Solar Irradiance Forecasting for Informed Solar Systems Design and Financing Decisions

Ronewa Mabodi and Jahvaid Hammujuddy

Abstract—This research presents the implementation and evaluation of machine learning models to predict solar irradiance (W/m²). The objective is to provide valuable insights for making informed decisions regarding solar system design and financing. A thorough exploratory data analysis was conducted on the Southern African Universities Radiometric Network (SAURAN) data collected at the University of Pretoria's station to gain insights into the patterns of solar irradiance over the past 10 years. Python's functions and libraries are utilized extensively for conducting exploratory data analysis, model implementation, model testing, forecasting, and data visualization. Random Forest (RF), k-Nearest Neighbors (KNN), Feedforward Neural Network (FFNN), Support Vector Regression (SVR), and eXtreme Gradient Boosting models (XGBoost) are implemented and evaluated. The KNN model was found to be superior achieving a relative Root Mean Squared Error (RMSE), relative Mean Absolute Error (MAE), and R-Squared (R^2) of 5.77%, 4.51% and 0.89 respectively on testing data. The variable importance analysis revealed that temperature (℃) exerted the greatest influence on predicting solar irradiance, accounting for 44% of the predictive power. The KNN model is suitable to inform solar systems design and financing decisions. Directions for future studies are identified and suggestions for areas of exploration are provided to contribute to the advancement of solar irradiance predictions.

***Index Terms* — *k*-Nearest Neighbors, Machine learning, Solar irradiance forecasting, Solar system design.**

I. INTRODUCTION

SEVERE loadshedding in South Africa disrupts commercial operations leading some businesses to suspend production. In response to the challenges posed by loadshedding, the banking sector is implementing programs to offer financial support to enable the implementation of solar projects [1]. The banking sector's keen interest in South Africa is commendable due to the country's geographic location, providing it with ample opportunities to harness substantial solar energy.

Stand-alone or off-grid solar systems have a potential to alleviate power challenges faced by South Africans. Although implementing these solar systems are costly, it will be worthwhile over a long period of time since they have minimal running costs. Designing and scaling solar systems (stand-alone or off-grid) requires an understanding of solar irradiance in the targeted location. Accurate solar forecasting, and precise solar systems, are essential factors in determining optimal financial models for solar projects.

The primary factor encountered when forecasting solar irradiance data is the presence of seasonality with fluctuations influenced by the Earth's axial tilt and orbit around the Sun [2]. This inherent seasonality in solar irradiance data necessitates the use of advanced analytical tools such as machine learning techniques and time series models to effectively capture patterns of solar irradiance variation over time [3, 4]. To address the existing challenges in forecasting solar irradiance, this study aims to adopt the machine learning approach wherein various machine learning models are implemented, and compared, based on their ability to grasp seasonal solar irradiance patterns, and forecasting future values.

A. *Objectives*

The objective of this study is to conduct exploratory data analysis on solar irradiance data obtained from the Southern African Universities Radiometric Network (SAURAN). The investigation aims to explore the relationship between solar irradiance and various weather factors. Additionally, the study seeks to implement and evaluate the performance of different machine learning models documented in the existing literature, focusing specifically on their suitability in predicting seasonal solar irradiance patterns in South Africa. The selected models include Random Forest (RF), k-Nearest Neighbors (KNN), Feedforward Neural Network (FFNN), Support Vector Regression (SVR), and eXtreme Gradient Boosting (XGBoost). The study further aims to forecast future monthly solar irradiance, comprehend how solar irradiance can inform solar systems design and financing, and identify potential directions for future research.

B. *Significance of The Study*

This research lays the groundwork for optimizing solar system designs, ensuring their accurate scaling to meet the energy consumption demands of customers. By providing valuable insights into solar energy forecasting and performance, the study contributes to informed investment and financing strategies within the solar energy sector. This not only reduces the risk of project delays and financial strain but also promotes long-term energy cost-savings and financial stability. The findings of this research have practical implications for enhancing the efficiency and sustainability of solar energy projects, ultimately supporting the growth and viability of the solar energy sector.

II. LITERATURE REVIEW

A. *Solar Irradiance Components*

Solar irradiance represents the total solar energy received at the Earth's surface and can be classified into three types: Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI) [4, 5]. GHI combines both direct normal and diffuse horizontal

This work was not supported by any institution.

Ronewa Mabodi is with the University of KwaZulu-Natal, Westville, Durban, 4000 (email: ronewamabodi@gmail.com).
Jahvaid Hammujuddy is with the University of KwaZulu-Natal, Westville, Durban, 4000 (email: Hammujuddy@ukzn.ac.za).

irradiance, accounting for the Sun's angle. Direct normal irradiance reaches the Earth's surface directly from the Sun, while diffuse irradiance results from scattering and reflection by atmospheric particles before reaching the surface.

B. Solar Irradiance Studies

There is a growing interest in understanding the solar irradiance patterns in South Africa [6]. Several studies have been presented with focus on modelling the behavior of solar irradiance. SAURAN has taken an initiative to provide researchers in the country with access to data and fostering a deep understanding in this field [7].

Several researchers have identified time series models and machine learning models as adequate modelling techniques to predict the seasonal nature of solar irradiance. Time series modelling is found to be a reliable tool in forecasting seasonal data, while machine learning models have proven to be a good tool in forecasting any non-linear data [8, 9, 10]. A study by [9] made a comparison of machine leaning models and time series models such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) in forecasting solar irradiance data [9]. The results of the study showed that modelling solar irradiance using machine learning models is more efficient than time series models.

Reference [3] conducted a study utilizing seasonal solar radiation data from three stations in Malaysia to forecast solar irradiance. The hybrid SARIMA and Artificial Neural Network (ANN) model were employed for this purpose. The overall performance between ANN and SARIMA was closely related across the three regions. Notably, in the dataset from Kluang, Malaysia, the ANN model outperformed SARIMA in both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Reference [6] conducted a recent study, marking the first use of Long Short-Term Memory (LSTM) networks in the South African context for forecasting seasonal solar irradiance. The study analyzed the seasonal nature of solar irradiance and compared the performance of LSTM with Support Vector Regression (SVR) and Feedforward Neural Network (FFNN) models. Variable selection employed the Least Absolute Shrinkage and Selection Operator (LASSO). The results revealed that the FFNN model produced the most accurate forecasts, exhibiting superior performance in terms of both MAE and RMSE.

Reference [12] used Brazil as a case study to predict solar irradiance using machine learning algorithms. The variables with high importance were found to be temperature, relative humidity, season, and cloud cover. The study implemented SVR and ANN which were evaluated in terms of rMAE, rRMSE and R^2.

Reference [13] made prediction of solar irradiance using XGBoost model. The important variables were found to be temperature, relative humidity, and cloud cover. The models were evaluated in terms of rMAE, rRMSE and R^2.

Reference [14] predicted hourly solar irradiance from satellite data using LSTM, FFNN and XGBoost model. Performance of different models were compared on data captured in seven cities in India covering four different climatic conditions. The results of the study show that climate conditions of a particular region could be a factor in deciding the appropriate model. The variables with high importance were found to be temperature, relative humidity, pressure, and wind speed. The models were evaluated in terms of rRMSE.

Reference [15] presented and compared the performance of different machine learning models for solar irradiance forecasting. The models used are the SVR, XGBoost, Categorical Boosting (CatBoost) and Voting-Average (VOA). Feature selection was based on Pearson coefficient, random forest, mutual information, and relief. Variables used in the study are temperature, relative humidity, wind speed, atmospheric pressure, and period (hour, day, and month). The models were evaluated in terms of rRMSE and R^2.

Relevant literature indicates that in predicting solar irradiance, essential variables include temperature, relative humidity, wind speed, pressure, and cloud cover. Variable importance is assessed through methods such as LASSO, Pearson coefficient, random forest, mutual information, and relief. Commonly considered machine learning models are RF, XGBoost, CatBoost, ANN, LTSM, FFNN, and SVR. Performance evaluation metrics include MAE, RMSE, and R^2, with R^2 values in the literature ranging from 0.81 to 0.93. Additionally, rRMSE values vary from 3.3% to 33.9% in the reviewed literature.

III. METHODOLOGY

The research commences by establishing clear goals and objectives, followed by an extensive literature review to identify gaps in the field. Hourly solar irradiance data from the reliable SAURAN station in Pretoria is used. Data preprocessing, including cleaning and normalization, is performed to enhance data quality. Machine learning models are chosen based on literature review results and recommendations. These models undergo training and testing phases, with performance evaluation metrics such as rRMSE, rMAE, and R^2. Model optimization is implemented for increased accuracy, and the results are thoroughly analyzed and interpreted in the context of solar irradiance forecasting and its impact on solar system design and financing decisions. Validation against existing literature ensures the reliability of the research findings, which are then documented comprehensively, providing a detailed report for reference and guiding future studies.

A. Data Overview

The SAURAN dataset provides comprehensive solar irradiance data on both daily and hourly averages, covering a significant timeframe dating back to September 20, 2013 [7]. This dataset offers valuable insights into solar irradiance patterns and trends over a substantial period, enabling in-depth analysis and understanding. The figure below depicts the geographical distribution of the SAURAN stations across South Africa.

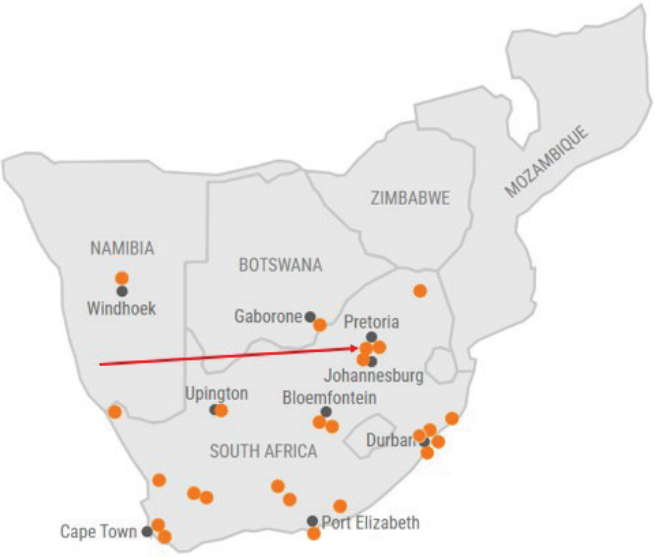

Fig. 1. Location of the dataset used in the study.

The dataset comprises of 20 variables, including 19 continuous variables and one date variable, totaling 82,847 records. The target variable is Global Horizontal Irradiance (GHI), with 44,861 records obtained between sunrise and sunset hours used in this study. Evening data are excluded as there is no direct solar radiation during this period, considered as noise that could introduce inaccuracies and inconsistencies into forecasting models, potentially reducing their reliability [14].

B. *Data Splitting*

The dataset is divided into a 70% training set and a 30% testing set, with the latter serving as a benchmark to evaluate model performance on new, unseen data. This split is conducted to prevent data leakage, where information from the testing dataset unintentionally influences the training of the machine learning model [16]. This precautionary measure ensures that the models are exposed only to relevant training data, enhancing their accuracy in generalizing to unseen solar irradiance data. The *train_test_split* function from the *sklearn.model*_selection library in Python is used to split the data.

C. *Missing Values*

Missing values contribute to less than 1% (908 records) of the dataset. All records with missing values are excluded from the dataset. Additionally, records with zero irradiance are removed, as it is expected that during daylight hours, GHI value is greater than zero [15]. Missing irradiance may signal a reset, or recalibration during that period. The *isnull()* function from the Pandas library was used to identify missing values.

D. *Handling Outliers*

This study makes use of the IQR method on variables to identify outliers on the response variable. The lower and upper bounds of the model are defined to be:

$$IQR=Q3\text{-}Q1 \tag{1}$$

$$Lower\colon Q1\text{-}1.5\times IQR \tag{2}$$

$$Upper\colon Q3+1.5\times IQR \tag{3}$$

No outliers were identified on the 44861 records captured during daylight hours implying there is no significant deviation on the response variable. The RF model would not have been impacted as much by outliers if they were present due to their ensemble nature [17, 18]. However, models such as KNN that calculates distances between points to make predictions would have been impacted by potential biases introduced [18]. The absence of outliers ensures a fair comparison of models as no bias is introduced.

E. *Conversions and Aggregations*

The hourly solar irradiance data was transformed into monthly averages by applying the *resample()* function in Pandas, followed by calculating the mean using the *mean()* function. The monthly data allows for easier identification of seasonal patterns which is more meaningful in the context of this research as described in the objectives and relevant literature. The final dataset contains 114 monthly records ranging from January 2014 to July 2023.

F. *Normalization*

Z-score normalization was employed to standardize features in the SAURAN dataset. This transformation ensures that features have a mean of 0 and a standard deviation of 1, preventing features from dominating the model due to differing ranges. Equation

$$Z = \frac{X-\mu}{\sigma} \tag{4}$$

where *X* represents an individual value in the dataset, μ is the mean and σ is the standard deviation of the data.

G. *Variables Selection*

Two approaches were employed to select variables for this study. Firstly, variables from previous studies [6, 12, 13] which are also available in the University of Pretoria SAURAN dataset were considered. Secondly, a correlation plot on the SAURAN dataset was utilized to visually summarize and analyze the strength and direction of relationships between the response variable and independent variables. Correlations are presented in the table below:

TABLE I

CORRELATION BETWEEN KEY VARIABLES

Variable	Correlation strength	Value Range
Temperature (℃)	0.46	0.45 - 37.47
Wind direction standard deviation (°)	0.36	1.16 -77.03
Wind direction (°)	0,17	0.04 - 360.00
Month	0.07	1-12
Wind speed (m/s)	0.06	0.04 - 360.00
Pressure (mbar)	-0.06	805.00 - 878.00
Relative Humidity (%)	-0.36	5.09 - 99.70

H. Exploratory Data Analysis

Exploratory data analysis was done in Python. The Pandas library was used for data manipulation, NumPy for numerical computing, Matplotlib for plotting, Seaborn for statistical visualization, and Statsmodels for statistical modeling and hypothesis testing.

1. The Response Variable

The Global Horizontal Irradiance (GHI) serves as the dependent variable for predicting solar irradiance. Monthly trends reveal a peak irradiance during the summer months, indicating their suitability for optimal solar energy generation. Conversely, winter months exhibit reduced solar irradiance. Analyzing these monthly trends is crucial for optimizing energy designs, influencing decisions regarding system size and required energy storage capacity.

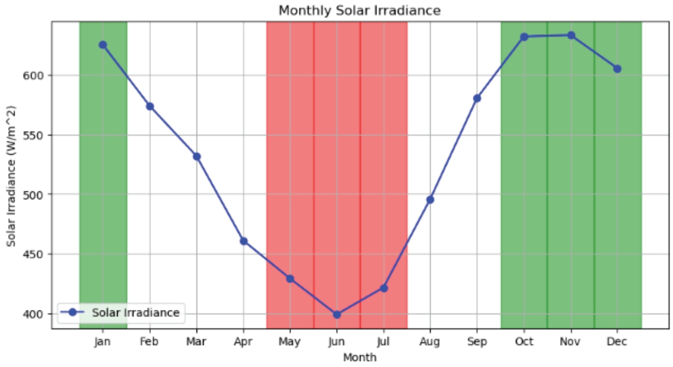

Fig. 2. Monthly solar irradiance.

The figure below shows that solar irradiance exhibits significant seasonal variations. This property exhibited by the solar irradiance influences features selection, how training and validation should be done, and which prediction models to be considered as discussed in the methodology section.

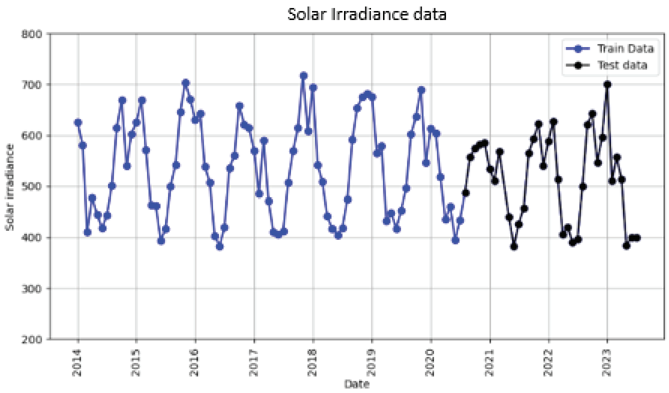

Fig. 3. Yearly solar irradiance.

2. Temperature

The following figure shows the relationship between temperature and solar irradiance. There is a direct and positive relationship between temperature and the amount of solar energy received at the Earth's surface. Temperature values are higher during the summer season. The air is warmer, less dense and allows for more efficient solar energy conversion in photovoltaic panels. The sunlight hours are longer during summer seasons and the angle of the Sun is more direct, leading to increased solar radiation and warmer [20].

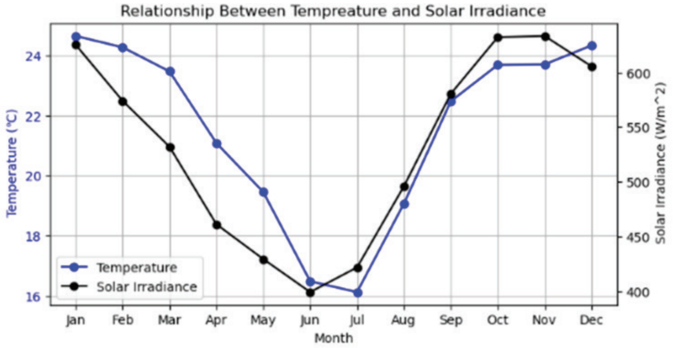

Fig. 4. Temperature vs Solar irradiance.

3. Wind Speed

The figure below reveals a significant correlation between monthly wind speed and solar irradiance, not apparent in the correlation table, suggesting a potential lead effect. Winter months (June, July, and August) show the lowest average wind speed during cooler temperatures. South Africa's transition from winter to spring brings a shift in pressure systems, resulting in more dynamic weather conditions, stronger winds, and the observed relationship in the figure [11].

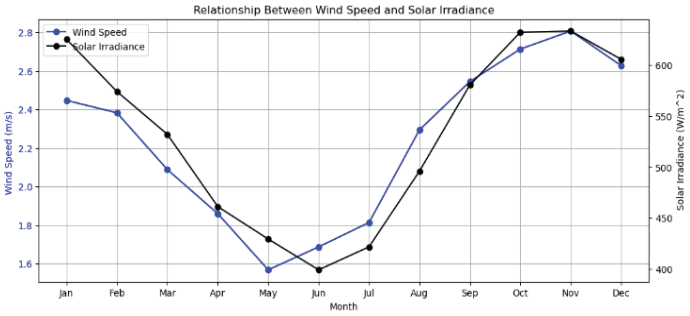

Fig. 5: Monthly Wind Speed vs Solar irradiance.

4. Wind Direction

The figure below shows the average monthly wind direction. There is no clear pattern between wind direction and patterns of solar irradiance. Studies by [19] mention that the two are interconnected in various ways, and their relationship can be influenced by other external geographical factors.

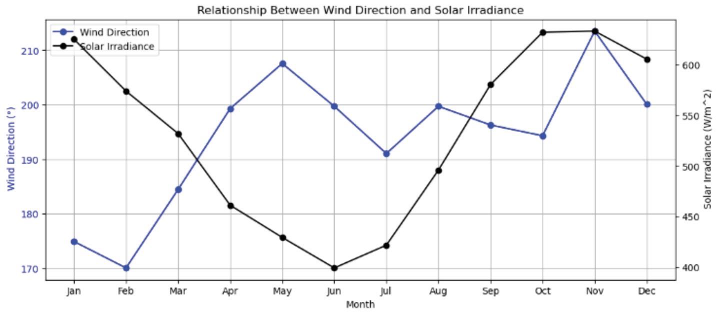

Fig. 6. Monthly Wind direction vs Solar irradiance.

5. *Wind Direction Standard Deviation*

The figure below indicates the correlation between wind direction standard deviation and the response variable. June has the lowest deviation, while March, April, and October show the highest values, suggesting increased atmospheric turbulence, impacting solar irradiance.

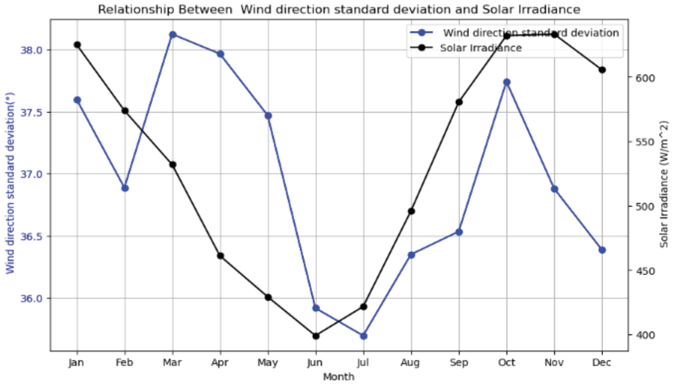

Fig. 7. Monthly Wind direction standard deviation vs Solar irradiance

6. *Relative Humidity*

The figure below shows the average relative humidity per month. Humidity is lower between June and October, associated. While a subtle correlation between solar irradiance and humidity is noted, the study suggests that relative humidity may not be a decisive factor in determining solar irradiance in the studied location. Despite high relative humidity in December, January, and February, the figure below indicates that June and October have similar humidity levels, yet October exhibits significantly higher solar irradiance, challenging the perceived importance of relative humidity in solar irradiance determination [12, 13, 21].

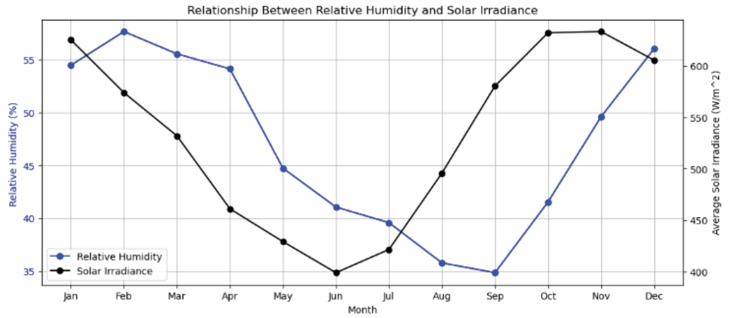

Fig. 8. Relative Humidity vs Solar irradiance.

7. *Barometric Pressure*

The graphical representation of monthly barometric pressure below shows an inverse correlation between barometric pressure and solar irradiance. An increase in barometric pressure tends to correspond to a reduction in solar irradiance, and conversely, a decrease in barometric pressure is associated with an increase in solar irradiance.

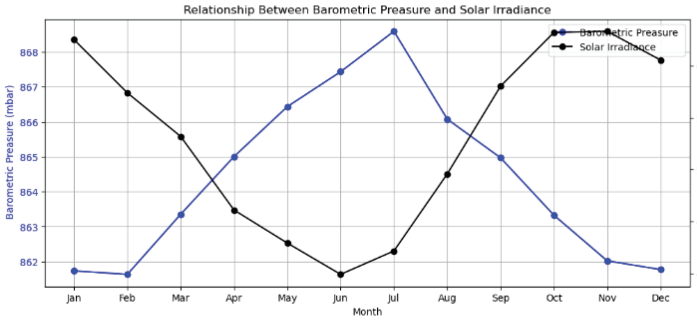

Fig. 9. Relative Humidity vs Solar irradiance.

IV. MODEL SELECTION

Five machine learning models were examined and are described below. The models were implemented using Python, leveraging libraries such as Scikit-learn, TensorFlow, NumPy, Pandas, Matplotlib, and Seaborn. These libraries are utilized to implement KNN, RF, XGBoost, FFNN, and SVM algorithms.

A. *Random Forest*

The Random Forest (RF) model is chosen to model solar irradiance data due to its capacity to handle non-linearity, rank feature importance, and utilize ensemble learning [17]. For continuous solar irradiance, RF forms forests by growing trees based on a random vector ϴ, where each tree predictor h(x, ϴ) yields numerical values [17]. The model provides feature importance scores by permuting or shuffling independent variables, assessing their impact on predictive accuracy. This process is repeated for each predictor, and rankings are determined based on the differences between accuracy using original and the shuffled data [17]. Implemented using *RandomForestRegressor* in *scikit-learn*, the RF model undergoes hyperparameter tuning, resulting in optimal parameters {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}. These parameters allow trees to grow without depth limits, prioritize precision per leaf, and ensure robust decision-making with a minimum of two samples required for splitting internal nodes. With 50 trees, the model aims for a balanced complexity and ensemble diversity.

1. *Handling Non-linearity*

Solar irradiance is influenced by non-linear relationships that exist between various weather conditions/environmental factors. Random forest method makes use of its inherent characteristics and the ensemble of decision trees to capture these non-linearities [22].

2. Addressing Limitations

While RF model is less prone to overfitting compared to other models, it can still overfit the training data if the number of trees is very high. This study makes a careful parameter tuning to prevent any potential overfitting.

B. k-Nearest Neighbors

The k-Nearest Neighbors (KNN) model, a lazy learner technique, retains the training dataset for predictions and waits until a test dataset is provided [18]. It predicts numerical targets based on the average response of a predefined number k of nearest neighbors. Cross-validation is employed in this study to determine the optimal k value, considering the trade-off between noise, stability, overfitting, and computational cost.

The KNN model, implemented in Python using *KNeighborsRegressor* from *scikit-learn*, optimizes the choice of k through *GridSearchCV*, resulting in optimal hyperparameters {neighbors: 5, p: 1, weights: distance}. This setting considers five nearest neighbors, uses the Manhattan distance metric (L1 norm), and employs a distance weighting scheme that gives higher importance to closer neighbors, emphasizing local patterns for precise solar irradiance predictions. The KNN model's advantage lies in its ability to make minimal assumptions about the underlying data distribution, making it suitable for solar irradiance data with a lack of consistent fixed patterns observed in time series data.

1. Non-linearity and Local Patterns

Solar irradiance involves non-linear and local patterns, where the current irradiance level is influenced by seasonal weather conditions and historical data. KNN becomes a good candidate model to predict solar irradiance as it makes predictions based on the k-nearest data points in the feature space.

2. Addressing Limitations

The KNN model is sensitive to outliers, impacting nearest neighbors significantly [18]. Addressing outliers, as discussed in Section 3, is crucial. While quick to implement, the model's prediction time increases as it searches through all training set points to find the nearest ones. For large datasets, KNN can be slower than other regressions, prompting the use of a monthly dataset instead of daily or hourly data to mitigate this limitation.

C. eXtreme Gradient Boosting

The XGBoost model, recognized for overcoming the limitations of single machine learning models, operates as an ensemble learning method. It combines multiple learners to create a single model that incorporates results from various models, employing a gradient boosting technique where trees are sequentially built to reduce errors. The base learners in XGBoost are weak, and the resulting model integrates them to form a robust learner that minimizes both bias and variance.

The XGBoost model undergoes hyperparameter tuning using GridSearchCV, resulting in optimal parameters: {learning_rate: 0.2, max_depth: 3, n_estimators: 50}. A learning rate of 0.2 ensures cautious model adjustments, preventing overshooting, while a depth limit of 3 maintains a balanced complexity to capture essential patterns without overfitting. With 50 estimators, the model efficiently balances computational resources and performance, making it valuable for effective generalization in solar irradiance predictions.

1. Regularization

XGBoost incorporates built-in L1 and L2 regularization to prevent overfitting by penalizing model complexity. This feature enhances robust predictions, particularly when dealing with a limited number of records in the dataset [23].

2. Addressing Limitations

XGBoost may overfit training data, especially with numerous trees [23]. Cross-validation tunes hyperparameters and evaluates generalization. XGBoost is sensitive to outliers, impacting performance. Outliers are handled before model fitting to address this weakness.

D. Feedforward Neural Network

In this study, the FFNN model is chosen for its effectiveness in capturing intricate, non-linear patterns in data and its adaptability to diverse data characteristics, enabling it to approximate any continuous function for enhanced flexibility [24].

The FFNN model consists of an input layer, hidden layers, and an output layer, processing data sequentially with neurons in each layer utilizing activation functions on weighted inputs to capture patterns in the response variable and its relationships within the independent variables [24].

During training on solar irradiance data, the FFNN model adjusts weights through backpropagation, minimizing the difference between predicted and actual solar irradiance for improved accuracy. Implemented in Python using TensorFlow's Keras API, the FFNN model is wrapped in a *scikit-learn* pipeline using the *make_pipeline* function. It is compiled using the Adam optimizer with a learning rate of 0.01, undergoing 100 epochs of training with a batch size of 32. The optimal parameters include an input layer matching the number of features, two hidden layers with 512 and 256 neurons and ReLU activation function, dropout applied in a layer with a 0.4 dropout rate, and an output layer with a linear activation function for regression.

1. Generalization Capability

The backpropagation in training FFNN models optimizes the neural network's weights to minimize the error or loss function. This optimization process enables the FFNN model to capture seasonal patterns of solar irradiance rather than memorizing specific instances. The likelihood of the model demonstrating effective generalization to new, unseen data is increased, whilst maintaining predictive consistency in both training and testing data [24].

2. Addressing Limitations

FFNN model does not possess memory of past inputs which might make it struggle with capturing the seasonal nature of solar irradiance as each input is processed independently. Considering the number of records available and the split between training and testing data, this model has the potential to rival other competing models.

E. Support Vector Regression

SVR is a type of support vector machine that is used for regression tasks. It finds a function that best predicts the continuous output value for a given input value. This research uses Radial Basis Function (RBF) over a linear kernel as it excels in capturing non-linear relationships, a crucial characteristic for modelling seasonal patterns presents in solar irradiance [25]. The SVR model is implemented in Python, and it uses the *GridSearchCV* for hyperparameter tuning. The identified optimal hyperparameters includes the regularization parameter(C) of 100 and an epsilon of 0.001. The 'C' parameter impacts the trade-off existing between a smooth decision boundary and accurate fitting to the training data. A higher 'C' value of 100 implies the model's preference for accurate fitting, while the low epsilon 0.001 implies narrow margin, which enhances precision in predicting solar irradiance values.

1. Handling Non-linearity

SVR focuses on reducing a combined measure of errors during training and a term that helps control complexity of the model [25]. The inclusion of the RBF kernel is proficient in capturing non-linear relationships which is crucial for modelling complex patterns in solar irradiance data.

2. Addressing Limitations

The SVR model is sensitive to outliers and the choice of kernels as their performance depends on the characteristics of the data. This research work addresses outliers in Section 3 and a comprehensive exploratory data analysis was done to understand the characteristics of the data and a suitable kernel.

V. RESULTS AND DISCUSSION

A. Model results

The five models implemented are evaluated in terms of R^2, rRMSE and rMAE. Examination will focus on both the model's sensitivity to data fluctuations (variance) and the error from a model's simplifying assumptions, causing deviations from true values (bias). A bias-variance trade-off is made when choosing a better model with an optimal level of complexity that captures underlying patterns without being overly influenced by noise.

1. R-squared (R^2)

R^2 is a statistical measure that depicts the proportion of the variance for a dependent variable that is explained by an independent variable or the percentage of variance for a dependent variable that is explained by independent variables, and it is given in the equation below [15].

$$R^2 = 1 - \frac{SSR}{SST} \tag{5}$$

where *SSR* is the sum of squared residuals and *SST* is the total sum of squares. For the three machine learning models, the study makes use of R^2 to understand the proportion of the variance in the response variable that is explained by the combination of predictor variables. This metric plays an important role in this study as it assesses how well the model captures the relationships between the response and predictor variables, considering both linear and non-linear patterns. The table below shows the performance of the model in terms of R^2 in both the training and testing data:

TABLE II
R^2 RESULTS

Metric	**Model**				
R-squared (R^2)	**RF**	**KNN**	**XGBoost**	**SVR**	**FFNN**
Training data	0.986	0.997	0.997	0.964	0.895
Testing data	0.823	0.891	0.885	0.861	0.777

The five models exhibit effective generalization of the training data well with KNN and XGBoost achieving the highest R^2 implying that the model explains the entire variability in the response variable using predictor variables when looking at the data provided for training. FFNN model has the lowest R^2, showing its inability to explain variability in the response variable using predictor variables when compared to other models. The SVR model demonstrates consistent performance as it achieves the lowest variance. The KNN maintain consistent performance when applied to new, unseen testing data as it also has a leading performance. KNN is considered the best model in terms of R^2 followed by XGBoost as they strike a good balance between bias and variance.

B. Relative RMSE

The rRMSE is a normalized measure of the accuracy of a predictive model and it is found by dividing the RMSE by the mean of the observed (actual) value in the dataset and multiplying by 100 to express it as a percentage [21].

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(F_t - O_t)^2} \tag{6}$$

$$Relative\ RMSE = \frac{RMSE}{Mean(O)} \times 100\% \tag{7}$$

where *n* is the number of observations, F_t is the predicted value, O_t is the observed (actual) value, and $Mean(O)$ is the mean of the observed values. The table below shows the performance of the model in terms of rRMSE in both the training and testing data:

TABLE III
RELATIVE RMSE RESULTS

Metric rRMSE	**Model**				
	RF	**KNN**	**XGBoost**	**SVR**	**FFNN**
Training data	2.12%	0.41%	0.95%	3.52%	5.70%
Testing data	6.95%	5.77%	5.91%	6.17%	8.22%

A model with the lowest values in terms of bias-variance trade-off is considered the best. KNN perfectly generalized the training data without an rRMSE error and achieves the best performance on the testing data. FFNN achieves the least favorable performance in both the training data, making it the least favorable model to be used in solar irradiance forecasting.

Although FFNN has the unfavorable performance when compared to other models, it is the most consistent model as it has the lowest variance. The KNN model appears to strike a good balance between bias and variance, as it performs well on both training and testing data in terms of rRMSE and it is voted the best model. XGBoost is considered the second-best model as it also demonstrates strong performance across training and testing datasets with.

C. Relative MAE

Mean Absolute Error (MAE) is a metric used to evaluate the accuracy of a model's predictions. It is calculated as the average of the absolute differences between the predicted and actual values and the formula is shown below [15].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} O_t - F_t \qquad (8)$$

where n is the number of observations, F_t is the predicted value, and O_t is the observed (actual) value. Relative Mean Absolute Error (rMAE) is a variant of MAE that is often used in the context of forecasting for easy comparison with previous studies [15]. It expresses the MAE as a percentage of the average of the actual values and is shown on the formula below:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|O_t - F_t|}{F_t} \times 100 \qquad (9)$$

TABLE IV
RELATIVE MAE RESULTS

Metric rMAE	**Model**				
	RF	**KNN**	**XGBoost**	**SVR**	**FFNN**
Training data	1.65%	0.29%	0.75%	2.31%	4.44%
Testing data	5.34%	4.51%	4.74%	4.79%	6.36%

The study evaluates the performance of machine learning models, particularly highlighting KNN and the XGBoost model as top performers in the testing data based on three metrics. Permutation importance from *scikit-learn* is employed to assess the significance of various weather factors in predicting solar irradiance. This method involves shuffling the values of individual weather features to observe their impact on the model's predictive accuracy. The results reveal that temperature emerges as the most influential variable in determining solar irradiance, consistent with correlation strength results presented in Section 3 and illustrated in Figure 5, emphasizing the relationship between temperature and solar irradiance.

TABLE V
FEATURE IMPORTANCE

Variables	**Importance**
Temperature	42.45%
Wind direction Std Dev	14.34%
Barometric Pressure	12.11 %
Month	10.75%
Wind Speed	10.24%
Relative Humidity	10.11%

D. Results Summary

The KNN model exhibited superior performance with an rRMSE, rMAE, and R^2 of 5.77%, 4.51% and 0.89 respectively on testing data. The KNN model's distinguishing feature in predicting solar irradiance lies in its adeptness in capturing localized patterns and adjusting to diverse spatial dependencies present in the SAURAN dataset. The predictions made are based on the similarity of data points in the feature space, considering the *k*-nearest neighbors to the query point. XGBoost model emerged as the second-best performing model. FFNN model was the lowest performing model. The figure below shows the graphical view of how the KNN performance in predicting solar irradiance.

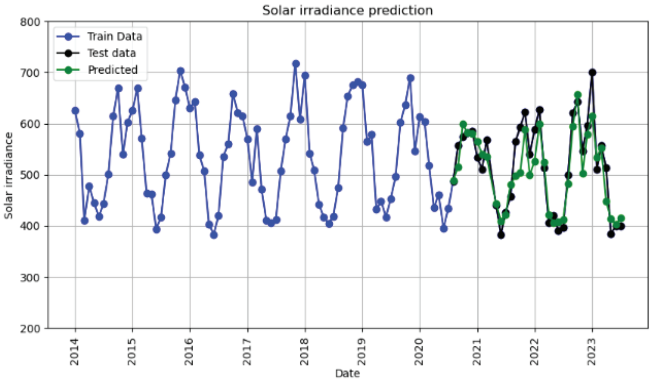

Fig. 10. KNN predictive performance.

E. Future Value Forecasting

The KNN model, identified as the most effective for predicting solar irradiance, is employed in forecasting future values for the next 13 months by initiating the sequence from the last month in the dataset. To determine future feature values, this study adopts an approach that considers both long-term patterns and short-term variations in solar irradiance. The method involves examining the midpoint between the averages of previous months and contrasting it with the deviation observed in the most recent month. This choice is informed by the need to account for both long-term trends and short-term fluctuations, as exemplified by a significant dip in solar irradiance observed in 2020 due to lockdown conditions. The dip was associated with a reduction in temperature between 2020 and 2021, influenced by decreased heat emissions during lockdown measures [26]. The KNN model is then fitted with the estimated feature values, and the forecasting results are presented below.

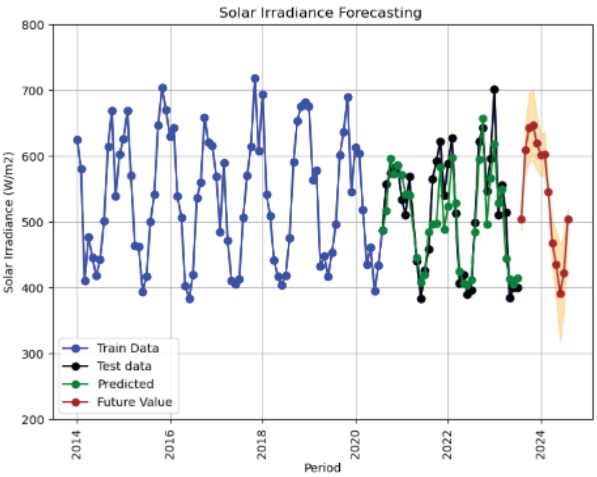

Fig. 11. Solar irradiance forecasting.

F. *Suitability For Solar System Design and Financial Decisions*

The study demonstrates strong suitability for informing solar system design and financial decisions, evident in the high R^2 value of 0.89, rMAE of 4.51%, and rRMSE of 5.77% on the testing data. The results, validated against the relevant literature, showcase competitiveness and reliability. The research potential to optimize energy production benefits both individual commercial customers and the national grid, aiding in loadshedding challenges and capacity planning for Eskom. Additionally, it supports precise financial planning for solar projects, offering insights into energy harnessing, facilitating accurate return on investment estimation, and guiding financial decisions for banks, businesses, and project developers.

VI. CONCLUSION

This study focused on implementing and evaluating machine learning models for predicting solar irradiance, aiming to inform solar system design and financing decisions. RF, KNN, FFNN, SVR, and XGBoost were assessed, with the KNN model outperforming others, exhibiting a relative RMSE, relative MAE, and R^2 of 5.77%, 4.51%, and 0.89, respectively, on testing data. The evaluation involved a trade-off between bias and variance to determine the best model among the five. Influential variables included temperature, wind direction standard deviation, and barometric pressure, contributing 42.5%, 14.34%, and 12.11%, respectively. The KNN model emerged as a reliable asset for solar energy system design and financial assessments in South Africa. Future research avenues may explore the KNN model's performance across multiple SAURAN datasets and investigate the feasibility of a hybrid approach combining machine learning and time series models for solar irradiance forecasting.

VII. DIRECTION FOR FUTURE STUDIES

Future studies in solar irradiance forecasting based on the findings of this study and identified gaps in the relevant literature can take two immediate directions. Firstly, focusing on the diverse stations within SAURAN across Inland, Coastal, and Desert areas, researchers can assess the performance of the five models explored in this study across datasets from these distinct regions. This examination aims to provide practical insights into model adaptability, robustness, and regional dependencies, aiding in model selection for specific regions and contributing to ongoing improvements in solar energy forecasting. Secondly, future studies can strategically compare the performance of various time series models with different machine learning models, exploring the potential of hybrid models that integrate both approaches. Hybrid models can leverage the strengths of time series modelling in forecasting seasonal patterns and machine learning models in interpreting complex non-linear data dynamics.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] N. Isibor, “Eskom News Top Banks Offering Solar Finance in South Africa.” 2023. Available: https://eskomnews.co.za/top-banks-offering-solar-finance-in-south-afria/.

[2] A. M. Soomar, A. Hakeem, M. Messaoudi, P. Musznicki, M. Iqbal and S. Czapp, “Solar Photovoltaic Energy Optimization and Challenges.” Frontiers in Energy Research, 10, Article 879985, 2022.

[3] M.Z. Mukaram and F. Yusof, “Solar radiation forecast using hybrid SARIMA and ANN model: A case study at several locations in Peninsular Malaysia.” UTM Press, 2017.

[4] H.W. Mkasi, M.B. Ayanna, L.E. Pratt and K.T. Roro, “Global irradiance on photovoltaic array.” 5th Southern African Solar Energy Conference, South Africa, 2018.

[5] M. K. Boutahir, Y. Farhaoui, M. Azrour and I. Zeroual, “Effect of Feature Selection on the Prediction of Direct Normal Irradiance.” Big Data Mining and Analytics, 2022.

[6] T. Mutavhatsindi, C. Sigauke and R. Mbuva, “Forecasting Hourly Global Horizontal Solar Irradiance in South Africa Using Machine Learning Models.” IEEE Access, 10, 1-1, 2020.

[7] Southern African Universities Radiometric Network (SAURAN). “Solar Radiometric Data for the Public.” 2023. Available at: https://sauran.ac.za/

[8] N. Azizi, M. Yaghoubirad, M. Farajollahi, and A. Ahmadi, “Deep learning based long-term global solar irradiance and temperature forecasting using time series with multi-step multivariate output.” Renewable Energy, 206(C), 135-147. Elsevier, 2023.

[9] H. Sharadga, S. Hajimirza and R.S. Balog, “Time series forecasting of solar power generation for large-scale photovoltaic plants.” Renewable Energy, Volume 149, Pages 1274-1286, 2019.

[10] I. Jebli, F.Z. Belouadha, M.I. Kabbaj and A. Tilioua, "Deep learning-based models for solar energy prediction" Advances in Science, 6, 349–355, 2021.

[11] L. Shikwambana, K. Xongo, M. Mashalane, and P. Mhangara, "Climatic and Vegetation Response Patterns over South Africa during the 2010/2011 and 2015/2016 Strong ENSO Phases. Atmosphere" 14(2), 416. 2023.

[12] G.F. Viscondi and S.N. Alves-Souza, "Solar Irradiance Prediction with Machine Learning Algorithms: A Brazilian Case Study on Photovoltaic Electricity Generation" 2021.

[13] P. S. Kamarouthu, "Solar Irradiance Prediction Using Xg-boost With the Numerical Weather Forecast." Utah State University, 2020.

[14] D. Toshniwal and P. Kumari, "Hourly Solar Irradiance Prediction From Satellite Data Using LSTM." 2019.

[15] E.S. Solano, P. Dehghanian and C.M. Affonso, "Solar Radiation Forecasting Using Machine Learning and Ensemble Feature Selection." Energies, 15, 7049. 2022.

[16] H. Weytjens and J. De Weerdt, "Creating Unbiased Public Benchmark Datasets with Data Leakage Prevention for Predictive Process Monitoring." Research Centre for Information Systems Engineering (LIRIS), KU Leuven, Leuven, Belgium. Preprint. 2021.

[17] L. Breiman, "Random Forests. Machine Learning" 45, 5–32, 2001.

[18] C. M. Bishop, "Pattern Recognition and Machine Learning." Springer New York, NY, 2006.

[19] K. Mohammadi, and N. Goudarzi, "Study of inter-correlations of solar radiation, wind speed and precipitation under the influence of El Niño Southern Oscillation (ENSO) in California." Renewable Energy, Volume number not provided, 2017.

[20] Time and Date. "Sun & moon times today" Pretoria, South Africa.Available:https://www.timeanddate.com/astronomy/south- africa/pretoria.

[21] S. Mujabar and R.C. Venkateswara, "Empirical models for estimating the global solar radiation of Jubail Industrial City." Solar Energy, 216, 603-612, 2021.

[22] J. Liu, M.Y. Cao, D. Bai, and R. Zhang, "Solar radiation prediction based on random forest of feature-extraction." China, 2019.

[23] C. Bentejac, A. Csorgo and G.A. Martinez-Munoz, "Comparative Analysis of XGBoost" Spain, 2020.

[24] M.H.A. Sazli, "Brief Review of Feed-Forward Neural Networks. Communications Faculty of Sciences University of Ankara." Series A2-A3 Physical Sciences and Engineering, 50(1), 11-17, 2006.

[25] D. Basak, S. Pal, and D.C. Patranabis, "Support Vector Regression. Statistics and Computing" November 2007. Neural Information Processing – Letters and Reviews, 11(10), 203, 2007.

[26] L. R. Ray and P. Singh, "What is the impact of COVID-19 pandemic on global carbon emissions?" science of The Total Environment, 816, 151503, 2022.

[27] X. Li, Y. Wang, and S.A. Basu, "Debiased MDI Feature Importance Measure for Random Forests." USA, 2019.

Ronewa Mabodi received a BSc Eng Electrical Engineering degree from the University of the Witwatersrand, Johannesburg, South Africa in 2018 and a Postgraduate Diploma in Data Science from the University of KwaZulu-Natal, Durban, South Africa in 2023.

In 2018, he was a student researcher in the School of Electrical and Information Engineering at the University of the Witwatersrand, Johannesburg, South Africa. From 2019 to 2022, Mr. Ronewa Mabodi worked as a Strategy and Technology consultant in several energy firms in South Africa. He is currently working as a Data Analyst within the banking industry in South Africa. His research interests lie in the design, financing, and optimization of renewable energy systems.

Jahvaid Hammujuddy received the BScHons and the MSc degrees in Statistics from the University of Natal, Durban, South Africa, in 2000 and the University of KwaZulu-Natal, Durban, South Africa, in 2004.

From 2000 to 2004, he was a Teaching Assistant in the School of Mathematical & Statistical Sciences at the University of Natal and University of KwaZulu-Natal, Durban, South Africa. From 2006 to 2008, he was employed as a Biostatistician at the Centre International de Développement Clinique Ltée, Mauritius. Since 2009, he has been a Lecturer with the School of Mathematics, Statistics & Computer Science, University of KwaZulu-Natal, Durban, South Africa. He is the co-author of three articles and two refereed conference proceedings. His research interests include statistical modelling and Public Health Data Science.

Mr. Jahvaid Hammujuddy is a member of the South African Statistical Association and a past member of the Institute of Certificated and Chartered Statisticians of South Africa.

CONTENTS PAGE