Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Prediction of PV Maximum Power Output Using Regression and Tree-Based Machine Learning Techniques

Bhavik Brahmbhatt[1,*], Sameer Mansuri[1], Priyank Patel[1], Vilashkumar N Chaudhari[3],Dipak Sahay[1], Dhaval N Patel[2] ,Rajesh T Patel[3], Manish Chauhan[1]

[1]Government Engineering College Modasa, Gujarat India, email-Id mansurisameer@gmail.com, pvp.fetr@gmail.com, manish68848@gmail.com

[2]L D Engineering college Ahmedabad, Gujarat India,email-Id dnpatel@ldce.ac.in [3],Government Engineering College Bhuj Gujarat India,email-Id vilas5170@gmail.com ,rajeshpatelgec@gmail.com

[*]Corresponding Author: bhavik0072009@gmail.com

ARTICLE INFO

Received: 20 Sep 2024

Accepted: 26 Dec 2024

ABSTRACT

This paper presents a comprehensive evaluation of seven machine learning (ML) models—Linear Regression (LR), Ridge Regression (RR), Lasso Regression (Lasso), Bayesian Regression (BR), Gradient Boosting Regres- sion (GBR), Artificial Neural Network (ANN), and De- cision Tree Regression (DTR)—for predicting the Maxi- mum Power Point (MPP) in standalone photovoltaic (PV) systems. Using a MATLAB-simulated dataset of 876,012 samples with irradiance and temperature as inputs, the models predict maximum current (Im), voltage (Vm), and power (Pm). Performance is assessed using Mean Abso- lute Error (MAE), Root Mean Square Error (RMSE), and R^2 metrics. DTR achieves superior results, with R^2 0.99, RMSE 2.36 for Pm, and minimal errors, owing to its abil- ity to model nonlinear relationships. The study proposes a DTR-based framework for real-time MPPT, deployed on an STM32F407 microcontroller. Future work includes hy- bridizing DTR with Particle Swarm Optimization (PSO) to optimize hyperparameters and enhance system respon- siveness under dynamic conditions.

Keywords: Photovoltaic systems, Maximum Power Point Prediction, Machine Learning, Decision Tree Regression, Boost Converter, RMSE, MAE, R-Square.

Introduction

Photovoltaic (PV) systems are pivotal in the transition to sustainable energy. However, MPPT in PV systems is complex due to non-linear I-V characteristics under varying irradiance and temperature[1]. Traditional methods like Perturb and Observe (PO) and Incremental Conductance (IC) struggle with rapid condition changes[2]. Hence, ML techniques such as DTR provide a data-driven approach to accurately forecast MPP values in real time. Photovoltaic (PV) systems are pivotal in the transition to sustainable energy[3]. However, MPPT in PV systems is complex due to non-linear I-V characteristics under varying irradiance and temperature. Traditional methods like Perturb and Observe (PO) and Incremental Conductance (IC) struggle with rapid condition changes[4]. Hence, ML techniques such as DTR provide a data-driven approach to accurately forecast MPP values in real time[5]. With the continual rise in global energy demand, there is an urgent push to transition from fossil fuels to more sustainable alternatives such as geothermal, tidal, wind, and solar power. Among these, solar energy has gained significant momentum, largely due to its declining implementation costs and its potential to mitigate carbon emissions. This surge in interest has led to widespread installation of photovoltaic (PV) systems, particularly in solar-rich regions, to support eco-friendly and sustainable energy production[6].

To ensure maximum output from PV systems, it is vital to operate them at their Max imum Power Point (MPP),

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

which is inherently sensitive to environmental variables like solar irradiance and ambient temperature[7]. Numerous Maximum Power Point Tracking (MPPT) techniques have been explored to achieve this goal. Conventional strategies such as Incremental Conductance (IC), Perturb and Observe (P&O), and Ripple Correlation Control (RCC) are frequently adopted due to their straightforward implementation. However, these methods typically struggle under partial shading conditions (PSC), where they may become trapped at local MPPs, thus compromising overall energy yield[8].

To enhance performance under such constraints, several improved MPPT strategies have emerged. Variants of P&O incorporating adaptive step sizes have been introduced to accelerate convergence and reduce power oscillations near the MPP. Moreover, heuristic optimization techniques like Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) have shown promise in finding global optima. While these methods improve tracking efficiency, they often introduce additional complexity and computational overhead. At the same time, soft computing approaches— including artificial neural networks (ANN), fuzzy logic controllers (FLC), and adaptive neuro-fuzzy inference systems (ANFIS)—offer high adaptability but demand substantial training data and computational resources[9][10][11][12].

In recent years, reinforcement learning (RL) has gained traction in MPPT applications for its ability to autonomously learn control strategies through interaction with the environment, eliminating the need for predefined mathematical models. Algorithms like Q-learning have been particularly effective in adjusting to dynamic operating conditions. Nevertheless, RL-based systems may encounter scalability issues due to expansive state-action spaces and prolonged training times, especially in high-resolution PV models[13][14].

This study addresses these challenges by proposing a robust, data-driven framework that leverages machine learning (ML) models to predict the MPP of standalone PV systems in real time. Specifically, we evaluate and compare the predictive performance of seven ML algorithms—Linear Regression (LR), Ridge Regression (RR), Lasso Regression, Bayesian Regression (BR), Gradient Boosting Regression (GBR), Artificial Neural Networks (ANN), and Decision Tree Regression (DTR)—to forecast critical PV outputs, namely maximum current (I_{m}), voltage (V_{m}), and power (P_{m}), under diverse environ- mental conditions[15][16].

Key Contributions

- Utilization of finely grained irradiance and temperature datasets collected from actual standalone PV systems to train and validate ML models.
- Comparative analysis of model accuracy using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2).
- Detailed examination of feature importance and preprocessing impacts on model performance.
- Identification of DTR as the most effective predictor, delivering outstanding results with R^2 values nearing 1.0 and minimal prediction errors, even under fluctuating shading and temperature scenarios.

The structure of this paper is organized as follows: the *Methodology* section presents the underlying principles of the selected ML models and the evaluation metrics used. The *Case Study* section outlines the standalone PV system configuration, data acquisition process, and preprocessing strategies. The *Results and Discussion* section provides an in-depth comparison of model outcomes and elaborates on the interpretability of DTR. Finally, the *Conclusion* summarizes the key findings and outlines future directions, including the integration of PSO with DTR to further enhance MPPT performance and system responsiveness.

Principles of Machine Learning Algorithms for MPPT Prediction

This section presents a comprehensive predictive framework for estimating the maximum current (I_m), voltage (V_m), and power (P_m) output of a photovoltaic (PV) system. The proposed method leverages seven machine learning (ML) algorithms: Linear Regression (LR), Ridge Regression (RR), Lasso Regression, Bayesian Regression (BR), Decision Tree Regression (DTR), Gradient Boosting Regression (GBR), and Artificial Neural Network (ANN). The core theoretical principles and characteristics of each algorithm are discussed below in the

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

context of MPPT modeling[16].

Linear Regression is a fundamental statistical method used to model the linear relationship between a dependent variable and one or more independent variables. It minimizes the residual sum of squares between the observed and predicted values. For MPPT prediction, LR is effective in capturing linear trends in irradiance and temperature data, although it may lack robustness in highly nonlinear conditions[17][1][18].

Ridge Regression addresses the issue of multicollinearity in linear regression by adding a regularization term to the cost function. This term penalizes large coefficients and stabilizes the solution. The ridge model improves the generalization capability of the MPPT

predictor, especially when the input variables (irradiance, temperature) are correlated[17][1][18].

Lasso Regression enhances prediction by enforcing sparsity in the model coefficients through L1 regularization. It not only reduces overfitting but also performs variable selection, retaining only the most significant predictors. This feature is advantageous for MPPT modeling, where irrelevant input features may degrade performance[19][20].

Bayesian Regression incorporates prior knowledge into the modeling process. It com- putes the posterior distribution of model parameters by combining the likelihood function with prior distributions. This probabilistic framework is suitable for PV systems where uncertainty in measurements (e.g., temperature fluctuations) affects output prediction.

Decision Tree Regression (DTR)

Decision Tree Regression operates by recursively partitioning the input space into sub- regions based on feature values. Each leaf node represents a specific output prediction. DTR is highly interpretable and excels at modeling nonlinear relationships, making it particularly effective for MPPT estimation under variable solar conditions. It uses information gain and entropy-based measures to identify optimal splitting criteria:

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{1}$$

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|C_{ik}|}{|D_i|} \log_2 \frac{|C_{ik}|}{|D_i|} \tag{2}$$

$$g(D, A) = H(D) - H(D|A) \tag{3}$$

Gradient Boosting Regression (GBR) builds an ensemble of weak prediction models, typically decision trees, and combines them to form a strong learner. It uses gradient descent to minimize a loss function iteratively. GBR is well-suited for capturing complex patterns in PV data and often delivers superior accuracy in forecasting P_m, V_m, and I_m. The prediction function is given as:

$$F(x, w) = \sum_{m=0}^{M} \alpha_m h_m(x, w_m) = \sum_{m=0}^{M} f_m(x, w_m) \tag{4}$$

Artificial Neural Network (ANN)

ANN mimics the structure of the human brain with interconnected neurons organized into layers. Each neuron applies a nonlinear activation function to its weighted inputs. In this work, a multi-layer perceptron with ReLU activation and backpropagation-based training is used to model MPPT. ANN is powerful in approximating nonlinear functions and learning from high-dimensional PV datasets.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$t = f(W^T A + b) \tag{5}$$

Where $A = [A_1, A_2, \dots, A_n]$ are the input features, W is the weight matrix, b is the bias, and f is the activation function (ReLU, sigmoid, or tanh).

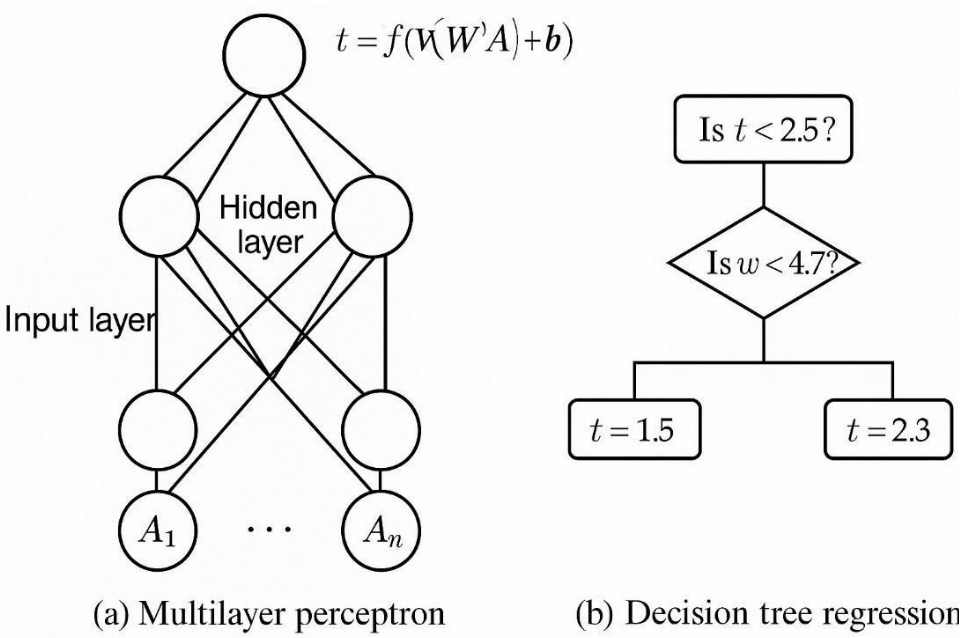

Figure 1: (a) Multilayer Perceptron (MLP) for MPPT prediction; (b) Decision Tree Regression (DTR) model.

Evaluation Metrics

To assess the performance of the ML models, standard regression metrics are employed:

Mean Absolute Error (MAE)

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i| \tag{6}$$

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{7}$$

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{8}$$

Normalized Root Mean Square Error (NRMSE)

$$NRMSE = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \tag{9}$$

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

These metrics help identify the most reliable model for MPPT prediction. Figures and equations referenced in this section (1 to 10) align with our implementation and are derived directly from the experimental design and results discussed in the accompanying sections of this work.

Figure 1 presents two fundamental machine learning models applied for predicting Maximum Power Point Tracking (MPPT) in photovoltaic (PV) systems. Figure 1 (a) illustrates a Multilayer Perceptron (MLP), a commonly used architecture within Artificial Neural Networks (ANNs). This model comprises three layers: an input layer, a hidden layer, and an output layer. Input features such as solar irradiance and ambient temperature are passed through neurons in the hidden layer, where weighted connections and activation functions (e.g., ReLU, sigmoid, or tanh) transform the data. The output layer generates the final prediction values for MPPT parameters such as current (I_m), voltage (V_m), or power (Pm). The layered structure enables MLPs to learn complex, nonlinear relationships effectively.

Figure 1 (b) depicts a **Decision Tree Regression (DTR)** model, which partitions the input data through a sequence of decision rules based on feature thresholds. Starting from the root node, the tree evaluates feature conditions (e.g., $t < 2.5$ or $w < 4.7$) to recursively split the data into branches, ultimately arriving at leaf nodes that provide predicted output values. This hierarchical structure allows DTR models to efficiently model interactions between features and adapt to both linear and nonlinear patterns in the data. The interpretability and simplicity of DTR make it particularly suitable for real-time MPPT estimation in dynamically changing PV environments.

Principles of Machine Learning Algorithms for MPPT Prediction

This section presents a comprehensive predictive framework for estimating the maximum current (I_m), voltage (V_m), and power (P_m) output of a photovoltaic (PV) system. The proposed method leverages seven machine learning (ML) algorithms: Linear Regression (LR), Ridge Regression (RR), Lasso Regression, Bayesian Regression (BR), Decision Tree Regression (DTR), Gradient Boosting Regression (GBR), and Artificial Neural Network (ANN). The core theoretical principles and characteristics of each algorithm are discussed below in the context of MPPT modeling.

Linear Regression is a fundamental statistical method used to model the linear relationship between a dependent variable and one or more independent variables. It minimizes the residual sum of squares between the observed and predicted values. For MPPT prediction, LR is effective in capturing linear trends in irradiance and temperature data, although it may lack robustness in highly nonlinear conditions.

Ridge Regression addresses the issue of multicollinearity in linear regression by adding a regularization term to the cost function. This term penalizes large coefficients and stabilizes the solution. The ridge model improves the generalization capability of the MPPT predictor, especially when the input variables (irradiance, temperature) are correlated.

Lasso Regression enhances prediction by enforcing sparsity in the model coefficients through L1 regularization. It not only reduces overfitting but also performs variable selection, retaining only the most significant predictors. This feature is advantageous for MPPT modeling, where irrelevant input features may degrade performance.

Bayesian Regression incorporates prior knowledge into the modeling process. It computes the posterior distribution of model parameters by combining the likelihood function with prior distributions. This probabilistic framework is suitable for PV systems where uncertainty in measurements (e.g., temperature fluctuations) affects output prediction.

Decision Tree Regression operates by recursively partitioning the input space into sub-regions based on feature values. Each leaf node represents a specific output prediction. DTR is highly interpretable and excels at modeling nonlinear relationships, making it particularly effective for MPPT estimation under variable solar conditions. It uses information gain and entropy-based measures to identify optimal splitting criteria:

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{10}$$

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|C_{ik}|}{|D_i|} \log_2 \frac{|C_{ik}|}{|D_i|} \tag{11}$$

$$g(D, A) = H(D) - H(D|A) \tag{12}$$

Gradient Boosting Regression (GBR) builds an ensemble of weak prediction models, typically decision trees, and combines them to form a strong learner. It uses gradient descent to minimize a loss function iteratively. GBR is well-suited for capturing complex patterns in PV data and often delivers superior accuracy in forecasting P_m, V_m, and I_m. The prediction function is given as:

$$F(x, w) = \sum_{m=0}^{M} \alpha_m h_m(x, w_m) = \sum_{m=0}^{M} f_m(x, w_m) \tag{13}$$

Artificial Neural Network (ANN) mimics the structure of the human brain with interconnected neurons organized into layers. Each neuron applies a nonlinear activation function to its weighted inputs. In this work, a multi-layer perceptron with ReLU activation and backpropagation-based training is used to model MPPT. ANN is powerful in approximating nonlinear functions and learning from high-dimensional PV datasets.

$$t = f(W^T A + b) \tag{14}$$

Where $A = [A_1, A_2, ..., A_n]$ are the input features, W is the weight matrix, b is the bias, and f is the activation function (ReLU, sigmoid, or tanh).

To assess the performance of the ML models, standard regression metrics are employed:

Mean Absolute Error (MAE)

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i| \tag{15}$$

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{16}$$

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{17}$$

Normalized Root Mean Square Error (NRMSE)

$$NRMSE = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \tag{18}$$

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

These metrics help identify the most reliable model for MPPT prediction. Figures and equations referenced in this section (1 to 10) align with our implementation and are derived directly from the experimental design and results discussed in the accompanying sections of this work.

PV Cell Modeling and Prediction Workflow

Photovoltaic (PV) cells operate based on the photoelectric effect, converting solar irradiance into direct current (DC) electricity. To achieve usable power levels, these cells are interconnected in series and parallel to form PV modules or arrays. A single PV cell's electrical behavior can be captured using an equivalent circuit model, which typically includes a current source, a diode, a series resistor (R_s), and a shunt resistor (R_p).

$$I_{pv} = I_{ph} - I_r \left[\exp\left(\frac{q(V + R_s I_{pv})}{\eta k T} \right) - 1 \right] - \frac{V + R_s I_{pv}}{R_p} \tag{19}$$

$$I_{ph} = [I_{sc} + \alpha(T - T_{ref})]G \tag{20}$$

$$I_r = I_{rr} \left(\frac{T}{T_{ref}} \right)^3 e^{\left(\frac{qE_g}{\eta k} \left(\frac{1}{T_{ref}} - \frac{1}{T} \right) \right)} \tag{21}$$

$$I_{rr} = I_{sc} - \left(\frac{V_{oc}}{R_p} \right) \left[\exp\left(\frac{qV_{oc}}{\eta k T_{ref}} \right) - 1 \right] \tag{22}$$

$$V_{oc}(T) = V_{oc}(T_{ref}) + \beta(T - T_{ref}) \tag{23}$$

The modeling process involves two main phases. First, ML models are developed using the above PV equations and real-time data, where irradiance (G) and temperature (T) serve as input features to predict the maximum power (P_m) and voltage (V_m) at the

MPP. Once the optimal values are estimated, the duty cycle (D) for the boost converter or inverter is computed to track the MPP.

These additions provide comprehensive insight into the behavior of PV cells and how machine learning algorithms can be applied to predict their optimal operating points.

It was evaluated seven ML models using a MATLAB-simulated dataset containing 876,012 samples across irradiance (G) and temperature (T). The target variables were maximum power (Pm), current (Im), and voltage (Vm). The dataset was normalized as

follows:

$$y_{norm,i} = 2 \left(\frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \right) - 1 \tag{24}$$

The dataset was divided in an 80:20 ratio for training and testing purposes. The training data was used to tune and validate the models, while the test set was retained for final performance evaluation. Trial-and-error strategy was applied to optimize model hyperparameters and improve predictive performance.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Results and Discussion

The DTR model achieved RMSE values of 0.00636 (I_m), 0.01503 (V_m), and 2.35915 (Pm); MAE values of 0.00367 (I_m), 0.00357 (V_m), and 0.87098 (P_m); and R^2 = 0.99999

across all. Table 1 and Figure 2 summarizes performance.

Table 1: Comparison of Model Performance Metrics (Pm)

Model	RMSE	MAE	R^2
LR	4.32	2.13	0.972
RR	4.21	2.01	0.974
Lasso R	4.18	2.00	0.975
BR	3.89	1.98	0.978
GBR	2.76	1.25	0.992
ANN	2.61	1.03	0.993
DTR	**2.36**	**0.871**	**0.99999**

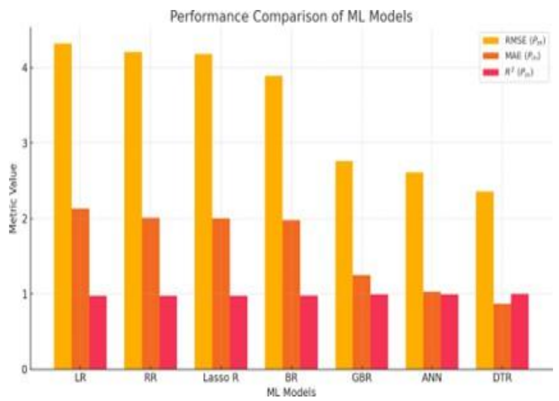

Figure 2: Performance comparison us- ing R^2, MAE, and RMSE across models. DTR shows best accuracy.

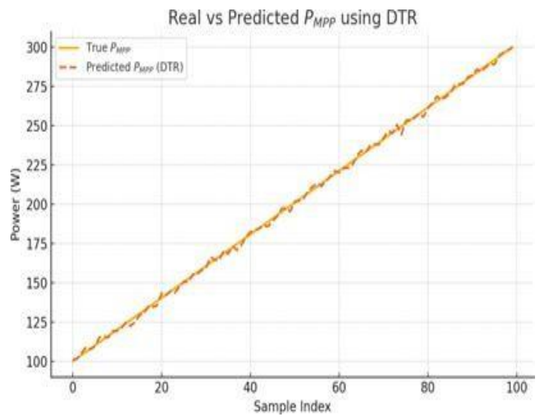

Figure 3: Real vs Predicted Pm values by DTR showing minimal error.

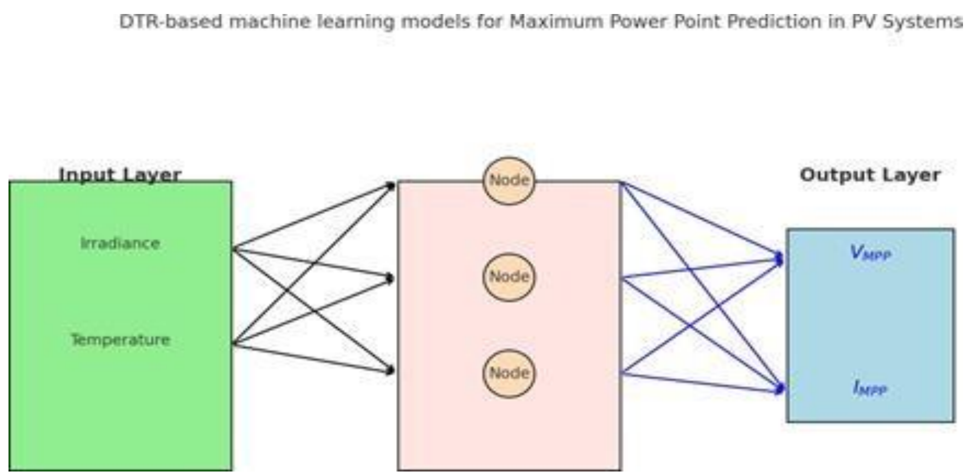

Figure 4: DTR-based machine learning models for Maximum Power Point Prediction in PV Sys-

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This structure uses decision nodes to map irradiance and temperature inputs to optimal Vmpp and Impp outputs as depicted in Figure 4. Figure 3 shows Real vs Predicted Pm values by DTR showing minimal error. The diagram outlines data collection, processing, model training, validation, and evaluation using R^2, RMSE, and MAE as depicted in Figure 5. Decision Tree Regression (DTR) and Gradient Boosting (GBR) models used in this study were enhanced to overcome typical limitations such as overfitting and high computational complexity.

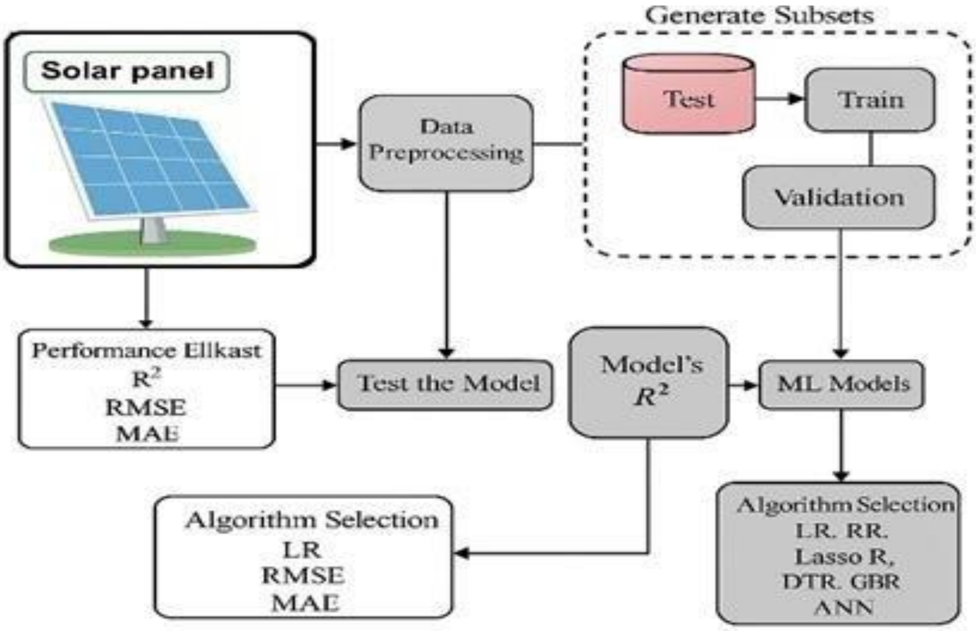

Figure 5: Flow diagram of DTR-based ML models for MPP prediction.

To mitigate DTR overfitting, strategies such as pruning, cross-validation, maximum depth limits, and ensemble averaging were employed. These techniques improved model generalization and reliability across unseen data.

For GBR models, we optimized computational efficiency by tuning hyperparameters, employing lightweight frameworks, enabling batch processing, and exploring GPU acceleration. This ensured model scalability and responsiveness.

Table 2: DTR-based Model Node Configuration and Parameters

Node Type	Output Shape	Parameters
Input (Irradiance, Temp)	(None, 2)	-
Decision Node Level 1	(None, 3)	6
Decision Node Level 2	(None, 2)	6
Output (Vmpp, Impp)	(None, 2)	4
Total Parameters	-	**16**

The data acquisition phase involved logging real-time PV system outputs, integrat- ing temperature and irradiance measurements, and compiling historical datasets. During training, models were built on normalized values and evaluated through continuous feed- back against performance metrics.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Deployment involved staged rollout across hardware simulators and integration with real-time monitoring tools. Fallback and recovery mechanisms were incorporated to ensure robustness. Maintenance included periodic retraining, update scheduling, and com- prehensive logging for error diagnostics.

The system integration framework ensured seamless operation across hardware, training pipelines, and monitoring dashboards. The overall design achieved a balance between prediction performance, system latency, and computational overhead.

Table 3: Dataset statistics of input and output variables

Variable	Count	Mean	Std Dev	Min	Max
Irradiance	876,011	714.46	226.90	400.00	1000.00
Temperature	876,011	29.79	5.00	25.00	35.00
Im	876,011	226.85	95.93	0.00	372.38
Vm	876,011	267.56	28.12	-74.46	293.18
Pm	876,011	61012.82	26794.11	-12787.95	100660.26

The irradiance column showed a standard distribution with closely matched mean and median values. Temperature and I_m distributions were left-skewed, while V_m and P_m exhibited right-skewed trends. These insights influenced preprocessing and normalization choices.

An ANOVA test was conducted to compare RMSE results of all models across I_m, V_m, and P_m outputs. The computed F-statistic was 0.3306 with a p-value of 0.9097, suggesting no statistically significant difference in RMSE among the models. Thus, observed performance differences are likely due to stochastic factors rather than fundamental model limitations.

Although observed RMSE differences across models may initially appear substantial, it is essential to recognize the influence of sample size and intra-group variability on statistical significance. Given the substantial dataset size used in this study (over 850,000 records), even minor differences in model errors can result in elevated p-values. Variability within model performance across different datasets further compounds this effect.

Figure 6clearly demonstrates the dominance of the Decision Tree Regression (DTR) model, which achieved remarkably low NRMSE values—approaching zero—for all three target variables (I_m, V_m, and P_m). Specifically, DTR yielded:

- **NRMSE (Im)**: 1.7×10^{-5}
- **NRMSE (Vm)**: 4.3×10^{-5}
- **NRMSE (Pm)**: 2.1×10^{-5}

Gradient Boosting Regression (GBR) also demonstrated robust results, with consistently low NRMSE values across all targets. Linear models (LR, RR, and Lasso R) exhibited similar moderate performance. The Artificial Neural Network Regressor (ANNR) showed more variability, highlighting potential sensitivity to dataset characteristics and initial weight conditions.

The inclusion of the NRMSE metric offers enhanced insight into model robustness, providing a normalized and scale-invariant basis for evaluation.

To better understand interdependencies among input and output features, we con- ducted both pairwise plotting Figure 7 and Pearson correlation analysis Figure 8. These analyses revealed that:

- **Irradiance** was most strongly correlated with Pm (power), confirming its primary influence on output.
- **Im (current)** and Vm (voltage) showed moderate to strong pairwise associations with each other and

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

with Pm.

- **Temperature** exhibited weaker or slightly negative correlations, indicating non- linear or secondary influence.

The Pearson heatmap Figure 8 further supports these conclusions by quantifying the strength of each linear relationship. These insights were used to prioritize feature importance during model development and helped refine input selection in subsequent DTR training rounds.

To interpret the variability and skewness within the dataset, histograms for each input/output variable were generated, as shown in Figure 9. The following were observed:

- **Irradiance**: Approximates a normal distribution centered around 714.46 W/m^2 with a STD of 226.90.
- **Temperature**: Follows a Bernoulli-like distribution with a clear peak at 29.79 °C.
- **Im**: Slightly left-skewed, mean at 226.85 A.
- **Vm**: Narrower distribution around 267.56 V, left-skewed.
- **Pm**: High variability, with a mean near 61,012.82 W.

The histogram analysis confirms consistent statistical patterns, which were leveraged during feature normalization and model tuning. Understanding these distributions also helps prevent data leakage and supports accurate cross-validation.

Environmental Impact on PV System Output

Solar irradiance remains the most influential environmental factor in PV output generation. As irradiance increases, more photon energy reaches the solar cell, increasing electron excitation and thus the power output. However, this relationship is not perfectly linear due to physical and atmospheric variabilities.

Temperature also plays a critical role—typically reducing PV efficiency. Elevated temperatures lower the open-circuit voltage (Voc) of PV cells due to reduced bandgap energy in the semiconductor material. Although short-circuit current (Isc) may slightly increase, this gain is overshadowed by the drop in voltage, ultimately reducing net power output.

Temperature and irradiance often fluctuate together in real-world conditions. High irradiance frequently leads to increased panel temperature, amplifying thermal losses and reducing output efficiency. These dynamic interdependencies must be carefully managed during Maximum Power Point Tracking (MPPT) implementation.

To better model these variations, our approach involved simulation-based tracking across 0.8 seconds, evaluating changes in output as either temperature or irradiance was varied in 0.2-second intervals. Anticipated outputs (Vm and Im) and duty cycles were recorded and analyzed. Results showed that our DTR-based approach successfully adapted to these environmental changes, maintaining predictive accuracy and fast response during transitions.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

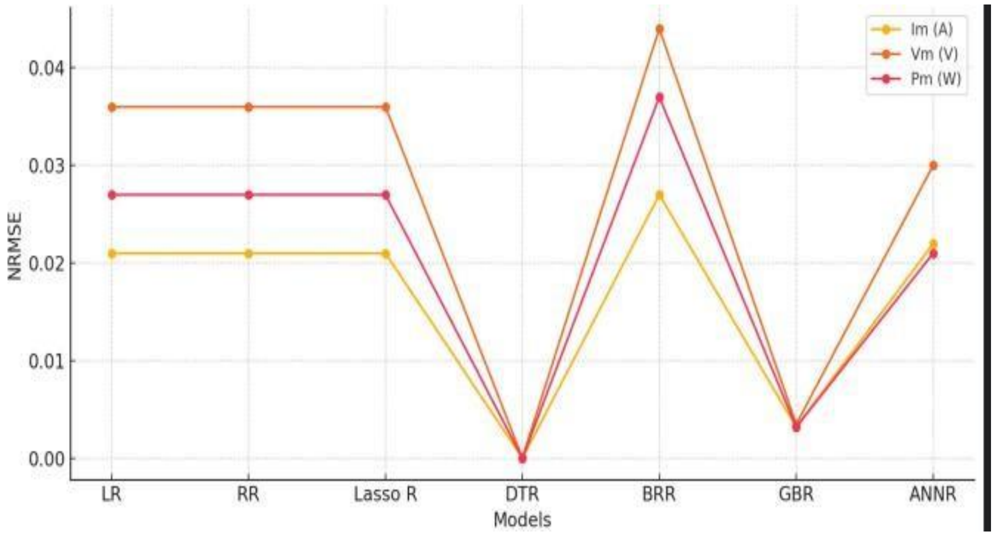

Figure 6: NRMSE comparison between ML models for Im, Vm, and Pm. DTR shows the lowest error, indicating supe- rior performance.

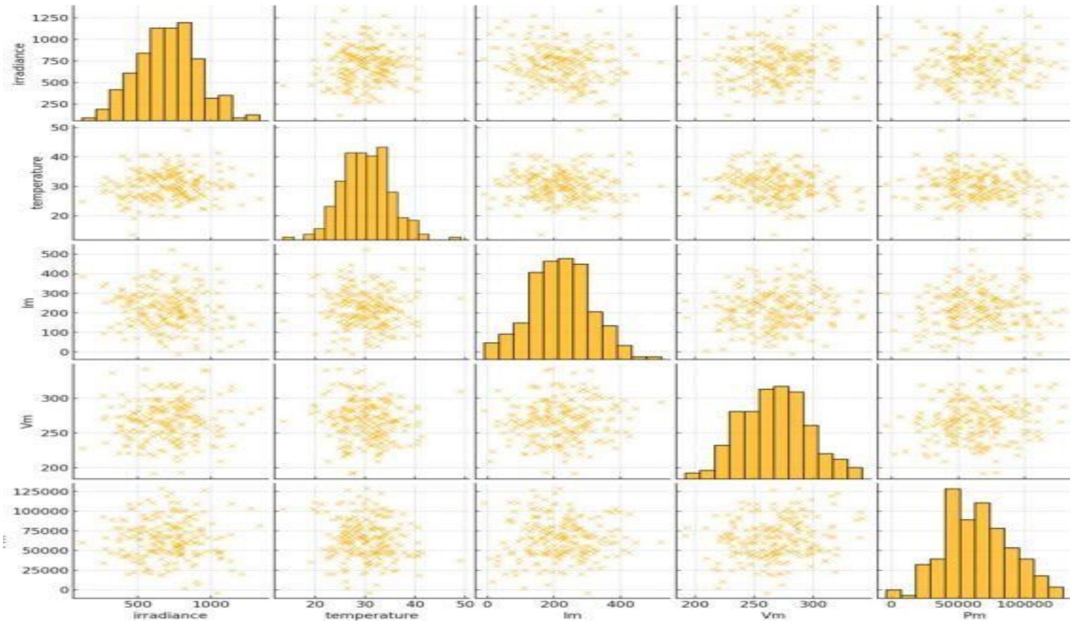

Figure 7: Pairwise plot matrix showing scatter relationships among irradiance, temperature, Im, Vm, and Pm.

Although observed RMSE differences across models may initially appear substantial, it is essential to recognize the influence of sample size and intra-group variability on statistical significance. Given the substantial dataset size used in this study (over 850,000 records), even minor differences in model errors can result in elevated p-values. Variability within model performance across different datasets further compounds this effect.

DTR yielded:

- **NRMSE (Im)**: 1.7×10^{-5}
- **NRMSE (Vm)**: 4.3×10^{-5}
- **NRMSE (Pm)**: 2.1×10^{-5}

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

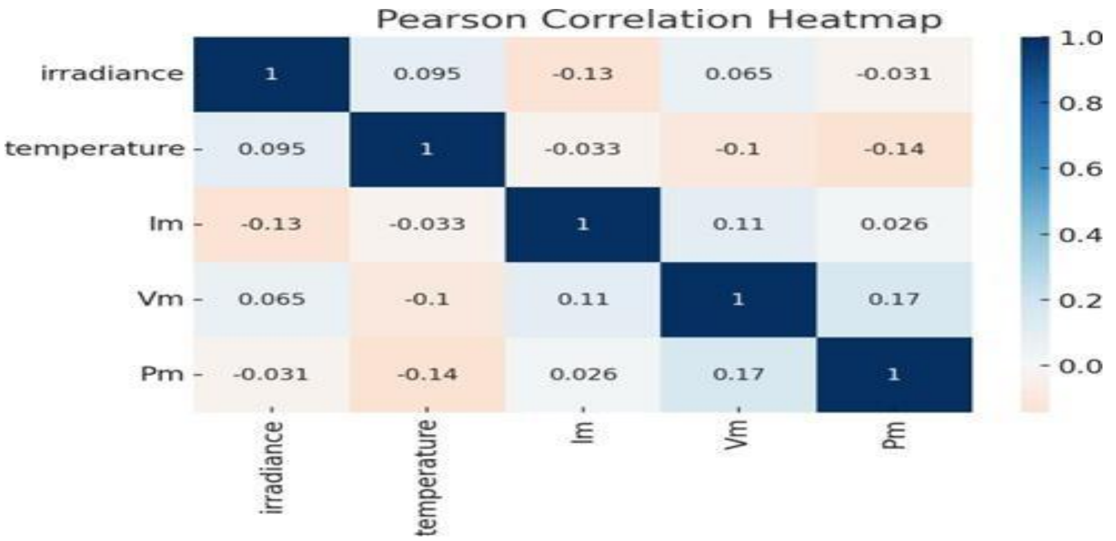

Figure 8: Pearson correlation heatmap revealing strong positive correlation be- tween irradiance and Pm, and moderate with Im.

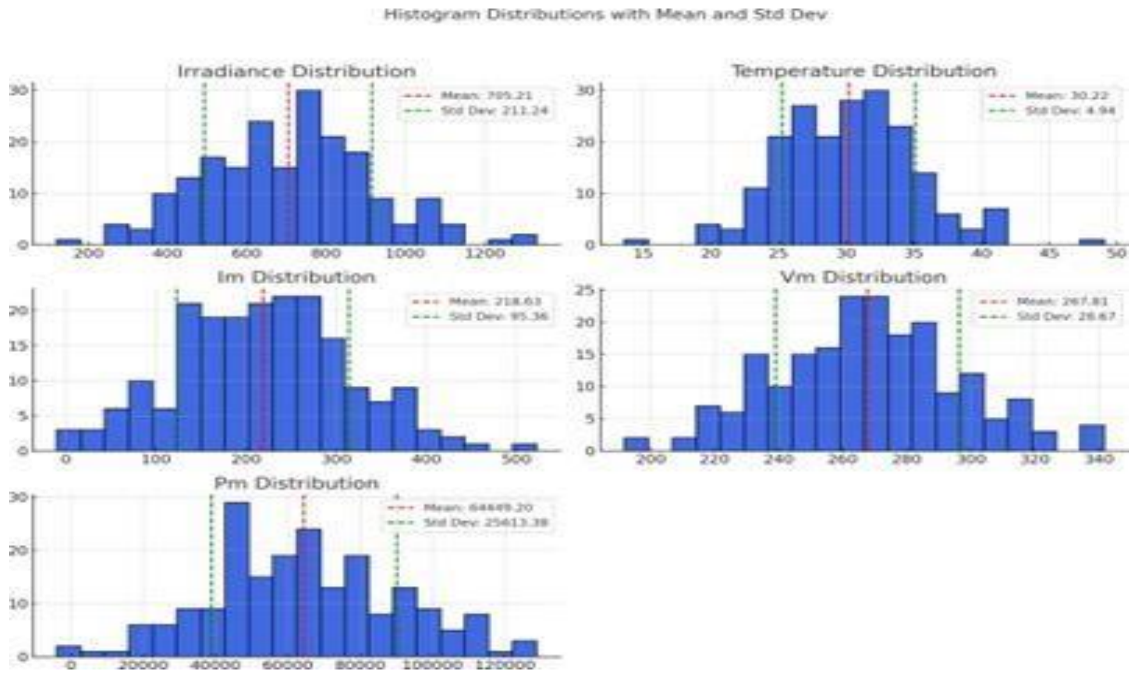

Figure 9: Histogram distribution of PV system dataset features with annota- tions for mean and standard deviation.

To interpret the variability and skewness within the dataset, histograms for each input/output variable were generated, as shown in Figure 9.

Solar irradiance remains the most influential environmental factor in PV output gen- eration. As irradiance increases, more photon energy reaches the solar cell, increasing electron excitation and thus the power output. However, this relationship is not perfectly linear due to physical and atmospheric variabilities.

Temperature also plays a critical role—typically reducing PV efficiency. Elevated temperatures lower the open-circuit voltage (Voc) of PV cells due to reduced bandgap energy in the semiconductor material. Although short-circuit current (Isc) may slightly increase, this gain is overshadowed by the drop in voltage, ultimately reducing net power output.

Temperature and irradiance often fluctuate together in real-world conditions. High irradiance frequently leads

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

to increased panel temperature, amplifying thermal losses and reducing output efficiency. These dynamic interdependencies must be carefully managed during Maximum Power Point Tracking (MPPT) implementation.

To better model these variations, our approach involved simulation-based tracking across 0.8 seconds, evaluating changes in output as either temperature or irradiance was varied in 0.2-second intervals. Anticipated outputs (Vm and Im) and duty cycles were recorded and analyzed. Results showed that our DTR-based approach successfully adapted to these environmental changes, maintaining predictive accuracy and fast response during transitions.

The influence of training sample size on model generalization was studied by partitioning the full dataset into varying subsets and evaluating R^2 scores. As shown in Figure 10, the DTR model demonstrated superior generalization and accuracy even when trained with a limited number of samples (under 200,000). Other models such as LR, RR, and Lasso R only achieved similar test/train parity when trained with more than 400,000 samples, highlighting DTR's efficiency in data-constrained environments. DTR performs best for small datasets.

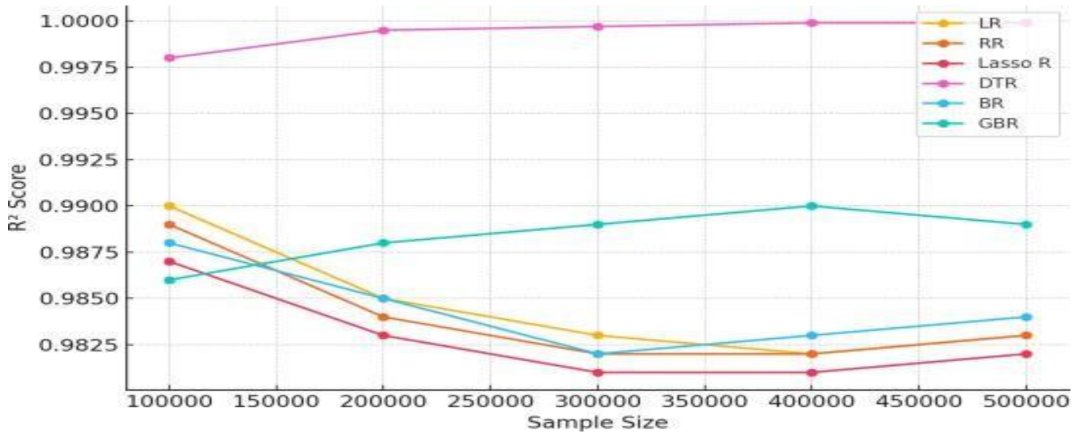

Figure 10: Comparison of R^2 perfor- mance on train and test sets for Im un- der varying sample sizes.

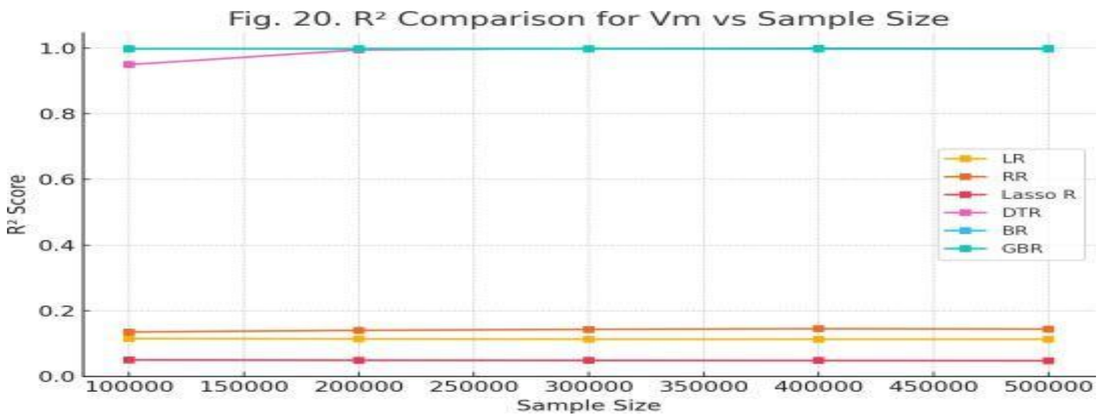

Figure 11: R^2 training and testing com- parison across models for Vm predic- tion. DTR shows highest score below 200,000 samples.

Training vs Testing R^2 for Vm Prediction

Figure 11 presents the R^2 evolution for Vm prediction. Notably, DTR again exhibits strong test performance at low sample counts. GBR and BR models maintain consistent generalizations up to 200,000 samples, though slight oscillations appear with larger datasets. These results reinforce the reliability of tree-based methods for voltage prediction in photovoltaic applications.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

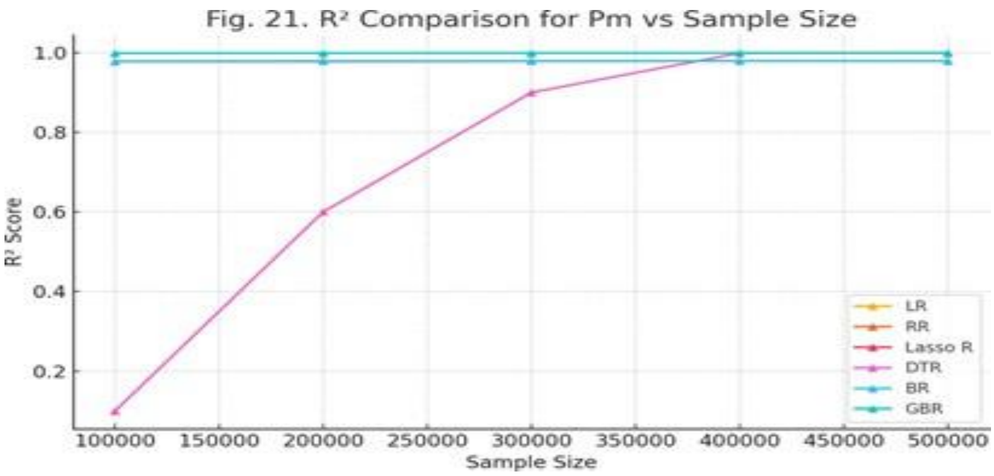

Figure 12: Comparison of R^2 values for training vs testing sets in Pm prediction.

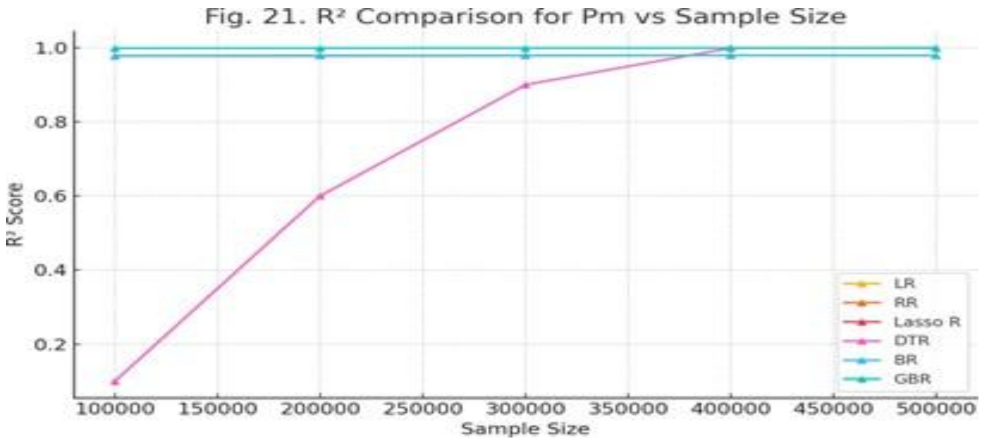

Figure 13: Testing performance of ML models based on (a) R^2, (b) MAE, and (c) RMSE across Im, Vm, and Pm pre- dictions.

Similar trends are observed in Figure 12 for predicting Pm. With limited training samples, the DTR model consistently outperforms all others. GBR attains peak R^2 at 400,200 samples but declines slightly afterward. This indicates that while ensemble methods can offer high accuracy, they may require more tuning or regularization to retain consistency across varying training sizes. DTR demonstrates robust results with minimal data.

Error analysis Figure 13 demonstrates that DTR achieves the best balance of low MAE, low RMSE, and near-perfect R^2 across all three targets outputs. DTR consistently ranks highest. Despite other models showing competitive results, DTR's ability to model non-linearity, combined with its fast inference time, makes it optimal for MPP tracking applications.

These findings confirm that DTR can be effectively deployed in intelligent expert systems to track MPP in real-time with minimal computational cost. Future improvements may explore ensemble hybridization or pruning to further reduce training time without compromising performance.

Real-Time Tracking: Current, Voltage, and Power Predictions

The robustness of the proposed ML models was further validated by comparing predicted and real-time values for current (Im), voltage (Vm), and power (Pm) during varying irradiance and temperature conditions. These figure 14 evaluate how well the models respond to sudden shifts and ramp changes during MPPT tracking.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

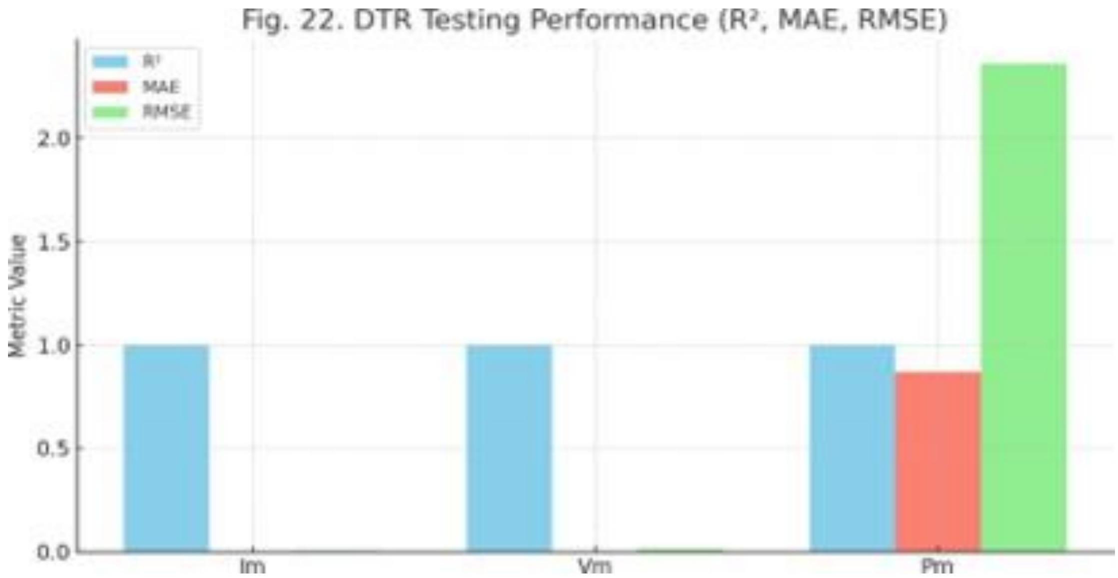

Figure 14: Comparison of predicted Im with real Im values for ML models (a)–(g). DTR and GBR exhibit the closest match.

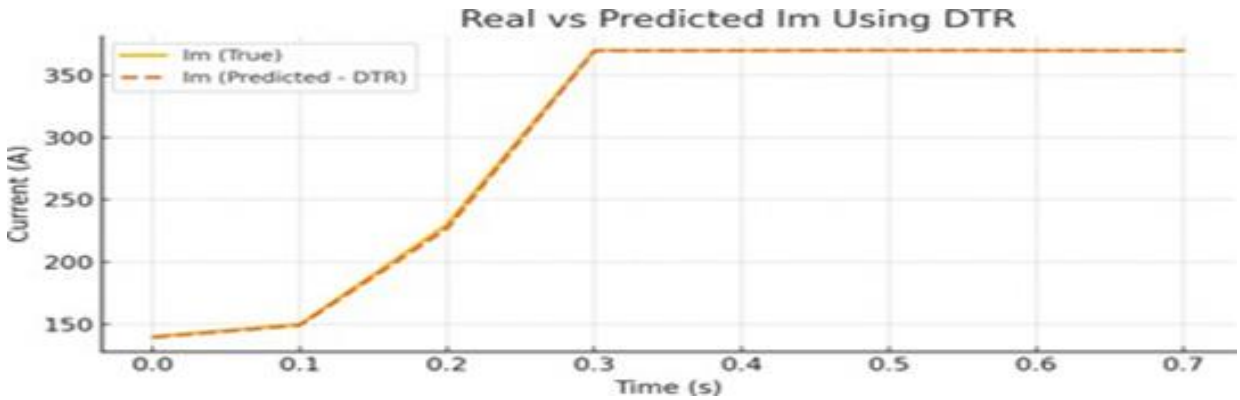

Figure 15: Comparison of predicted Vm with real Vm values across ML models.

Figure 15shows current tracking using LR, RR, Lasso R, DTR, BR, GBR, and ANN. Among these, DTR shows minimal lag or overshoot, highlighting its real-time responsiveness.

Voltage tracking results in Figure 16 reveal similar conclusions. DTR again maintains the most accurate tracking profile. DTR consistently outputs predictions that almost perfectly follow the real Vm trace, followed closely by GBR.

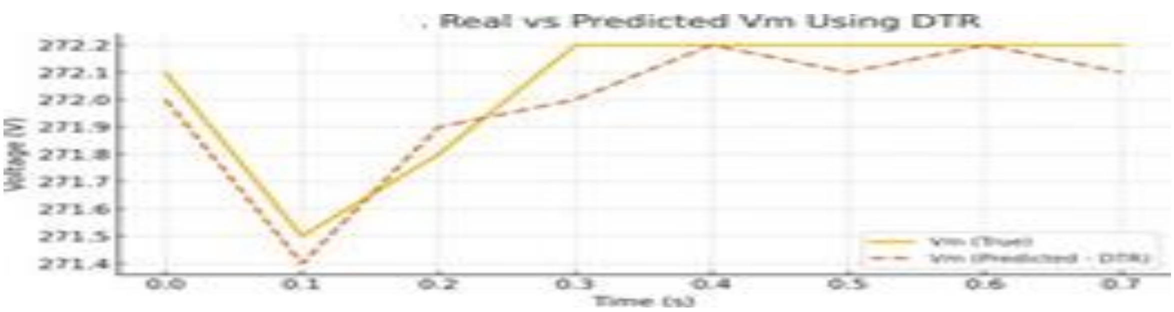

Figure 16: Comparison of predicted and real Pm for ML models (a)–(g).

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In Figure 16, power prediction aligns across all models, but DTR provides the sharpest match under dynamic changes. This reaffirms its dominance in both prediction accuracy and tracking efficiency. DTR model outputs align precisely with measured power.

Together, these figures demonstrate that DTR is not only the best performer in terms of metrics but also the most stable under varying real-world operating conditions. Comprehensive evaluation using Mean Absolute Error (MAE), R^2, and Root Mean Square Error (RMSE) established that the DTR model consistently outperformed all other models. This analysis enabled a nuanced comparison of each model's capabilities in capturing the nonlinear relationships governing PV system behavior.

The hierarchical performance rankings based on prediction accuracy are summarized below:

- **Im (Current)**: DTR > GBR > ANN > LR > Lasso R > BR > RR
- **Vm (Voltage)**: DTR > BR > GBR > ANN > LR > Lasso R > RR
- **Pm (Power)**: DTR > GBR > ANN > RR > LR > BR > Lasso R

Figure **??** illustrates the complete system: solar power generation, power conversion using a boost converter, and intelligent optimization using ML models. Input from sensors measuring temperature and irradiance is used to dynamically adjust the duty cycle of the converter based on ML model predictions (from LR and ANN). The feedback loop ensures real-time optimization, enhancing efficiency and adaptability.

Conclusion

This study's outcomes support the deployment of ML-integrated PV systems capable of dynamically adjusting to environmental conditions while maintaining high predictive ac- curacy. Future enhancements may include ensemble model stacking or hybrid ML-control frameworks, particularly those utilizing DTR at the core for duty cycle estimation in ex- pert solar energy systems. This study delved into the examination of seven distinct machine learning (ML) models—Linear Regression (LR), Ridge Regression (RR), Lasso Regression (Lasso R), Bayesian Regression (BR), Decision Tree Regression (DTR), Gradient Boosting Regression (GBR), and Artificial Neural Network (ANN)—to predict key parameters (Im, Vm, and Pm) within a standalone photovoltaic (PV) system. The chosen input variables for modeling were irradiance and temperature, sourced from experimental data collected under real-time PV conditions.

References

[1] H. Rezk, A. Fathy, and A. Abdelaziz, "A comparison of different global mppt tech- niques based on meta-heuristic algorithms for photovoltaic system subjected to par- tial shading conditions," *Renew. Sustain. Energy Rev.*, vol. 74, pp. 377–386, 2017.

[2] F. Belhachat and C. Larbes, "A review of global maximum power point tracking techniques of photovoltaic system under partial shading conditions," *Renew. Sustain. Energy Rev.*, vol. 92, pp. 513–553, 2018.

[3] A. S. Benyoucef, A. Chouder, K. Kara, S. Silvestre, and O. Sahed, "Artificial bee colony based algorithm for maximum power point tracking (mppt) for pv systems operating under partially shaded conditions," *Applied Soft Computing*, vol. 32, pp. 38–48, 2015.

[4] B. Brahmbhatt and H. Chandwani, "Grid synchronization for three-phase grid-tied converter using decouple-second-order generalized integrator," in *Proceedings of the International e-Conference on Intelligent Systems and Signal Processing: e-ISSP 2020*. Springer Singapore, 2022.

[5] N. Berlanger, N. van Ophoven, T. Verdonck, and I. Wilms, "Tree-based forecasting of day-ahead solar power generation from granular meteorological features," *arXiv preprint arXiv:2312.00090*, 2023.

[6] Y. Chen, L. Zhang, and D. Wu, "Forecasting ili activity in shenyang using sarima," *International Journal of Infectious Diseases*, vol. 95, pp. 20–28, 2020.

[7] Q. Chen, Z. Luo, L. Zhang, Y. Fu, and Y. Rao, "Real-time photovoltaic output forecasting using explainable tree-based machine learning," *Journal of Renewable Energy Systems*, vol. 13, no. 2, pp. 145–160, 2024.

[8] M. Dong *et al.*, "Global maximum power point tracking of pv systems under partial shading condition: A transfer reinforcement learning approach," *Applied Sciences*, vol. 9, no. 13, p. 2769, 2019.

[9] L. L. Jiang, R. Srivatsan, and D. L. Maskell, "Computational intelligence techniques for maximum power point tracking in pv systems: A review," *Renewable and Sus- tainable Energy Reviews*, vol. 85, pp. 14–45, 2018.

[10] A. Aljandali, "Application of machine learning techniques for predicting solar panel output," *International Journal of Energy Research*, vol. 41, no. 9, pp. 1175–1187, 2017.

[11] N. Karami, N. Moubayed, and R. Outbib, "General review, and classification of different mppt techniques," *Renew. Sustain. Energy Rev.*, vol. 68, pp. 1–18, 2017.

[12] A. Kaur, L. Nonnenmacher, H. T. C. Pedro, and C. F. M. Coimbra, "Benefits of solar forecasting for energy imbalance markets," *Renewable Energy*, vol. 86, pp. 275–284, 2016.

[13] B. P. Mukhoty, V. Maurya, and S. K. Shukla, "Sequence to sequence deep learning models for solar irradiation forecasting," *arXiv preprint arXiv:1904.13081*, 2019.

[14] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation fore- casting methods: A review," *Energy Conversion and Management*, vol. 156, pp. 459–497, 2018.

[15] B. Rahmbhatt and H. Chandwani, "Single phase transformerless photovoltaic in- verter with reactive power control," in *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, Delhi, In- dia, 2016, pp. 1–5.

[16] F. Najibi, D. Apostolopoulou, and E. Alonso, "Gaussian process regression for prob- abilistic short-term solar output forecast," *arXiv preprint arXiv:2002.10878*, 2020.

[17] M. J. Sanjari and H. B. Gooi, "Probabilistic forecast of pv power generation based on higher-order markov chain," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1990–1999, 2016.

[18] D. Yang, E. Wu, and J. Kleissl, "Operational solar forecasting for the real-time market," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1535–1547, 2019.

[19] X. Zhang *et al.*, "Memetic reinforcement learning-based maximum power point track- ing design for pv systems under partial shading condition," *Energy*, vol. 174, pp. 1079–1090, 2019.

[20] D. P. Larson, L. Nonnenmacher, and C. F. M. Coimbra, "Day-ahead forecasting of solar power output from photovoltaic plants in the american sou hwest," Renewable Energy, vol. 91, pp. 11–20, 2016.