

A Data-Driven Machine Learning Approach to Solar Energy Generation Forecasting Using Meteorological Variables

Tharushika D.D.H.,* Napagoda N.A.D.N., and Ekanayake E.M.P.

Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka

Received June 2, 2026; Accepted July 6, 2026; Published August 11, 2026

The production of solar power is highly dependent on weather conditions, including temperature, humidity, wind velocity, wind direction, cloud coverage, and solar radiation. These relations must be well understood to develop photovoltaic systems with high efficiency and reliability. A solar power prediction model is proposed in this paper based on meteorological data with more than 30,000 samples gathered at La Trobe University, Victoria, Australia. The purpose is to determine the most significant atmospheric variables and to improve prediction performance by modeling linear and non-linear relationships between solar output and meteorological parameters. Three machine learning (ML) algorithms: Ridge Regression (RR), Elastic Net Regression (ELNET), and Kernel Ridge Regression (KRR) have been employed and compared based on Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-Squared evaluation measure. SHapley Additive exPlanations (SHAP) was utilized to explain model outcomes and the significance of the variables. Among all the exemplified approaches, kernel ridge regression provides the best forecasting results. Air temperature, wind direction, wind speed, and relative humidity were the top four important features affecting solar power generation identified from SHAP analysis. The results show the potential of using machine learning techniques for accurate solar energy prediction, leading to enhanced grid management, renewable energy integration, and sustainable energy planning.

Keywords: Solar Generation; Meteorological Variables; Ridge Regression; Elastic Net Regression; Kernel Ridge Regression

Introduction

A wide range of meteorological factors including rainfall, snowfall, air pressure, temperature, humidity, wind direction and speed, cloud cover, and precipitation directly impact the natural environment, daily operations, and human safety. Accurate information on meteorological data is known to be critical for the aviation and transportation industries while it is essential for disaster management, generation of solar and wind energy, and growing and harvesting agricultural crops.

The significant impact of meteorological variables on renewable energy generation, particularly solar energy, is considered in this study as it is a versatile renewable energy source with a wide range of applications in electricity generation for households, businesses, and industries, as well as heating, lighting, and power for automobiles and EV charging stations. Agricultural support includes irrigation solutions

*Corresponding author: tharushikadampahala71@gmail.com

and efficient crop drying systems while portable power choices are part of it as well. Further, specialized support comes in the form of improving water pH levels and maintaining hygiene requirements. Solar energy is also utilized in industrial processes, mitigating disasters, environmental monitoring, satellite power, and smart grid integration to enhance energy storage and distribution efficiency. Environmental sustainability and global energy transitions are promoted by solar power, with greenhouse gas emissions being lowered and energy individuality being increased. Solar power generation is affected by a variety of meteorological factors. Therefore, understanding and optimizing the interaction between meteorological variables are important for the improvement of reliability and efficiency in solar energy systems.

As the aim of this work is to develop a robust forecasting model that uses lasso regression modeling to improve accuracy and reliability across a variety of climate settings, it is imperative to explore previous research in solar energy forecasting while focusing on data mining approaches and regression techniques used in them. These studies emphasize the significance of meteorological conditions in making accurate solar generation predictions and the use of lasso regression approaches, among other models, to improve predictive accuracy in environmental and agricultural forecasts. Daily weather data and leave-one-out cross validation within the Ridge Regression–Best Linear Unbiased Prediction (RR-BLUP) framework have been employed to predict wheat genotype performance in Switzerland, achieving high prediction ability, but the volatility of the predictions was influenced by data unpredictability [1]. The effect of meteorological factors on surface ozone in China from 2015 to 2019 has been further studied using ML methods, where the Random Forest Regression outperformed linear model-based approaches, emphasizing the need to capture the nonlinearity of climatic influences [2]. In recent years, axial data on weather conditions covering 21 years have been used to predict the yield of rice in Chhattisgarh, India, wherein more promising results were yielded by ANNs depending on regions, whereas LASSO and ELNET (Elastic Net) were concluded as more powerful; while the accuracy of prediction was enhanced by RF ensembles, overfitting and the paucity of long-term data remained as primary concerns [3]. Ridge Regression Models have been proposed to combine snow water equivalent information at multiple spatial scales to yield more accurate but still uncertain estimates subject to the linearity assumptions [4]. In the meantime, temperature prediction in Indonesia has been conducted using Ridge Regression with Principal Component Analysis (PCA), leading to a reduction in prediction errors compared to that of conventional Numerical Weather Prediction (NWP) models, although model inputs have been reduced through the application of PCA, making the model easier to interpret [5].

The analysis of time series data of historical solar intensities was conducted by means of clustering, and daily profiles of sun intensity were clustered at different time scales for the enhancement of the training datasets in both Artificial Neural Networks (ANN) and Support Vector Machines (SVM), resulting in improved prediction accuracy; however, the determination of the best clustering parameters remains as a challenge to be further addressed [6]. The prediction of solar farm power generation in California has been studied with Decision Tree–based models, such as Light Gradient Boosting Machine, but limited success has been noted due to a strong dependence on weather and site-specific features, suggesting the need for the development of more robust predictive schemes to facilitate the grid integration of renewable energy [7]. Prediction of solar power generation in Ghana’s Bui Solar Generating Station (BSGS) with Random Forest

(RF) and Gradient Boosting Regression revealed better performance by RF with a mean absolute error of 1.18%, but difficulties in describing the real-time weather fluctuations were also faced [8]. A two-stage modeling technique with anticipated and unanticipated weather elements was included in the meteorological based solar power prediction method for South Korea, with an R^2 value of 70.5 % obtained by RF; however, forecast errors of 3 %–38 % appeared due to uncertainties from local meteorological conditions, implying that supplemental flux was needed for use in complicated surroundings [9]. Weather data has been used in ridge regression to address multicollinearity in linear regression models for the meteorological data, resulting in lower errors in estimation and more accurate predictions than Ordinary Least Squares; however, issues related to linearity assumptions and guessing multicollinearity still remained [10]. Prediction of solar power generation based on machine learning, such as Support Vector Regression, Random Forest Regression, and Linear Regression, has revealed that the best prediction accuracy of 94.01% was achieved by Random Forest, but deterioration in the performance of the model occurred in severe weather [11].

Despite all these works reported in the literature, new insights remain to be explored using advanced regularized regression techniques in order to forecast solar energy generation based on a standard meteorological database compiled at La Trobe University in Victoria, Australia, which include apparent temperature, air temperature, dew point temperature, relative humidity, wind speed, and wind direction. Unlike previous studies, which focused on specific geographical regions or relied heavily on solar energy data, this investigation was carried out to ascertain how commonly available meteorological parameters can be effectively used for solar forecasting, making the findings more widely applicable to data-limited environments. Ridge Regression, Elastic Net, and Kernel Ridge Regression were rigorously compared under identical data settings, and new empirical insights into the models' performance and behavior were provided by the study. Further, advanced methods like SHAP analysis and L-curve optimization were used to increase model interpretability and robustness, an approach that is not extensively used in the existing literature. The scientific worth of this study could be augmented by the incorporation of both predictive accuracy and theoretical explanations based on regularization principles, providing a transparent and flexible framework for solar energy modelling in a variety of environmental contexts.

The analysis of solar power generation prediction using meteorological data was thus carried out in this work to achieve the above objectives based on a dataset of over 30,000 observations which contained apparent temperature, air temperature, dew point temperature, relative humidity and wind speed, with direction. Since a large number of meteorological data were to be investigated, a comprehensive big data analysis was required for the predictive power generation of solar energy. Data mining techniques were utilized to extract valuable insights from large data sets, which would result in more accurate estimates of solar power generation.

More specifically, a predictive model for solar power generation shall be constructed in the process using meteorological data for determining the most significant climatic parameters influencing solar energy output. Accurate solar forecasting found from this study would be useful in power grid management as it could help optimize power distribution and minimize the reliance on fossil fuels. Stakeholders could benefit from well informed decisions on strategic planning and investment in solar infrastructure with genuine estimates. Further, once the impact of meteorological conditions on solar generation is understood it could help maximize the use of renewable energy and

promote sustainable development. The integration of this renewable energy source into the power grid would also be enhanced by the findings, paving the way for a more sustainable and efficient future for solar energy use.

Materials and Methods

A collection of photovoltaic (PV) solar energy generation and meteorological data from a campus site at La Trobe University in Victoria, Australia was selected for this study, which contained 31,229 instances of seven variables namely apparent temperature, air temperature, dew point temperature, relative humidity, wind speed with wind direction, and the solar generation.

Data Preprocessing

Data preprocessing is regarded as an important phase in data analysis that incorporates numerous strategies for improving data quality and model performance. A correlation matrix was used to identify correlations between variables and aid in feature selection. Data standardization assures that all features contribute equally by scaling them to a mean of zero and a standard deviation of one. Z-score standardization is commonly employed for normally distributed data [12]. Handling missing values is regarded critical, with approaches like deletion, imputation, or regression-based filling used to ensure data integrity. Outliers that may skew the analysis are emphasized by outlier identification with boxplots, and the Interquartile Range (IQR) method is considered a reliable way for detecting and removing them by setting lower and upper bounds. Model generalization is ensured by dividing the dataset into training and testing sets, with an average split of 80% for training and 20% for testing to assess model performance on previously unseen data.

Lasso Regression Modeling

LASSO Regression, an L1 regularization technique, is characterized by the penalization of absolute values of coefficients, which allows for variable selection by having some coefficients reduced to zero and thus removed. Ridge Regression is characterized by using an L2 penalty to decrease coefficients rather than eliminate them, making it beneficial for dealing with multicollinearity. Elastic Net is characterized by the combination of L1 and L2 penalties, making use of LASSO's feature selection and Ridge's multicollinearity handling, thus making it especially useful when predictors are highly correlated or outnumber observations.

Ridge Regression

Ridge Regression, also known as L2 regularization, is employed as a linear regression approach for preventing over fitting by penalizing unnecessary coefficient values. Multicollinearity is effectively handled by it, making it appropriate for models with strongly linked predictors. Prediction stability and accuracy are improved by Ridge Regression through the inclusion of a penalty term in the loss function, even if some underperformance occurs on the training set. The L2 regularization approach is utilized in Ridge Regression, where the loss is defined as:

$$L_{Ridge}(\hat{\beta}) = \sum_{i=1}^n (y_i - x_i' \hat{\beta})^2 + \lambda \sum_{j=1}^m \beta_j^2 = y - X\hat{\beta}^2 + \lambda \hat{\beta}^2 \quad (01)$$

In the equation, the variable vector for the i^{th} observation is represented by x_i , while the actual value of the dependent variable for the i^{th} observation is denoted by y_i . The dataset comprises n samples and M predictor variables. An intercept term β_0 is included in the model, and the coefficient of x_j is represented by β_j . The regularization strength is controlled by the parameter λ , which balances the trade-off between the data fitting and the penalization of large coefficients. The loss function is made up of two parts: the residual sum of squares $\sum_{i=1}^n (y_i - x_i' \hat{\beta})^2$, which measures the error of the model, and the regularization, $\lambda \sum_{j=1}^m \beta_j^2$, which constrains the magnitude of the coefficients to prevent over fitting..

Elastic Net Regression

Multicollinearity is addressed by Elastic Net by combining L1 (Lasso) and L2 (Ridge) penalties while allowing feature selection too. The difficulties of Lasso with highly correlated variables are addressed, making it useful for high-dimensional data. Model robustness, predictability, and bias reduction are improved by Elastic Net by using Ridge Regression first, then Lasso shrinkage, and coefficient rescaling [13]. ELNET equation is defined as,

$$\hat{\beta}_{EN} = \arg \min \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \left(\alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right) \right\} \quad (02)$$

In the ELNET equation, the estimated coefficients for the Elastic Net Regression model are represented by $\hat{\beta}_{EN}$. The regularization parameter is denoted by λ while the mixing parameter is indicated by α . The value of the j^{th} variable for the i^{th} observation is represented by x_{ij} . The dataset comprises of n samples. An intercept term β_0 is included in the model and the coefficient of x_j is represented by β_j . The L1 penalty is defined as $\sum_{j=1}^p |\beta_j|$ and the L2 penalty as $(1 - \alpha) \sum_{j=1}^p \beta_j^2$.

Kernel Ridge Regression

Ridge Regression is expanded by Kernel Ridge Regression (KRR) to capture non-linear correlations by moving input characteristics into a higher-dimensional space with kernel functions such as polynomials or RBF. Unlike traditional Ridge Regression, which is used to manage over fitting in linear models with an L2 penalty, complicated patterns can be learned by KRR while retaining model regularization. The objective function for KRR is,

$$L_{KRR}(\hat{\beta}) = \sum_{i=1}^n (y_i - y(x_i))^2 + \lambda \|y\|^2 \quad (03)$$

where y is the function to be found, represented in terms of kernel. Using the kernel trick, the solution to kernel Ridge Regression is,

$$\hat{y} = (K + \lambda I)^{-1}y \quad (04)$$

Here, the kernel matrix is represented by K , the identity matrix by I , and the regularization parameter λ . The target vector y is defined as $y = [y_1, \dots, y_N]^T$, where each y_i corresponds to an observed value in the dataset.

Hyperparameter Tuning

K-Fold Cross-Validation is considered a robust technique for parameter selection in Lasso Regression models [14]. This procedure involves repeating k times, with each time having a different fold acting as the test set. The average performance across all folds is used to provide a reliable estimate of the model's predictive power.

Parameter selection for Kernel Ridge Regression (KRR) involves the tuning of essential hyper parameters with approaches such as cross-validation, grid search, randomized search, and L-curve.

L-curve

The norm of a regularized solution plotted against the norm of the associated residual norm is called an L-curve. It is a useful graphical tool that shows, as the regularization parameter changes, the trade-off between the size of a regularized solution and its fit to the given data. Information on characteristics of the underlying regularization technique and the best regularization value for the given data is provided by the L-curve [15]. Figure 1 is shown to illustrate the graph of an L-curve.

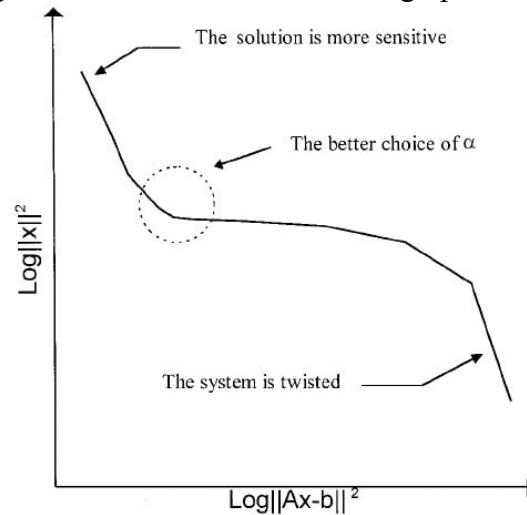


Figure 1. Graph of an L-curve [15]

SHapley Additive exPlanations (SHAP) Summary Plot

The SHAP (SHapley Additive ExPlanations) summary plot is a visualization for interpreting machine learning models, which shows the contribution of each feature to the model's predictions. Insights into how individual features affect the output are provided in a way that is consistent with game theory (Shapley values). An example of a SHAP summary plot is illustrated in Figure 2 [16].

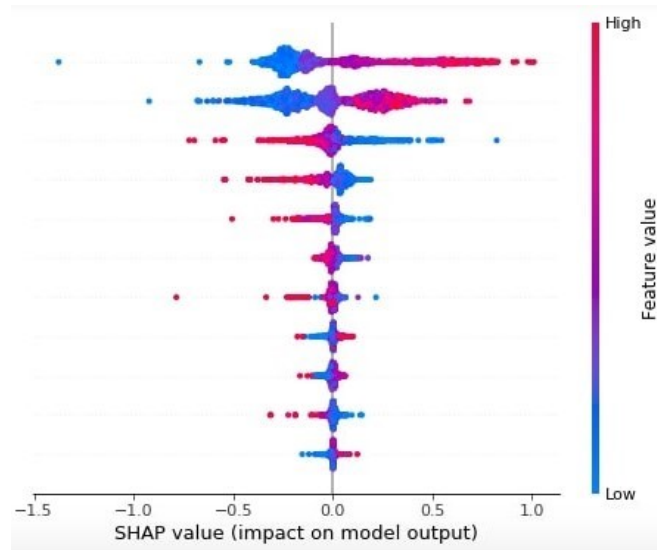


Figure 2. SHAP Summary Plot [16]

Evaluation Metrics

Prediction errors are determined by the Root Mean Square Error (RMSE), where a lower RMSE indicates better model fit. The square root of the average squared discrepancies between the actual and anticipated values is calculated.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (04)$$

Prediction accuracy is also estimated by the Mean Absolute Percentage Error (MAPE). A more accurate model corresponds to a lower MAPE value.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \quad (05)$$

Goodness-of-fit is assessed by calculating the proportion of variance in the dependent variable explained by the independent variables, which is determined by the Coefficient of Determinant (R-squared) whose higher values indicate better model performance.

$$R^2 = \left(\frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sigma_y \sigma_{\hat{y}}} \right)^2 \quad (06)$$

In the above evaluation matrices, n is the total number of observations and the actual observed value of the dependent variable for the i^{th} observation is represented by y_i and its predicted value $\hat{y}_i$. The mean of the observed values y_i is denoted by $\bar{y}$, and the mean (average) of the predicted values $\hat{y}_i$ by $\bar{\hat{y}}$. σ_y is the standard deviation of the observed values y_i and that of the predicted values $\hat{y}_i$ is $\sigma_{\hat{y}}$.

Results and Discussion

Data Preprocessing

The dataset contains 25701 observations for apparent temperature, air temperature, dew point temperature, and relative humidity. Wind speed and wind direction each have 19,778 instances, while the target variable, solar generation, consists of 14,527 instances. As shown in Figure 3, there is no significant relationship between any meteorological variable and solar generation. Apparent temperature, air temperature, dew point temperature, and wind speed exhibit a positive relationship with solar generation, whereas relative humidity and wind direction show a negative relationship with solar generation.

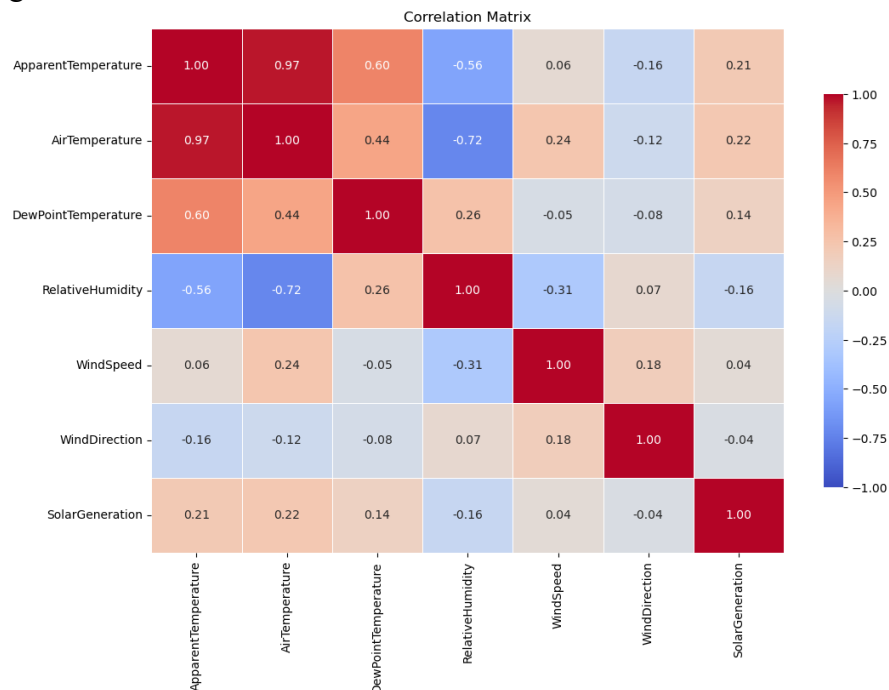


Figure 3. Correlation Matrix

In this analysis, only the independent variables were standardized using the standard scaling technique, while the target variable, solar generation, was left unstandardized. The Interquartile Range (IQR) approach was used to identify and remove the outliers in each variable and accordingly, 86 outliers were discovered in apparent temperature, 220 in air temperature, 249 in dew point temperature, and 248 in wind speed. In contrast, no outliers were found in relative humidity, wind direction, or the solar generation. After the outliers were addressed and missing values were removed, the dataset simplified to 9,715 instances. This pre-processed dataset has been prepared for further analysis, ensuring that the data utilized for modeling is more accurate and revealing. To proceed with analysis, the dataset was divided into training and testing sets with a ratio of split of 0.2, resulting in 20% for testing and 80% for training. As a result of this splitting, the training dataset was allocated 7,772 instances while 1,943 instances were chosen for testing.

Ridge Regression Analysis

The Ridge Regression model was trained using split data, with a random choice of 1.0 for the λ value to represent the model training process. The ideal regularization

parameter (λ) was identified as 0.82 using RidgeCV technique and it is illustrated in Figure 4.

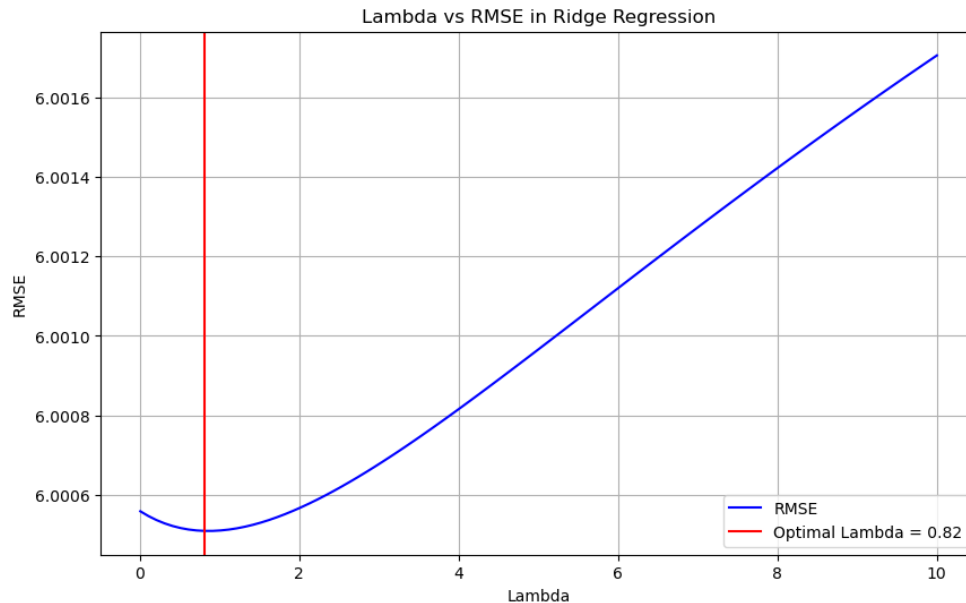


Figure 4. Optimal regularization parameter (λ)

After the Ridge Regression model was fitted with $\lambda = 0.82$, its performance was evaluated using the above evaluation metrics: RMSE, MAPE and the R- Squared. It was observed that, on average, the model predictions differed from actual values by 6.03 units. The MAPE of the model predictions was found to be 3.1537%, with an average deviation from actual values of roughly 3.15%. The R^2 score was 0.921, indicating that 92.1% of the variance in the target variable can be explained by the independent factors.

$$\begin{aligned} \text{The Ridge Regression Equation emerged as,} \\ y = 8.8989 + (-5.4906 * \text{Apparent Temperature}) \\ \quad + (4.1991 * \text{Air Temperature}) \\ \quad + (2.8470 * \text{Dew Point Temperature}) \\ \quad + (-2.3548 * \text{Relative Humidity}) + (-1.1964 * \text{Wind Speed}) \\ \quad + (0.0958 * \text{Wind Direction}) \end{aligned} \quad (07)$$

where y is the target variable.

Elastic Net Regression Analysis

The split data are used to train an Elastic Net Regression model with random values of 0.1 and 0.5 for the regularization parameter (λ) and a mixing parameter (α) respectively. However, the ideal regularization value (λ) was identified as 0.01, and the mixing parameter (α) was determined to be 0.1 by the ElasticNetCV approach, as shown in Figure 5.

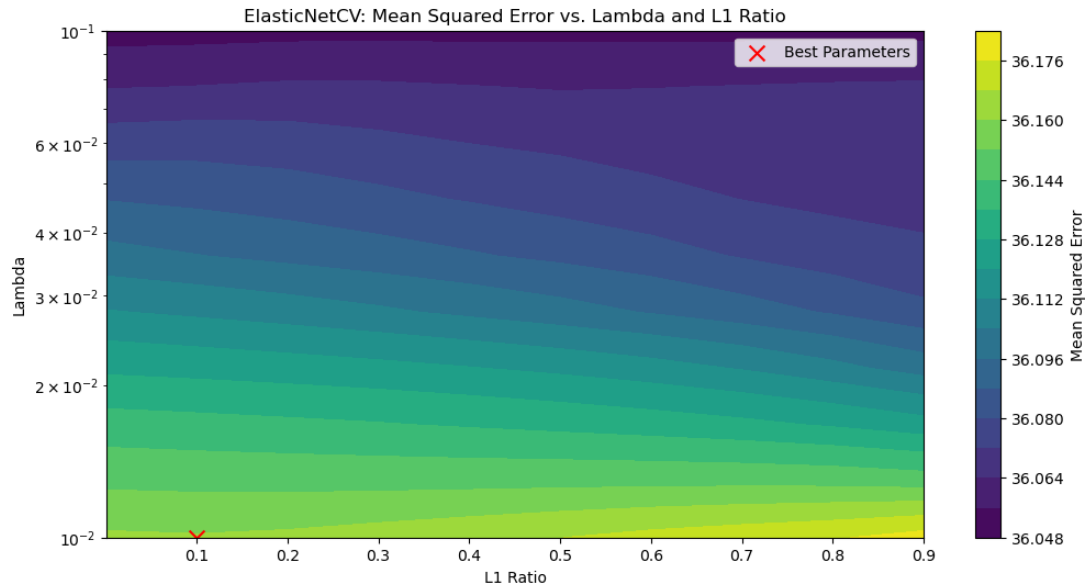


Figure 5. Optimal ELNET parameters

An average difference of about 6.056 units from the actual values was indicated by the RMSE, suggesting a moderate amount of error in model predictions. A MAPE of 3.1735% is displayed, indicating that model predictions are different from actual values by about 3.17% on average. The R^2 value of 0.857 implied, with 85.7% of the variance in the target variable being explained, indicating strong explanatory power.

The Elastic Net model equation is,

$$y = 8.9571 + (-0.4912 * \text{Apparent Temperature}) + (0.2911 * \text{Air Temperature}) + (1.5416 * \text{Dew Point Temperature}) + (-1.7532 * \text{Relative Humidity}) + (-0.2333 * \text{Wind Speed}) + (0.0960 * \text{Wind Direction}) \quad (08)$$

where y is the dependent variable.

Table 1. Comparison between Ridge Regression and Elastic Net Regression

Evaluation Matrices	Ridge Regression	Elastic Net Regression
RMSE	6.0347	6.056
MAPE	3.15%	3.17%
R^2 Score	0.921	0.857

Table 1 shows that the lowest RMSE and MAPE and the highest R^2 score are achieved by the Ridge Regression model. Accordingly, the Ridge Regression model outperformed the Elastic Net Regression model in terms of accuracy and deviation from actual values.

Kernel Ridge Regression Analysis

The split data were used to train a Kernel Ridge Regression (KRR) model with a randomly chosen regularization parameter (λ) of 1.0 and a gamma value of 1.0, using a radial basis function (RBF) kernel, during the initial training phase. The L-curve is shown using the RBF kernel, with the norm of the residuals vs. the norm of the coefficients in Figure 6, revealing that the ideal regularization parameter (λ) is 1.668 when gamma is equal to 0.2.

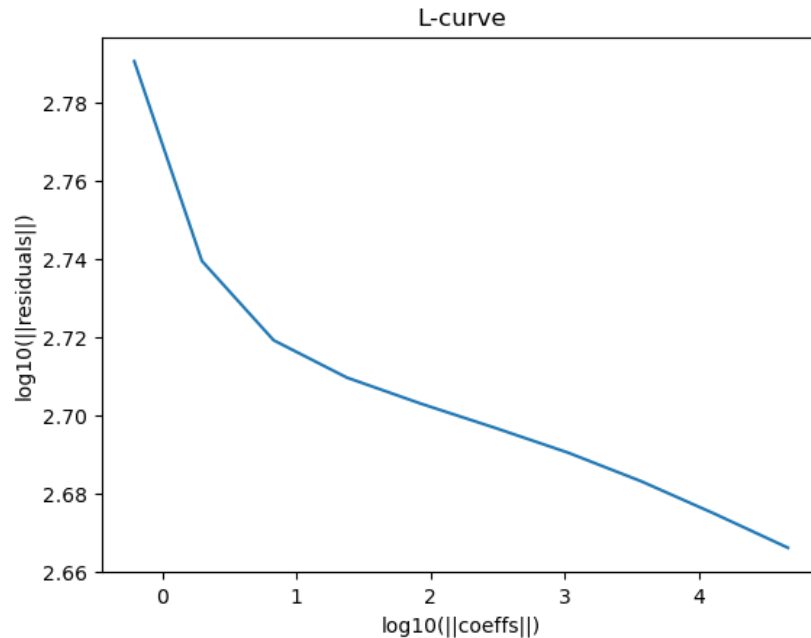


Figure 6. L-curve

An average difference of approximately 5.696 units from the actual values was recorded by the RMSE. A MAPE of 2.9679% meant that the predictions of the Kernel Ridge model differ from the actual values by an average of approximately about 2.97%. The R^2 score of 0.952 indicated that a well-performing model has been captured describing nearly all the variance in the target variable.

The SHAP summary plot for the optimal Kernel Ridge Regression model with a sample size of 1,000 is illustrated in Figure 7.

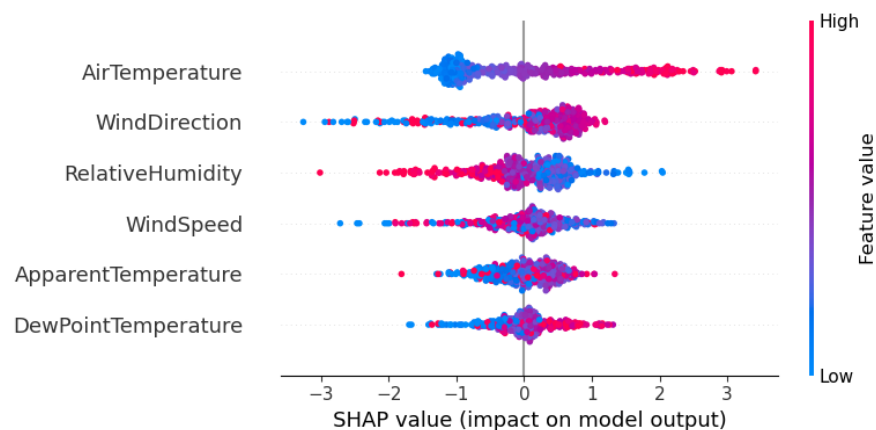


Figure 7. SHAP summary plot

- According to the SHAP summary plot, the following observations could be made:
- Air Temperature: Most red dots are located on the positive side of the SHAP value, indicating that as air temperature increases, solar generation also increases. That is, air temperature positively impacts solar generation.
 - Wind Direction: Most of the red dots lie on the right side, implying the positive impact of wind direction on solar generation.

- Wind Speed: Red dots are mainly found on the left side for wind speed, indicating that wind speed has a negative effect on solar generation predictions.
- Relative humidity: Higher humidity values (red dots) correspond to negative SHAP values, suggesting that as humidity increases, solar generation decreases.
- Apparent Temperature: A mixture of SHAP values lying on either side of zero with color variation implies that apparent temperature may have a more complex relationship with solar generation, potentially depending on other factors.
- Dew Point Temperature: Most of the red dots lying on the right side suggest that dew point temperature may have a more complex relationship with solar generation.

Table 2. Comparison between Ridge Regression and Kernel Ridge Regression Models

Evaluation Matrices	Ridge Regression	Kernel Ridge Regression
RMSE	6.0347	5.696
MAPE	3.15%	2.97%
R ² score	0.921	0.952

It is evident from Table 2 that the Kernel Ridge Regression model has resulted in the lowest RMSE and the lowest MAPE percentage values, and highest R² score value when compared to the Ridge and Elastic Net models. These statistics confirm that a more accurate and robust prediction of solar energy generation can be obtained from Kernel Ridge Regression, underscoring its dexterity for forecasting needs.

Discussion

Researchers have used various machine learning techniques to predict solar power generation using meteorological data, most relevant of which carried out lately are summarized in Table 3. Most of them have used Random Forest Regression (RFR) technique in the analysis and it has been found to outperform the other techniques like Support Vector Machine Regression, Linear Regression, and Gradient Boosting Regression. However, while estimating solar production from meteorological data, RFR has several limitations too. The computational expense and time consumption of the task could increase, especially with large datasets. Over-fitting could be an issue when the number of trees is too large, or the data includes excessive unrelated features. Moreover, RFR Models are sometimes denoted as "black-box" models, suggesting a lack of interpretability, making it challenging to understand the relationship between meteorological parameters and solar generation. This may limit the value to stakeholders who prefer straightforward and comprehensible models. Kernel Ridge Regression (KRR) typically outperforms Random Forest Regression (RFR) in predicting solar generation because complicated nonlinear relationships between meteorological variables and solar energy generation can be captured by KRR. Unlike RFR that suffers from high-dimensional data and complicated interactions, this study has shown that such issues could be overcome by merging Ridge Regression and kernel approaches in Kernel Ridge

Regression. Further, smoother predictions were produced by KRR avoiding the over-fitting reported with RFR, when working with large datasets and numerous trees.

Table 3. Summary of past studies on solar power generation

Author	Country	Methods	Findings
Anuradha et al., (2021) [11]	Sri Lanka	Support Vector Machine Regression Random Forest Regression Linear Regression	Random Forest Regression emerged as the most accurate, achieving 94.01% accuracy. This study highlighted the significance of environmental conditions such as solar irradiance, cloud cover, and temperature in influencing solar power generation.
Malakouti et al., (2023) [7]	California	Decision Tree Light Gradient Boosting Machine Extra Tree Algorithms.	This study demonstrated that all three models achieved high accuracy, with R-squared values close to 1.
Abdulai et al., (2023) [8]	Ghana	Random Forest Regression Gradient Boosting Regression	In this study, the Random Forest model excelled in achieving accurate predictions, with a Mean Absolute Percentage Error of 1.18%.
Kim et al., (2019) [9]	South Korea	Random Forest Regression	The Random Forest Regression model provided the highest accuracy, reaching an R-squared value of 70.5%.

CONCLUSIONS

Solar power production can be effectively predicted by machine learning models using weather data. Among the models considered, Kernel Ridge Regression (KRR) obtains the best forecasting performance with RMSE = 5.696, MAPE = 2.9679, $R^2 = 0.952$, outperforming the other two models, i.e. Ridge Regression and Elastic Net Regression. Air temperature, wind direction, wind speed, and relative humidity were the key driving variables for the solar electricity generation. The introduction of SHAP analysis and the hyperparameter tuning enhanced the transparency, interpretability and dependability of the models. Future studies could focus on extending dataset coverage to different regions and time scales, as well as on integrating real-time meteorological data and advanced ML or hybrid solutions for further enhancement of forecasting performance. Energy companies can apply such precise forecasting models to better manage their grids, energy storage, maintenance planning, and resource use. Analogous frameworks could enable even greater integration of renewable energy and development of sustainable energy systems.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper

REFERENCES

- [1] Herrera, J. M., Häner, L. L., Holzkämper, A., & Pellet, D. (2018). Evaluation of ridge regression for country-wide prediction of genotype-specific grain yields of wheat. *Agricultural and Forest Meteorology*, 252, 1-9.
- [2] Weng, X., Forster, G. L., & Nowack, P. (2022). A machine learning approach to quantify meteorological drivers of ozone pollution in China from 2015 to 2019. *Atmospheric Chemistry and Physics*, 22(12), 8385-8402.
- [3] Satpathi, A., Setiya, P., Das, B., Nain, A. S., Jha, P. K., Singh, S., & Singh, S. (2023). Comparative analysis of statistical and machine learning techniques for rice yield forecasting for Chhattisgarh, India. *Sustainability*, 15(3), 2786.
- [4] Shao, D., Li, H., Wang, J., Hao, X., Che, T., & Ji, W. (2022). Reconstruction of a daily gridded snow water equivalent product for the land region above 45° N based on a ridge regression machine learning approach. *Earth System Science Data*, 14(2), 795-809.
- [5] Qona'ah, N., Ferawati, K., & Nirwana, M. B. (2020, April). Temperature Forecast Using Ridge Regression as Model Output Statistics. In *Proceeding International Conference on Science and Engineering* (Vol. 3, pp. 383-388).
- [6] Pinto, T., Santos, G., Marques, L., Sousa, T. M., Praça, I., Vale, Z., & Abreu, S. L. (2015). Solar intensity characterization using data-mining to support solar forecasting. In *Distributed Computing and Artificial Intelligence, 12th International Conference* (pp. 193-201). Springer International Publishing.
- [7] Malakouti, S. M., Menhaj, M. B., & Suratgar, A. A. (2023). The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. *Cleaner Engineering and Technology*, 15, 100664.
- [8] Abdulai, D., Gyamfi, S., Diawuo, F. A., & Acheampong, P. (2023). Data analytics for prediction of solar PV power generation and system performance: A real case of Bui Solar Generating Station, Ghana. *Scientific African*, 21, e01894
- [9] Kim, S. G., Jung, J. Y., & Sim, M. K. (2019). A two-step approach to solar power generation prediction based on weather data using machine learning. *Sustainability*, 11(5), 1501.
- [10] Ahmed, L. (2020). Using ridge regression to analysis the meteorological data in sulaimani. *Iraqi Journal of Statistical Sciences*, 17(2), 51-57.
- [11] Anuradha, K., Erlapally, D., Karuna, G., Srilakshmi, V., & Adilakshmi, K. (2021). Analysis of solar power generation forecasting using machine learning techniques. In *E3S Web of Conferences* (Vol. 309, p. 01163). EDP Sciences.
- [12] Blessing, E., & Klaus, H. (2023). Normalization and Standardization: Methods to preprocess data to have consistent scales and distributions. 2237. 10.
- [13] Nur, A. R., Jaya, A. K., & Siswanto, S. (2023). Comparative Analysis of Ridge, LASSO, and Elastic Net Regularization Approaches in Handling Multicollinearity for Infant Mortality Data in South Sulawesi. *Jurnal Matematika, Statistika dan Komputasi*, 20(2), 311-319.
- [14] Manorathna, R. (2020). *k-fold cross-validation explained in plain English (For evaluating a model's performance and hyperparameter tuning)*.

- [15] Frasso, G., & Eilers, P. H. (2012, September). Smoothing parameter selection using the L-curve. In *27th International Workshop on Statistical Modelling, Prague, Czech Republic*
- [16] learner101. (2024). *SHAP Summary Plot and Mean Values displaying together*. Stack Overflow. <https://stackoverflow.com/questions/72087719/shap-summary-plot-and-mean-values-displaying-together>
- [17] CDAClab. (2022, November 9). Unisolar Solar Power Generation Dataset. Kaggle. <https://www.kaggle.com/datasets/cdaclab/unisolar>

Article Copyright: © 2026 Tharushika D.D.H., Napagoda N.A.D.N., Ekanayake E.M.P. This is an open access article distributed under the terms of the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and distribution provided the original author and source are credited.

