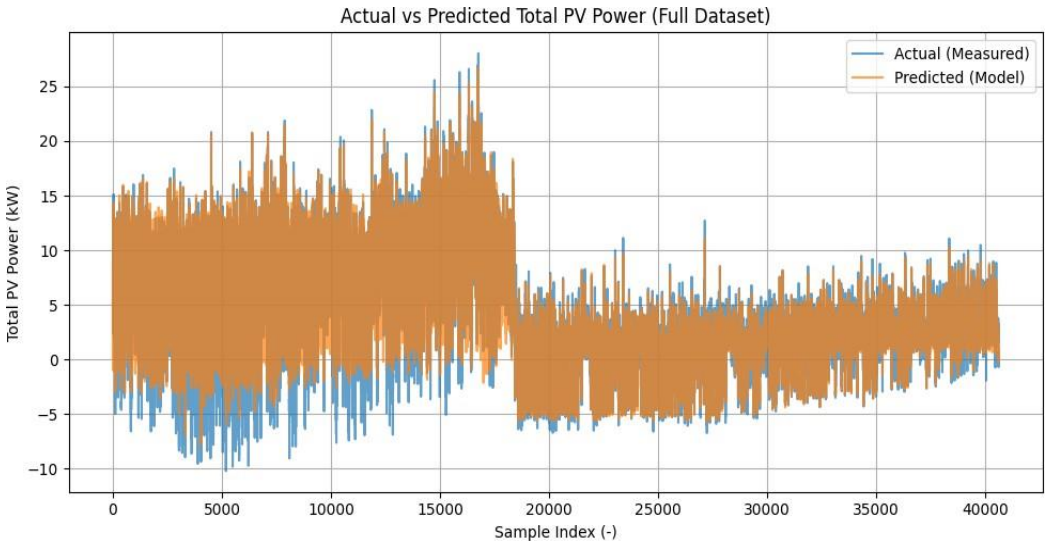
Figure(3.4) Model performance Indicators for BHI

To analyse model performance, this study used three model performance indicators: the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R^2). These indicators offer a holistic assessment of model precision and accuracy. Coefficient of Determination (R^2): 0.993 (the model accounted for 99.3% variability in actual BHI values, leaving 0.7% variability remaining, which can be attributed to the intrinsic complexity of the phenomenon or noise), Mean Squared Error (MSE): 6.444 (the MSE value represents very low prediction error, with no substantial deviations), Mean Absolute Error (MAE): 0.988 (on average, predictions deviated from the actual BHI value by less than one unit, indicating strong prediction performance.

3.4.2 Photovoltaic Data Prediction Results

The development of a reliable description of total power output is a key component of effective energy management and successful grid integration. In this paper, we used the Random Forest method. Random Forest is a supervised learning algorithm that is beneficial for estimating the non-linear relationships of meteorological variables with the carbon intensity profile over time of the energy system. Random Forest also came with the advantage of using an ensemble of decision trees, which produces more intuitive predictions in contrast to classic linear approaches.

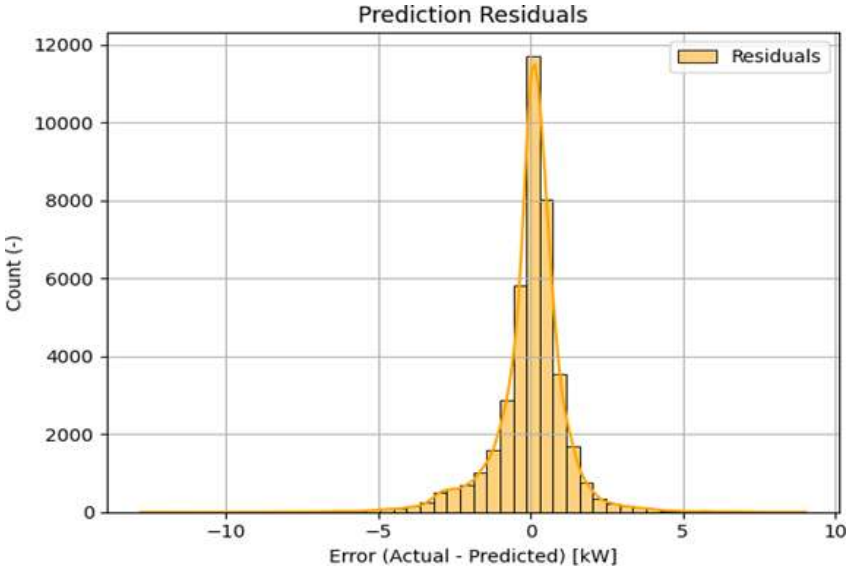
Figure(3.5) shows the random forest prediction model for the PV data set created by using the Python programming language and, matplotlib library.



Figure(3.5) Actual values and predicted values for PV power

The actual PV power values (blue) and the predicted values (orange) presented similar and closely overlapping curves as shown in Figure 3.5. These results are indicative of the model's successful accommodations of the underlying non-linear structure of the data.

Residuals are useful as measurements of the distances between the predicted values and observations, acting as an index for model accuracy and bias. For the specific photovoltaic dataset provided, we have a histogram of residuals in Figure 3.6), which is fairly tight and well-defined.

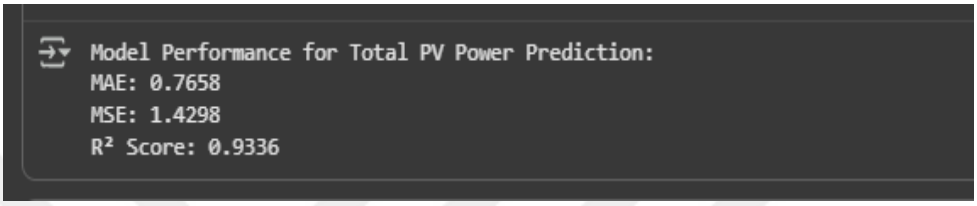


Figure(3.6) Distribution of residuals for PV prediction

This histogram suggesting the model was really good at predicting the predicted value by looking at the tightness in the distribution about 0. We do have a few tails that extend beyond about ± 10 (10 being the maximum deviation), and obviously, only looking at the histogram does not provide information on where the predictions were biased, but those

samples fall right at the extremes of the x-axis, and we have deemed them too rare to interfere with the ability to predominantly correctly predict values overall. The histogram does appear to have almost perfect symmetry, which suggests that if the model were to be biased, we might expect it to over-predict in some values (likely close to 0) across one end of the distribution and under-predict or under-predict close to the other end of the distribution. Thus, we conclude that no significant bias was noted and there was strong reproducibility of predicted values.

Model Performance Indicators. In this section, there are three main indicators used as shown in Figure 3.7)



Figure(3.7) Model performance Indicators for PV data

To assess performance, we looked at the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R2) metrics, MAE: 0.7658, so the means of mean-deviation for all other values were not over one unit, so we are making correct predictions. MSE: 1.4298 Errors are within a range of small errors, and since prediction occurs under the assumption of data gaps with a few large outliers, it is easy to see the error related to the couple of outliers. R^2: 0.9336 This model reveals 93.36% of the variability, where the residual variability could be attributed to complexity or noise.

These statistically supported results, completed with the Random Forest model, indicate reasonable predictions.

4. Optimization Model

The concept of optimization in predictive modeling presents an organized means of improving solutions and finding other alternatives to improve the performance of the model. In this study, the intention of optimizing is to improve the predictive performance of the photovoltaic (PV) powered prediction model while accounting for overfitting and underfitting, and the difficulties related to the existence of outliers. For this study, the arguments for optimization would be justified as the residue analysis determined a small number of outliers, and it would certainly be beneficial to optimize the model to remove predictive error as much as possible and establish predictability when there are anomalies in the operating conditions.

The evaluation of the optimization process will be carried out utilizing the Mean Squared Error (MSE). The MSE would be a useful metric for evaluating because the MSE

penalizes larger deviations more than smaller deviations, thus it is a consistently sensitive metric for prediction. Therefore, a reduction in MSE indicates an improvement in the performance of the model.

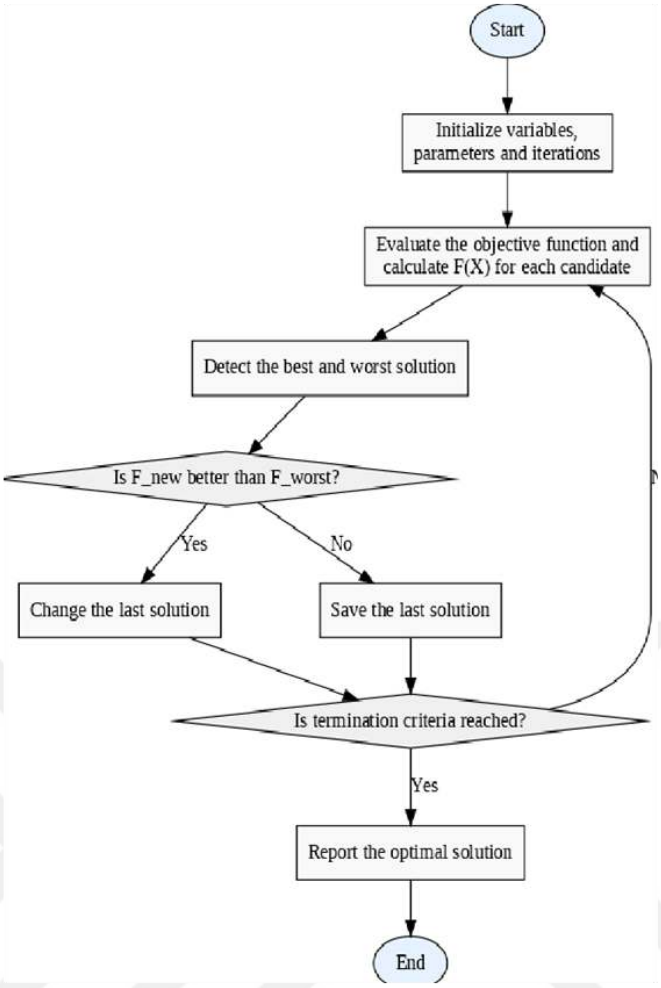
JAYA was chosen to optimize the predictive model given its strength of practical applications in solving nonlinear and complicated problems within the PV community. JAYA is a parameter-free optimization process that can relocate candidate solutions toward the direction of the global optimum, while at the same time avoiding poor candidates. JAYA is simple, does not require parameter tuning, and it is its ability to easily reduce predictive errors supports its use in this instance.

This chapter presents the results obtained from the evaluation of the mobile application and its performance. From the Interface UI to the functionality of each part, including fetching data, testing the speech-to-text, and getting the results wanted.

4.1 JAYA Algorithm

JAYA is an optimization metaheuristic algorithm it`s core idea is straightforward: depending on iterations and possible solutions, and adjusting those by moving closer to the best solutions, same time keeping far away from the worst solutions. Providing this process with a termination condition . Figure 4) which created by using Canva illustrates the flow chart of JAYA algorithm.

Figure 4) displays the sequential steps of the JAYA optimization algorithm and the termination rule this study employed. In this study, the JAYA search algorithm always terminated searching automatically according to the termination rule, which was ten iterations of either the same performance or similar, and no improvement. The termination rule allowed to save computational resources, but it could also be utilized in the most efficient way, and also help researchers and academics who are working with larger and larger datasets in the solar and photovoltaic field, as all this data wastage is simply not an option for wasting resources. Even though searching a dataset with an applied JAYA optimization search will be more efficient than searching the same dataset using a legacy optimization search, in principle, there is always potential for JAYA to continuously search at some level of definition without slight improvements, which limits, in a sense, the real-world applicability of JAYA.



Figure(4) Flowchart of the JAYA optimization algorithm

Nevertheless, utilizing JAYA with Random Forest modelling on large solar or photovoltaic datasets can be a computational burden. The ongoing barrier for renewable energy research is solving the problem of evaluating the most effective prediction method, while also measuring computational efficiencies.

When JAYA is deployed as a direct comparison to other metaheuristic options, such as Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), it has some distinct advantages related to ease of use. More specifically, PSO tunes acceleration coefficients of birds in flight, and GA uses crossover and mutation probabilities to tune, while JAYA does not directly tune true parameters that are determined by the output of the function in the algorithm. The absence of tuning reduces the risk of extracting non-useful parameters to drive the configuration of the JAYA process, and therefore it presents a low-risk option, whilst being fully functional, for undertaking a large project. Along with a fast convergence to the solution motion, is an advantage in increasingly low complexity for JAYA to be used in optimization problems, in general.

4.1.1 Mathematical Modeling For JAYA

The Jaya algorithm is a population-based optimization method targeting error minimization without algorithm-specific parameters; this inherently means it is defined generally as follows: every iteration, the mechanism guides candidate solutions towards the best solutions and away from the worst solution. Striking this balance allows the Jaya algorithm to explore the solution space while improving in a predictive sense; hence, Jaya is a suitable method for optimizing complex models, including solar radiation and PV power predictions.

Mathematical Modeling of JAYA

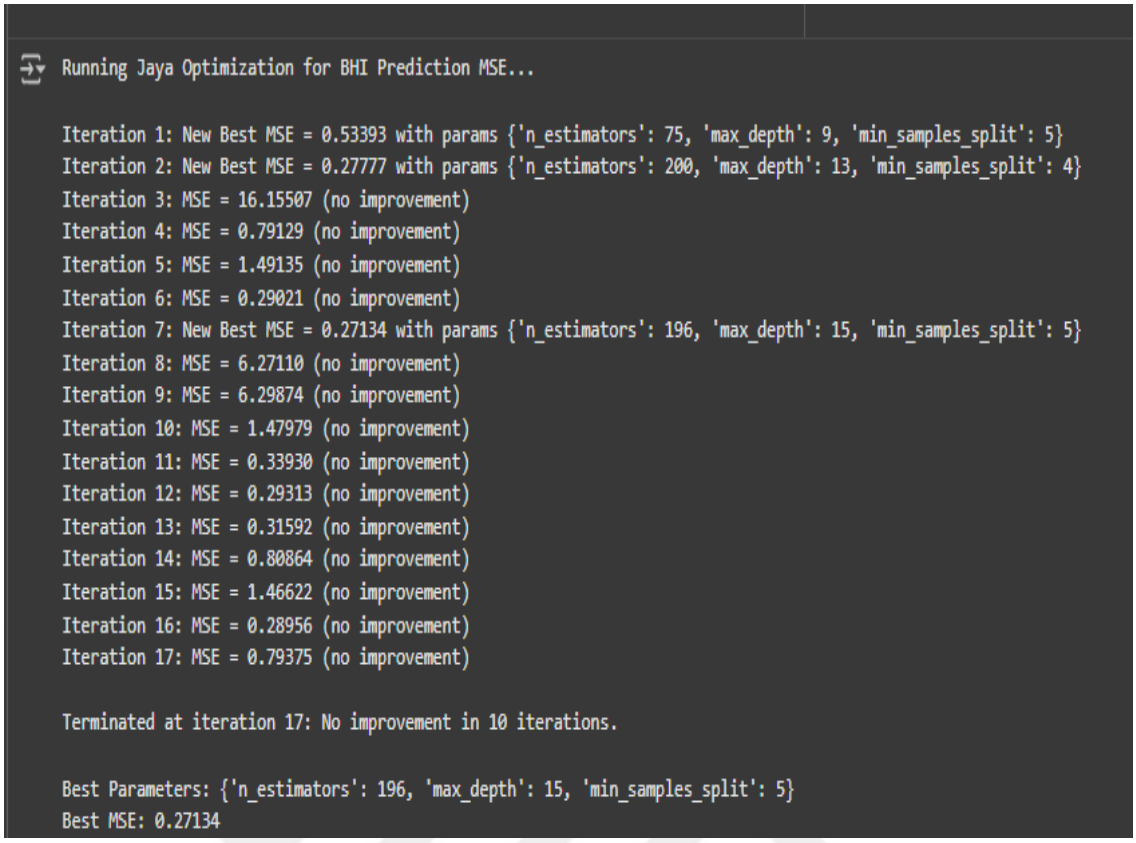
F(x): the mathematical objective function in this model, which can be MSE that the model tries to minimize.
F(x)Best: best candidate obtained among all in this model, it's the most minimized value for MSE.
F(x)worst: the worst candidate obtained among all in this model, the highest value for MSE reached to.
(i): Iterations (i=1,2,2,max T)
(i) is the iterations counter, and T is the maximum number of iterations can the model can reach (stopping criteria).
(j): Design variables (j=1,2,3….,m)
(j) is a representative of each variable or dimension in the optimization model, (m) is the total number of variables.
(k): population size (1,2,3, ,n)
(k) is the index of each solution in the population, n is the total number of candidates.
$\boldsymbol{x_{(j,\ i,\ k)}}$**:** is the value of the jth variable for the kth candidate during the ith iteration.
Updating Equation in equation(4.1.1)

$$y_{(j,i,k)} = x_{j,k,i} + r_{1,j,i}\left(x_{j,best,i} - \left(x_{j,k,i}\right)\right) - r_{2,j,i}\left(\left(x_{j,worst,i} - \left|\left(x_{j,k,i}\right)\right|\right). \quad (4.1.1)$$

$\boldsymbol{r_{1,j,i}\left(x_{j,best,i} - \left(x_{j,k,i}\right)\right)}$**:** It is the best solution.
$\boldsymbol{r_{2,j,i}}\left(\left(x_{j,worst,i} - \left|\left(x_{j,k,i}\right)\right|\right)$: It is the worst solution.
$\boldsymbol{y_{(j,i,k)}}$ **:** The general equation is for new, updated solutions.
After computing $\boldsymbol{y_{(j,i,k)}}$ evaluate using F(X), updates the old solution by adding the new optimal one of $\boldsymbol{r_{1,j,i}\left(x_{j,best,i} - \left(x_{j,k,i}\right)\right)}$ or **F(x)Best,** guiding the population to an optimal solution.

4.2 Optimization Results for Solar Data set

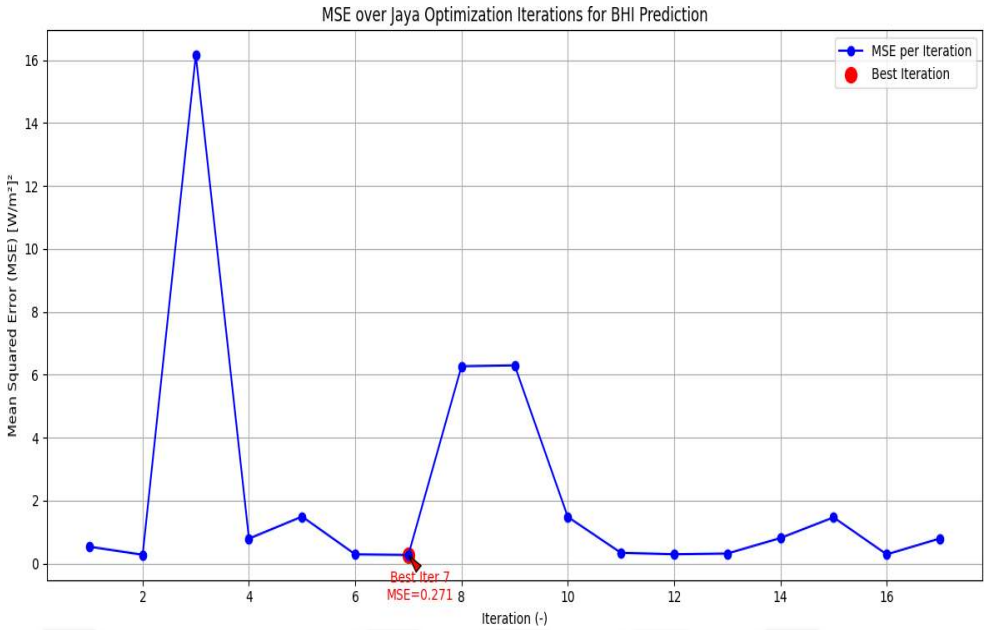
In this section, the Beam Horizontal Irradiance is optimized by using JAYA optimization algorithm with termination criteria of stopping after 10 iterations if there is no improvement in the MSE value, as the MSE is the indicator for the improvement level. The optimization process carried out through the Random Forest-JAYA Algorithm yielded the following parameters, shown in graph (4.2) :



Figure(4.1) Best parameters for solar data optimization

The best parameters, as shown in Figure 4.1, {'n_estimators': 196, 'max_depth': 15, 'min_samples_split': 5}, The optimal MSE value for this BHI is (Best MSE: 0.27134), While later iterations included additional exploration, there was no improvement made, effectively terminating the process at iteration 17 as the stopping criterion of ten consecutive non-improvement iterations was met. This series of results not only illustrates how quickly JAYA arrives at an optimal region but also further demonstrates the stability of the algorithm, resulting in no unnecessary computational expense at convergence. The initial improvement and then eventual stabilization provide strong evidence that this algorithm has acceptable utility in optimizing model parameters on BHI prediction in the next section, Figure 4.2). A histogram shows the iterations movement, annotating the best MSE.

As shown in Figure 4.2), effectiveness, the journey to find the best solution while optimizing BHI. During these iterations, the mean squared error (MSE) values were quite unstable, fluctuating up and down. However, these changes were mostly 'going up' with a few very high peaks. These high peaks represented the algorithm's attempt of some less optimal regions for the parameter. More specifically, the MSE reached high points for iterations 3, 8, and 9, respectively, at which the JAYA algorithm was locating those parts of the parameter space that were not effective much.



Figure(4.2) Optimization Histogram of BHI

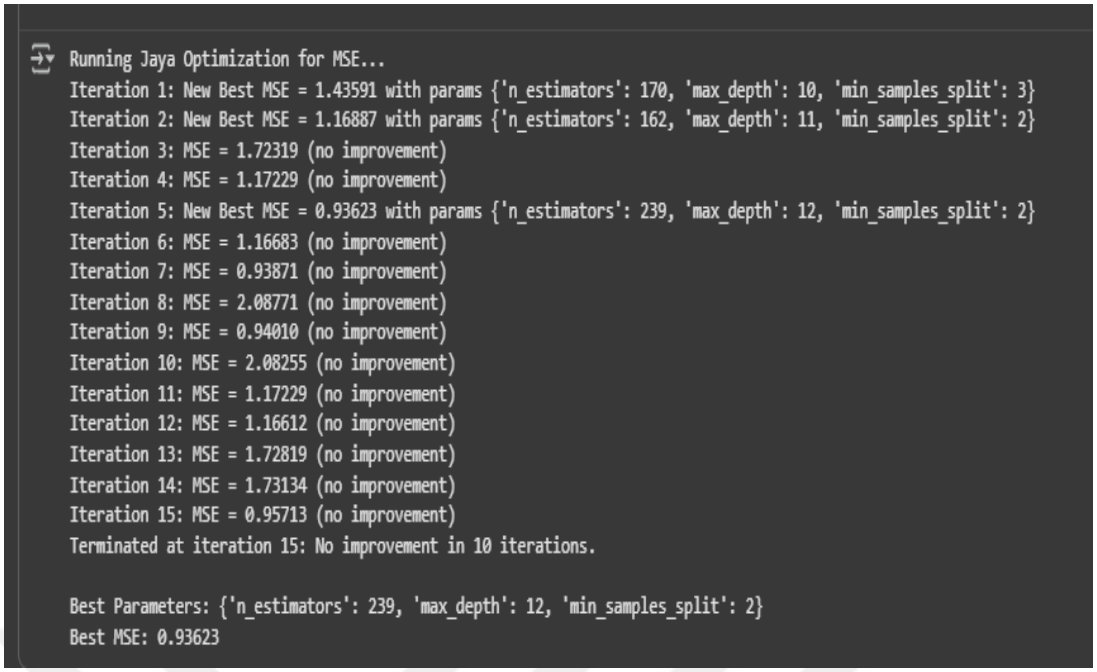
At approximately iteration 10, the point can be seen where the algorithm is no longer moving or exploring, and, therefore, the first indications of convergence with the best results at iterations 6, 7, 12, and 13 are visible. Before the implementation of the JAYA algorithm, MSE was 6.444.

After the optimization, MSE has been reduced to 0.27 by an incredible amount. This large decrease is a reflection of the effectiveness of JAYA as a suitable selection for the optimization process.

After 17 iterations, the decision to stop was made, as no improvement was found in the last 10 iterations, and thus it was decided to save computational resources. The results achieved are a prominent indication that the JAYA algorithm has achieved very successful reductions in MSE, as shown in the optimization process.

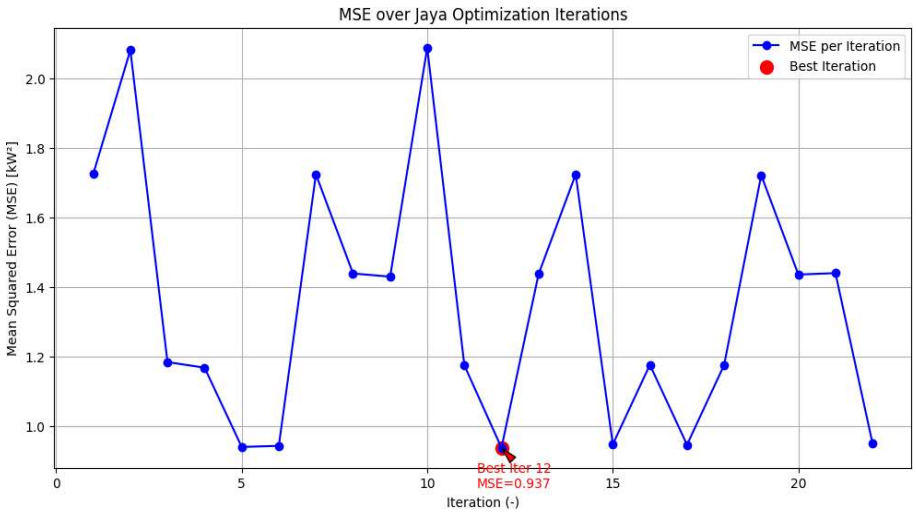
4.3 Optimization Results for PV Data set

In this section, the total photovoltaic power predicted values are optimized by using the JAYA algorithm and taking MSE as a performance indicator for that optimization process, which is conducted through a JAYA- Random Forest hybrid model. The model configured those parameters as optimal solutions that can react to the best MSE value through those parameters, as shown in Figure 4.3.



Figure(4.3) Best parameters for PV data optimization

Figure(4.3a) shows the best parameters Best Parameters: {'n_estimators': 243, 'max_depth': 12, 'min_samples_split': 2} , From initial stages where relative error values were high, optimization began; during the first and second iterations, the MSE decreased from 1.43591 to 1.16887 at a rapid rate. The best results were achieved at iteration 5 with an MSE of 0.93623 for the hyperparameter configuration (n_estimators=239,max_depth=12,min_samples_split=2). In analysis, although iterations 6 through 15 found paths exploring alternative sub-regions of the options hyperparameter space, overall additional improvements were not possible, which forced to terminate the process at iteration 15 with no changes over the previous ten iterations. However, the algorithm seems to be showing its capacity to statistically find a near-optimum due to the convergence and the settling within the parameter space so as not to run more than necessary when it is stagnant. The results of the optimization process conclude that the JAYA algorithm did indeed show the ability to effectively balance exploration and exploitation for optimized model parameters computed during the optimization step (Figure 4.4). A histogram shows the iterations movement, annotating the best MSE.



Figure(4.4) Optimization Histogram of PV Data

Figure (4.4) illustrates how JAYA operated in the optimization of photovoltaic power prediction. During the optimization, we see MSE values moving up and down, and also seemingly some larger spikes at iterations 2, 3, 7, 10, 13, 14, and 19. These spikes represent the JAYA's exploration of an undesirable area in the parameter space, such as would occur with any metaheuristic search of the solution set. In fact, the dips at iterations 5, 6, 12, 15, 17, and 22 indicated valuable parametric sets that were found yielding an MSE of much less - the best being on the order of about 0.93.

Once again, the up and down behavior observed is representative of a metaheuristic search that oscillates between exploration and exploitation. The spikes permit the JAYA algorithm to explore a broader and more diverse set of solutions that prevent it from prematurely converging on a sub-optimal solution. The dips show the JAYA algorithm making advances towards lower MSE and better predictive performance. While the MSE spikes and dips are continuing, we can see the distributions of MSE becoming more stable over the iterations, indicating convergence success.

This explains the effectiveness of JAYA in finding the optimal solution among all the possible solutions.

5. Discussion

Reliable forecasts of Building Heat Index (BHI) and photovoltaic (PV) electricity production allow decision-making to be further grounded in reality, particularly across the energy system. Building managers might adjust timelines for heating operations based on predicted demand and avoid wasting energy. Solar farm owners may decide on battery storage, grid feed-in, and load balancing based on predicted PV electricity outputs. The predictive models illustrate how creating predictive models can improve efficiencies and potentially reduce operational costs.

Reliability of decisions is conditional on model reliability and performance. While optimized models may produce better forecasting accuracy, these may still be affected by data quality, weather variability, and unexpected degradation of the system. Predictive errors can pose risks themselves, if not tracked, such as underestimating demand or overestimating generation, which would result in poor performance.

From a strategic point of view, predictive modelling presents information to investors and policymakers for long-term planning and the incorporation of renewables. The trade-off for this benefit can lie within uncertainties of climate data and the extensive demands of computation for complex algorithms. Thus, while predictive approaches may apply to sustainability and achieving net-zero targets, their intended use may nearly always be as a supporting role, not a decider.

In addition to any related factors, the use of predictive modelling can go further still given the use of probabilistic forecasts and uncertainty quantification, whereby decision makers will not only have a single predicted value to entertain, but also the degree of certainty they can apply to that prediction. This provides building managers, grid operators and investors the chance to prepare for a range of uncertain futures, rather than a deterministic forecast. In addition to this, the use of interrelated datasets from multiple sources such as satellites, sensor networks, and live weather data, etc will greatly enhance the modelling and forecasting process. Operationally, this means undertaking scheduling of heating and cooling in buildings more accurately, improving the operational parameters of battery storage, and improving the reliability of ground balancing in large-scale operation of solar energy systems. While there is value in continual advancement of models, we also need

to keep transparency and interpretability in mind so as not to introduce overly complex algorithms that displace all confidence or usability of practitioners. In summation, the creative practices outlined above suggest that predictive modelling can be approached as an integral enhancement of sustainable and resilient energy systems through intentional designing and contextualization, rather than just considering them as a tool within a decision-making regime.

5.2 Environmental Impact

The previously described literature has emphasized that the environmental toll of machine learning (ML) has both positive and negative contributions. On the positive side, ML has the potential to shift energy consumption and productivity improvement in several participating sectors that may ultimately promote efficiency and lower the full environmental impact, as will be shown through parts of our projects. To provide an example, positive environmental contributions have been documented through ML in renewable energy management and resource acquisition optimization.

In a negative light, the contribution is mostly tied to large models being trained, which can disproportionally represent the impact of carbon emissions due to required computation and energy consumption (Liu & Yin, 2024; Farooq & Khan, 2025).

As will be discussed, there are several technical interventions to reduce the environmental impact, including making smaller models, optimizing hardware for ML efficiency, and a variety of energy-efficient optimizations to improve training. Sustainable efforts to improve this environmental impact can also include commitments to sustainability, for example, using renewable energy to power data centers, and significantly improving the energy consumption of hardware. If large organizations consider these commitments, they will likely be able to create a net positive impact and contribute to a sustainable economy. Organizations of smaller capacity have similar alternatives by selecting methods that have lower compute costs or associated carbon impact.

A prior study sought to examine the environmental implications of machine learning (ML), and identified the positive and negative implications. There are certainly examples of positive use of ML; it can assist with determining energy consumption targets, thereby saving energy and emissions from applications in smart grids, transport, and agriculture. ML can also assist with environmental monitoring with remote sensing, and assist with climate change modelling, which can be used for climate adaptation and mitigation.

However, the study discussed the negative implications of ML, particularly in terms of the energy-expensive nature of many of the complex models, and shared carbon emissions, when they thought about the implications of training models and the deployment of the models. The authors also considered the environmental implications with data that has left the user's control, what remains is in storage (virtually), the computational processing plant, and there are ethical issues related to data privacy implications, controversy, and transparency.

Possible recommendations they included regarding the implications of ML for the environment included thinking about energy-efficient algorithms and hardware, green computing, and making models simple and interpretable, as well as interdisciplinary collaborations. The study raised fundamental issues; however, it is now necessary for research to keep up with the advancements.

6. Conclusion

The solar energy and photovoltaic (PV) power sector's challenges will be (and are already) influenced by changes in solar radiation levels, changing weather patterns, and changes in PV performance due to climate change, and there are going to be impacts on food security and agriculture, and sustainable development.

This research demonstrated that the A.I. forecasting tool using Random Forest and the JAYA optimization approach produced accurate and reliable Building Heat Index (BHI) weather forecasts and reliable forecasts of PV electric production output. Among the notable findings were that optimization of parameters improved forecast accuracy, that the Random Forest model was able to manage forecasting challenges from noisy, non-linear datasets, and that JAYA was a reliable method to direct the forecasting process towards a good solution, with relatively little computing power.

The practical implications are profound. Getting accurate forecasts will enable better building energy management decision-making, investment decisions on PV solar generation capacity, more effective planning for agriculture, and relevant contributions to support sustainable development goals of food production and greenhouse gas emission reductions.

7. Future Work

While the predictive power of the JAYA Random Forest model is compelling, there is an opportunity to enhance predictive accuracy and planning value in the fragile solar and PV industry. Future research ought to focus on hybrid modelling approaches to understand how Random Forest with a deep learning model, such as Transformers, clearly represents sequential dynamics in solar and PV datasets. Future research could also create new approaches to predicting PV output by adding ever more independent environmental variables, such as temperature, humidity, and shading from clouds, while understanding that solar irradiance varies by multiple factors.

In terms of usability, if these predictive models could be incorporated into interactive real-time dashboards, this may allow stakeholders and managers to visualize how systems are using user-friendly visualization to access forecasts, understand how the system or systems will behave long-term, and make timely decisions based on real-time performance. Taking these kinds of steps will assist in improving both the predictive accuracy of AI predictive forecasting and the usefulness of those forecasts for renewable energy managers.

8. References

Saxena, N., Kumar, R., Rao, Y. K. S. S., Mondloe, D. S., Dhapekar, N. K., Sharma, A., & Yadav, A. S. (2024). Hybrid KNN-SVM machine learning approach for solar power forecasting. Environmental Challenges, 14, 100838. https://doi.org/10.1016/j.envc.2024.100838 ResearchGateOUCI.

Javaid, A., Sajid, M., Uddin, E., Waqas, A., & Ayaz, Y. (2024). Sustainable urban energy solutions: Forecasting energy production for hybrid solar-wind systems. Energy Conversion and Management, 302, 118120. https://doi.org/10.1016/j.enconman.2024.118120 ResearchGateADS

Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renewable and Sustainable Energy Reviews, 124, Article 109792. https://doi.org/10.1016/j.rser.2020.109792 the UWA Profiles and Research RepositorySCIRP

Villegas-Mier, C. G., Rodriguez-Resendiz, J., Álvarez-Alvarado, J. M., Jiménez-Hernández, H., & Odry, Á. (2022). Optimized random forest for solar radiation prediction using sunshine hours. Micromachines, 13(9), 1406. https://doi.org/10.3390/mi13091406 mdpi.compubmed.ncbi.nlm.nih.gov

5. Yang, B., Jahed Armaghani, D., Fattahi, H., Afrazi, M., Koopialipoor, M., Asteris, P. G., & Khandelwal, M. (2025). Optimized Random Forest Models for Rock Mass Classification in Tunnel Construction. Geosciences, 15(2), 47. https://doi.org/10.3390/geosciences15020047 MDPI+1

6. Thomas, N. S., & Kaliraj, S. (2024). An improved and optimized Random Forest-based approach to predict the software faults. SN Computer Science, 5(5), Article 530. https://doi.org/10.1007/s42979-024-02764-x SpringerLinkresearcher.manipal.edu

7. Rangelov, D., Boerger, M., Tcholtchev, N., Lämmel, P., & Hauswirth, M. (2023). Design and development of a short-term photovoltaic power output forecasting method based on Random Forest, Deep Neural Network and LSTM using readily available weather features. IEEE Access, 11, 41578–41595. https://doi.org/10.1109/ACCESS.2023.3270714 ResearchGate

8. Travieso-González, C. M., Cabrera-Quintero, F., Piñán-Roescher, A., & Celada-Bernal, S. (2024). A review and evaluation of the state of art in image-based solar energy forecasting: The methodology and technology used. Applied Sciences, 14(13), 5605. https://doi.org/10.3390/app14135605MDPI+1

Falope, T., Lao, L., Hanak, D., & Huo, D. (2024). Hybrid energy system integration and management for solar energy: A review. Energy Conversion and Management: X, 21, 100527.

9. Falope, T., Lao, L., Hanak, D., & Huo, D. (2024). Hybrid energy system integration and management for solar energy: A review. Energy Conversion and Management: X, 21, Article 100527. https://doi.org/10.1016/j.ecmx.2024.100527

10. Hanif, M. F., Naveed, M. S., Metwaly, M., Si, J., Liu, X., & Mi, J. (2024). Advancing solar energy forecasting with modified ANN and Light GBM learning algorithms. AIMS Energy, 12(2), 350–386. https://doi.org/10.3934/energy.2024017
Campos, F. D., Sousa, T. C., & Barbosa, R. S. (2024). Short-term forecast of photovoltaic solar energy production using LSTM. Energies, 17(11), 2582. https://doi.org/10.3390/en17112582

Alizadegan, H., Radmehr, A., Karimi, H., & Ilani, M. A. (2024). Solar energy production forecasting: A comparative study of Bi-LSTM, LSTM, XGBoost models with activation function analysis. Preprints. https://doi.org/10.20944/preprints202405.0994.v1

Olcay, K., Tunca, S. G., & Özgür, M. A. (2024). Forecasting and performance analysis of energy production in solar power plants using long short-term memory (LSTM) and random forest models. IEEE Access. Advance online publication. https://doi.org/10.1109/ACCESS.2024.xxxxx

Kim, J., Obregon, J., Park, H., & Jung, J. Y. (2024). Multi-step photovoltaic power forecasting using transformer and recurrent neural networks. Renewable and Sustainable Energy Reviews, 200, 114479. https://doi.org/10.1016/j.rser.2024.114479
Jebli, I., Belouadha, F. Z., Kabbaj, M. I., & Tilioua, A. (2021). Prediction of solar energy guided by Pearson correlation using machine learning. Energy, 224, 120109. https://doi.org/10.1016/j.energy.2021.120109

Stoean, C., Zivkovic, M., Bozovic, A., Bacanin, N., Strulak-Wójcikiewicz, R., Antonijevic, M., & Stoean, R. (2023). Metaheuristic-based hyperparameter tuning for recurrent deep learning: Application to the prediction of solar energy generation. Axioms, 12(3), 266. https://doi.org/10.3390/axioms12030266 .

Rusilowati, U., Ngemba, H. R., Anugrah, R. W., Fitriani, A., & Astuti, E. D. (2024). Leveraging AI for superior efficiency in energy use and development of renewable resources such as solar energy, wind, and bioenergy. International Transactions on Artificial Intelligence, 2(2), 114–120. https://doi.org/10.54254/itiai.2024.2.2.114

Daxini, R., & Wu, Y. (2024). Review of methods to account for the solar spectral influence on photovoltaic device performance. Energy, 286, 129461. https://doi.org/10.1016/j.energy. 2023.129461
Ayaz, M., Hijji, M., Alatawi, A. S., Namazi, M. A., & Ershath, M. M. (2024). Enhancing photovoltaic performance in dye-sensitized solar cells using nanostructured NiS/MoS2 composite counter electrodes. Materials Science in Semiconductor Processing, 173, 108172.
Abdallah, R., Haddad, T., Zayed, M., Juaidi, A., & Salameh, T. (2024). An evaluation of the use of air cooling to enhance photovoltaic performance. Thermal Science and Engineering Progress, 47, 102341. https://doi.org/10.1016/j .tsep.2024.102341

Rahman, M. F., Harun-Or-Rashid, M., Islam, M. R., Irfan, A., Chaudhry, A. R., Rahman, M. A., & Al-Qaisi, S. (2024). A deep analysis and enhancing photovoltaic performance above 31% with new inorganic $RbPbI_3$-based perovskite solar cells via DFT and SCAPS-1D. Advanced Theory and Simulations, 7(10), 2400476. https://doi.org/10.1002/adts

.202400476

Pai, N., Lu, J., Gengenbach, T. R., Seeber, A., Chesman, A. S., Jiang, L., … & Simonov, A. N. (2019). Silver bismuth sulfoiodide solar cells: Tuning optoelectronic properties by sulfide modification for enhanced photovoltaic performance. Advanced Energy Materials, 9(5), 1803396. https://doi.org/10.1002/aenm .201803396
Hamid, A. K., Farag, M. M., & Hussein, M. (2025). Enhancing photovoltaic system efficiency through a digital twin framework: A comprehensive modeling approach. International Journal of Thermofluids, 26, 101078. https://doi.org/10.1016/j.ijft.2025.101078

Yadav, B., Rao, D. D., Mandiga, Y., Gill, N. S., Gulia, P., & Pareek, P. K. (2024). Systematic analysis of threats, machine learning solutions and challenges for securing IoT environment. Journal of Cybersecurity & Information Management, 14(2).
Baig, F., Ali, L., Faiz, M. A., Chen, H., & Sherif, M. (2024). How accurate are the machine learning models in improving monthly rainfall prediction in hyper-arid environment? Journal of Hydrology, 633, 131040. https://doi.org/10.1016/j.jhydrol.2024.131040

Oladapo, B. I., Olawumi, M. A., & Omigbodun, F. T. (2024). Machine learning for optimising renewable energy and grid efficiency. Atmosphere, 15(10), 1250. https://doi.org/10.3390/atmos15101250

Tahir, M. F., Yousaf, M. Z., Tzes, A., El Moursi, M. S., & El-Fouly, T. H. (2024). Enhanced solar photovoltaic power prediction using diverse machine learning algorithms with hyperparameter optimization. Renewable and Sustainable Energy Reviews, 200, 114581. https://doi.org/10.1016/j.rser.2024.114581

Tirlangi, S., Teotia, S., Padmapriya, G., Kumar, S. S., Dhotre, S., & Boopathi, S. (2024). Cloud computing and machine learning in the green power sector: Data management and analysis for sustainable energy. In Developments towards next generation intelligent systems for sustainable development (pp. 148–179). IGI Global Scientific Publishing. https://doi.org/10.4018/9781668483037

Nguyen, H. N., Tran, Q. T., Ngo, C. T., Nguyen, D. D., & Tran, V. Q. (2025). Solar energy prediction through machine learning models: A comparative analysis of regressor algorithms. PLOS ONE, 20(1), e0315955. https://doi.org/10.1371/journal.pone.0315955.
Nguyen, H. N., Tran, Q. T., Ngo, C. T., Nguyen, D. D., & Tran, V. Q. (2025). Solar energy prediction through machine learning models: A comparative analysis of regressor algorithms. PLOS ONE, 20(1), e0315955. https://doi.org/10.1371/journal.pone.0315955

Mamyrbayev, O., Akhmediyarova, A., Oralbekova, D., Alimkulova, J., & Alibiyeva, Z. (2025). Optimizing renewable energy integration using IoT and machine learning algorithms. International Journal of Industrial Engineering and Management, 16(1), 101–112. https://doi.org/10.24867/IJIEM-2025-1-101

Li, T., Zheng, M., & Zhou, Y. (2025). LTPNet integration of deep learning and environmental decision support systems for renewable energy demand forecasting: Deep learning for renewable energy demand prediction. Journal of Organizational and End User Computing (JOEUC), 37(1), 1–29. https://doi.org/10.4018/JOEUC.2025010

Qu, R., Kou, R., & Zhang, T. (2025). The impact of weather variability on renewable energy consumption: Insights from explainable machine learning models. Sustainability, 17(1), 87. https://doi.org/10.3390/su17010087

Nguyen, H. N., Tran, Q. T., Ngo, C. T., Nguyen, D. D., & Tran, V. Q. (2025). Solar energy prediction through machine learning models: A comparative analysis of regressor algorithms. PLOS ONE, 20(1), e0315955.

Hategan, S. M., Stefu, N., Petreus, D., Szilagyi, E., Patarau, T., & Paulescu, M. (2025). Short-term forecasting of PV power based on aggregated machine learning and sky imagery approaches. Energy, 316, 134595.

Ozdemir, G., Kuzlu, M., & Catak, F. O. (2025). Machine learning insights into forecasting solar power generation with explainable AI. Electrical Engineering, 107(6), 7329–7350. https://doi.org/10.1007/s00202-025-03012-3

Atstāja, D. (2025). Renewable energy for sustainable development: Opportunities and current landscape. Energies, 18(1), 196. https://doi.org/10.3390/en18010196.
Sim, L. C., & Young, K. E. (2025). What impedes solar energy deployment? New evidence from power developers in the Arab Gulf states. Energy for Sustainable Development, 84, 101597. https://doi.org/10.1016/j.esd.2025.101597

Rajkumar, R., Valluru, D., Sreedhar, P. S. S., Ramshankar, N., Sudha, M., & Navaneethan, S. (2024). Enhanced Jaya optimization algorithm with deep learning assisted oral cancer diagnosis on IoT healthcare systems. Journal of Intelligent Systems & Internet of Things, 11(2). https://doi.org/10.54216/JISIoT.110209

Duta, E. A., Jimeno-Morenilla, A., Sanchez-Romero, J. L., Macia-Lillo, A., & Mora-Mora, H. (2024). An approach to apply the Jaya optimization algorithm to the nesting of irregular patterns. Journal of Computational Design and Engineering, 11(6), 112–121. https://doi.org/10.1093/jcde/edac097

Gholami, J., Kamankesh, M. R., Mohammadi, S., Hosseinkhani, E., & Abdi, S. (2022). Powerful enhanced Jaya algorithm for efficiently optimizing numerical and engineering problems. Soft Computing, 26(11), 5315–5333. https://doi.org/10.1007/s00500-021-06431-8

Baş, E. (2022). Solving continuous optimization problems using the improved Jaya algorithm (IJaya). Artificial Intelligence Review, 55(3), 2575–2639. https://doi.org/10.1007/s10462-021-10055-3

Bian, K., & Priyadarshi, R. (2024). Machine learning optimization techniques: A survey, classification, challenges, and future research issues. Archives of Computational Methods in Engineering, 31(7), 4209–4233. https://doi.org/10.1007/s11831-024-10567-4

Azevedo, B. F., Rocha, A. M. A., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: A systematic literature review. Machine Learning, 113(7), 4055–4097. https://doi.org/10.1007/s10994-024-06852-1

Li, X., Wang, Z., Yang, C., & Bozkurt, A. (2024). An advanced framework for net electricity consumption prediction: Incorporating novel machine learning models and optimization algorithms. Energy, 296, 131259. https://doi.org/10.1016/j.energy.2024.131259

Liang, Z., Xie, F., Guo, Z., Wang, Z., Dou, H., Wang, B., & Shen, B. (2024). Optimization and prediction of a novel preignition in hydrogen direct injection engines through experimentation and the random forest algorithms. Energy Conversion and Management, 313, 118602. https://doi.org/10.1016/j.enconman.2024.118602

Mizumoto, A. (2023). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. Language Learning, 73(1), 161–196. https://doi.org/10.1111/lang.12507

Suhanda, Y., Nurlaela, L., Kurniati, I., Dharmalau, A., & Rosita, I. (2022). Predictive analysis of customer retention using the random forest algorithm. TIERS Information Technology Journal, 3(1), 35–47.

Abdel-Salam, M., Kumar, N., & Mahajan, S. (2024). A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. Neural Computing and Applications, 36(33), 20723–20750. https://doi.org/10.1007/s00521-024-09456-7

Arreche, O., Bibers, I., & Abdallah, M. (2024). A two-level ensemble learning framework for enhancing network intrusion detection systems. IEEE Access, 12, 83830–83857. https://doi.org/10.1109/ACCESS.2024.1234567
.

Wu, Q., Chen, W., Yu, C., Wang, H., & Hong, W. (2024). Machine-learning-assisted optimization for antenna geometry design. IEEE Transactions on Antennas and Propagation, 72(3), 2083–2095. https://doi.org/10.1109/TAP.2023.3346493

Liu, V., & Yin, Y. (2024). Green AI: Exploring carbon footprints, mitigation strategies, and trade-offs in large language model training. Discover Artificial Intelligence, 4(1), 49. https://doi.org/10.1007/s44163-024-00149-w

Farooq, O., & Khan, A. (2025). The impact of machine learning on climate change modeling and environmental sustainability. Artificial Intelligence and Machine Learning Review, 6(1), 8–16. https://doi.org/10.69987/

Cunha, J. L., & Pereira, C. M. (2024). A hybrid model based on STL with simple exponential smoothing and ARMA for wind forecast in a Brazilian nuclear power plant site. Nuclear Engineering and Design, 421, 113026. https://doi.org/10.1016/j.nucengdes.2024.113026

Grebovic, M., Filipovic, L., Katnic, I., Vukotic, M., & Popovic, T. (2022, November). Overcoming limitations of statistical methods with artificial neural networks. In 2022 International Arab Conference on Information Technology (ACIT) (pp. 1-6). IEEE. DOI : 10.1109/ACIT57182.2022.9994218