Acta Scientiarum
http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v47i1.70900

Improving prediction of anti-reflective film-coated photovoltaic solar panel efficiency by integrating bayesian with machine learning algorithms

Durairaj Sankaran[1*], Balakrishnan Pappan[2] and Sathiya Selvaraj[3]

[1*]Department of Electrical and Electronics Engineering, Annaporana Engineering College, Salem, Tamil Nadu, India. [2]Department of Electrical and Electronics Engineering, J. J. College of Engineering & Technology, Ammapettai, Tiruchirappalli, Tamil Nadu, India. [3]Department of Chemistry, Selvam College of Technology, Namakkal, Tamil Nadu, India. *Author for correspondence. E-mail: durairajeeebe@gmail.com

ABSTRACT. This manuscript explores machine learning models for predicting the efficiency of anti-reflective film-coated photovoltaic solar panels in the South Indian climate. Three models—standard artificial neural network (ANN), random forest algorithm, and multilinear regression—were developed and compared. 80% of the dataset was used for training and 20% for testing. The random forest model demonstrated superior effectiveness with a lower prediction error. Bayesian optimization refined both ANN and random forest models. Experiments yielded an average solar panel efficiency of 16.79%, with performance indicators (coefficient of determination) of 0.96966, 0.93466, 0.98419 and mean absolute percentage errors of 7.518, 10.658, 5.089%. Bayesian optimization improved the traditional ANN model by up to 36.2%. Random forest exhibited lower sensitivity to hyperparameters compared to ANN. Two important parameters such as coating thickness and solar insolation were identified by feature sensitivity analysis. In terms of accuracy and robustness, random forest outperformed in predicting anti-reflective film-coated photovoltaic solar panel efficiency.

Keywords: Machine learning; bayesian algorithm; random forest; multi regression algorithm.

Received on January 4, 2024.
Accepted on March 15, 2024.

Introduction

Amidst current concerns, the push for lower emissions prompts the energy industry to embrace abundant, cleaner solar power (Boluk & Mert, 2014). Use of solar panels for electricity generation has gained importance as they are cost efficient, easy to install and use (Ureña-Sánchez, 2012). It is a green energy source, requiring minimum maintenance & can decrease carbon footprint. Solar panels, situated outdoors in direct sunlight, face varied atmospheric challenges like smoke, dust, rain, snow, shading, and dirt (Liu et al., 2020). Photovoltaic solar panel performance relies on environmental factors: solar insolation, air density, rainfall, surface temperature, wind conditions (De Jong et al., 2019). Anti-reflective coatings significantly enhance solar panel efficiency by reducing reflection, resulting in a 3 to 5% increase (Baker-Finch & McIntosh, 2011). Advancements include organo-metallic and hybrid organic-inorganic photovoltaic cells (Goh et al., 2007). Cooling systems and water spraying improve performance (Laseinde & Ramere, 2021). Sol-gel processed ARC coatings, particularly PTFE-modified silica hydrosols, exhibit superior properties (Foorginezhad & Zerafat, 2019). Coating thickness, density, and porosity impact effectiveness (Osorio et al., 2011). Various methods, including spray coating and thermal spraying (Syafiq et al., 2018), are employed, with room temperature spray coating offering cost-effective, even coverage. Various solar panel performance models are suggested, including polynomial differential equations (Raisee et al., 2015), regression models (Song et al., 2021), and AI and ML integration (Hong et al., 2020). ANN, simulating human brain neurons (Maind & Wankar, 2014), forecasts global solar insolation (Bilgili & Ozgoren, 2011). Random Forest, a supervised ML algorithm (Kulkarni et al., 2022), amalgamates diverse decision trees, enhancing predictive performance (Ahmad et al., 2018). As multilinear regression is easy to understand, straightforward, and predicated on linearity, it is a good option for forecasting solar panel efficiency in particular situations. It is computationally efficient and provides a clear picture of how each parameter affects efficiency. For small datasets, the decreased risk of overfitting is advantageous (Maulud &

Abdulazeez, 2020). Application domains include engineering (Farooq et al., 2020)] and nanoscience (Basheer et al., 2019). Regression is underutilized in PV predictions; ML improves accuracy (Agbulut et al., 2021). In order to solve sensor failure problems in data centre cooling systems, Wang et al. (2024) developed a successful Hybrid Multi-Label Random Forest and Bayesian Inference technique. For a more concise and accurate diagnosis of diabetes, Chen et al. (2024) proposed an approach combining a modified random forest with a deep neural network. Using a hybrid GA-ANN and correlation technique, Udaybhanu et al. (2024) estimated laminar burning velocity for isooctane blends with a high degree of accuracy. Machello et al. (2024) used tree-based machine learning to model tensile strength retention in Fibre Reinforced Polymer composites exposed to high temperatures. This study employs ML to forecast anti-reflective film-coated solar panel efficiency and enhances hyperparameter adjustment models. ANN, RF, and MLR algorithms were compared. For this purpose, anti reflective film coated solar panel test kit was installed on the roof top of Department of Mechanical Engineering at University College of Engineering, Tiruchirappalli, Tamil Nadu, India (Latitude - 10.6581° N, Longitude - 78.7439° E, 88 m approx. above sea level). Experiments were conducted during May and June (60 days) (summer season in south India). The prediction accuracy of the developed models for solar panel efficiency was evaluated based on the measured experimental values. A 5 fold cross validation procedure was adopted to train the prediction models with 20% of the experimental data. The predicted values were compared with the experimental values and the prediction accuracy of the models was ascertained.

Photovoltaic solar panel, anti reflective film and spray coating solar panel test kit

The solar panel test kit from Vinamara Enterprises was used, along with sensors for temperature, humidity, wind, and solar irradiation. Anemometer, precipitation, and indigenously fabricated mounting structures were employed. Dust cleaners and air blowers ensured panel cleanliness. Measurements included open circuit voltage and short circuit current using a digital multi-meter. The experiments were conducted on a terrace, strategically placed away from shade, and under South Indian climatic conditions. The equipment used included a digital temperature sensor (Evelta), humidity measurement tools, a solar irradiation sensor (Sims Instrumentation), and a tipping rain gauge sensor (Balaji Hydromet). The study aimed to enhance solar panel efficiency.

Anti-reflective coating preparation

The following chemicals were used for preparing antireflective PTFE (Polytetrafluoroethylene) modified silica hydrosols. 1.35 mm thick soda lime heat resistance glass sheets were bought from M/s. Akshar Exim Company Private Limited, West Bengal, India. (99% pure) Ammonium hydroxide and poly tetra fluro-ethylene (PTFE) were purchased from M/s. Ecokem Technologies Private Limited, Navi Mumbai, India. (99% pure) Tetra ethyl ortho silicate chemical was purchased from M/s. Tritech Catalyst & Intermediate, Pune. (99.9% pure) Silicon dioxide nanopowder and (98% pure) Hexamethyldisiloxane (HMDS) were bought from M/s. Supreme Silicones, Maharashtra, India. 99.9% pure ethanol and de-ionized water were purchased from M/s. Sigma Aldrich Chemicals, Bangalore, India. By using Stober method (Tadanaga et al., 2013), the pre-cursor solution for silica sol was prepared. In molar ratio of 1:0.17:38:0.17, Tetra ethyl ortho silicate, ethanol, ammonium hydroxide & de-ionized were mixed. The mixture was stirred using a magnetic stirrer at 25°C for 3h. Then, the mixture was transferred into a clean airtight glass container and stored in a dark place for one week. After one week, polytetrafluoroethylene was added to it and stirred for 3h using a magnetic stirrer at 25°C. Again, the solution was sealed to be air tight and stored in a cool and dark place for one week (Sun et al., 2020).

Glass surface preparation and spray coating

The solar panel's glass substrate underwent a thorough three-step cleaning process with de-ionized water, acetone, and hydrochloric acid to ensure optimal coating adhesion. Glass sheets were immersed in ethanol, heated to 80°C, and coated with a polytetrafluoroethylene-modified silica hydrosol solution using a Gautham Kit spraying equipment. The nozzle diameter was 1.25 mm, and spraying occurred at 40 psi pressure with a 100 mm distance from the substrate. After spraying, the glass dried in a portable furnace at 80°C for 10 minutes, resulting in a 75 to 90 nm thick coating per spray. The coating thickness was measured using a nano coat meter (Presice-UTM09), calibrated with standard nano films. This process iterated until

achieving the desired 350 nm coating thickness. The thickness of the anti reflective coating as 350 nm was selected from precious literatures and trial experiments (Oudir & Bourguig, 2024), (Sathya & Ponraj, 2024). On using thicker coatings, a significant decrease in surface transparency reduced sunlight transmittance, which in turn reduced solar panel output. On the other hand, panel performance did not significantly increase with thinner coatings. The 350 nm thickness optimizes solar panel output by matching surface transparency and anti-reflective efficacy, based on trends seen in experiments. Before panel use, the glass immersed in Hexamethyldisiloxane solution for 2 days enhanced hydrophobic characteristics (Huh et al., 2019). The solar panel test setup is shown in Figure 1. The efficiency of the solar panel was calculated using the formula

$$\eta = \frac{OutputPower}{InputPower} = \frac{P_{max}}{P_{inp}} \tag{1}$$

$$P_{max} = V_{max} \text{ x } I_{max} \tag{2}$$

$$P_{inp} = \text{Input solar Irradiance x Area of solar cell} \tag{3}$$

In the above equations η is the efficiency of the solar panel, P_{max} is the maximum power of the solar panel, P_{inp} is the power generated in the solar panel, V_{max} is the maximum voltage of the solar panel, I_{max} is the maximum current of the solar panel

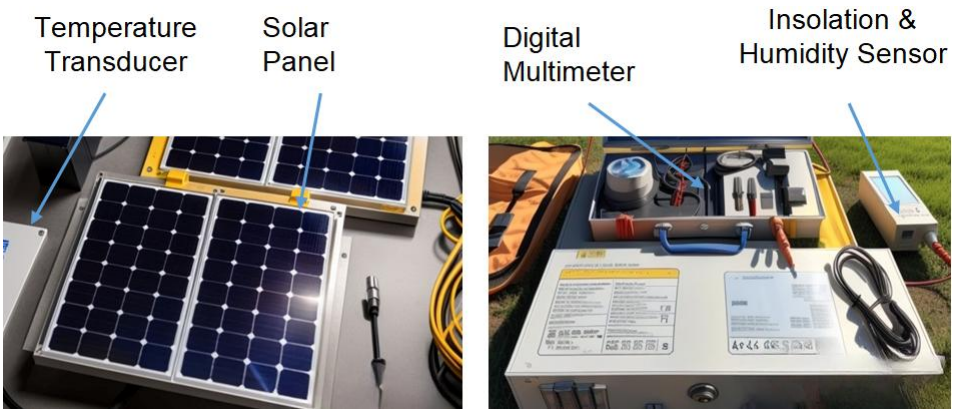

Figure 1. Solar panel test kit used for conducting the experiments

The important technological parameters affecting the output efficiency of the solar panel were identified to be coating thickness in nm (CT), solar irradiation at the panel surface in Wh m^{-2} (SIS), temperature in the upper surface of the solar panel in °C (T_{Top}), temperature in the bottom side of the solar panel in °C (T_{rear}), wind velocity in m/s (WV), relative humidity in % (RH). The output (panel efficiency) was calculated from the experimental data and the daily average efficiency has been indicated as η.

Machine learning methodologies

The data set of the anti reflective film coated solar panel is given as

$$D = \{A, b\}_{1:n} \tag{4}$$

In the above equation, b is the objective value (coated PV solar panel efficiency) correlated to the coated PV solar panel input parameter A. A denotes CT, SIS, T_{Top}, T_{rear}, WV or RH, n denotes the total number of experimental data samples. For K number of input parameters, the data set is given as follows.

$$D \in R^K \tag{5}$$

For construction of ML algorithms using python language packages (Raschka & Mirjalili, 2019), the following details were used.

Random forest model, design & development

Leveraging decision trees, the CART algorithm enhances prediction accuracy, combining with Random Forest in this research for superior results (Antipov & Pokryshevskaya, 2012). CART (Classification and Regression Trees) was chosen over Gradient Boosting Machines (GBM), CHAID (Chi-squared Automatic

Interaction Detection), and MARS (Multivariate Adaptive Regression Splines) for its binary tree structure and versatility in handling both classification and regression tasks. For error correction, MARS creates piecewise linear functions, CHAID uses chi-squared tests for categorical data, and GBM builds trees one after the other. CART emphasizes a balance between interpretability and performance in decision tree algorithms, making it the preferred method for the particular characteristics of the data and prediction goals due to its simplicity and efficacy (Al-Janabi, 2015). When it comes to solar panel efficiency prediction, the Random Forest algorithm outperformed Support Vector Machines (SVM) and Gradient Boosting Machines (GBM). Its advantages include managing huge feature sets effectively, using ensemble learning, avoiding over fitting by aggregating predictions across trees, and handling non-linear relationships. On the other hand, complex interactions, over fitting, and feature management would be difficult for SVM and GBM (Ghafouri-Kesbi et al., 2016). The three step process involving the combined CART and RF process is shown in Figure 2.

Manipulation of dataset

Initially, from dataset D, s-tree sets of data were generated randomly by using bootstrap re-sampling technique (Liu et al., 2005).

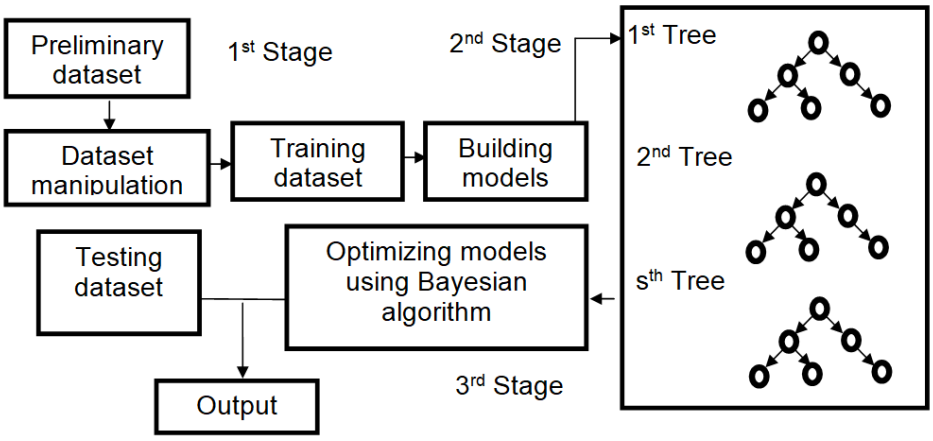

Figure 2. RF sequence used in this investigation.

Data were split into testing and training sets. Scikit-learn clustering normalized data to 0 and 1 for improved convergence and reduced variance. There are several benefits of using Scikit-learn clustering for data normalization over other approaches. With the use of tools like {StandardScaler} and `MinMaxScaler}, Scikit-learn guarantees better convergence and reduces the influence of varied scales on clustering methods. Together with log transformation, the adaptability of choices like `RobustScaler`, `PowerTransformer`, and `QuantileTransformer` meets the demands of different clustering algorithms and data characteristics. With the help of this all-inclusive suite, users can easily experiment with various normalization algorithms, which make it easier to make an informed decision based on certain dataset properties and to optimize clustering results (Akhatov et al., 2017).

Building Models

Post-data manipulation, s-tree datasets spawned unique decision trees through distinct random algorithms. Ensuring diverse performance, s decision trees were averaged for the final random forest model, enhancing accuracy and reducing variance.

Optimizing models using Bayesian technique

Bayesian optimization is superior for enhancing solar panel efficiency prediction models due to its efficient exploration-exploitation tradeoff, sample efficiency, and ability to handle noisy objective functions. It is highly efficient in global optimization and automatic hyperparameter adjustment while avoiding local optima. Its continuous improvement pattern is beneficial in high-dimensional hyperparameter spaces (Abdolrasol et al., 2021). Bayesian optimization automatically adjusted hyperparameters, preventing data overfitting while maintaining prediction accuracy in 5-fold cross-validation on training sets (Mahendran et al., 2022). The highly accurate RF model validated, assessed

attribute significance using the Gini index to identify contamination, aiding impurity reduction (Blanco et al., 2000). The Gini index is favored over other impurity measures like misclassification error and variance reduction due to its effectiveness in decision trees and Random Forests (Gwetu et al., 2014). In order to encourage the development of pure nodes, the Gini index measures the likelihood of incorrectly classifying a randomly selected element within a node. Its formula makes evaluating node impurity simple and effective. Scikit-learn defaults to the Gini index, highlighting its performance and wide applicability, even if other metrics like entropy and misclassification error are legitimate. On the basis of particular problem features and modelling objectives, it is advised to experiment with different criteria (De'ath & Fabricius, 2000). They were calculated using the equations given below

$$S_e = \sum_l GN_e -_{before}(t) - GN_{e-after}(t) \quad (6)$$

In the above equation, S_e indicates the significance of feature e, $GN_{e\text{-}before}$ (t) is the Gini index of node t before node division using descriptor e, $GN_{e\text{-}after}$ (t) is the Gini index of node t after node division using descriptor e. The value of Gini index is calculated according to the following equation

$$GN(t) = 1 - \sum_{m \in M} \gamma_m^2 \quad (7)$$

In the above equation GI(t) indicates the Gini index of node t and γ indicates the relative frequency of class m in node t

Artificial neural network model

ANN uses probability encumbered associations with no presumptions, for developing prediction models. Hence ANN helps in reliable model development for non-linear systems (Sajikumar & Thandaveswara, 1999). A schematic representation of the ANN model developed in this investigation is shown in Figure 3. The ANN architecture consists of the initial input layer, intermediate hidden layers and the final output layer. 6 nodes were used in the input layer which corresponds to the six important input technological features of the anti reflective film coated solar panel (CT, SIS, T_{Top}, T_{rear}, WV and RH). The objective value of the solar panel output was represented in the output layer as η. ANN established a non-linear relationship between input parameters and output efficiency through activation functions and hidden layer interconnections (Ozkan & Erbek, 2003). The following sequence was adopted for training the ANN model

The dataset D was proportionally divided into testing data sets and training data sets. Initially a random value was assigned for the ANN weights. Then, the training data set was transmitted throughout the network and the value of solar panel efficiency η was acquired.

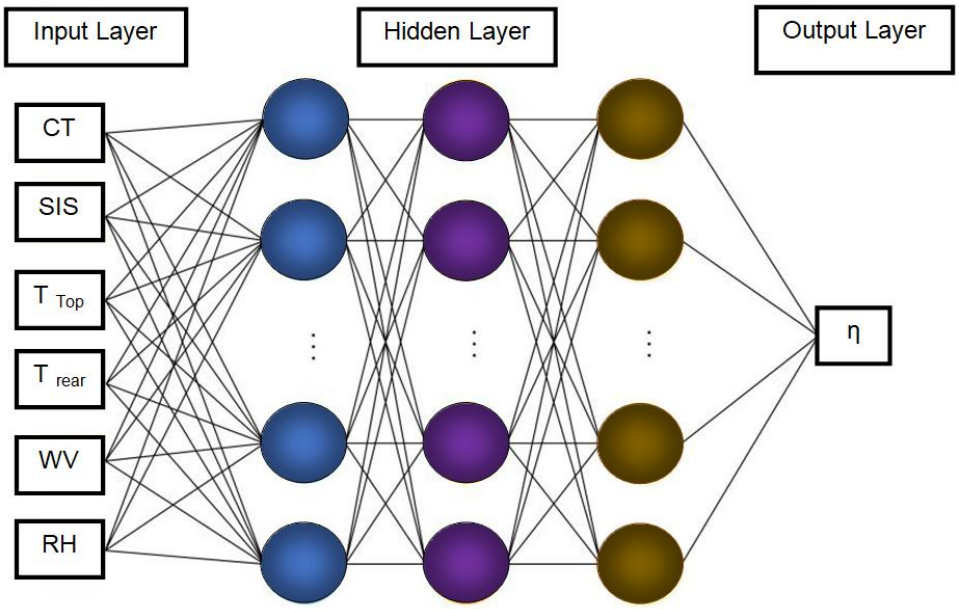

Figure 3. Schematic representation of the developed ANN model.

The following equation was used for calculating the error $\mathscr{E}$ in a data sample (a_t, b_t)

$$E_t = \frac{1}{2}\left(b_0 - b_t\right)^2 \tag{8}$$

In the above equation, the value of the final layer is indicated as b_0.

The weights between hidden - output and input – hidden layer was assumed as g_{eb} and h_{af}, respectively. The quantity of neurons in the hidden layer was e and f. Further, back propagation method was implemented to circulate the $\mathscr{E}$ error all over the network (Chang & Chao, 2006). When it comes to convergence acceleration, backpropagation outperforms Mini-batch Gradient Descent and Stochastic Gradient Descent (SGD) techniques. By effectively descending the error surface using gradient descent and adjusting to intricate patterns in the training data, it optimizes neural network weights (Haji & Abdulazeez, 2021). Backpropagation acts as a universal approximator, allowing the network to extract meaningful representations from buried layers and learn non-linear mappings. Backpropagation is preferable in terms of convergence speed for a variety of neural network topologies and applications due to its systematic approach and adaptability, while other approaches such as SGD and Mini-batch Gradient Descent introduce stochasticity (Premalatha & Valan Arasu, 2016). It was done for adjusting weights within the nodes. Using the following equation, the value of g_{eb} was calculated.

$$g_{eb\_new} = -\nabla \frac{\partial E_t}{\partial g_{eb}} + g_{eb} \tag{9}$$

$$\frac{\partial E_t}{\partial g_{eb}} = \frac{\partial E_t}{\partial b_0} \times \frac{\partial b_o}{\partial \omega_e} \tag{10}$$

In the above equation, the input value of the output layer is indicated by ω_e.

Using the following equation, the value of h_{af} was calculated.

$$h_{af\_new} = -\nabla \frac{\partial E_t}{\partial \tau_f} \times \frac{\partial \tau_f}{\partial \delta_f} + h_{af} \tag{11}$$

In the hidden layer, the output of the f^{th} neuron is indicated as τ_f, whereas the input of the f^{th} neuron is indicated as δ_f. Hidden layers are essential for feature transformation, non-linearity, and hierarchical abstraction in artificial neural networks (ANNs). Specific patterns are recorded by each neuron, enabling the encoding of complicated relationships (Agatonovic-Kustrin & Beresford, 2000). Compression and dimensionality reduction are made possible by hidden layers, which result in a lower-dimensional representation that keeps all relevant data. This condensed form facilitates generalization to new data while improving efficiency. The flexibility of the layer to adjust to complex dependencies is essential for tasks such as natural language processing and picture recognition. Choosing the right learning rate ($\nabla\epsilon$) affects convergence accuracy and speed, which is essential for effective training and the best avoidance of solutions (Arif et al., 2024). The learning rate of ANN is indicated as $\nabla\epsilon(0,1)$. High ∇ values imply rapid convergence, lost in localized optimization. Low ∇ values signify slow convergence, enhancing accuracy.

The sequence 1 and 2 was repeated till the final state was attained.

The activation function, hidden layers (k_t), and neurons (k_r) impact ANN model complexity. Increasing layers and neurons heightens complexity, risking overfitting; overly simple models risk underfitting (Wei, 2021). The hidden layer of artificial neural networks (ANNs) is essential for hierarchical abstraction, non-linearity, and feature transformation, which allow ANNs to adapt to challenging tasks. The outputs of neurons affect the learning rate ($\nabla\epsilon$), which affects the rate of convergence. While a low rate prioritizes precision and slows convergence, a high rate speeds up convergence but increases the chance of entrapment (Attoh-Okine, 1999). High expertise is required for evaluating the values of ∇, k_t and k_r. Bayesian optimization method was used for identifying the final values of ∇, k_t and k_r, as 0.0021, 2.76, (63, 221, 726) respectively. In this experimental process, Relu function was used as the activation function (Schmidt-Hieber, 2020). Due to its non-linearity, the Rectified Linear Unit (ReLU) activation function is preferred in neural networks for modelling complex interactions. ReLU accelerates training compared to other functions because of its computational efficiency. It ensures efficient gradient flow by mitigating the vanishing gradient issue in deep networks (Banerjee et al., 2020).

Bayesian optimization

Bayesian algorithm identified the global optimum performance of RF and ANN models by searching for the most suitable hyperparameters. This aided in choosing the best model for predicting solar panel efficiency (Wang & Jegelka, 2017). The technique minimizes evaluation costs of objective functions, utilizing a Gaussian model to fit examined observations and predict rear sample positions (Brochu et al., 2010). In this study, the 5-fold cross-validation output served as the Bayesian optimization objective function, mitigating overfitting errors. The choice of 5-fold cross-validation over LOOCV and Stratified Cross-Validation is advantageous for managing computational efficiency and reliable model assessment. Because LOOCV evaluates the model for each data point separately, it is computationally intensive. While it preserves the class distribution, stratified cross-validation might not be required for datasets that are balanced. By dividing data into five subsets, 5-fold CV, on the other hand, reduces variability and offers a fair compromise for model validation in a variety of machine learning applications while balancing computational cost and robust evaluation (Molinaro et al., 2005).

Bayesian optimization assigned input variables (tree quantity, neurons, hidden layers, and learning rate) as hyperparameters for RF and ANN models. The input values were indicated as A = q_1, q_2, q_3 ..... q_e, whereas the output variable was b. The output exhibited 5 fold cross validation level of accuracy. The acquisition function in Bayesian algorithm balances scrutiny and non-utilization, identifying consecutive maxima points. The algorithm employed the Upper Confidence Bound (UCB) acquisition function (Berk et al., 2022). Because it can efficiently balance exploration and exploitation, the Upper Confidence Bound (UCB) acquisition function is a recommended option in Bayesian optimization. UCB provides a well-adjusted trade-off between investigating points with high uncertainty and exploiting points with high anticipated mean. It does this by taking into account both the predicted mean and uncertainty (variance) of the surrogate model. In comparison to alternative acquisition functions such as Probability of Improvement (PI), Expected Improvement (EI), Lower Confidence Bound (LCB), Thompson Sampling, Entropy Search, and Probability Matching Acquisition (PMA), this balance improves the optimization process and makes UCB a strong option (Bian et al., 2021). Bayesian optimization process was executed according to the following steps.

By using the following equation, the Gaussian model (GM) was developed.

$$GM \sim K\left(O,U\right) \tag{12}$$

In the above equation, U is the kernel matrix. The kernel matrix is shown in the following equation

$$U = \begin{bmatrix} U\left(a_1,a_1\right),\ldots U\left(a_1,a_n\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots.. \\ U\left(a_n,a_1\right)\ldots U\left(a_n,a_n\right) \end{bmatrix} \tag{13}$$

$$K(a_e,a_f) = \exp\left\{-\frac{1}{2}\left[a_{e-}a_f\right]^2\right\} \tag{14}$$

Second, the next location b_{n+1} to sample, according to acquisition function, is determined at which the observation property M was expected to be the best (the accuracy of 5-fold cross-validation). Subsequently, the Gaussian process model M was updated by including the new observation b_{n+1}.

The two steps were repeated until the terminal condition is reached.

Prediction accuracy indicators

The prediction accuracy of the developed models was ascertained by using statistical and mathematical evaluation criteria. In this research, for different performance indicators were used. They were coefficient of determination (R^2), mean absolute error (MAE), mean square error (MSE) and mean absolute percentage error (MAPE) (Chicco et al., 2021).

Equation for calculating Mean absolute error (MAE) is shown below

$$MAE = \frac{1}{n}\sum_{e=1}^{n}\mathrm{ADV_e} - \mathrm{FDV_e} \tag{15}$$

Mean absolute percentage error indicates the percentage of deviation from the actual output [65].

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\mathrm{ADV_t} - \mathrm{FDV_t}}{\mathrm{A_t}}\right)\times 100\% \tag{16}$$

In the above equations, ADV_t denotes the actual data value, FDV_t denotes the forecasted data value and n is the number of true values.

Coefficient of determination (R^2) is used for prediction of future outcomes. For independent variables, R^2 is ascertained by using variance proportion.

$$b' = \frac{1}{n}\sum_{e=1}^{n}\mathrm{b_i} \tag{17}$$

In the above equation b' is the mean of data set b_i, n is the number of values in dataset b_i. SST is the total sum of squares and its equation is shown below

$$SST = \sum_{e=1}^{n}\left(\mathrm{by_i} - \mathrm{b'}\right)^2 \tag{18}$$

SSE is the error sum of squares, which is shown in the equation below

$$SSE = \sum_{e=1}^{n}\left(\mathrm{by_i} - \mathrm{f_i}\right)^2 \tag{19}$$

In the above equation, the predicted data value is f_i.

The error sum of squares and total sum of squares is used to determine the R^2 value. The equation of R^2 is shown below

$$R^2 = 1 - \frac{SSE}{SST} \tag{20}$$

R^2 gauges predicted data proximity to the regression line, ranging from 0% (no closeness) to 100% (all close). Mean Absolute Error (MSE) measures average error value.

$$MSE = \frac{1}{n}\sum_{e=1}^{n}\left(B_e - A_e\right)^2 \tag{21}$$

Result and discussion

Experimental results

By using the experimental data, recorded during the 60 days, the prediction models were developed. The important parameters measured were η (P_{max}, P_{inp}, V_{max}, I_{max}), CT, SIS, $T_{Top,}$ T_{rear}, WV and RH. The average values of the (4 per day) recorded observations are shown in Figure 4. The total number of samples, maximum, minimum, mean and standard deviation has been evaluated and shown in Table 1. Figure 4 reveals reduced coating thickness after exposure to the external environment for a few months. Solar insolation peaked on the 17th day and hit a minimum on the 59th day. The top surface of the solar panel consistently had higher temperatures than the rear side, except during continuous precipitation. Figure 5 displays observed efficiency variations (14.01 to 16.97%). Daily measurements of the coated efficiency of the solar panel were made for a period of two months. The average efficiency fluctuated, peaking at 16.99% on May 31, 2022, and falling to 14.012% on June 10, 2022. These variations reveal information about the panel's operation and suggest possible operational or environmental influences on its effectiveness.

Prediction results of machine learning models

Based on experimental observations, prediction models were developed using 80% of the dataset for training and 20% for testing, employing a 5-fold cross-validation technique. Figure 6 compares the constructed prediction performances of all models. Incorporating Bayesian algorithm for hyperparameter search yielded R^2 values close to 1 for both random forest (RF) and artificial neural network (ANN) models.

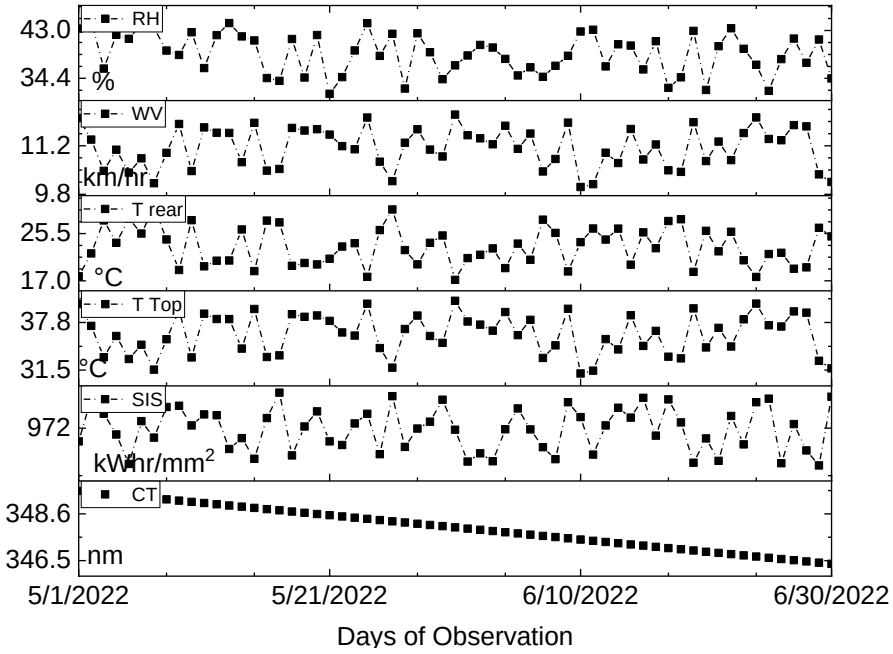

Figure 4. Daily observations of the coated solar panel process parameters.

Table 1. Evaluation of the observed data.

Input variables	Sample quantity	Maximum	Minimum	Mean	Standard Deviation
Coating Thickness (nm)	300	346.35	349.63	347.99	0.971
Solar Insolation (kwhr m^{-2})	300	1062.83	900.69	969.22	55.504
T at solar panel top surface (°C)	300	40.67	31.04	36.308	2.68
T at solar panel rear surface (°C)	300	21.21	27.16	23.11	3.47
Wind Velocity (km hr^{-2})	300	12.09	10.00	11.15	0.581
Relative humidity (%)	300	44.84	31.61	38.85	3.682
Panel efficiency	300	16.97	14.01	15.64	0.831

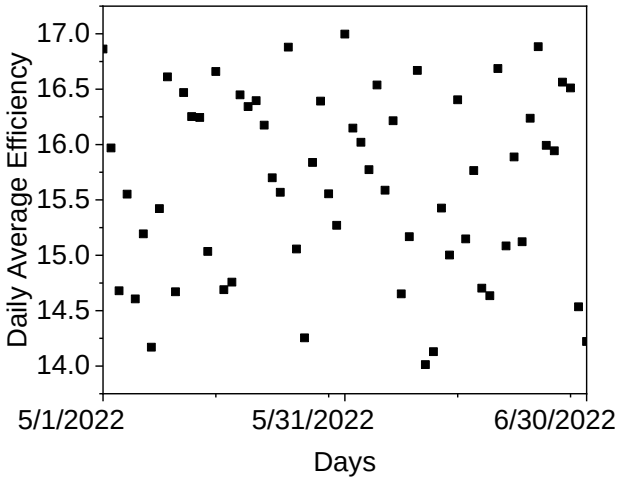

Figure 5. Daily average coated solar panel efficiency.

This proximity signifies the closeness between simulated and target results, indicating better generalization when using Bayesian optimization in both RF and ANN models. Minimal variations were observed in the results of the developed random forest model and Bayesian optimization incorporated RF model. Conversely, variations in the results of the developed ANN prediction model and Bayesian optimization incorporated ANN model were higher, indicating greater sensitivity to hyperparameters. Bayesian optimization helped identify maxima and minima in the machine learning with a black box model. The correlation between the predicted and actual values can be established by the prediction accuracy indicator values. For identifying the prediction performance of the developed models, the variations in the values of performance accuracy indicators (MAE, MAPE, MSE, R^2) are shown in Figure 7.

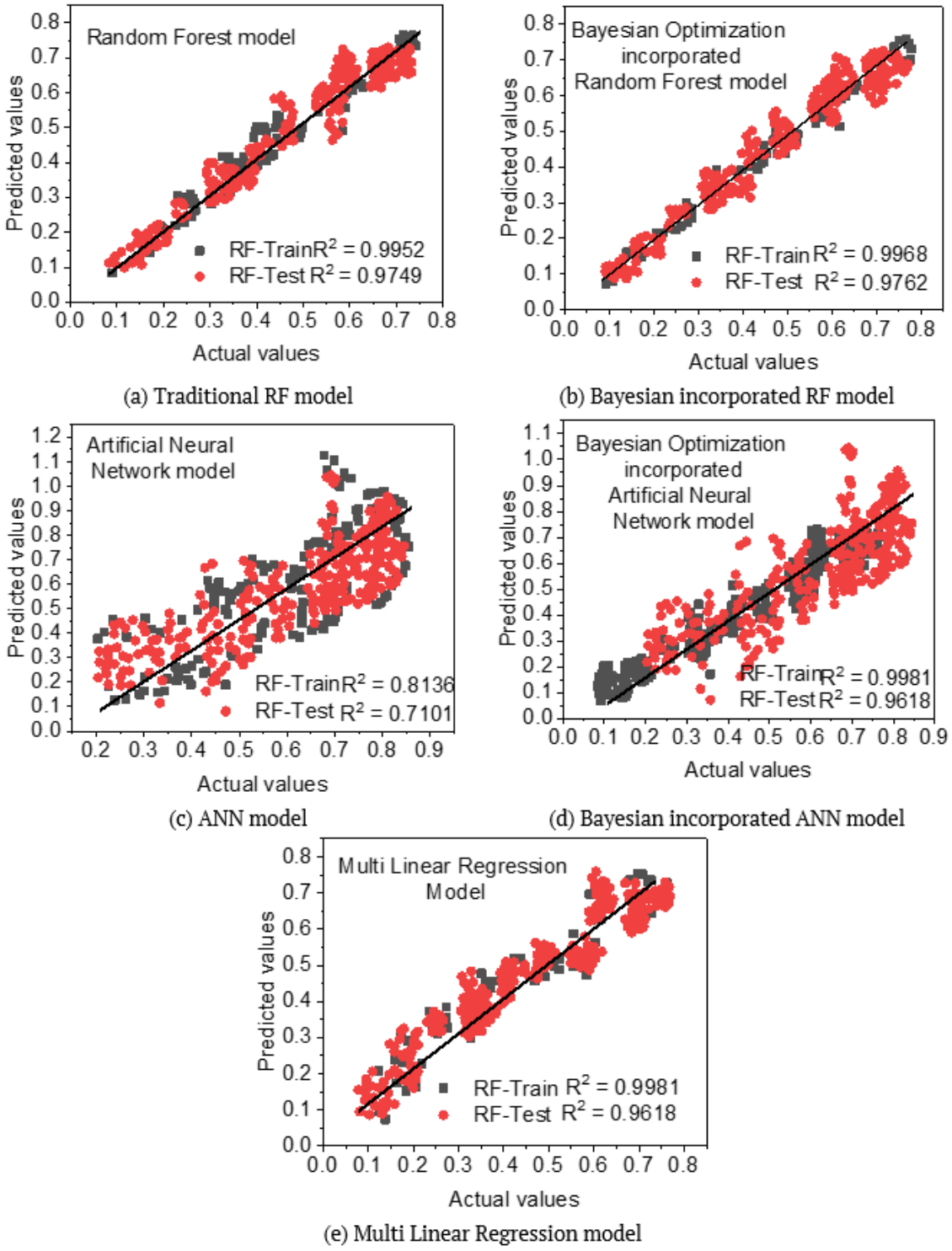

Figure 6. Co-relation between the predicted and actual results of different models.

Incorporating Bayesian optimization in the RF model showed minimal change in MAPE values for training and testing sets compared to the original RF model. Conversely, considerable variations were observed in MAPE values on incorporating Bayesian optimization in the ANN model. The performance of the Bayesian optimization-incorporated RF model surpassed that of the Bayesian optimization-incorporated ANN model. Evaluating R2 values, the Bayesian optimization-incorporated RF model exhibited values very close to 1, indicating high predictability. Studies on solar panel efficiency improvement using machine learning algorithms were scarce. ANN prediction models showed R^2 values between 0.91 to 0.96, signifying a high predictability level. Figure 7 depicted similar variation patterns in MSE, MAPE, and R^2, indicating high generalization performance and accurate prediction with minimal overfitting.

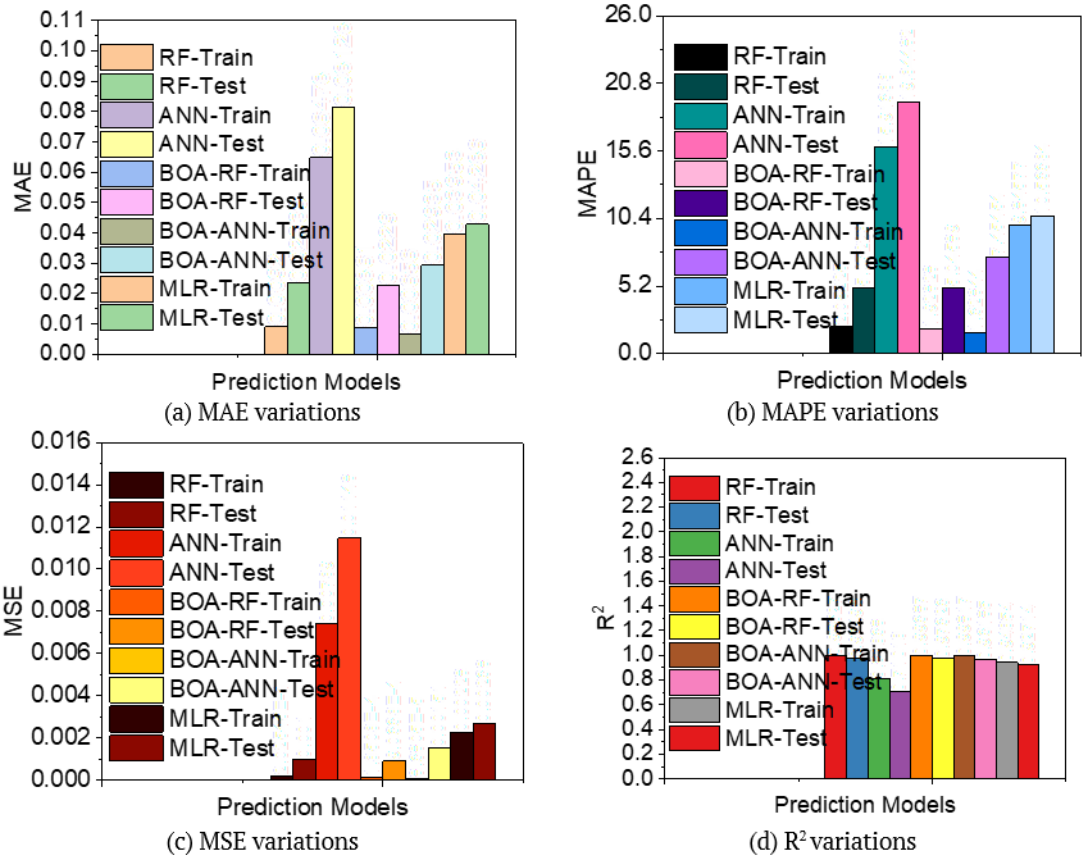

Figure 7. Variations in prediction accuracy indicators for various models.

Effect of the size of dataset

Due to dataset size limitations, variations in prediction accuracy were examined for test sizes (0.1, 0.2, and 0.3). Figures 8 demonstrate the prediction-actual correlation for Bayesian optimization-incorporated random forest and artificial neural network models. Increasing test sizes showed no remarkable difference between the models, indicating sufficient dataset convergence.

Prediction results of multi linear regression model

Multi-linear regression model was used for establishing relation between the independent input factors such as (CT, SIS, T_{Top}, T_{rear}, WV and RH) and output (coated solar panel efficiency). Using the same observation data used for other machine learning algorithms, multi linear regression model was developed for predicting the efficiency of photo voltaic solar panel (Table 2).

Table 2. Pearson Co-relation matrix for the multi linear regression model.

	η	CT	SIS	T_{Top}	T_{rear}	WV	RH
η	1	0.0217509	-0.100561	1	-0.949397	1	0.0774466
CT	0.0217509	1	0.0638014	0.0217509	0.0630666	0.0217509	0.278496
SIS	-0.100561	0.0638014	1	-0.100561	0.0947841	-0.100561	-0.192107
T_{Top}	1	0.0217509	-0.100561	1	-0.949397	1	0.0774466
T_{rear}	-0.949397	0.0630666	0.0947841	-0.949397	1	-0.949397	-0.121327
WV	1	0.0217509	-0.100561	1	-0.949397	1	0.0774466
RH	0.0774466	0.278496	-0.192107	0.0774466	-0.121327	0.0774466	1

Pearson Co-relation matrix for the multi linear regression model is shown in Table 3 and the developed regression equation is $\eta = 2.4048 + 0.04673\ CT - 0.001563\ SIS + 0.00128\ T_{Top} - 0.00631\ T_{rear} + 0.01258\ WV - 0.01236\ RH$ Equation 22.

Feature importance

From this investigation, it was observed that the prediction accuracy of random forest model was better than the other two prediction models such as artificial neural network model and multi linear regression model. The significance of the six factors (CT, SIS, T_{Top}, T_{rear}, WV and RH) on the output efficiency was evaluated. The effect of the six factors on the output has been indicated in descending order in Figure 9. The

most important input variables were identified as solar insolation and coating thickness. This is because they significantly affect the efficiency.

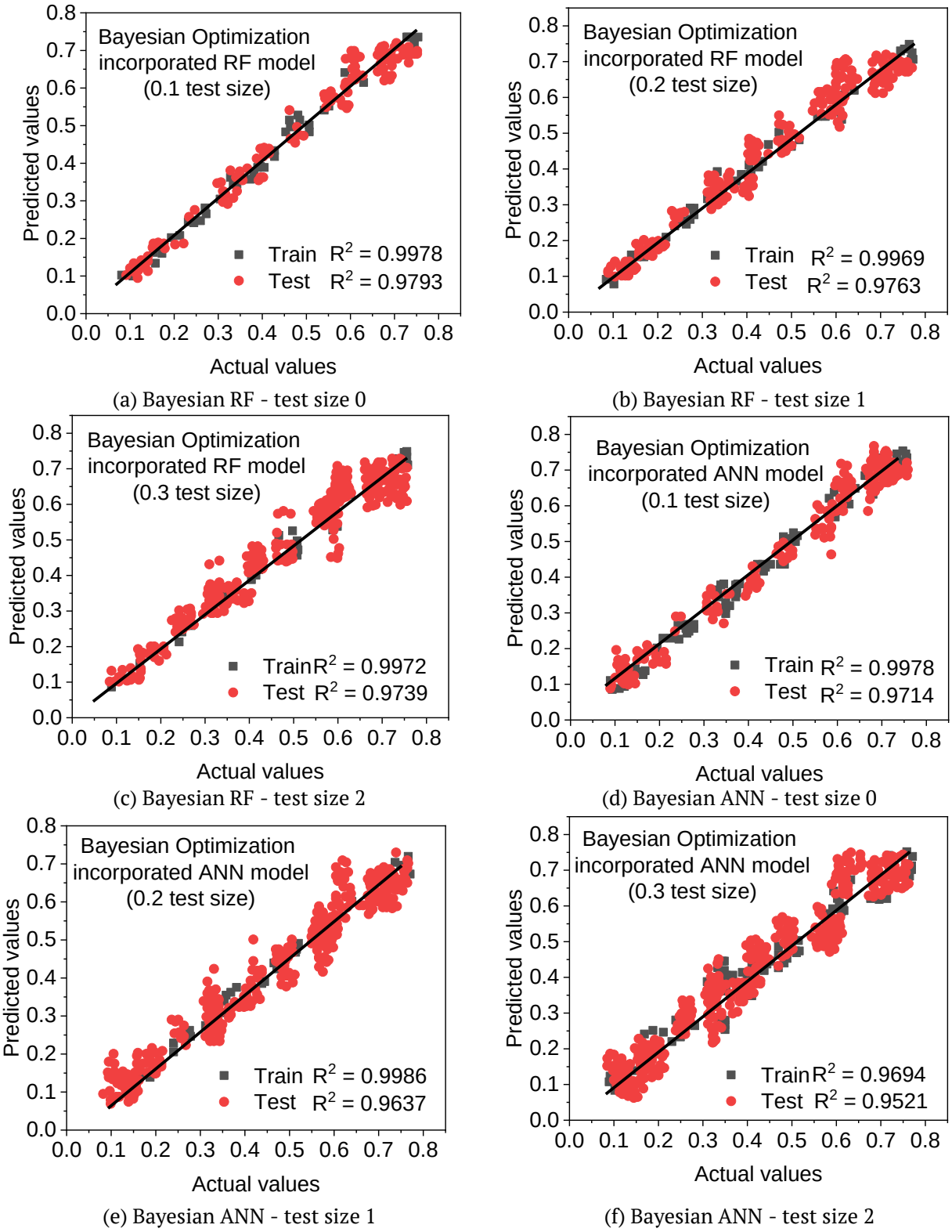

Figure 8. Predicted vs actual co-relation for the prediction models with various test sizes.

Temperature at the top side of the solar panel (T_{Top}), which comes in direct contact with sunlight, was found to have a greater significance in modifying the output efficiency, than the temperature at the rear side (T_{rear}) of the solar panel. Randomizing the data sets by altering the sequence resulted in very less modification of the output (1.1%).

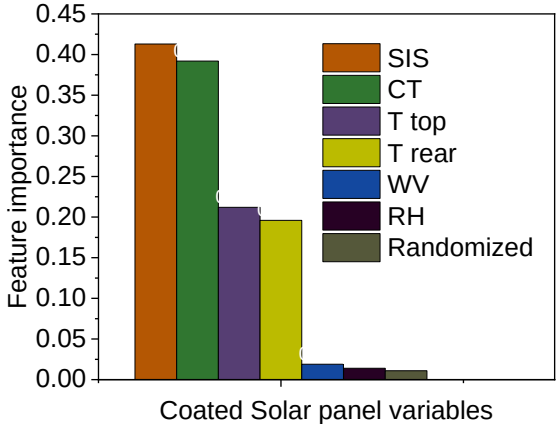

Figure 9. Sensitivity of the output to different input variables.

Conclusion

This investigation used random forest, artificial neural network, and multi-linear regression models for predicting the efficiency of photovoltaic solar panels coated with anti-reflective film. Bayesian optimisation was used to fine-tune the hyperparameters. Two months of climate measurements in South India produced an average daily efficiency of 16.79%. Strong accuracy was demonstrated by machine learning models throughout testing and training. Model performance was highlighted by mean square error and mean absolute percentage error analysis. Among the developed models, random forest model demonstrated better stability and reliability for predicting solar panel output efficiency.

Acknowledgements

The authors thank the assistance rendered by R & D Division, Vikram Engineering Industry, Vazhavandankottai, Tamil Nadu, India for conducting solar panel experiments.

References

Abdolrasol, M. G., Hussain, S. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., & Milad, A. (2021). Artificial neural networks based optimization techniques: A review. *Electronics*, *10*(21), 2689. https://doi.org/10.3390/electronics10212689

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717-727. https://doi.org/10.1016/S0731-7085(99)00272-1

Agbulut, Ü., Gürel, A. E., & Biçen, Y. (2021). Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renewable and Sustainable Energy Reviews*, *135*, 110114. https://doi.org/10.1016/j.rser.2020.110114

Ahmad, M. W., Reynolds, J., & Rezgui, Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of Cleaner Production*, *203*, 810-821. https://doi.org/10.1016/j.jclepro.2018.08.207

Akhatov, A. R., Renavikar, A., Rashidov, A. E., & Nazarov, F. M. (2017). Optimization Of The Number Of Databases In The Big Data Processing. *Databases*, *48*, 399-420. https://doi.org/10.24412/2073-0667-2023-1-33-47

Al-Janabi, S. (2015). A novel agent-DKGBM predictor for business intelligence and analytics toward enterprise data discovery. *Journal of Babylon University/Pure and Applied Sciences, 23*(2), 482-507.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772-1778. https://doi.org/10.1016/j.eswa.2011.08.077

Arif, U., Zhang, C., Chaudhary, M. W., & Khalid, H. H. (2024). Optimizing lung cancer prediction: leveraging Kernel PCA with dendritic neural models. *Computer Methods in Biomechanics and Biomedical Engineering, 2024*, 1-14. https://doi.org/10.1080/10255842.2024.2374949

Attoh-Okine, N. O. (1999). Analysis of learning rate and momentum term in backpropagation neural network algorithm trained to predict pavement performance. *Advances in Engineering Software*, *30*(4), 291-302. https://doi.org/10.1016/S0965-9978(98)00071-4

Baker-Finch, S. C., & McIntosh, K. R. (2011). Reflection of normally incident light from silicon solar cells with pyramidal texture. *Progress in Photovoltaics: Research and Applications*, *19*(4), 406-416. https://doi.org/10.1002/pip.1050

Banerjee, C., Mukherjee, T., & Pasiliao, E. (2020). Feature representations using the reflected rectified linear unit (RReLU) activation. *Big Data Mining and Analytics*, *3*(2), 102-120. https://doi.org/10.26599/BDMA.2019.902002

Basheer, S., Mariyam Aysha Bivi, S., Jayakumar, S., Rathore, A., & Jeyakumar, B. (2019). Machine learning based classification of cervical cancer using k-nearest neighbour, random forest and multilayer perceptron algorithms. *Journal of Computational and Theoretical Nanoscience*, *16*(5-6), 2523-2527. https://doi.org/10.1166/jctn.2019.7925

Berk, G., Pouya, P. N., Alireza, M., Yusuf, A., & Cagatay, B. (2022). An adaptive admittance controller for collaborative drilling with a robot based on subtask classification via deep learning. *Mechatronics*, 86, 102851. https://doi.org/10.1016/j.mechatronics.2022.102851

Bian, C., Wang, X., Liu, C., Xie, X., & Haitao, L. (2021). Impact of exploration-exploitation trade-off on UCB-based Bayesian Optimization. *In IOP Conference Series: Materials Science and Engineering* (Vol. 1081, No. 1, p. 012023). IOP Publishing. https://doi.org/10.1088/1757-899X/1081/1/012023

Bilgili, M., & Ozgoren, M. (2011). Daily total global solar radiation modeling from several meteorological data. *Meteorology and Atmospheric Physics*, *112*(3-4), 125-138, http://dx.doi.org/10.1007/s00703-011-0137-9

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., & Pages, J. (2000). NIR calibration in non-linear systems: different PLS approaches and artificial neural networks. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 75-82. https://doi.org/10.1016/S0169-7439(99)00048-9

Boluk, G., & Mert, M. (2014). Fossil & renewable energy consumption, GHGs (greenhouse gases) and economic growth: Evidence from a panel of EU (European Union) countries. *Energy*, *74*, 439-446. https://doi.org/10.1016/j.energy.2014.07.008

Brochu, E., Brochu, T., & De Freitas, N. (2010, July). A Bayesian interactive optimization approach to procedural Animation Design. *Communities and Collections*, 103-112. https://doi.org/10.5555/1921427.1921443

Chang, T. C., & Chao, R. J. (2006). Application of back-propagation networks in debris flow prediction. *Engineering Geology*, *85*(3-4), 270-280. https://doi.org/10.1016/j.enggeo.2006.02.007

Chen, T. C. T., Wu, H. C., & Chiu, M. C. (2024). A deep neural network with modified random forest incremental interpretation approach for diagnosing diabetes in smart healthcare. *Applied Soft Computing*, *152*, 111183. https://doi.org/10.1016/j.asoc.2023.111183

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peer Journal Computer Science*, *7*, e623-e623. https://doi.org/10.7717/peerj-cs.623

De Jong, P., Barreto, T. B., Tanajura, C. A., Kouloukoui, D., Oliveira-Esquerre, K. P., Kiperstok, A., & Torres, E. A. (2019). Estimating the impact of climate change on wind and solar energy in Brazil using a South American regional climate model. *Renewable energy*, *141*, 390-401. https://doi.org/10.1016/j.renene.2019.03.086

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178-3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

Farooq, F., Nasir Amin, M., Khan, K., Rehan Sadiq, M., Javed, M. F., Aslam, F., & Alyousef, R. (2020). A comparative study of random forest and genetic engineering programming for the prediction of compressive strength of high strength concrete (HSC). *Applied Sciences*, *10*(20), 7330. https://doi.org/10.3390/app10207330

Foorginezhad, S., & Zerafat, M. M. (2019). Fabrication of stable fluorine-free superhydrophobic fabrics for anti-adhesion and self-cleaning properties. *Applied Surface Science*, *464*, 458-471. https://doi.org/10.1016/j.apsusc.2018.09.058

Ghafouri-Kesbi, F., Rahimi-Mianji, G., Honarvar, M., & Nejati-Javaremi, A. (2016). Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. *Animal Production Science*, *57*(2), 229-236. https://doi.org/http://dx.doi.org/10.1071/AN15538

Goh, C., Scully, S. R., & McGehee, M. D. (2007). Effects of molecular interface modification in hybrid organic-inorganic photovoltaic cells. *Journal of Applied Physics*, *101*(11). https://doi.org/10.1063/1.2737977

Gwetu, M. V., Tapamo, J. R., & Viriri, S. (2014). Segmentation of retinal blood vessels using normalized Gabor filters and automatic thresholding. *South African Computer Journal*, 55(1), 12-24. http://dx.doi.org/10.18489/sacj.v55i0.228

Haji, S. H., & Abdulazeez, A. M. (2021). Comparison of optimization techniques based on gradient descent algorithm: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *18*(4), 2715-2743.

Hong, Y., Hou, B., Jiang, H., & Zhang, J. (2020). Machine learning and artificial neural network accelerated computational discoveries in materials science. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 10*(3), e1450. https://doi.org/10.1002/wcms.1450

Huh, J. Y., Ma, S. H., Kim, K., Choi, E. H., Lee, H. J., Lee, H. U., & Hong, Y. C. (2019). Facile, rapid, one-pot synthesis of hydrogenated TiO2 by using an atmospheric-pressure plasma jet submerged in solution. *Scripta Materialia*, *162*, 9-13. https://doi.org/10.1016/j.scriptamat.2018.10.028

Kulkarni, P., Sreekanth, V., Upadhya, A. R., & Gautam, H. C. (2022). Which model to choose? Performance comparison of statistical and machine learning models in predicting PM2. 5 from high-resolution satellite aerosol optical depth. *Atmospheric Environment*, 282, 119164. https://doi.org/10.1016/j.atmosenv.2022.119164

Laseinde, O. T., & Ramere, M. D. (2021). Efficiency Improvement in polycrystalline solar panel using thermal control water spraying cooling. *Procedia Computer Science*, *180*, 239-248. https://doi.org/10.1016/j.procs.2021.01.161

Liu, F., Tait, S., Schellart, A., Mayfield, M., & Boxall, J. (2020). Reducing carbon emissions by integrating urban water systems and renewable energy sources at a community scale. *Renewable and Sustainable Energy Reviews*, *123*, 109767, https://doi.org/10.1016/j.rser.2020.109767

Liu, H., Li, G., Cumberland, W. G., & Wu, T. (2005). Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping. *Journal of Data Science*, 3(3), 257-278. https://doi.org/10.6339/JDS.2005.03(3).206

Machello, C., Baghaei, K. A., Bazli, M., Hadigheh, A., Rajabipour, A., Arashpour, M., & Hassanli, R. (2024). Tree-based machine learning approach to modelling tensile strength retention of Fibre Reinforced Polymer composites exposed to elevated temperatures. *Composites Part B: Engineering*, *270*, 111132. https://doi.org/10.1016/j.compositesb.2023.111132

Mahendran, M., Lizotte, D., & Bauer, G. R. (2022). Describing intersectional health outcomes: an evaluation of data analysis methods. *Epidemiology*, *33*(3), 395-405. https://doi.org/10.1097/ede.0000000000001466

Maind, S. B., & Wankar, P. (2014). Research paper on basic of artificial neural network. International *Journal on Recent and Innovation Trends in Computing and Communication*, *2*(1), 96-100. https://doi.org/10.17762/ijritcc.v2i1.2920

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140-147. https://doi.org/10.38094/jastt1457

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), 3301-3307. https://doi.org/10.1093/bioinformatics/bti499

Osorio, E., Urteaga, R., Acquaroli, L. N., García-Salgado, G., Juaréz, H., & Koropecki, R. R. (2011). Optimization of porous silicon multilayer as antireflection coatings for solar cells. *Solar Energy Materials and Solar Cells*, *95*(11), 3069-3073. https://doi.org/10.1016/j.solmat.2011.06.036

Oudir, A., & Bourguig, R. (2024). Method for designing a broadband and omnidirectional hybrid antireflection coating for organic solar cells using the quarter-wavelength rule. *Organic Electronics*, *127*, 107001. https://doi.org/10.1016/j.orgel.2024.107001

Ozkan, C., & Erbek, F. S. (2003). The comparison of activation functions for multispectral Landsat TM image classification. *Photogrammetric Engineering and Remote Sensing*, *69*(11), 1225-1234. https://doi.org/10.14358/PERS.69.11.1225

Premalatha, N., & Valan Arasu, A. (2016). Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms. *Journal of Applied Research and Technology*, *14*(3), 206-214. https://doi.org/10.1016/j.jart.2016.05.001

Raisee, M., Kumar, D., & Lacor, C. (2015). A non-intrusive model reduction approach for polynomial chaos expansion using proper orthogonal decomposition. *International Journal for Numerical Methods in Engineering, 103*(4), 293-312. https://doi.org/10.1002/nme.4900

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.

Sajikumar, N., & Thandaveswara, B. S. (1999). A non-linear rainfall–runoff model using an artificial neural network. *Journal of Hydrology*, *216*(1-2), 32-55. https://doi.org/10.1016/S0022-1694(98)00273-X

Sathya, R. A., & Ponraj, C. (2024). Superhydrophobic route of fabricating antireflective, self-cleaning, and durable coatings for solar cell applications. *Journal of Coatings Technology and Research*, *21*(1), 1-30. https://doi.org/10.1007/s11998-023-00843-x

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, *48*(4), 1875-1897. https://doi.org/10.1214/19-AOS1875

Song, Y., Thatcher, D., Li, Q., McHugh, T., & Wu, P. (2021). Developing sustainable road infrastructure performance indicators using a model-driven fuzzy spatial multi-criteria decision making method. *Renewable and Sustainable Energy Reviews*, *138*, 110538. https://doi.org/10.1016/j.rser.2020.110538

Sun, X., Li, L., Xu, X., Song, G., Tu, J., Yan, P., Zhang, W., & Hu, K. (2020). Preparation of hydrophobic SiO2/PTFE sol and antireflective coatings for solar glass cover. *Optik*, *212*, 164704. https://doi.org/10.1016/j.ijleo.2020.164704

Syafiq, A., Pandey, A. K., Adzman, N. N., & Abd Rahim, N. (2018). Advances in approaches and methods for self-cleaning of solar photovoltaic panels. *Solar Energy*, *162*, 597-619. https://doi.org/10.1016/j.solener.2017.12.023

Tadanaga, K., Morita, K., Mori, K., & Tatsumisago, M. (2013). Synthesis of monodispersed silica nanoparticles with high concentration by the Stöber process. *Journal of Sol-Gel Science and Technology*, *68*, 341-345. https://doi.org/10.1007/s10971-013-3175-6

Udaybhanu, G., & Reddy, V. M. (2024). A hybrid GA-ANN and correlation approach to developing a laminar burning velocity prediction model for isooctane/blends-air mixtures. *Fuel*, *360*, 130594. https://doi.org/10.1016/j.fuel.2023.130594

Ureña-Sánchez, R., Callejón-Ferre, Á. J., Pérez-Alonso, J., & Carreño-Ortega, Á. (2012). Greenhouse tomato production with electricity generation by roof-mounted flexible solar panels. *Scientia Agricola*, *69*(4), 233-239. https://doi.org/10.1590/S0103-90162012000400001

Wang, Z., & Jegelka, S. (2017, July). Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning* (pp. 3627-3635). PMLR. https://doi.org/10.48550/arXiv.1703.01968

Wang, J., Tian, Y., Qi, Z., Zeng, L., Wang, P., & Yoon, S. (2024). Sensor fault diagnosis and correction for data center cooling system using hybrid multi-label random Forest and Bayesian Inference. *Building and Environment*, *249*, 111124. https://doi.org/10.1016/j.buildenv.2023.111124

Wei, Z. (2021). Forecasting wind waves in the US Atlantic Coast using an artificial neural network model: Towards an AI-based storm forecast system. *Ocean Engineering*, *237*, 109646. https://doi.org/10.1016/j.oceaneng.2021.109646