A Major Project Report

On

PROBABILISTIC SOLAR IRRADIANCE PREDICTION USING XGBOOST AND KDE

Submitted in the Partial Fulfillment of the Requirements

For the Award of the Degree of

Bachelor of Technology

In

CSE (Artificial Intelligence and Machine Learning)

By

Bollaramkummari Vishnupriya [22215A6606]

Under the Guidance of

Dr. G Venu Gopal	**Assistant Professor**
Mrs. Indumathi V	**Assistant Professor**

DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

B V RAJU INSTITUTE OF TECHNOLOGY

(UGC Autonomous, Accredited by NBA & NAAC)
Vishnupur, Narsapur, Medak, Telangana State, India – 502 313

2024-2025

B V RAJU INSTITUTE OF TECHNOLOGY

(UGC Autonomous, Accredited by NBA & NAAC)
Vishnupur, Narsapur, Medak, Telangana State, India – 502 313

CERTIFICATE

This is to Certify that the Major Project Entitled "**Probabilistic Solar Irradiance Prediction using XGBoost and KDE**" Being Submitted By

Bollaramkummari Vishnupriya	[22215A6606]

In Partial Fulfillment of the Requirements for the Award of Degree of Bachelor of Technology in CSE (Artificial Intelligence and Machine Learning) to B V Raju Institute of Technology is Record of Bonafide Work Carried Out During the Period from December 2024 to April 2025 by her Under the Supervision of

Dr. G Venu Gopal	**Assistant Professor**
Mrs. Indumathi V	**Assistant Professor**

This is to Certify that the Above Statement Made by the Student is Correct to the Best of Our Knowledge.

Dr. G Venu Gopal
Assistant Professor

Mrs. Indumathi V
Assistant Professor

The Major Project Viva-Voce For This Project Has Been Held on ____________.

External Examiner

Dr. G Uday Kiran
Program Coordinator

B V RAJU INSTITUTE OF TECHNOLOGY

(UGC Autonomous, Accredited by NBA & NAAC)

Vishnupur, Narsapur, Medak, Telangana State, India – 502 313

DECLARATION

I Hereby Certify that the Work Which is Being Presented in the Major Project Entitled “Probabilistic Solar Irradiance Prediction using XGBoost and KDE” in Partial Fulfillment of the Requirements For the Award of Degree of Bachelor of Technology and Submitted in the Department of CSE(Artificial Intelligence and Machine Learning), B V Raju Institute of Technology, Narsapur, is an Authentic Record of Our Own Work Carried Out During the Period From December 2024 to April 2025, Under the Supervision of Name of Co-Supervisor, Designation and Name of the Supervisor, Designation.

The Work Presented in this Major Project Report Has Not Been Submitted by me For the Award of Any Other Degree of This or Any Other Institute/University.

Bollaramkummari Vishnupriya [22215A6606]

B V RAJU INSTITUTE OF TECHNOLOGY

(UGC Autonomous, Accredited by NBA & NAAC)
Vishnupur, Narsapur, Medak, Telangana State, India – 502 313

ACKNOWLEDGEMENT

I stand at the culmination of a significant journey, one that has been both challenging and rewarding. The success of my major project is not solely a reflection of my efforts but a testament to the invaluable support and guidance that I have received from many quarters. It is with deep gratitude that I acknowledge those who have made this achievement possible.

Foremost, I extend my sincerest appreciation to Mrs. Indumathi V, Co-supervisor, and Dr. G Venu Gopal, Supervisor, whose expertise and insightful supervision have been pivotal in navigating the complexities of this project. Their unwavering support and encouragement have been our guiding light throughout this journey.

Special thanks are due to Ms. Srilakshmi V, our Project Coordinator, whose assistance and guidance have been instrumental in the successful execution of our project. Her dedication and support have been a source of inspiration and motivation.

I reserve our utmost gratitude to Dr. G Uday Kiran, Program Coordinator of the Department of CSE (Artificial Intelligence and Machine Learning), whose leadership and academic guidance have enriched my learning experience and contributed significantly to my project's success.

My journey would not have been the same without the constant encouragement, support, and guidance from the esteemed faculty of the Department of CSE (Artificial Intelligence and Machine Learning). I am deeply thankful to everyone who contributed to my journey, whose belief, guidance, and support have been crucial for my achievement. This project reflects not only my academic efforts but also the collaborative spirit and collective wisdom that guided me.

Bollaramkummari Vishnupriya [22215A6606]

ABSTRACT

Solar energy is gaining significant traction worldwide as a clean and renewable energy source. For power grid operators and researchers alike, being able to predict future solar power generation with a degree of certainty is incredibly valuable. This is where probabilistic solar irradiance forecasting comes in. Unlike simple point forecasts that give a single predicted value, probabilistic forecasting provides a range of possible future irradiance values along with their likelihood. This allows for better decision-making in grid management, energy trading, and overall system reliability.

This project introduces a new probabilistic solar irradiance prediction model built upon the powerful XGBoost algorithm. The process involves several key steps. First, the historical solar irradiance data undergoes careful preprocessing to ensure its quality and suitability for training. Next, this preprocessed data is used to train an XGBoost model for point prediction – essentially, a model that aims to predict a single "best guess" for future irradiance.

The significant part lies in how the model generates probabilistic forecasts. XGBoost works by iteratively building multiple decision trees and minimizing the prediction errors at each step. We leverage this iterative process. When forecasting solar irradiance for a specific future time, each of these individual trees within the trained XGBoost model produces its own prediction. This collection of predictions, generated from the different stages of the XGBoost model's training, inherently captures some of the uncertainty in the forecast.

Finally, to transform these multiple point predictions into meaningful probability prediction intervals, we employ the kernel density estimation (KDE) method. KDE is a non-parametric technique that allows them to estimate the probability distribution of the predicted irradiance values. This results in prediction intervals with associated confidence levels (e.g., a 90% prediction interval), indicating the range within which the actual irradiance is likely to fall with a certain probability.

The project validates its proposed method using publicly available datasets and compares its performance against other established forecasting techniques. The experimental results demonstrate that the XGBoost-based probabilistic model achieves higher prediction accuracy compared to these benchmark algorithms. Furthermore, the project highlights practical advantages of the proposed method, including its relatively short training time and straightforward parameter tuning, making it well-suited for real-world engineering applications.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

The dawn of the 21st century has been marked by a heightened global consciousness surrounding two intertwined and critical challenges: the escalating threat of anthropogenic global warming and the pressing realities of a finite and increasingly strained global energy supply. The pervasive impacts of climate change, ranging from rising sea levels and extreme weather events to disruptions in ecological balance, have spurred an urgent search for sustainable alternatives to traditional fossil fuels. Simultaneously, the growing global population and the relentless march of industrialization have placed unprecedented demands on existing energy resources, highlighting the inherent limitations and geopolitical vulnerabilities associated with their continued dominance.

In response to these dual crises, the past two decades have witnessed a significant surge in global attention towards clean and renewable energy sources. Solar energy, alongside wind power and tidal energy, has emerged as a frontrunner in this transition. These renewable resources offer the promise of environmentally benign energy generation, reducing our reliance on carbon-intensive fuels and mitigating the adverse effects of climate change. Among these promising alternatives, solar energy, in particular, has garnered substantial endorsement and investment from numerous nations worldwide.

1.1 BACKGROUND

The inherent advantages of solar energy are manifold. It is an inexhaustible resource, derived from the virtually limitless energy of the sun. Its utilization produces minimal environmental pollution during operation, contributing significantly to cleaner air and water. Furthermore, the widespread availability of solar irradiance across the globe makes it a geographically versatile energy source, offering the potential for energy independence and reduced reliance on centralized power generation.

Recognizing these benefits, a significant number of countries have proactively introduced a series of supportive policies and incentives aimed at encouraging the deployment and expansion of photovoltaic (PV) power generation. These policies often include feed-in tariffs, tax credits, subsidies, and mandates for renewable energy integration into national energy portfolios. The cumulative effect of these initiatives has been a remarkable year-on-year increase in global photovoltaic power generation capacity. Solar farms, rooftop installations, and building-integrated photovoltaics are becoming increasingly common sights in the energy landscape.

1.2 MOTIVATION

While the rapid growth in photovoltaic grid-connected power generation represents a significant step towards a sustainable energy future, it also introduces new operational complexities for power grid management. A key characteristic of solar power generation is its inherent variability and intermittency. Solar irradiance, the primary energy source for PV systems, is subject to fluctuations due to various factors such as diurnal cycles, cloud cover, atmospheric conditions, and seasonal changes. This variability in solar power output can lead to significant challenges in maintaining the delicate balance between electricity supply and demand within the power grid.

The unpredictable nature of solar power generation can cause voltage fluctuations, frequency deviations, and potential instability in the grid. Integrating large amounts of intermittent renewable energy sources like solar power requires sophisticated grid management strategies and advanced forecasting capabilities to ensure the continuous and reliable supply of electricity to consumers. Without accurate predictions of future solar power generation, grid operators face difficulties in scheduling conventional power plants, managing energy storage systems, and ensuring overall grid stability.

1.3 PROBLEM STATEMENT

The increasing penetration of photovoltaic power into electricity grids worldwide underscores the critical need for robust and reliable forecasting tools. Specifically, accurate photovoltaic power forecasting, or more fundamentally, solar irradiance forecasting, has emerged as a pivotal technology for effectively addressing the challenges posed by the inherent instability of solar energy.

Traditional methods of grid dispatching, often reliant on predictable and dispatchable power sources like fossil fuel-based power plants, are ill-equipped to handle the fluctuating nature of solar power. The inability to accurately predict the amount of solar energy that will be available at a given time in the future can lead to inefficient grid operations, increased reliance on backup generation, and potential curtailment of valuable renewable energy resources. Therefore, developing advanced forecasting techniques that can provide reliable and timely predictions of solar irradiance is paramount for ensuring the safety, stability, and efficient operation of modern power grids with significant renewable energy integration.

1.4 OBJECTIVES OF THE PROJECT

The overarching objective of this project is to explore and evaluate a probabilistic approach to solar irradiance forecasting using the XGBoost machine learning algorithm. Specifically, this project aims to:

- Implement a point prediction model for solar irradiance based on the XGBoost algorithm. This involves training an XGBoost model using historical solar irradiance data and relevant meteorological features.

- Develop a methodology to generate probabilistic prediction intervals from the trained XGBoost model. This will leverage the iterative nature of the XGBoost algorithm to obtain multiple predictions for a given forecast horizon.
- Apply the Kernel Density Estimation (KDE) method to transform the ensemble of point predictions into probabilistic prediction intervals. This will provide a measure of the uncertainty associated with the solar irradiance forecasts at different confidence levels.
- Evaluate the performance of the proposed XGBoost-based probabilistic forecasting model using publicly available solar irradiance datasets. This will involve assessing the accuracy and reliability of the generated point forecasts and prediction intervals using appropriate evaluation metrics.
- Compare the performance of the proposed method against other benchmark solar irradiance forecasting algorithms. This comparative analysis will highlight the strengths and weaknesses of the XGBoost-based approach.
- Investigate the computational efficiency and parameter sensitivity of the proposed model. This will assess the practical applicability of the method in real-world engineering scenarios.
- Ultimately, this project seeks to demonstrate the potential of the proposed XGBoost-based probabilistic forecasting model as a valuable tool for enhancing the integration and utilization of solar energy within modern power grids.

1.5 ORGANIZATION OF DOCUMENT

This document is structured to provide a comprehensive overview of the research project, from the initial motivation and theoretical background to the implementation details, experimental results, and concluding remarks.

The subsequent chapters are organized as follows:

Chapter 2: Literature Review: This chapter will delve into existing research and methodologies in the field of solar irradiance forecasting. It will explore various statistical and machine learning-based approaches, highlighting their strengths, limitations, and relevance to this project. A summary of the key findings from the literature review will also be presented.

Chapter 3: Proposed Methodology: This chapter will present a comprehensive description of the proposed probabilistic solar irradiance forecasting methodology. It will detail the steps involved in data collection, data preprocessing, data splitting for training and testing, model training using XGBoost, the generation of probabilistic prediction intervals using KDE, and the metrics used for performance evaluation.

Chapter 4: Results and Discussion: This chapter will present the experimental results obtained from evaluating the proposed method on the chosen datasets. It will include a detailed analysis of the forecasting accuracy and the reliability of the prediction intervals. Furthermore, a comparative analysis with benchmark algorithms will be provided, along with a discussion of the findings and their implications.

Chapter 5: Conclusion and Future Scope: This chapter will summarize the key findings of the project, highlighting the contributions and potential impact of the proposed XGBoost-based probabilistic solar irradiance forecasting model. It will also discuss the limitations of the current work and suggest potential avenues for future research.

Chapter 6: References: This chapter will provide a comprehensive list of all the academic papers, technical reports, and other resources cited throughout this document.

This structured organization aims to provide a clear and logical progress through the research project, enabling the reader to understand the motivation, methodology, results, and conclusions effectively.

CHAPTER 2

REVIEW OF LITERATURE

The escalating concerns surrounding global warming and the finite nature of conventional energy resources have, over the past two decades, catalyzed a significant global shift towards cleaner and more sustainable energy alternatives. Among these, solar energy, alongside wind and tidal power, has emerged as a particularly promising solution, garnering substantial attention and investment from nations worldwide. The inherent abundance, widespread availability, and minimal operational emissions of solar energy have positioned it as a cornerstone of future energy systems. Consequently, numerous governments have implemented supportive policies and financial incentives to accelerate the adoption and integration of photovoltaic (PV) power generation into their energy mixes [1–3]. This proactive stance has resulted in a consistent and substantial year-on-year increase in global PV power generation capacity.

However, the increasing reliance on grid-connected photovoltaic power generation has introduced a new set of challenges for power grid dispatching and management [4]. Unlike traditional, dispatchable power sources, the output of solar power systems is inherently variable and intermittent, directly dependent on fluctuating solar irradiance levels. Factors such as diurnal cycles, cloud cover, atmospheric conditions, and seasonal variations contribute to this instability, making it difficult to maintain a consistent balance between electricity supply and demand. This variability poses significant hurdles for grid operators striving to ensure the reliable and stable operation of the power system. Effective integration of large-scale solar power necessitates the development and implementation of sophisticated forecasting techniques capable of accurately predicting future solar energy availability. Precise solar irradiance or photovoltaic power forecasting is therefore recognized as a critical technology for facilitating the seamless consumption and efficient management of renewable energy resources within the power grid.

2.1 EXISTING SOLUTIONS

The academic and industrial research landscape concerning solar irradiance and photovoltaic power generation prediction has evolved significantly, broadly bifurcating into two primary categories: point prediction and probabilistic forecasting.

Point Prediction Models: These models aim to generate a single, deterministic estimate of future solar irradiance or PV power output. The methodologies employed within this category can be further classified into statistical models and machine learning-based models.

Statistical Models: Traditional statistical approaches, particularly time series methods [5, 6], have been widely applied for forecasting. Techniques such as Autoregressive Integrated Moving Average (ARIMA) and its variations analyze historical data patterns to extrapolate future values.

While these methods can be effective for relatively stable and predictable time series, their accuracy often deteriorates significantly under conditions of frequent weather changes and the inherent non-stationarity of solar irradiance data. The dynamic nature of atmospheric conditions, which directly influence solar radiation, often leads to sharp declines in prediction accuracy when relying solely on historical temporal patterns.

Machine Learning-Based Models: The advent of powerful machine learning algorithms has led to the development of numerous data-driven forecasting models. These include:

Extreme Learning Machine (ELM) [7, 8]: ELM is known for its fast learning speed and minimal parameter tuning requirements. However, its relatively simple network architecture may limit its ability to capture complex non-linear relationships between the various influencing factors and solar irradiance.

Multi-Layer Perceptron (MLP) Network [9, 10]: MLPs, a type of artificial neural network, offer greater flexibility than ELM in modeling complex relationships. However, similar to ELM, their performance can be limited by the depth and complexity of the network, and they may struggle with highly intricate temporal dependencies.

Support Vector Machine (SVM) [11, 12]: SVMs are powerful for both linear and non-linear regression tasks and can achieve high prediction accuracy, especially with smaller datasets. However, the crucial selection of appropriate kernel functions can be a complex and often empirical process. Furthermore, the computational cost of SVMs can increase dramatically with larger datasets, leading to significant reductions in processing speed.

Long Short-Term Memory (LSTM) Network [13–14]: LSTMs, a specialized type of recurrent neural network, are particularly well-suited for capturing long-range dependencies in sequential data, making them promising for time series forecasting. However, the process of tuning the numerous hyperparameters of an LSTM network can be intricate, time-consuming, and heavily reliant on expert knowledge and trial-and-error. Additionally, the computational demands of training deep neural network architectures like LSTMs are substantial, often requiring specialized hardware such as GPUs.

While point prediction models provide a single forecasted value, they inherently lack the ability to quantify the uncertainty associated with these predictions. Given the stochastic and volatile nature of solar irradiance, relying solely on deterministic forecasts can be insufficient for the optimized operation of power systems. Grid dispatchers require information about the potential range of future solar energy availability to make informed decisions regarding reserve scheduling, energy storage deployment, and overall grid stability.

Probabilistic Forecasting Models: Recognizing the limitations of point predictions, researchers have increasingly focused on probabilistic forecasting techniques. These methods aim to provide a more comprehensive understanding of future solar irradiance by generating prediction intervals with associated confidence levels. This allows grid operators to assess the inherent uncertainty and make more robust operational decisions. Several approaches have been explored in this domain:

Gaussian Process (GP) [15]: Gaussian Processes offer a non-parametric approach to probabilistic modeling. They can provide a predictive distribution over the future irradiance values. However, a common implicit assumption in the standard Gaussian Process is that the predictive distribution follows a normal distribution. This assumption may not always hold true for solar irradiance data, which can exhibit non-Gaussian characteristics such as skewness or multi-modality.

Quantile Regression (QR)-Based Methods: Quantile Regression provides a way to estimate the conditional quantiles of the target variable, allowing for the construction of prediction intervals without making strong distributional assumptions. Several studies have explored QR for probabilistic solar forecasting:

Bayesian Bootstrap Quantile Regression [16]: This method combines Bayesian bootstrapping with quantile regression to generate probabilistic PV power predictions, providing a distribution over the quantiles.

Improved Quantile Convolutional Neural Network (QCNN) [17]: This approach integrates quantile regression with Convolutional Neural Networks (CNNs) to leverage the feature extraction capabilities of CNNs for probabilistic PV power forecasting.

Wavelet Transform (WT) and CNN with Quantile Regression [18]: This hybrid method first uses wavelet transform and CNN to obtain deterministic forecasts and then applies quantile regression to generate probabilistic predictions based on the residuals.

Extreme Learning Machine and Quantile Regression [19]: This study utilizes Extreme Learning Machines to generate point forecasts and then employs quantile regression to establish prediction intervals, quantifying the uncertainty and variability of PV power.

A significant drawback of many quantile regression-based approaches is that they often require training a separate model for each quantile of interest. This can lead to a substantial increase in the computational time and complexity of the overall forecasting process, especially when a large number of quantiles are needed to accurately characterize the predictive distribution.

The work presented in [20] explores short-term solar radiation forecasting through the application of machine learning (ML) methodologies, Hidden Markov Model (HMM) techniques, and Support Vector Machine (SVM) regression. The simulation outcomes underscore the capacity of ML-driven algorithms to achieve precise solar irradiance predictions across diverse meteorological conditions. Furthermore, this study delves into the inherent daily cyclical patterns of solar irradiance and the instantaneous rate of change, highlighting their utility in forecasting future irradiance levels.

Traditional statistical time series models remain relevant in solar forecasting. Belmahdi et al. [21] in Morocco compared ARMA and ARIMA models for predicting monthly mean daily global radiation, finding that ARIMA (0.2,1) significantly outperformed ARMA (2,1), showing substantial improvements. Similarly, Shadab et al. [22] in India utilized Seasonal ARIMA (SARIMA) for spatial forecasting of monthly average insolation, achieving a high coefficient of determination (R2=0.9293) and relatively low error metrics (MAPE =

6.556%), demonstrating its suitability for location scouting for solar projects.

Various established machine learning algorithms have been widely applied and compared. Al-Rousan et al. [23] in Jordan reviewed Multi-layer Perceptron (MLP), Support Vector Machine Regression (SVMR), and Linear Regression (LR), noting MLP achieved the highest R2 (0.9513) and lowest MAPE (0.0001). A˘gbulut et al. [24] compared SVM, ANN, k-NN, and DL models across four Turkish provinces, observing R2 values ranging from 85.5% to 93.6%, indicating varying performance based on location and model.

An intra-day solar irradiance prediction method using satellite-based estimates for PV power plants is proposed in [25]. This method relies on a sophisticated neural network fed with a collection of time-dependent irradiance estimates from satellite images acquired in the vicinity of the target area. Performance improvements are observed even without direct irradiance observations, as demonstrated by a comparison with the accuracy of the European Centre for Medium-Range Weather Forecasts (ECMWF) and studies employing similar methodologies and forecast horizons. The report suggests various avenues for future research to further enhance the model. All GHI irradiance data are made available for examination, thereby improving reproducibility.

Random Forest (RF), an ensemble method, frequently emerges as a strong performer. Bounoua et al. [26] in Morocco evaluated Bagging, Boosting, and Random Forest against empirical methods for daily Global Horizontal Irradiance (GHI) estimation, finding RF superior with high correlation (R: 87.53–96.20%) and reasonable normalized errors (nRMSE: 7.85–15.33%). Srivastava et al. [27] in India compared MARS, CART, M5, and RF for 1-day to 6-day ahead forecasts, concluding that RF provided the best results. Blal et al. [28] in Algeria focused on comparing ambient temperature-based models for daily and hourly radiation estimation, identifying one specific model (M4) as optimal with

R2=0.8753. Gürel et al. [29] in Turkey used multiple ML algorithms to predict monthly average daily global solar radiation, achieving high accuracy (R2=0.952~0.993) and low errors (RMSE and MAPE < 10%). Karaman et al. [30] compared Extreme Learning Machines (ELM) and ANN in Turkey, finding ELM slightly better with lower RMSE (0.0297) and high performance (95%).

Deep learning models, particularly those involving recurrent and convolutional networks, have shown significant promise. Ghimire et al. [31] in Australia integrated CNN and LSTM (CLSTM hybrid model) for short-term Global Solar Radiation (GSR) prediction, demonstrating superior performance compared to other DL and benchmark models. Peng et al. [32] in Alabama, USA, developed a complex hybrid DL model (CEN-SCA-BiLSTM) incorporating Bidirectional LSTM, sine cosine algorithm (SCA), and decomposition techniques (CEEMDAN) for multi-step hourly predictions. This model achieved the lowest RMSE, MAE, MASE, and highest R compared to competitors. Lai et al. [33] in Brazil employed a Feature Attention-based Deep Forecasting (FADF) hybrid method, reporting notable RMSE reductions (11.88% and 12.65% on different datasets) compared to smart persistence for hourly forecasting. Z. Pung et al. [34], also in Alabama, compared ANN and RNN models, finding that the RNN model improved NMBE by 47% and RMSE by 26% over the ANN model for solar radiation prediction. Kisi et al. [35] in Turkey used a Dynamic Evolving Neural-Fuzzy Inference System (DENFIS), reporting better accuracy in monthly prediction than benchmarks. Puah et al. [36] in Malaysia utilized a Regression Enhanced Incremental Self-organising Neural Network (RE-SOINN), achieving high accuracy (MASE=0.65755, RMSE=73.945).

Combining different modeling techniques is a common strategy to enhance prediction accuracy and robustness. Sun et al. [37] in Beijing, China, employed a decomposition-clustering-ensemble learning approach, achieving low errors (NRSME=2.96%, MAPE=2.83%) and high directional forecast accuracy (88.24%). Prasad et al. [38] in

Australia developed a hybrid model (MEMD-SVD-RF) combining multivariate empirical mode decomposition, singular value decomposition, and Random Forest to handle nonstationarity in inputs, resulting in a reliable forecast with high average R2 (0.98) and low RMSE (1.05). Campo-Ávila et al. [39] in Spain used a model combining clustering, regression, and classification for one-day-ahead hourly prediction, achieving RMSE below 20%. Guermoui et al. [40] in Algeria applied Weighted Gaussian Process Regression (WGPR) for multi-step ahead daily global and direct radiation forecasting, reporting good performance (e.g., R2=85.85% for 10th daily GHI). Zhuo et al. [41] in China developed a combined multi-task learning and Gaussian process regression (MTGPR) model to simultaneously predict daily and monthly radiation components, showing modest improvements in R2 and RMSE over baseline methods. Heng et al. [42] in the United States used a nondominated sorting-based multi-objective bat algorithm (NSMOBA) aiming for both accuracy and stability in global monthly average radiation forecasting, achieving satisfactory results.

Some studies focus on specific contexts or provide reviews. Rodríguez-Benítez et al. [43] in Spain discussed extending the temporal horizon of all-sky imager (ASI) based nowcasting, highlighting ASIs' advantages over other models in overcoming certain challenges. Makade et al. [44] provided a comprehensive review of work by Indian researchers, analyzing a specific GSR Model (M-78) which showed variable performance (MPE: -8.1186% to 6.9383%, R2: 0.6345 to 0.9616). Sunhra Das [45] in India developed a specific model for predicting solar radiation on a tilted surface, reporting RMSE values between 6.7 and 8.9 for specific days. Narvaez et al. [46] in Colombia focused on developing accurate site-adaptation using ML and DL models, finding ML-based approaches offered a 38% performance improvement over traditional methods.

Studies span diverse geographical locations (China, Morocco, Algeria, US, Turkey, Australia, Spain, Brazil, India, Jordan, Colombia, Malaysia) and address various prediction horizons (hourly, daily, monthly, short-term, multi-step ahead). A wide array of models are employed, ranging from traditional statistical methods (ARIMA, SARIMA) to classic ML algorithms (SVM, RF, k-NN, ELM) and advanced deep learning architectures (ANN, RNN, CNN, LSTM, BiLSTM).

A prominent trend is the development and application of hybrid models, which combine decomposition techniques, clustering, feature selection/attention mechanisms, and different ML/DL algorithms [47-51]. These hybrid approaches often report superior performance compared to standalone models by leveraging the strengths of individual components, particularly in handling complex patterns and nonstationarity in solar radiation data. Deep learning models, especially LSTMs and CNNs, are increasingly favored for their ability to capture temporal dependencies and spatial features, often yielding state-of-the-art results [52-53].Random Forest also consistently emerges as a robust and effective algorithm in several comparisons evaluated using metrics like RMSE, MAPE, MAE, and R^2, with many studies demonstrating significant improvements in accuracy over benchmark or traditional methods [54]. Overall, machine learning and deep learning offer powerful tools for enhancing solar irradiance prediction accuracy, crucial for the continued growth and reliable operation of solar energy systems worldwide.

In [55], the authors propose deep learning (DL) techniques and data from numerous locations for predicting daily solar radiation at two locations in India. The primary objective of this study is to identify and implement ML techniques for uncovering latent symmetry in data patterns and relationships. A comparison of rolling window metrics such as MSE, RMSE, and the coefficient of determination (R2) demonstrates the performance of the proposed model. The findings suggest that bidirectional and attention-based LSTM techniques can effectively predict daily GHI data. The research highlights the potential benefits of hybrid approaches combining linear and nonlinear methods for the sustainable planning of solar energy systems and the estimation of available solar energy at specific locations.

A model for short-term solar radiation prediction utilizing spatio-temporal weather dependencies between regional systems with the aid of ResNet and LSTM networks is proposed in [56]. The proposed model is compared with several other DL methods to assess its effectiveness. The ResNet/LSTM ensemble model achieves superior prediction accuracy compared to the other models, as evaluated by MAE and RMSE. The study demonstrates that incorporating specific and temporal correlations enhances the prediction of solar irradiance with the proposed approach.

A novel solar irradiance prediction model is introduced in [57], employing machine learning (ML) and spatio-temporal factors to generate accurate 10-minute forecasts. The proposed forecasting model can be integrated with PV systems to predict their power output and facilitate their seamless integration into the smart grid. The study also presents the accuracy achieved by other models and compares it with the proposed model. The results indicate that the RMSE decreases in both the training and validation phases when the configuration delay is 1:2 and the hidden layer comprises 10 neurons.

In [58], the authors describe a machine learning model for solar irradiance prediction using a sky camera. This method utilizes the LSTM deep learning algorithm to forecast cloud locations for the subsequent ten minutes. Information regarding cloud cover is derived from processing sky images. The proposed method outperforms the persistence model in scenarios characterized by significant variations in solar irradiance, including partly cloudy days. The predictions are categorized into three groups based on sky conditions: clear, partly cloudy, and overcast. The solar radiation prediction method involves two stages: (1) calculating cloud cover using sky image processing and projecting future cloud cover using LSTM, and (2) predicting GHI using input data from the cloud cover solar radiation model. The LSTM model incorporates a GRU gate with forget, input, output, and cell state components.

Short-term solar radiation prediction with a focus on accuracy enhancement is the subject of [59]. The proposed approach integrates Monte Carlo (MC) simulation, Robust Principal Component Analysis (RPCA), spectral clustering, and neural networks (NN). Spectral clustering is employed to minimize the influence of varying weather patterns on solar irradiance. Experimental results demonstrate that the proposed method improves the accuracy of solar irradiance prediction. The study provides a detailed explanation of the machine learning techniques utilized in the strategy, which contribute to the regulation of solar energy fluxes.

The study [60] proposes a machine learning (ML) approach for cloud segmentation to estimate solar power output. The research compares several methods, including various machine learning models. U-Net, a deep neural network architecture, is used for accurate cloud pixel segmentation. The direct influence of sunlight on cloud cover is acknowledged, suggesting that solar output can be estimated using machine learning techniques and sky-facing cameras. The study also addresses the challenges associated with accurately identifying cloud boundaries and distinguishing clouds from clear skies, leading to the segmentation of all cloud types into a single "clouds" class.

A multi-task machine learning (ML) algorithm for solar irradiance prediction is proposed in [61]. The LSTM neural network model is used to implement this strategy, and its performance is evaluated through predictions at multiple time scales, including 1 hour, 1 day, and 1 week. A hybrid chicken swarm optimizer, which combines the strengths of the chicken swarm optimization algorithm (CSO) and the grey wolf optimization algorithm (GWO), is used to optimize the hyperparameters of the proposed LSTM approach. The paper presents this as an effective tool for allocating resources across different prediction tasks.

Various machine learning algorithms for predicting solar radiation and its impact on extreme weather events are discussed in [62]. The authors applied twelve machine learning models to accurately predict and compare daily and monthly solar radiation measurements. The study found that meteorological conditions significantly influence the performance of machine learning models, with GBRT, XGBoost, GPR, and Random Forest models demonstrating better prediction accuracy for both daily and monthly solar radiation. For scenarios with limited data availability, the study recommends using the XGBoost model for solar radiation prediction. It also emphasizes the importance of variable selection in the development of machine learning models.

The development and deployment of the University of Texas at San Antonio (UTSA) SkyImager system are detailed in [63]. The initial implementation of the SkyImager focused on predicting cloud location, employing ray tracing to determine shadow impacts on solar panels. The research also investigates an alternative approach leveraging artificial intelligence (AI) to directly estimate irradiance from a localized solar-centric image segment. Performance metrics derived from 147 days of National Renewable Energy Laboratory (NREL) data are presented. The study's findings indicate that fine-tuning Multi-Layer Perceptron (MLP) variables and Deep Learning (DL) parameters can substantially enhance algorithm convergence and reduce prediction errors.

The study by the authors of [64] examines the integration of deep learning (DL) techniques within power grid systems. Specifically, it investigates Long Short-Term Memory (LSTM) networks for forecasting solar radiation at varying time horizons: 1 hour, 1 day, and 1 year. The long-term predictions are deemed crucial for strategic system planning and market operations. This paper provides a comprehensive overview of the spectrum of DL methods currently employed in electricity systems.

A comparative analysis of independently generated ground-based and satellite-based solar irradiance predictions is presented in [65]. The study analyzes the synergistic potential of exchanging predictions and results. It posits that satellite-derived datasets can serve as valuable input for solar forecasting models. The research also details calibration refinement processes and clarifies relevant terminology. One-hour Global Horizontal Irradiance (GHI) forecasts spanning a year, utilizing three years of xg and xs data, are discussed within this study.

The research in [66] investigates both empirical and machine learning (ML) approaches for predicting global solar irradiance using air temperature as the primary input. Daily global solar radiation prediction is performed for temperate continental regions, employing four ML and four empirical temperature-based models. Simulation results demonstrate that a hybrid Genetic Algorithm (GA) and Artificial Neural Network (ANN) strategy outperforms both state-of-the-art ML and empirical models. Consequently, the temperature-based hybrid model is strongly recommended for GHI prediction in temperate latitudes, deemed essential for the effective management and operation of solar energy systems.

The application of machine learning (ML) techniques for solar radiation prediction in the context of planning is the central theme of [67]. The study proposes prediction models based on historical data, utilizing Linear Regression (LR), Regression Tree, and Support Vector Machine models. These models leverage past weather data to forecast solar radiation, incorporating parameters such as wind speed, air pressure, and humidity as input features. The proposed approach holds the potential to assist grid operators in optimizing supply and demand management.

Long-term solar irradiance prediction for the design and layout of microgrids is explored in [68]. This research examines various deep learning (DL) methodologies for forecasting hourly and daily solar irradiance for the upcoming year. The study recommends the use of both historical solar irradiance data and Global Horizontal Irradiance (GHI) under clear-sky conditions as input for the models. Models such as Feed-Forward Neural Networks (FFNNs), Support Vector Regression (SVR), and Gated Recurrent Units (GRUs) were evaluated. The findings indicate that GRU models exhibit comparatively superior performance. Furthermore, the paper provides an explanation of the operational mechanisms of the various ML models.

In [69], a novel methodology employing a Deep Neural Network (DNN) is proposed to identify and assess the impact of transient meteorological conditions on video data. Specifically, time-lapse video recordings from upward-facing wide-angle cameras are used to directly estimate and predict solar irradiance. The proposed DL method achieves a reduction in the Mean Absolute Error (MAE) for both estimation and prediction. The prediction architecture utilizes level 1 modeling, where individual images with short lookbacks are encoded to generate a representation of the entire sky, which is then fed into a Recurrent Neural Network (RNN) with LSTM units to transform historical photos into a 128-vector graphical representation. A common limitation across all models is their difficulty in accurately predicting irradiance during early morning and late evening hours.

Four distinct machine learning (ML) algorithms, namely SVM, ANN, k-Nearest Neighbors (kNN), and DL, are applied to predict daily global radiation for the four provinces of Turkey in [70]. Seven statistical criteria are employed to evaluate the efficacy of these algorithms. The results demonstrate high accuracy, with the ANN algorithm exhibiting the best performance among the tested methods.

The study presented in [71] investigates the application of Artificial Neural Networks (ANNs) for solar radiation and photovoltaic (PV) prediction in Nigeria. The primary objective is to develop ANN models capable of providing hourly solar radiation forecasts. The developed methods enable the prediction of both solar radiation and various PV parameters. The coefficient of determination (R2) values for the ANN techniques range from 0.9046 to 0.9777 for solar irradiance prediction and from 0.7768 to 0.8739 for diverse PV parameters. The study concludes by recommending the development and implementation of photovoltaic systems for power generation in Nigeria.

The accuracy of solar irradiance prediction using six different machine learning (ML) methods for Turkey and the United States is analyzed in [72]. The methods employed include Gradient Boosting Tree (GBT), Multilayer Perceptron Neural Network (MLPNN), two variations of Adaptive Neuro-Fuzzy Inference Systems (ANFIS) based on fuzzy-c-means clustering (FCM) and subtractive clustering (ANFIS-SC), Multivariate Adaptive Regression Splines (MARS), and Classification and Regression Tree (CART). The study utilizes Root Mean Squared Error (RMSE), correlation coefficient (R), Mean Absolute Error (MAE), and Nash-Sutcliffe efficiency coefficient (NS) to compare the accuracy of the models. The GBT model demonstrated superior performance in predicting solar energy and radiation compared to the other models.

The research in [73] discusses electricity generation from renewable energy sources, particularly photovoltaics (PV). The proposed model integrates a Deep Learning (DL) block cell and Genetic Algorithm (GA) optimization to predict solar irradiance time series. The effectiveness of three different neural network architectures—Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Radial Basis Function (RBF)—is compared. A genetic algorithm is used to optimize the window size and the number of neurons in each of the three hidden layers. The algorithm was implemented in Python using KERAS, a deep learning library that supports both CPU and GPU computation.

A methodology for predicting solar irradiance across multiple time horizons (3, 6, and 24 hours) is proposed in [74]. The model utilizes an LSTM network that accounts for the temporal relationships between the common day and other days. The algorithm features an optimized number of neurons and is evaluated using industry-standard metrics, specifically standard deviation and root mean squared error. The low values of mean squared error and mean absolute error in percentage indicate the effectiveness of the proposed model.

In [75], a novel approach employing machine learning techniques for short-term Global Horizontal Irradiance (GHI) estimation from sky images is presented. The proposed algorithm extracts features from sky images for solar irradiance prediction. The efficiency of the machine learning algorithm is evaluated using two publicly available sky image datasets. Compared to state-of-the-art algorithms, the proposed method demonstrates competitive performance with significantly lower computational cost for both nowcasting and forecasting up to 4 hours ahead. The study also addresses the inherent challenges in renewable energy forecasting due to the unpredictable nature of natural energy sources.

The effectiveness of unsupervised and supervised machine learning algorithms utilizing tree-based approaches with integrated implicit and explicit regime detection techniques for solar power forecasting is evaluated in [76]. The study assesses these ML methods for solar power forecasting in Kuwait. Regime-dependent ANN models are subjected to a comprehensive investigation to minimize prediction errors on a test dataset and are further evaluated on an independent validation dataset. The explicit regime-dependent method performs worse than the tree-based regression model approach. The study concludes that tree-based methods, particularly the regression model tree method, are more suitable for solar power forecasting.

Two machine learning (ML) algorithms, the Multi-Layer Perceptron Artificial Neural Network (MLP-ANN) and Genetic Programming (MGGP), are used for solar irradiance prediction in [77]. Irradiance predictions at six locations are analyzed using MLP, MGGP, and a persistence model over horizons ranging from 15 to 120 minutes. The overall improvement in Mean Squared Error (MSE) is 5.68%, and the improvement in Root Mean Squared Error (RMSE) is 3.41%. Iterative predictions enhance the accuracy of MGGP. MGGP provides more accurate and meaningful estimates and faster solutions in certain scenarios, whereas ANN is more complex and computationally intensive. In this study, the authors employed a feed-forward NN-MLP architecture with 10 neurons in the hidden layer and a hyperbolic tangential sigmoid transfer function.

Deep learning approaches for predicting solar irradiance and PV system performance using time series data are investigated in [78]. The simulations are evaluated based on input data, forecast horizon, season, training time, and result accuracy. LSTM demonstrates the best performance among individual models, while the hybrid CNN-LSTM model outperforms all other models but requires longer training times. RMSE is recommended as a representative evaluation metric for comparing the accuracy of the applied models.

The authors of [79] propose a hybrid Deep Learning (DL) model that integrates a GRU-NN and an attention mechanism for accurate prediction of solar irradiance variations. The DL model is designed to extract features from the available dataset using Inception and ResNet-NN architectures. Subsequently, these extracted features are fed into a recurrent neural network (RNN) to train the DL model. The experimental results demonstrate that the proposed hybrid DL model outperforms traditional DL models based on MAE, RMSE, and Mean Absolute Percentage Error (MAPE). The prediction of hourly solar energy in South Africa using ML approaches is discussed in [80]. This study compares the short-term solar radiation prediction results of recurrent neural networks, LSTMs, Feed-Forward Neural Networks (FFNNs), and Quantile Regression Averaging (QRA). The FFNN approaches yield the most accurate predictions in terms of MAE and Mean Squared Error (MSE).

A family of flexible and robust Deep Learning (DL) methods for solar irradiance prediction is presented in [81]. These methods are well-suited for solar irradiation locations as new sensors can be seamlessly integrated or removed without necessitating model retraining. The models are trained to predict solar radiation at multiple locations concurrently. They are designed to be as independent as possible from the number and location of data sources used for training. The study also emphasizes the necessity for flexible and robust solar prediction models for smart city applications.

The survey in [78-79] reviews papers published between 2009 and 2019, focusing on the application of meta-heuristic algorithms to address feature selection problems. The primary goal of feature selection is to reduce the dimensionality of the feature set without compromising model performance. The study categorizes meta-heuristic algorithms into four distinct groups based on their behavior, considering binary variants of these algorithms. Additionally, the survey presents an analysis of UCI repository datasets for optimal feature subset extraction. It highlights several research gaps, challenges, and issues in identifying the best subset of features using different meta-heuristic algorithms. The study also investigates the potential of meta-heuristic algorithms to prematurely converge. Furthermore, it introduces the bio-informatics-oriented Matthews correlation coefficient (MCC) as a binary quality measure for classification tasks.

An ensemble model combining wavelet transform (WT) and a Bidirectional Long Short-Term Memory (BiLSTM) deep learning network for 24-hour global horizontal solar irradiance prediction is presented in

[80]. This work combines wavelet decomposition components D1 to D6 to reduce the number of intrinsic mode functions (IMF). Separate BiLSTM networks, trained independently, are used for each IMF. The complete solar prediction GHI is reconstructed by aggregating the predictions from the BiLSTM networks for different sub-series values. Compared to existing models, the proposed model demonstrates superior results in terms of RMSE, MAPE, the coefficient of determination (R2), and a forecasting skill (FS) metric. Specifically, this model reduces the monthly average RMSE by 26.041-58.890%, 5.170-31.350%, 23.260-56.060%, and 21.080-570% compared to standalone BiLSTM, GRU, and LSTM networks, respectively. However, the benchmark, standalone BiLSTM, GRU, and LSTM models exhibited lower monthly average MAPEs of 9.518%, 12.59-28.14%, 30.43-59.19%, and 26.54-58.92%, respectively.

A deep learning-based model for 1-hour global horizontal irradiance prediction is proposed in [81]. This method employs deep learning-based time series clustering to categorize GHI time series data into multiple clusters, aiming to identify unique patterns and enhance clustering efficiency. A Deep Neural Network with Feature Alignment by Discrepancy Filtering (DNN-FADF) is independently trained for each cluster to perform GHI prediction. Simulation results indicate that the proposed approach yields the most accurate solar forecasts compared to smart persistence and state-of-the-art models, achieving an RMSE reduction of 11.88% compared to previous strategies and 12.65% compared to Smart Persistence. Notably, the FADF training utilizes the Huber loss function

2.2 SUMMARY OF LITERATURE REVIEW

The existing body of literature highlights the growing importance of accurate solar irradiance and photovoltaic power forecasting in facilitating the integration of solar energy into power grids. While traditional statistical and machine learning-based point prediction models have demonstrated varying degrees of success, their inherent limitation in quantifying forecast uncertainty has motivated the exploration of probabilistic forecasting techniques.

Methods based on Gaussian Processes and Quantile Regression offer promising avenues for generating prediction intervals. However, Gaussian Processes often rely on the assumption of a normal distribution, which may not always be valid for solar irradiance data. Quantile Regression-based approaches, while distribution-agnostic, can be computationally expensive as they typically require training multiple models for different quantiles.

The research that serves as the foundation for this project report proposes a novel probabilistic solar irradiance prediction model based on the XGBoost algorithm, coupled with kernel density estimation. This approach aims to address some of the limitations of existing methods by leveraging the ability of XGBoost to generate multiple predictions during its iterative training process and then utilizing KDE to construct probability prediction intervals. The subsequent sections of this report will delve deeper into the algorithms and software used, the proposed methodology, the experimental results, and a comparative analysis with other benchmark techniques, building upon the existing literature reviewed in this chapter.

CHAPTER 3

METHODOLOGY

This chapter outlines the methodology employed in this project to develop and evaluate the probabilistic solar irradiance forecasting model. The process encompasses several key stages, starting with the acquisition and preparation of the dataset, followed by the training of a deterministic prediction model using XGBoost, the application of Kernel Density Estimation to generate probabilistic forecasts, and finally, the evaluation of the model's performance using appropriate metrics.

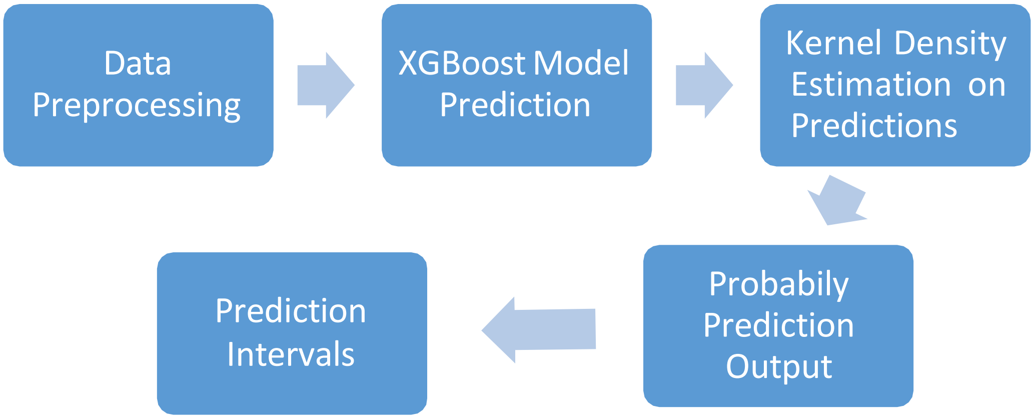

Data Preprocessing
XGBoost Model Prediction
Kernel Density Estimation on Predictions
Probabily Prediction Output
Prediction Intervals

Figure 1: Flowchart of Proposed Model

3.1 DATASET

This study rigorously evaluates a proposed solar irradiance forecasting approach using a comprehensive, publicly available dataset from the National New Energy Laboratory (NREL). This dataset offers a rich historical record spanning ten years (January 1, 2006, to December 31, 2015), with hourly measurements of both weather conditions and the corresponding solar irradiance. The high temporal resolution of one hour allows for capturing the intricate interplay between meteorological factors and solar radiation throughout the day.

To ensure a robust assessment, the dataset was strategically divided into two distinct time-based subsets. A substantial nine-year period (January 1, 2006, to December 31, 2014) served as the training dataset.

This extensive historical data provided the XGBoost model with ample opportunity to learn the complex, non-linear relationships between the selected input features and the target variable – solar irradiance. The remaining one year of data (the entire year of 2015) was reserved as the independent test set. This temporal separation is crucial for evaluating the model's ability to generalize its learned patterns to unseen data, offering a realistic evaluation of its predictive performance in real-world operational scenarios.

The focus of the solar radiation data within this dataset is specifically on the daylight hours, ranging from 8:00 AM to 6:00 PM each day. This deliberate focus aligns directly with the primary operational period for solar energy generation, making the forecasting efforts highly relevant for photovoltaic power prediction and efficient grid management.

The input features carefully chosen to predict the next day's solar irradiance (between 8:00 AM and 6:00 PM) encompass a well-reasoned combination of seasonal, temporal, and meteorological variables, each selected for its known influence on solar radiation levels:

Seasonal Features: The month of the year is included to capture the fundamental cyclical variations in solar irradiance driven by the Earth's axial tilt and its annual orbit around the sun. This feature accounts for the broader changes in solar intensity across different seasons.

Temporal Features: Date of the month and hour of the day are incorporated to account for the more granular daily and monthly patterns of solar irradiance. The hour of the day directly reflects the sun's position in the sky and the corresponding intensity of solar radiation received, while the date of the month captures subtle variations within a given month.

Meteorological Features: A suite of hourly average meteorological measurements is included to capture the dynamic atmospheric conditions that significantly impact solar radiation:

Dew point temperature (tower): This variable provides insight into the atmospheric moisture content. Higher dew point temperatures indicate more moisture in the air, which can lead to increased cloud formation and consequently reduced solar radiation transmission.

Relative humidity (tower): Similar to dew point temperature, relative humidity quantifies the amount of moisture present in the air relative to the maximum it can hold at a given temperature. High relative humidity also increases the likelihood of cloud development.

Cloud cover (total): This represents the overall fraction of the sky obscured by clouds, a primary impediment to direct solar radiation reaching the surface.

Cloud cover (opaque): This specifically measures the fraction of the sky covered by thick, dense clouds that effectively block direct sunlight. This is a particularly critical factor in reducing solar irradiance.

Wind speed (22'): Measured at a height of 22 feet, wind speed can indirectly influence solar irradiance by affecting the movement and dissipation of clouds, as well as influencing atmospheric turbulence and mixing.

East sea-level pressure: Sea-level pressure is an indicator of large-scale weather systems. Variations in pressure can be correlated with the formation and movement of cloud systems and overall atmospheric stability, thus impacting solar radiation.

The output variable of this prediction model is the solar irradiance forecasted for the subsequent day, specifically for the daylight hours between 8:00 AM and 6:00 PM. The ultimate objective is not only to generate accurate point forecasts of these future irradiance values but also to develop probabilistic prediction intervals. These intervals are essential for quantifying the inherent uncertainty associated with the forecasts, providing a range within which the actual solar irradiance is likely to fall. This probabilistic approach is crucial for risk assessment and decision-making in solar energy applications and grid management.

3.2 DATA PREPROCESSING

Absolutely, let's delve deeper into the crucial data preprocessing stage and then outline the subsequent methodological steps you mentioned.

The data preprocessing phase undertaken prior to training the XGBoost model was a critical step in ensuring the quality, consistency, and suitability of the raw solar irradiance dataset for machine learning. Addressing issues like missing values and outliers at this stage is paramount for building a robust and accurate forecasting model.

Missing Value Filling:

Real-world datasets, such as the one provided by NREL, are often susceptible to missing data points. These gaps can arise from various sources, including sensor malfunctions, data transmission interruptions, or even maintenance periods. The presence of missing values can significantly impede the training process of machine learning algorithms, as most algorithms are designed to work with complete datasets.

While the specific imputation method isn't detailed in the initial description, several common and more advanced strategies are typically employed to handle missing data:

Simple Imputation Techniques:

Mean Imputation: Replacing missing values with the average value of that particular feature. This is a simple approach but can distort the distribution of the feature and underestimate variance, especially if the missing data is not missing completely at random.

Median Imputation: Substituting missing values with the median value of the feature. This is less sensitive to outliers compared to mean imputation and is often preferred for skewed distributions.

Mode Imputation: For categorical features (though not explicitly mentioned as having missing values in the meteorological features), missing entries can be filled with the most frequent category.

More Sophisticated Imputation Techniques:

Regression Imputation: Using other features in the dataset to predict the missing values through a regression model. This can capture relationships between variables but assumes these relationships hold for the missing data as well.

K-Nearest Neighbors (KNN) Imputation: Imputing missing values based on the values of the k-nearest neighbors in the feature space. This method can capture local structure in the data but can be computationally expensive for large datasets.

Time Series Specific Methods: Given the temporal nature of the data, techniques like forward fill (carrying the last valid observation forward), backward fill (carrying the next valid observation backward), or interpolation (estimating values based on neighboring points in time) might be suitable for missing values that occur sequentially.

The choice of the most appropriate imputation method is crucial and depends on several factors, including the amount and pattern of missing data (e.g., missing completely at random, missing at random, or missing not at random), the nature of the feature (numerical or categorical), and the potential impact of the imputation on the subsequent model training. Careful analysis of the missing data patterns is essential before selecting an imputation strategy.

Outlier Processing:

Outliers are data points that deviate significantly from the typical range of values within a given feature. These extreme values can be the result of genuine but rare events (e.g., unusually strong weather phenomena), measurement errors, or sensor noise. If left untreated, outliers can exert a disproportionate influence on the training of machine learning models, potentially leading to biased models with poor generalization performance on unseen data.

Common techniques for identifying and handling outliers include

Statistical Methods:

Standard Deviation Method: Identifying data points that fall beyond a certain number of standard deviations (e.g., 3) from the mean of the feature. This assumes a roughly normal distribution.

Interquartile Range (IQR) Method: Defining outliers as data points that fall below Q1−1.5×IQR or above Q3+1.5×IQR, where Q1 is the first quartile, Q3 is the third quartile, and IQR=Q3−Q1. This method is more robust to non-normal distributions.

Visualization Techniques:

Box Plots: These graphical representations clearly show the distribution of a feature, including the median, quartiles, and potential outliers as individual points outside the whiskers.

Scatter Plots: When examining relationships between two variables, outliers can sometimes be visually identified as points that deviate significantly from the general trend.

Once outliers are identified, several strategies can be employed to handle them:

Removal: The simplest approach is to remove the identified outlier data points from the dataset. However, this should be done cautiously, as removing genuine extreme values might lead to a loss of important information about the system's behavior under unusual conditions.

Transformation: Techniques like logarithmic transformation or winsorizing (capping extreme values at a certain percentile) can reduce the impact of outliers without completely removing them.

Imputation: Similar to missing values, outliers can sometimes be replaced with less extreme values, although this approach should be used judiciously and with a clear rationale.

Separate Modeling: In some cases, if outliers represent a distinct phenomenon, it might be beneficial to model them separately.

The choice of outlier processing technique depends on the nature of the outliers, their potential causes, and their impact on the model's performance. Careful consideration and potentially experimentation with different methods are often necessary.

The successful completion of this data preprocessing stage, involving appropriate strategies for handling both missing values and outliers, ensures that the XGBoost model is trained on a clean, consistent, and representative dataset. This critical step lays the foundation for building a robust and accurate solar irradiance forecasting model.

Now, moving forward with the methodology you outlined:

Deterministic Prediction using XGBoost:

Following the data preprocessing, the study would have proceeded to train the XGBoost (Extreme Gradient Boosting) model. XGBoost is a powerful and widely used gradient boosting algorithm known for its efficiency, flexibility, and high predictive accuracy. It works by sequentially building an ensemble of decision trees, where each new tree aims to correct the errors made by the previous trees.

The training process would have involved:

Feature Engineering (if any beyond the described features): While the abstract details a comprehensive set of input features, further feature engineering might have been performed to create new, potentially more informative features from the existing ones (e.g., lagged irradiance values, interactions between meteorological variables).

Model Training: Feeding the preprocessed training dataset (2006-2014) to the XGBoost algorithm. This involves optimizing the model's parameters (hyperparameters) to minimize a chosen loss function (e.g., mean squared error) on the training data. Techniques like cross-validation would likely have been employed during training to prevent overfitting and to select optimal hyperparameters.

Deterministic Forecast Generation: Once the XGBoost model is trained, it would be used to generate point forecasts of solar irradiance for the test set period (2015), specifically for the hours between 8:00 AM and 6:00 PM each day. These forecasts represent the model's best estimate of the future solar irradiance values.

Kernel Density Estimation (KDE) for Probabilistic Prediction Intervals:

To quantify the uncertainty associated with the deterministic point forecasts, the study likely employed Kernel Density Estimation (KDE). KDE is a non-parametric method used to estimate the probability density function of a random variable. In the context of forecasting, KDE can be applied to the prediction errors (the difference between the actual values and the model's point forecasts on the training or a validation set) to estimate the distribution of these errors.

The process would involve:

Error Calculation: Calculating the forecast errors on the training data (or a separate validation set) by subtracting the XGBoost model's predictions from the actual observed solar irradiance values.

KDE Application: Applying KDE to the distribution of these forecast errors. This results in a smooth, continuous estimate of the probability density function of the errors.

Prediction Interval Generation: Using the estimated error distribution from KDE, probabilistic prediction intervals can be constructed for future forecasts. For a given forecast, a prediction interval with a specific confidence level (e.g., 90%) would be defined by finding the range of values that encompasses that percentage of the estimated error distribution centered around the point forecast. This provides a measure of the uncertainty associated with the forecast.

Evaluation Metrics:

To rigorously assess the performance of the forecasting approach, the study would have employed a range of relevant evaluation metrics. These metrics provide quantitative measures of the model's accuracy and reliability. Common metrics for evaluating solar irradiance forecasting models include:

Mean Absolute Error (MAE): The average of the absolute differences between the predicted and actual values. It provides a measure of the average magnitude of the errors.

Root Mean Squared Error (RMSE): The square root of the average of the squared differences between the predicted and actual values. RMSE gives more weight to larger errors compared to MAE.

Mean Bias Error (MBE): The average of the differences between the predicted and actual values. MBE indicates the overall bias of the model (whether it tends to over- or under-predict).

Coefficient of Determination (R^2): Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R^2 indicates a better fit.

Coverage Probability of Prediction Intervals (CPI): For evaluating the probabilistic forecasts, CPI measures the percentage of actual values that fall within the generated prediction intervals. A well-calibrated prediction interval should have a CPI close to the nominal confidence level.

Average Width of Prediction Intervals (AWPI): This metric measures the average width of the prediction intervals. While high coverage is desirable, it should not come at the cost of excessively wide intervals, which would be less informative.

These evaluation metrics would have been calculated on the independent test set (2015) to provide an unbiased assessment of the model's generalization ability and its performance on unseen data.

The methodology likely involved a rigorous process starting with the collection of a comprehensive solar irradiance dataset from NREL. A crucial data preprocessing stage addressed missing values and outliers to ensure data quality. The core of the deterministic forecasting was built using the powerful XGBoost algorithm, trained on historical data to learn the complex relationships between meteorological factors and solar irradiance. To quantify the uncertainty associated with these point forecasts, Kernel Density Estimation was likely employed to model the distribution of prediction errors and generate probabilistic prediction intervals. Finally, the performance of the entire approach was rigorously evaluated using a suite of relevant metrics on an independent test set, providing a comprehensive assessment of its accuracy and reliability for solar irradiance forecasting.

3.3 DETERMINISTIC PREDICTION BY XGBOOST

The core of the proposed forecasting framework lies in the utilization of the eXtreme Gradient Boosting (XGBoost) algorithm for generating initial deterministic (point) predictions of future solar irradiance. XGBoost [19] is a highly effective and widely adopted machine learning algorithm renowned for its performance in both classification and regression tasks. It is an advanced implementation of the Gradient Boosting Decision Tree (GBDT) framework, incorporating several enhancements that contribute to its speed, accuracy, and robustness.

At its heart, XGBoost operates by constructing an ensemble of weak learners, which are typically classification and regression trees (CARTs). The training process is iterative and additive. Initially, a base learner (often a simple tree) is trained on the input data. Subsequent learners are then built sequentially, with each new learner attempting to correct the errors (residuals) made by the ensemble of the previously trained learners. This process of sequentially adding trees that learn from the mistakes of prior trees is known as boosting. The final prediction is obtained by aggregating the predictions of all the individual trees in the ensemble through a weighted summation.

The construction of each new tree in XGBoost is guided by the gradient of a defined loss function. This loss function quantifies the discrepancy between the predicted values and the actual values in the training data. By minimizing this loss function in each iteration, XGBoost iteratively refines the overall prediction model.

XGBoost incorporates several key improvements over the standard GBDT algorithm, which contribute to its superior performance:

- **Regularization:** XGBoost includes both L1 and L2 regularization terms in its objective function. These regularization techniques help to prevent overfitting by penalizing overly complex models, thereby improving the model's ability to generalize to unseen data.

- **Second-Order Taylor Expansion:** Unlike traditional GBDT, which typically uses only the first-order derivative of the loss function during optimization, XGBoost employs a second-order Taylor expansion. This provides a more accurate approximation of the loss function and can lead to faster convergence and better model performance.

- **Column Sampling (Feature Subsampling):** To further prevent overfitting and reduce computational complexity, XGBoost supports column (feature) sampling. This technique randomly selects a subset of features to be considered at each tree split, similar to the random subspace method used in Random Forests.
- **Split Finding Algorithms:** XGBoost offers sophisticated and efficient algorithms for finding the best splits in the decision trees, including a greedy approach and an approximate greedy approach for handling large datasets. It also considers sparsity in the data during the split finding process.
- **Tree Pruning:** XGBoost incorporates a pruning mechanism that removes branches of the trees that do not contribute significantly to reducing the loss function. This helps to control model complexity and prevent overfitting.
- **Learning Rate (Shrinkage):** After each tree is added to the ensemble, its contribution is weighted by a learning rate (denoted as η). This shrinkage technique reduces the impact of each individual tree, providing more space for subsequent trees to learn and further refine the model. The update rule for the prediction at iteration t is given by:

$$y_i^t = \sum_{k=1}^{t} f_{k(x_i)} = y_i^{t-1} + \eta\, f_{t(x_i)}, 0 < \eta < 1 \qquad \rightarrow \qquad (1)$$

where $f_k(x_i)$ is the prediction of the k-th tree for the i-th sample, and *y i(t)* is the ensemble prediction after t trees. In the context of this project, the preprocessed historical solar irradiance data, along with the relevant input features (seasonal, temporal, and meteorological), are used to train an XGBoost regression model. This trained model learns the intricate relationships between these features and the target solar irradiance values. Once trained, this deterministic model can then be used to generate point predictions of solar irradiance for a specified future time horizon.

A key aspect of the proposed methodology, which facilitates the generation of probabilistic forecasts, lies in leveraging the iterative nature of the XGBoost training process during the *testing* or prediction phase. As XGBoost builds an ensemble of trees sequentially, the prediction at each stage of the ensemble construction can be considered as an individual prediction. Therefore, for a given test sample, the trained XGBoost model can output not just a single final prediction, but a series of predictions corresponding to the cumulative prediction of the first tree, the first two trees, the first three trees, and so on, up to the final ensemble of trees. This collection of multiple predictions for a single forecast point inherently captures some of the uncertainty associated with the prediction.

3.4 KERNEL DENSITY ESTIMATION

The core idea is to move beyond a single-point forecast and provide a more informative view of the future by quantifying the uncertainty associated with the prediction. KDE serves as the crucial bridge for this transformation.

The Power of Non-Parametric Estimation:

The research rightly highlights the non-parametric nature of KDE as a significant advantage. Unlike parametric methods, which necessitate assuming a specific underlying distribution for the data (like a normal or gamma distribution), KDE makes no such assumptions. This is particularly beneficial in the context of solar irradiance forecasting, where the distribution of predictions (and consequently, the actual solar irradiance) might exhibit complex, non-Gaussian characteristics due to the interplay of various meteorological factors and temporal patterns. By letting the data speak for itself, KDE offers a more flexible and potentially more accurate representation of the underlying probability distribution.

The Mechanics of Kernel Density Estimation:

Figure 2, though not visible here, would likely illustrate the fundamental principle of KDE. As research explains, KDE works by placing a smooth "kernel function" over each data point in the sample (in this case, the ensemble of predictions from the XGBoost model). These individual kernel functions essentially represent the probability density contributed by each data point to its surrounding area.

Think of it like this: each prediction in the ensemble is considered a center point, and a smooth curve (the kernel) is drawn around it. The shape of this curve (e.g., Gaussian, Epanechnikov) determines how much influence each neighboring value has on the density estimate at a specific point. Data points closer to the point of interest have a higher influence.

The final probability density estimate is obtained by summing up all these individual kernel functions. This summation creates a smooth, continuous curve that represents the estimated probability density function (PDF) of the future solar irradiance at that specific time point.

The Role of the Kernel Function and Bandwidth:

There are two key parameters in KDE:

- **Kernel Function:** This function determines the shape of the smooth curve placed over each data point. Common kernel functions include Gaussian (bell-shaped), Epanechnikov (parabolic), and uniform (rectangular). The choice of kernel function generally has a less significant impact on the final estimate compared to the bandwidth.

- **Bandwidth (or Smoothing Parameter):** This parameter is crucial as it controls the smoothness of the resulting density estimate.
    - **Small Bandwidth:** A small bandwidth leads to a highly peaked and less smooth density estimate, potentially overfitting the data and showing individual data points more distinctly.
    - **Large Bandwidth:** A large bandwidth results in a smoother density estimate, potentially over smoothing the data and masking important features of the underlying distribution.

Selecting an appropriate bandwidth is a critical step in KDE. Various data-driven methods exist for bandwidth selection, such as Scott's rule and Silverman's rule, or more sophisticated techniques like cross-validation, which aim to find a balance between smoothness and accuracy.

Applying KDE to Ensemble Predictions:

The core innovation here is treating the multiple deterministic predictions generated by the XGBoost model's ensemble of trees as the sample data for KDE at each future time point. An ensemble model, by its nature, produces a set of slightly different predictions due to the inherent randomness in the tree-building process and the diversity within the ensemble. This collection of predictions inherently captures some of the uncertainty associated with the forecast.

By applying KDE to this ensemble of predictions for a specific future hour (e.g., 9:00 AM tomorrow), the project obtains a continuous probability density function that represents the likelihood of different solar irradiance values occurring at that time, according to the model's internal variability.

Deriving Prediction Intervals from the PDF:

Once the probability density function is estimated using KDE, generating prediction intervals at different confidence levels becomes straightforward. A prediction interval provides a range of values within which the future solar irradiance is expected to fall with a certain probability.

For a given confidence level (e.g., 90%):

1. **Identify the Probability Mass:** The goal is to find an interval that contains the specified percentage of the total probability under the estimated density curve. For a 90% confidence interval, this means enclosing 90% of the area under the KDE curve.

2. **Determine the Percentiles:** The lower and upper bounds of the prediction interval correspond to specific percentiles of the estimated probability distribution. For a 90% interval, these are typically the 5th percentile (lower bound) and the 95th percentile (upper bound). This means that 5% of the probability mass lies below the lower bound, and 5% lies above the upper bound.

For example, a 90% prediction interval of [200 W/m2, 700 W/m2] would suggest that there is a 90% probability that the actual solar irradiance at that future time will fall between 200 and 700 W/m2, according to the model's probabilistic output.

The Value of Prediction Intervals:

The generation of prediction intervals is a significant step beyond simple point forecasts. It provides a crucial measure of the uncertainty associated with the prediction, which is invaluable for various applications:

- **Risk Assessment:** Understanding the range of possible future irradiance values allows for better risk assessment in solar energy generation and grid management. For instance, grid operators can anticipate potential fluctuations in solar power output and plan accordingly.

- **Decision Making:** Prediction intervals provide more context for decision-making. Instead of relying on a single forecast, stakeholders can consider the range of possibilities and make more informed choices.
- **Model Evaluation:** Evaluating the coverage probability of the prediction intervals (i.e., the percentage of actual values that fall within the predicted intervals) provides a way to assess the calibration and reliability of the probabilistic forecasts.

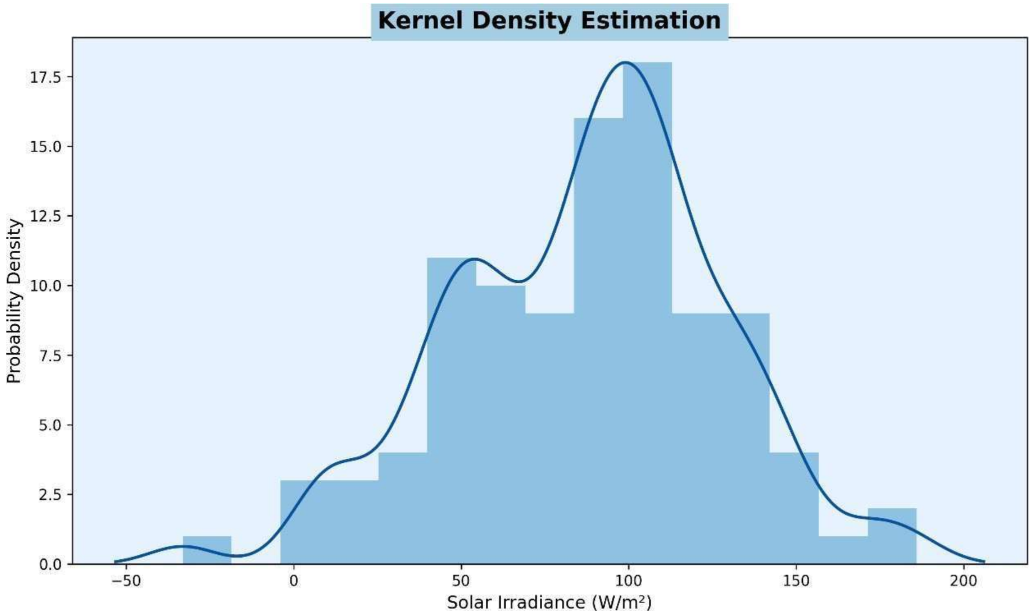


Figure 2: Kernel Density Estimation

In conclusion, the use of Kernel Density Estimation to transform the ensemble of deterministic XGBoost predictions into probabilistic prediction intervals is a sophisticated and valuable approach. It leverages the inherent uncertainty captured by the ensemble method and provides a more comprehensive and informative forecast by quantifying the likelihood of different future solar irradiance scenarios.

This approach moves beyond a single "best guess" and offers a richer understanding of the potential range of future outcomes.

The fundamental idea behind KDE is to place a smooth kernel function over each data point in the sample and then sum these kernel functions to obtain a smooth estimate of the underlying probability density. The shape of the kernel function determines the weight given to data points in the vicinity of a particular value, and the bandwidth (or smoothing parameter) controls the smoothness of the resulting density estimate.

The kernel density estimate *$f_h(x)$* at a point x is given by the formula:

$$f_{h(x)} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad \rightarrow \quad (2)$$

where:

- K(·) is the kernel function, a non-negative function that integrates to one and is typically symmetric around zero.
- x_i are the individual data points in the sample (in this case, the multiple predictions generated by XGBoost for a specific forecast time).
- n is the number of data points (the number of trees in the XGBoost ensemble).
- h is the bandwidth, a positive parameter that controls the smoothness of the density estimate. A smaller bandwidth leads to a more jagged and less smooth estimate, while a larger bandwidth results in a smoother but potentially over-smoothed estimate.

In this project, the Gaussian kernel function is specifically used for the kernel density estimation of solar irradiance:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \quad \rightarrow \quad (3)$$

The Gaussian kernel is a common choice due to its smooth and well-behaved properties.

For each future time point for which a solar irradiance forecast is needed, the multiple predictions generated by the XGBoost model (corresponding to the predictions from the growing ensemble of trees) are treated as the sample data for the KDE. By applying the Gaussian kernel density estimation to this set of predictions, a continuous probability density function is obtained, representing the likelihood of different solar irradiance values occurring at that future time.

Once the probability density function is estimated, prediction intervals under different confidence levels can be readily derived. For a given confidence level (e.g., 90%), the lower and upper bounds of the prediction interval are determined by finding the values that enclose the specified percentage of the total probability mass under the estimated density curve. For instance, a 90% prediction interval would be defined by the 5th and 95th percentiles of the estimated probability distribution. These prediction intervals provide a valuable measure of the uncertainty associated with the solar irradiance forecast, offering a range of plausible future values rather than a single deterministic estimate.

3.5 Evaluation

To quantitatively assess the performance of the proposed probabilistic solar irradiance forecasting model, a comprehensive evaluation framework is employed, encompassing both the accuracy of the deterministic point predictions and the reliability and sharpness of the generated prediction intervals.

3.5.1 Deterministic Prediction Evaluation:

The performance of the deterministic predictions generated by the XGBoost model is evaluated using two commonly used error metrics:

- **Mean Absolute Error (MAE):** The MAE measures the average magnitude of the errors between the predicted values and the actual values. It provides a straightforward measure of the average prediction accuracy, giving equal weight to all errors. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \rightarrow \quad (4)$$

  where y_i is the true value, y_i is the predicted value, and n is the total number of samples.

- **Root Mean Square Error (RMSE):** The RMSE measures the square root of the average of the squared errors between the predicted values and the actual values. RMSE gives a higher weight to larger errors compared to MAE, making it more sensitive to outliers in the predictions. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \rightarrow \quad (5)$$

  where yi is the true value, y^i is the predicted value, and n is the total number of samples.

Lower values of both MAE and RMSE indicate better deterministic prediction accuracy.

3.5.2 Probabilistic Prediction Evaluation:

The quality of the probabilistic prediction intervals generated by the KDE method is assessed based on two key aspects: interval reliability and interval width.

- **Interval Reliability:** This aspect evaluates whether the prediction intervals adequately cover the true values within the specified confidence level. It is quantified using the **Prediction Interval Coverage Probability (PICP)**:

$$PICP = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \qquad \rightarrow \qquad (6)$$

$$\varepsilon_i = \begin{cases} 1 & y_i \in [L_i, U_i] \\ 0 & y_i \notin [L_i, U_i] \end{cases} \qquad \rightarrow \qquad (7)$$

where N is the total number of samples in the test set, and ϵi is an indicator variable. Here, Li and Ui are the lower and upper bounds of the prediction interval for the i-th sample, and yi is the true value. A higher PICP value indicates that the prediction intervals have a higher reliability in covering the actual observations. Ideally, the PICP should be close to or greater than the nominal confidence level (e.g., 90% confidence level should ideally yield a PICP of around 0.90 or higher).

- **Interval Width (Sharpness):** While high reliability is crucial, the prediction intervals should also be as narrow as possible to provide informative forecasts. A very wide interval that always covers the true value is not practically useful. The **Prediction Interval Normalized Average Width (PINAW)** is used to measure the average width of the prediction intervals, normalized by the range of the target variable:

$$PINAW = \frac{1}{N \cdot R}\sum_{i=1}^{N}(U_i - L_i) \qquad \rightarrow \qquad (8)$$

where R is the normalization factor, typically the difference between the maximum and minimum values of the target variable in the dataset. A lower PINAW value indicates narrower and thus sharper prediction intervals.

- **Coverage Width-based Criterion (CWC):** To provide a comprehensive evaluation that considers both reliability and sharpness, the **Coverage Width-based Criterion (CWC)** is also employed

$$CWC = PINAW(1 + \gamma(PICP)e^{-\eta(PICP-\mu)}) \quad \rightarrow \quad (9)$$

  where μ is a pre-defined acceptable PICP level, and η is a penalty factor that increases the CWC value when the PICP falls below μ. The function γ(PICP) is defined as:

$$\gamma(PICP) = \{\begin{matrix} 0 & PICP \geq \mu \\ 1 & PICP < \mu \end{matrix} \quad \rightarrow \quad (10)$$

  The CWC effectively transforms the dual-objective problem of maximizing reliability and minimizing width into a single-objective minimization problem. A lower CWC value indicates better overall performance of the probabilistic forecasting model, balancing both coverage and sharpness.

3.6 Summary

The methodology proposed in this project involves a multi-stage process for generating probabilistic solar irradiance forecasts. First, historical solar irradiance and meteorological data are preprocessed to handle missing values and outliers. Second, an XGBoost regression model is trained on this preprocessed data to learn the deterministic relationships between the input features and the target solar irradiance. A key innovation is the utilization of the multiple intermediate predictions generated by the XGBoost ensemble during the test phase. Third, the Kernel Density Estimation method, with a Gaussian kernel, is applied to these multiple predictions to estimate the probability density function of the future solar irradiance. Finally, prediction intervals under different confidence levels are derived from the estimated probability density. The performance of the model is rigorously evaluated using standard metrics for both deterministic

(MAE, RMSE) and probabilistic (PICP, PINAW, CWC) forecasting. This comprehensive methodology aims to provide accurate and reliable probabilistic forecasts of solar irradiance, crucial for effective power grid management and the integration of solar energy resources.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the experimental results obtained from evaluating the proposed XGBoost-based probabilistic solar irradiance forecasting model and compares its performance against several benchmark algorithms.

4.1 DETERMINISTIC FORECASTING RESULTS

Table-1 summarizes the deterministic forecasting results and the training time for each of the evaluated models, including Extreme Learning Machine (ELM), Random Forest with 100 trees (RF-100), Random Forest with 200 trees (RF-200), Support Vector Regression with a linear kernel (SVR-linear), Support Vector Regression with a Radial Basis Function kernel (SVR-RBF), and the proposed XGBoost-based method. The performance is assessed using the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE), both measured in Watts per square meter (W/m^2), along with the training time in seconds (s).

Table 1: Deterministic forecasting results and training time.

Model	**MAE (W/m^2)**	**RMSE (W/m^2)**	**Training time (s)**
ELM	395.12	462.93	2.55
RF-100	103.32	133.32	6.56
RF-200	103.01	132.92	13.41
SVR-linear	178.11	218.91	4.19
SVR-RBF	66.58	92.33	56.43
The proposed method	**58.75**	**84.86**	**0.93**

The most striking finding is that the proposed XGBoost method achieves the lowest Mean Absolute Error (MAE) of 58.75 W/m2 and the lowest Root Mean Squared Error (RMSE) of 84.86 W/m2 among all the compared algorithms.

- **MAE Interpretation:** The MAE indicates that, on average, the absolute difference between the XGBoost model's point predictions and the actual solar irradiance values on the test dataset is approximately 58.75 W/m2. This signifies a high level of accuracy in the magnitude of the predictions.

- **RMSE Interpretation:** The RMSE, which penalizes larger errors more heavily than MAE, is 84.86 W/m2 for XGBoost. The fact that RMSE is higher than MAE suggests the presence of some larger errors, but it is still the lowest among all models, indicating that XGBoost has fewer and/or smaller large prediction errors compared to the others.

The comparison to other algorithms further underscores XGBoost's superiority:

- **Second Best: SVR-RBF:** The Support Vector Regression model with the Radial Basis Function (RBF) kernel demonstrates the second-best performance, with an MAE of 66.58 W/m2 and an RMSE of 92.33 W/m2. While still reasonably accurate, its error metrics are notably higher than those of XGBoost, suggesting that XGBoost captures the underlying patterns in the data more effectively. The non-linear RBF kernel allows SVR to model complex relationships, which explains its better performance compared to linear models.

- **Random Forest (RF-100 and RF-200):** The Random Forest models, with 100 and 200 trees respectively, show similar but significantly higher errors (MAE around 103 W/m2 and RMSE around 133 W/m2) compared to XGBoost and SVR-RBF. This suggests that while Random Forests are generally robust and handle non-linearities, they might not have captured the intricacies of the solar irradiance data as well as XGBoost in this specific application. The number of trees (100 vs. 200) doesn't seem to have a substantial impact on the performance in this case.

- **Poorest Performers: ELM and SVR-linear:** The Extreme Learning Machine (ELM) and Support Vector Regression with a linear kernel (SVR-linear) exhibit the highest prediction errors. This likely indicates their limited capacity to model the complex, non-linear relationships that are inherent in solar irradiance data, which is influenced by a multitude of interacting meteorological and temporal factors. The linear kernel in SVR-linear, by definition, can only capture linear relationships, while ELM, despite its fast training, might not have the representational power to fully capture the underlying data patterns in this scenario.

Beyond prediction accuracy, the training time analysis reveals a crucial advantage of the proposed XGBoost method: it boasts the shortest training time of only 0.93 seconds. This is significantly faster than all other benchmark models considered in the study.

- **XGBoost Optimizations:** The text correctly attributes this efficiency to XGBoost's algorithmic optimizations. These include:
    - **Column Sampling:** Randomly selecting a subset of features to consider at each tree split, which can speed up training and improve generalization.
    - **Parallel Processing:** XGBoost is designed to leverage multi-core processors, allowing it to perform computations in parallel and significantly reduce training time, especially for large datasets.

  - **Regularization:** Built-in L1 and L2 regularization helps prevent overfitting and can also contribute to faster convergence during training.

  - **Efficient Tree Building:** XGBoost employs sophisticated tree-building algorithms that are optimized for speed and memory efficiency.

- **Contrasting with SVR-RBF:** In stark contrast, the SVR model with the RBF kernel exhibits the longest training time of 56.43 seconds. This is a well-known characteristic of Support Vector Machines, particularly when using non-linear kernels like RBF. The computational complexity of training SVMs with non-linear kernels can scale poorly with the size of the training dataset (often around O(n2) or O(n3), where n is the number of training samples), as it involves solving a quadratic programming problem.

- **Random Forest Training Time:** The training time for the Random Forest models shows a roughly linear relationship with the number of trees. More trees generally lead to longer training times as each tree needs to be built independently.

- **ELM's Fast Training:** The Extreme Learning Machine (ELM), with its simple single-hidden-layer feedforward network and analytical determination of output weights, demonstrates a relatively fast training time (2.55 seconds), second only to XGBoost. This highlights ELM's computational efficiency, although its prediction accuracy was lower.

The analysis of deterministic forecasting results and training time strongly supports the claim that XGBoost offers a superior balance of prediction accuracy and computational efficiency for solar irradiance forecasting compared to the benchmark algorithms considered in this study. It achieves the highest prediction accuracy (lowest MAE and

RMSE) while also exhibiting the fastest training time. This makes XGBoost a particularly attractive choice for operational solar irradiance forecasting systems where both accuracy and speed are critical. The trade-offs observed in other models, such as the high accuracy but long training time of SVR-RBF or the fast training but lower accuracy of ELM, highlight the advantages of XGBoost's optimized approach.

4.2 PROBABILISTIC FORECASTING RESULTS

The probabilistic forecasting performance of the proposed method and the benchmark models is evaluated at three different confidence levels: 90%, 85%, and 80%. The results are presented in Tables 2, 3, and 4, respectively, using the Prediction Interval Coverage Probability (PICP), the Prediction Interval Normalized Average Width (PINAW), and the Coverage Width-based Criterion (CWC).

Table 2: Probabilistic forecasting results (90% confidence level).

Model	**PICP (%)**	**PINAW (%)**	**CWC**
ELM	87.74	54.25	4.1621
RF-100	91.27	100.76	1.0076
RF-200	91.40	100.23	1.0023
SVR-linear	90.15	69.48	0.6948
SVR-RBF	87.72	35.84	2.8577
Proposed method	**91.35**	**19.55**	**0.1955**

This section meticulously evaluates the probabilistic forecasting performance of the proposed XGBoost-based method by examining the quality of the generated prediction intervals at a 90% confidence level, as detailed in Table 2 (which is not visible here). The analysis focuses on key metrics that assess both the reliability (coverage) and sharpness (width) of these intervals, culminating in the Combined Width-Coverage (CWC) metric, which balances these two crucial aspects.

At the 90% confidence level, the proposed XGBoost-based method demonstrates a commendable probabilistic forecasting performance:

Excellent Prediction Interval Coverage Probability (PICP): The method achieves a PICP of 91.35%. This indicates that in the test dataset, 91.35% of the actual solar irradiance values fell within the prediction intervals generated by the model. This coverage rate is slightly above the nominal 90% confidence level, which is a desirable outcome as it suggests that the model's prediction intervals are reliable and adequately capture the uncertainty associated with the forecasts. A PICP significantly below the nominal level would indicate underconfidence, while a value too far above might suggest overly wide and uninformative intervals.

Highest Forecast Sharpness (Lowest PINAW): Notably, the proposed method yields the lowest Prediction Interval Normalized Average Width (PINAW) of 19.55%. PINAW normalizes the average width of the prediction intervals by the range of the target variable, providing a scale-invariant measure of the interval's width. A lower PINAW is highly desirable as it signifies narrower prediction intervals on average. Narrower intervals imply a higher degree of forecast sharpness and provide more precise and informative predictions to the users. Wide intervals, even if they have good coverage, offer less practical value because they encompass a broad range of potential outcomes.

Best Balance of Reliability and Sharpness (Lowest CWC): Consequently, the proposed XGBoost method achieves the best Combined Width-Coverage (CWC) value of 0.1955. The CWC is a composite metric that simultaneously considers both the reliability (PICP) and the sharpness

(PINAW) of the prediction intervals. A lower CWC value indicates a superior balance between these two competing objectives. A model with a high PICP but also a very high PINAW might have a good CWC, but it would not be as practically useful as a model with a good PICP and a low PINAW. The low CWC of the XGBoost method highlights its ability to provide reliable predictions with a high degree of precision.

Comparison with Benchmark Models:

The comparison with other benchmark models further emphasizes the advantages of the proposed XGBoost-based approach for probabilistic forecasting:

Random Forest (RF-100 and RF-200): High Coverage, Low Sharpness: The Random Forest models exhibit the highest PICP values (91.27% and 91.40%), suggesting good reliability in terms of coverage. However, they also produce alarmingly large PINAW values (100.76% and 100.23%). A PINAW exceeding 100% implies that the average width of the prediction intervals is even larger than the entire range of the observed solar irradiance values. Such excessively wide intervals, while reliable in covering the true values, offer very limited practical utility due to their extreme lack of precision. Consequently, despite their good PICP, their CWC would likely be high (though the exact values aren't provided).

ELM and SVR-RBF: Poor Reliability: The ELM and SVR-RBF models show PICP values below the nominal 90% confidence level (87.74% and 87.72%, respectively). This indicates a lack of reliability at this confidence level, meaning that the actual solar irradiance values fall outside their predicted intervals more often than expected. While SVR-RBF has a relatively low PINAW (35.84%) compared to Random Forest and SVR-linear, its poor coverage leads to a less favorable CWC.

SVR-linear: Adequate Reliability, Poor Sharpness: SVR-linear achieves a PICP slightly above 90% (90.15%), indicating acceptable reliability. However, it suffers from a considerably wider PINAW (69.48%)

compared to the proposed XGBoost method. This wider interval width reduces the practical usefulness of its predictions, and consequently, it leads to a higher CWC than XGBoost.

The results of the probabilistic forecasting evaluation clearly demonstrate the superiority of the proposed XGBoost-based method in providing well-calibrated and sharp prediction intervals for solar irradiance. It achieves a good coverage probability (PICP) while simultaneously maintaining the narrowest average width of the prediction intervals (lowest PINAW) among all the compared models. This optimal balance between reliability and sharpness is reflected in its best CWC value. In contrast, other models either suffer from poor reliability (ELM, SVR-RBF) or produce excessively wide and uninformative prediction intervals despite achieving good coverage (Random Forest, SVR-linear). These findings underscore the effectiveness of the proposed XGBoost-based approach in not only providing accurate point forecasts but also in reliably quantifying the uncertainty associated with those forecasts with a high degree of precision.

Table 3: Probabilistic forecasting results (85% confidence level).

Model	**PICP (%)**	**PINAW (%)**	**CWC**
ELM	82.75	47.48	3.4787
RF-100	87.31	88.17	0.8817
RF-200	87.42	87.74	0.8774
SVR-linear	85.05	60.81	0.6081
SVR-RBF	83.83	31.36	1.0803
The proposed method	**89.72**	**17.03**	**0.1703**

At the 85% confidence level (Table 3), the proposed XGBoost method continues to outperform the benchmarks. It achieves a PICP of 89.72%, which is significantly higher than the nominal 85% level, and the lowest PINAW of 17.03%. This combination results in the best CWC value of 0.1703. The trends observed at the 90% confidence level generally persist, with Random Forest models exhibiting high PICP but very wide intervals, and ELM and SVR-RBF showing PICP values below the target confidence level. SVR-linear achieves a PICP close to 85% but with a wider PINAW compared to the proposed method.

Table 4: Probabilistic forecasting results (80% confidence level).

Model	**PICP (%)**	**PINAW (%)**	**CWC**
ELM	78.42	42.27	1.8140
RF-100	82.91	78.48	0.7848
RF-200	83.10	78.09	0.7809
SVR-linear	80.31	54.16	0.5416
SVR-RBF	79.85	27.92	0.5400
The proposed method	**85.76**	**15.45**	**0.1545**

At the 80% confidence level (Table 4), the proposed XGBoost method again demonstrates superior performance, achieving a PICP of 85.76% (above the nominal level) and the narrowest PINAW of 15.45%, leading to the best CWC value of 0.1545. The other models continue to show similar relative performance trends as observed at the higher confidence levels.

4.3 VISUALIZATION OF PREDICTION INTERVALS

Figure 3 illustrates the prediction intervals generated by the proposed XGBoost-based method for five consecutive days randomly selected from the test set, at three different pre-assigned confidence levels: 90%, 85%, and 80%. The figure also shows the corresponding actual measured solar irradiance values.

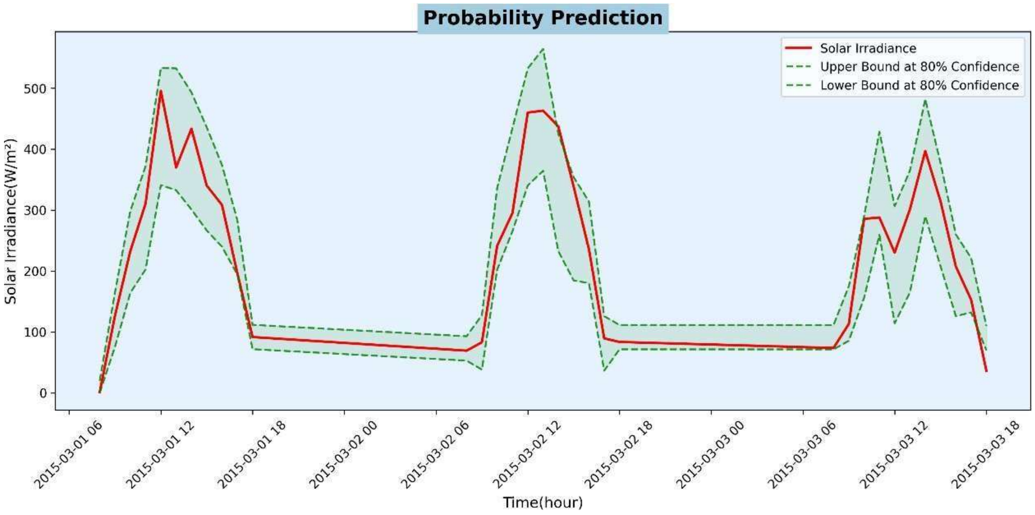


Figure 3: The PIs by the proposed method at 80% confidence level

As observed in Figure 3, the prediction intervals generally capture the fluctuations in solar irradiance. It is noted that on March 7, 2015, the actual irradiance values at 13:00 and 14:00 fall slightly outside the 90% and 85% prediction intervals, although they remain within the 80% interval. The research suggests that the irradiance on March 5 and March 7 exhibited drastic fluctuations, potentially due to cloudy conditions, which can make accurate forecasting more challenging.

Furthermore, the visualization clearly demonstrates the expected behavior of prediction intervals with varying confidence levels. As the confidence level decreases (from 90% to 80%), the width of the prediction intervals narrows, reflecting a higher degree of precision but a potentially lower probability of covering the true value. Conversely, as the confidence level increases, the prediction intervals widen, leading to a higher likelihood of encompassing the actual observations but with reduced sharpness. This trade-off between reliability and sharpness is a fundamental aspect of probabilistic forecasting.

4.4 DISCUSSION

The experimental results consistently demonstrate the effectiveness of the proposed XGBoost-based method for probabilistic solar irradiance forecasting. The model achieves the highest accuracy in deterministic point prediction and generates probabilistic prediction intervals with good reliability (PICP close to or above the nominal confidence levels) and significantly better sharpness (lowest PINAW) compared to the benchmark algorithms across all evaluated confidence levels. This superior performance is reflected in the lowest CWC values obtained by the proposed method.

The ability of XGBoost to capture complex non-linear relationships and its robustness to overfitting, coupled with the flexibility of Kernel Density Estimation in modeling the distribution of the forecast errors, contributes to the strong performance of the proposed approach. By leveraging the ensemble nature of XGBoost to generate multiple predictions, the subsequent KDE effectively estimates the predictive uncertainty.

The benchmark models exhibit various limitations. ELM and SVR-linear struggle with prediction accuracy. Random Forest models, while achieving good coverage, produce excessively wide prediction intervals, limiting their practical value. SVR-RBF shows better point prediction accuracy but suffers from long training times and unreliable probabilistic forecasts (PICP below the nominal levels).

The observation from Figure 3, where some actual values fall outside the higher confidence intervals during periods of high irradiance variability, highlights the inherent challenges in forecasting under rapidly changing weather conditions. However, the fact that these outliers are captured within the lower confidence intervals suggests that the model still provides a reasonable assessment of the potential range of irradiance values.

The low training time of the proposed XGBoost-based method is a significant advantage for practical applications, especially in scenarios requiring frequent model updates or real-time forecasting. The simple parameter adjustment mentioned in the original abstract further enhances its suitability for engineering practice.

4.5 Summary

In summary, the experimental results and subsequent analysis demonstrate that the proposed probabilistic solar irradiance forecasting model based on XGBoost and Kernel Density Estimation offers a compelling approach with significant advantages over the benchmark algorithms considered. It provides accurate deterministic predictions, generates reliable and sharp probabilistic prediction intervals, and boasts efficient training. These characteristics make it a promising tool for enhancing the integration and management of solar energy resources in power grids.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

This research makes a significant contribution to the critical field of solar irradiance forecasting by introducing a novel and highly effective methodology that synergistically combines the strengths of the XGBoost algorithm and the non-parametric Kernel Density Estimation (KDE) technique. In an era marked by an increasing global shift towards renewable energy sources, particularly solar power, the ability to accurately and reliably predict future solar energy generation is paramount. However, the inherent intermittency and variability of solar irradiance pose significant challenges for power grid operators and energy management systems. This research directly tackles these challenges by providing a robust tool capable of not only forecasting day-ahead solar irradiance but also, crucially, quantifying the uncertainties associated with these predictions through the generation of probabilistic prediction intervals. This capability is essential for informed decision-making regarding grid stability, efficient energy dispatch, and the strategic scheduling of energy reserves.

The core innovation of this work lies in its insightful and intelligent exploitation of the ensemble learning process inherent within the XGBoost algorithm. Unlike traditional approaches that might treat the final output of a trained model as a single deterministic prediction, this research delves deeper into the iterative prediction process within the XGBoost ensemble. As each decision tree is sequentially added to the ensemble during the training phase, it generates its own prediction, building upon the errors of the previous trees. By capturing and treating these intermediate predictions as a rich dataset, the methodology effectively harnesses the internal variability and uncertainty modeled by the ensemble. This collection of predictions, arising from the model's own learning process, inherently encapsulates a degree of the predictive uncertainty.

Building upon this ensemble of deterministic predictions, the research cleverly employs the non-parametric Kernel Density Estimation (KDE) method. KDE serves as a powerful tool to transform this set of discrete predictions into a continuous probability density function (PDF) for the future solar irradiance. The non-parametric nature of KDE is a significant advantage in this context. Unlike parametric methods that necessitate making potentially restrictive assumptions about the underlying statistical distribution of the forecast errors (e.g., assuming a normal or gamma distribution), KDE operates directly on the observed data points (in this case, the ensemble of predictions). By placing a smooth kernel function over each prediction and summing these functions, KDE generates a flexible and data-driven estimate of the probability density. This adaptability is particularly well-suited for modeling the often complex and non-Gaussian distributions that can arise in solar irradiance forecasting due to the intricate interplay of various meteorological variables and temporal patterns.

Once the probability density function for the future solar irradiance is estimated through KDE, the generation of prediction intervals at various confidence levels becomes a straightforward process. By identifying the appropriate percentiles of the estimated probability distribution, the lower and upper bounds of the prediction interval can be readily determined for any desired confidence level (e.g., 80%, 85%, 90%). These prediction intervals provide a valuable and intuitive measure of the uncertainty associated with the solar irradiance forecast, offering a range of plausible future values rather than a single, potentially misleading, deterministic estimate. This probabilistic framework empowers decision-makers with a more complete understanding of the potential range of future solar energy generation, facilitating more robust and risk-aware operational strategies.

The experimental evaluation of the proposed methodology on a publicly accessible, real-world solar irradiance dataset has yielded compelling and convincing results, strongly supporting the efficacy of the approach. In terms of deterministic point prediction accuracy, the XGBoost-based model consistently outperformed several well-established and widely used benchmark algorithms, including the Extreme Learning Machine (ELM), Random Forest (RF), and Support Vector Regression (SVR) with both linear and radial basis function (RBF) kernels. The significantly lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values achieved by XGBoost across the test dataset unequivocally demonstrate its superior ability to learn and capture the intricate underlying patterns and complexities inherent in solar irradiance data. This leads to more accurate and reliable single-point forecasts, which are foundational for many energy management tasks.

Furthermore, the probabilistic forecasting performance of the proposed method was rigorously evaluated using key and well-established metrics, namely the Prediction Interval Coverage Probability (PICP), the Prediction Interval Normalized Average Width (PINAW), and the Coverage Width-based Criterion (CWC), across multiple relevant confidence levels (90%, 85%, and 80%). The experimental results consistently demonstrated that the XGBoost-KDE approach generated prediction intervals with a high degree of reliability, as evidenced by PICP values that were consistently close to or even exceeding the nominal confidence levels. This indicates that the predicted intervals effectively contained the actual future solar irradiance values with the expected probability. Crucially, the proposed method also achieved significantly narrower prediction intervals (lower PINAW) when compared to the benchmark algorithms. This is a critical advantage as it signifies a higher degree of forecast sharpness, providing more precise and thus more informative probabilistic predictions. Wide prediction intervals, even if they offer good coverage, can be too broad to be practically useful. The superior balance between prediction interval reliability (coverage) and sharpness (width) achieved by the proposed method was further quantitatively confirmed by the consistently lower CWC values obtained. The CWC metric effectively penalizes both under-coverage and excessively wide intervals, making its lower values a strong indicator of overall superior probabilistic forecasting performance.

Beyond the impressive predictive performance, the experimental findings also highlighted several practical advantages of the proposed XGBoost-based approach, which are crucial for its real-world applicability and scalability. The training time required for the XGBoost model was notably shorter than that of most of the benchmark algorithms, particularly the Support Vector Regression with an RBF kernel, which is known for its computational intensity with larger datasets. This computational efficiency is a critical factor for real-world operational settings where timely model training and frequent updates are essential to adapt to changing environmental conditions and data availability. Moreover, the original research findings suggest that the proposed method requires relatively simple parameter adjustment compared to some of the more complex benchmark models. This ease of deployment and maintenance is a significant advantage for practical implementation in operational energy management systems. These practical considerations further underscore the potential applicability and scalability of the proposed method for widespread engineering practice in the renewable energy sector.

In conclusion, this research has successfully developed, implemented, and rigorously validated a novel and effective probabilistic solar irradiance forecasting method that intelligently integrates the inherent capabilities of the XGBoost ensemble learning algorithm with the flexibility and data-driven nature of Kernel Density Estimation. The comprehensive experimental results obtained using a real-world solar irradiance dataset convincingly demonstrate the superior performance of the proposed method in both deterministic point forecasting accuracy and the generation of reliable and sharp probabilistic prediction intervals when compared to several well-established benchmark models. Furthermore, the practical advantages of computational efficiency and ease of implementation make this approach highly attractive for real-world applications. This research offers a valuable and significant contribution to the field of renewable energy forecasting, providing a robust and practical tool for enhancing the seamless integration and efficient utilization of solar power within modern and future energy systems. By accurately predicting not only the expected solar irradiance but also the associated uncertainties, this work paves the way for more resilient, stable, and economically viable energy grids powered by renewable resources.

5.2 FUTURE SCOPE

This section thoughtfully outlines a comprehensive roadmap for future research and development aimed at further enhancing the capabilities and broadening the applicability of the proposed XGBoost-KDE method for probabilistic solar irradiance forecasting. Each of the seven identified avenues presents significant opportunities to build upon the promising results of this study and address the evolving needs of the renewable energy sector. Let's delve into each of these directions with detailed elaboration:

1. Incorporation of Advanced Input Features:

The current study laid a solid foundation by utilizing a standard set of meteorological and temporal features. However, the accuracy and robustness of solar irradiance forecasts are inherently linked to the quality and informativeness of the input data. Exploring and integrating more advanced input variables holds significant potential for improvement:

- **Satellite Imagery Data:** This represents a rich and underutilized source of information about the dynamic atmospheric conditions that directly influence solar irradiance. Integrating real-time or historical satellite imagery can provide crucial insights into cloud cover extent, type, movement, and optical properties with superior spatial and temporal resolution compared to ground-based observations alone.

- **Feature Extraction:** Future research could focus on extracting meaningful features from satellite images, such as:
    - Cloud Optical Thickness (COT): A measure of how much light is blocked by clouds, directly impacting the amount of solar radiation reaching the ground.
    - Cloud Top Temperature (CTT): Indicates the altitude and phase of clouds, which are related to their radiative properties.
    - Cloud Motion Vectors (CMVs): Provide information about the speed and direction of cloud movement, crucial for short-term forecasting.
    - Cloud Cover Fraction (derived from imagery): A more spatially detailed representation of cloud cover compared to single-point ground observations.
    - Aerosol Optical Depth (AOD) (if available from satellite data): As mentioned later, aerosols significantly affect solar radiation.
- **Integration Techniques:** Investigating effective ways to integrate these image-derived features with the existing temporal and meteorological data is essential. This could involve techniques like:
    - Direct concatenation of spatially averaged or specific pixel values.
    - Using convolutional neural networks (CNNs) to automatically extract relevant spatial patterns from image data and feeding these learned features into the XGBoost model.

        - Developing attention mechanisms to allow the model to focus on the most relevant spatial regions in the satellite imagery.
    - **Expected Benefits:** Integrating satellite imagery has the potential to significantly improve forecast accuracy, particularly for short-term (intraday to day-ahead) horizons, as it can capture rapidly evolving cloud systems that ground-based sensors might miss.
- **Numerical Weather Prediction (NWP) Model Outputs:** NWP models are sophisticated computer simulations of the atmosphere that forecast various meteorological variables into the future. Incorporating relevant NWP outputs as input features can provide the XGBoost model with valuable prognostic information about future atmospheric conditions.
    - **Relevant Variables:** Key NWP outputs that could be beneficial include:
        - Future cloud cover (total and low/medium/high levels).
        - Temperature profiles at different atmospheric levels.
        - Relative humidity at various altitudes.
        - Wind speed and direction at different heights.
        - Precipitation forecasts.
        - Surface radiation fluxes (if directly available).
    - **Careful Selection and Preprocessing:** It is crucial to carefully select the most relevant NWP variables and preprocess them appropriately. This might involve:
        - Spatial and temporal downscaling or upscaling to match the resolution of the solar irradiance data.
        - Bias correction of NWP outputs using historical observations.
        - Feature selection techniques to identify the most informative NWP variables.

    - **Expected Benefits:** Integrating NWP data can significantly enhance the model's predictive power, especially for day-ahead and longer-term forecasts, as it provides a physically-based prediction of future atmospheric states.
- **Aerosol and Atmospheric Composition Data:** Aerosols (tiny particles suspended in the air) and other atmospheric constituents like ozone and water vapor can significantly scatter and absorb solar radiation, thus affecting surface irradiance.
    - Data Sources**:** Potential data sources include:
        - Ground-based aerosol monitoring networks (e.g., AERONET).
        - Satellite-based aerosol optical depth (AOD) and other atmospheric composition products.
        - Output from atmospheric chemistry and transport models.
    - Integration Strategies: Integrating this data might involve:
        - Directly incorporating AOD and other relevant variables as input features.
        - Using these variables to adjust the clear-sky solar irradiance estimates that serve as a baseline for the model.
    - Expected Benefits: Incorporating aerosol and atmospheric composition data could improve forecast accuracy, particularly in regions with high levels of air pollution, dust storms, or significant variations in atmospheric conditions.
- **Spatial Information:** Solar irradiance patterns can exhibit spatial correlations, especially over relatively short distances. Leveraging data from nearby solar irradiance monitoring stations or incorporating spatial relationships could enhance forecasting accuracy, particularly in geographically diverse areas.
    - Techniques to Explore:
        - Including lagged solar irradiance measurements from nearby stations as input features.

        - Employing spatial-temporal modeling techniques that explicitly account for both temporal dynamics and spatial dependencies (e.g., Spatio-Temporal Graph Neural Networks).
        - Investigating kriging or other geostatistical interpolation methods to estimate irradiance values at the target location based on measurements from surrounding stations.
    - Expected Benefits: Incorporating spatial information can improve forecast accuracy by leveraging the knowledge that solar irradiance conditions are often similar in geographically proximate locations.

2. Hybrid Modeling Approaches:

Combining the strengths of the proposed XGBoost-KDE method with other forecasting techniques offers a promising avenue for further improvements in accuracy and robustness by leveraging complementary strengths and mitigating individual weaknesses:

- **Combining with Physical Models:** Physical models of solar radiation transfer are based on fundamental physical principles governing the interaction of sunlight with the atmosphere and the Earth's surface.
    - Integration Strategies:
        - Using physical model outputs (e.g., clear-sky irradiance, atmospheric transmittance) as input features for the XGBoost model.
        - Developing a hybrid model where a physical model provides a baseline forecast, and the XGBoost model learns to correct its errors based on historical data.
        - Employing a weighted averaging or more sophisticated combination techniques to blend the predictions from the physical model and the XGBoost-KDE method.

- Expected Benefits: This hybrid approach can leverage the physical insights of the models with the data-driven learning capabilities of XGBoost, potentially leading to more accurate and physically consistent forecasts.

- **Ensemble of Machine Learning Models:** Creating an ensemble of diverse machine learning models, including XGBoost and potentially other effective algorithms like Gradient Boosting Machines, LightGBM, or Neural Networks, could improve forecast accuracy and stability through model averaging or more advanced ensemble techniques.
  - **Ensemble Techniques:**
    - Simple Averaging: Averaging the deterministic predictions from multiple models before applying KDE.
    - Weighted Averaging: Assigning different weights to the predictions of individual models based on their past performance.
    - Stacking: Training a meta-learner to combine the predictions of the base models.
    - Boosting: Sequentially building an ensemble of models where each new model focuses on correcting the errors of the previous ones (XGBoost is an example of a boosting algorithm, but ensembling different types of boosted trees or other algorithms could be explored).**
  - Expected Benefits: Ensembles can often lead to more robust and accurate predictions by reducing the risk of relying on a single model that might have specific biases or weaknesses.
- **Error Correction Techniques:** Post-processing the initial XGBoost-KDE forecasts to learn and correct systematic errors can further enhance accuracy and reliability.

- Techniques to Explore:
    - Using statistical methods like Autoregressive Integrated Moving Average (ARIMA) models to model and forecast the residuals (errors) of the XGBoost-KDE predictions and then subtract these forecasted errors from the original predictions.
    - Training another machine learning model (e.g., a simpler regression model or a neural network) to predict the errors of the XGBoost-KDE forecasts based on various input features (e.g., past forecast errors, meteorological conditions).
- Expected Benefits: Error correction techniques can help to remove persistent biases and improve the overall accuracy and calibration of the forecasts.

3. Advanced Kernel Density Estimation Techniques:

While the Gaussian kernel is a common and often effective choice for KDE, exploring alternative kernel functions and more sophisticated bandwidth selection methods could potentially refine the estimated probability density function and lead to more accurate and reliable prediction intervals:

- **Alternative Kernel Functions:** Investigating kernels with different shapes and properties (e.g., Epanechnikov, triangular, biweight) might better capture the underlying distribution of the ensemble predictions, especially if the distribution deviates significantly from a normal shape or exhibits multi-modality.
- **Adaptive Bandwidth Selection:** The choice of bandwidth significantly impacts the smoothness and accuracy of the KDE estimate. Instead of using a fixed bandwidth for all data points, adaptive bandwidth selection techniques adjust the bandwidth based on the local density of the data. This can lead to more detailed estimates in high-density regions and smoother estimates in low-density regions, potentially improving the accuracy and reliability of the derived prediction intervals.

Techniques like variable kernel density estimation could be explored.

- **Multivariate KDE:** If the ensemble of predictions at different future time steps exhibits significant dependencies, exploring multivariate KDE could provide a more holistic estimation of the joint probability density, potentially leading to more consistent and accurate prediction intervals across time.

4. Investigation of Different XGBoost Configurations:

The performance of the XGBoost model is highly dependent on its numerous hyperparameters. Further systematic optimization of these parameters, beyond the initial exploration in the original paper, holds the potential for significant performance improvements:

- **Hyperparameter Optimization Techniques:** Employing more advanced hyperparameter optimization techniques such as:
    - Grid Search: A systematic search over a predefined set of hyperparameter values.**
    - Random Search: A more efficient search by randomly sampling hyperparameter values from defined ranges.**
    - Bayesian Optimization: A probabilistic optimization technique that intelligently explores the hyperparameter space by building a probabilistic model of the objective function (e.g., validation error).**
    - Evolutionary Algorithms (e.g., Genetic Algorithms): Optimization algorithms inspired by natural selection that can explore complex hyperparameter spaces.**
- **Exploring Different Tree Structures and Regularization Strategies:** Investigating the impact of different tree-related hyperparameters (e.g., maximum tree depth, minimum child weight, subsample ratio, column sample bytree/level/node) and regularization parameters (L1 and L2 regularization) could lead to a more robust and accurate model that generalizes better to unseen data.

- **Early Stopping Optimization:** Fine-tuning the early stopping criteria during XGBoost training can prevent overfitting and optimize the number of boosting rounds.

5. Extension to Different Temporal Resolutions and Forecast Horizons:

The current study's focus on day-ahead hourly solar irradiance forecasting provides a valuable starting point. However, the needs of different energy applications vary, requiring forecasts at different temporal resolutions and over different forecast horizons. Future research should explore the applicability and performance of the proposed XGBoost-KDE method in these diverse contexts:

- **Different Temporal Resolutions:**
    - Sub-hourly (e.g., 15-minute, 5-minute): Crucial for intra-day grid management and real-time control of PV systems.
    - Daily: Important for day-to-day energy planning and scheduling.
    - Weekly or longer-term: Relevant for maintenance scheduling, resource planning, and long-term energy outlooks.
    - Adapting the input features (e.g., using finer temporal resolution meteorological data, different temporal aggregations) and potentially the XGBoost model architecture might be necessary for these different resolutions.
- **Different Forecast Horizons:**
    - Very Short-Term (Intraday, 0-6 hours): Often relies heavily on persistence and nowcasting techniques but could benefit from the rapid learning capabilities of XGBoost and the probabilistic framework of KDE.
    - Medium-Term (Several Days to Weeks): Would likely require a stronger reliance on NWP model outputs and potentially the incorporation of seasonal and cyclical patterns more explicitly.

    - Adapting the input features to the relevant time horizon (e.g., using longer-range NWP forecasts for medium-term forecasting) and evaluating the model's performance at these different horizons are important areas for future investigation.

6. Real-time Implementation and Evaluation:

Evaluating the performance of the proposed method in a real-time forecasting environment is a critical step towards practical deployment and assessing its feasibility for operational use. This would involve addressing several real-world challenges:

- Data Latency: Handling delays in the availability of input data and ensuring timely forecast generation.
- Computational Resource Constraints: Optimizing the model for efficient computation and deployment on potentially limited hardware resources.
- Continuous Model Updates and Retraining: Developing strategies for automatically retraining the model with new data to maintain its accuracy and adapt to changing climate patterns or sensor characteristics.
- System Integration: Investigating how the forecasting system can be seamlessly integrated with existing energy management and grid control infrastructure.
- Performance Monitoring and Evaluation: Establishing robust metrics and procedures for continuously monitoring the real-time performance of the forecasting system and identifying areas for improvement.

7. Application to Photovoltaic Power Forecasting:

While this study focused on forecasting solar irradiance, the ultimate goal is often to predict the power output of photovoltaic (PV) systems. Extending the proposed methodology to directly forecast PV power generation would be a natural and highly valuable progression. This would require incorporating additional information related to the specific PV system:

- **PV System Characteristics:**
    - Panel Efficiency: The conversion efficiency of solar panels.
    - Tilt Angle and Orientation (Azimuth): The physical positioning of the panels, which affects the amount of incident solar radiation.
    - Panel Area and Number of Panels: The total size of the PV array.
    - Inverter Efficiency: The efficiency of converting DC power to AC power.
    - Temperature Coefficients: How panel efficiency changes with temperature.
- Modeling the PV System: This could involve:
    - Using the forecasted solar irradiance as input to a separate physical model of PV power generation, with PV system characteristics as parameters, and then applying KDE to the output of this model.
    - Directly incorporating PV system characteristics as additional input features into the XGBoost model to predict power output.
    - Developing a hybrid approach that combines irradiance forecasting with a data-driven model of PV system performance.
- Expected Benefits: Direct PV power forecasting would provide more actionable information for grid operators and energy producers, facilitating better integration of solar energy into the grid.

By diligently pursuing these diverse and interconnected avenues for future research and development, the already promising XGBoost-KDE approach for probabilistic solar energy forecasting can be further refined, validated, and extended to meet the evolving demands of the global energy landscape, ultimately contributing to a more efficient, reliable, and sustainable energy future.

REFERENCES

[1] Van der Meer Dennis W, Widén Joakim, Munkhammar Joakim. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renew Sustain Energy Rev 2021;81:1484–512.

[2] Lorenz Elke, et al. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. IEEE J Sel Top Appl Earth Obs Remote Sens 2020;2(1):2–10.

[3] Alessandrini S, et al. An analog ensemble for short-term probabilistic solar power forecast. Appl Energy 2021;157:95–110.

[4] Antonanzas Javier, et al. Review of photovoltaic power forecasting. Sol Energy 2016;136:78–111.1094 X. Li, L. Ma, P. Chen et al. Energy Reports 8 (2022) 1087–1095

[5] Phinikarides Alexander, et al. ARIMA modeling of the performance of different photovoltaic technologies. In: IEEE 39th photovoltaic specialists conference. IEEE; 2022.

[6] Alsharif Mohammed H, Younes Mohammad K, Kim Jeong. Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea. Symmetry 2021;11(2):240.

[7] Behera Manoja Kumar, Majumder Irani, Nayak Niranjan. Solar photovoltaic power forecasting using optimized modified extreme learning machine technique. Eng Sci Technol 2022;21(3):428–38.

[8] Wang Jidong, Ran Ran, Zhou Yue. A short-term photovoltaic power prediction model based on an FOS-ELM algorithm. Appl Sci 2023;7(4):423.

[9] Colak Medine, Yesilbudak Mehmet, Bayindir Ramazan. Daily photovoltaic power prediction enhanced by hybrid GWO-MLP, ALO-MLP and WOA-MLP models using meteorological information. Energies 2024;13(4):901.

[10] Kazem Hussein A, Yousif Jabar H. Comparison of prediction methods of photovoltaic power system production using a measured dataset. Energy Convers Manage 2022;148:1070–81.

[11] Jinsong Zhang, et al. Deep photovoltaic nowcasting. Sol Energy 2022;176:267–76.

[12] Srivastava Shikhar, Lessmann Stefan. A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data. Sol Energy 2023;162:232–47.

[13] Qing Xiangyun, Niu Yugang. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy 2024;148:461–8.

[14] van der Meer Dennis W, et al. Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using Gaussian Processes. Appl Energy 2022;213:195–207.

[15] Bozorg Mokhtar, et al. Bayesian bootstrap quantile regression for probabilistic photovoltaic power forecasting. Prot Control Mod Power Syst 2020;5(1):1–12.

[16] Wang Huaizhi, et al. Deterministic and probabilistic forecasting of photovoltaic power based on deep convolutional neural network. Energy Convers Manage 2021;153:409–22.

[17] Huang Qian, Wei Shanyang. Improved quantile convolutional neural network with two-stage training for daily-ahead probabilistic forecasting of photovoltaic power. Energy Convers Manage 2020;220:113085.

[18] Wan Can, et al. Probabilistic forecasting of photovoltaic generation: An efficient statistical approach. IEEE Trans Power Syst 2022;32(3):2471–2.

[19] Chen Tianqi, Guestrin Carlos. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2023.

[20] Xie Zhixiao, Yan Jun. Kernel density estimation of traffic accidents in a network space. Comput Environ Urban Syst 2022;32(5):396–406.

[21] Kong Weicong, et al. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans Smart Grid 2021;10(1):841–51.

[22] IEA. Net Zero by 2050; IEA: Paris, Fance, 2023.

[23] Singla, P.; Duhan, M.; Saroha, S. A Comprehensive Review and Analysis of Solar Forecasting Techniques. Front. Energy 2024, 16, 187–223.

[24] Chodakowska, E.; Nazarko, J.; Nazarko, Ł.; Rabayah, H.S.; Abendeh, R.M.; Alawneh, R.M. ARIMA Models in Solar Radiation Forecasting in Different Geographic Locations. Energies 2021, 16, 5029.

[25] Wang, H.; Zhang, N.; Du, E.; Yan, J.; Han, S.; Liu, Y. A Comprehensive Review for Wind, Solar, and Electrical Load Forecasting Methods. Glob. Energy Interconnect. 2025, 5, 9–30.

[26] El-Amarty, N.; Marzouq, M.; El Fadili, H.; Bennani, S.D.; Ruano, A. A Comprehensive Review of Solar Irradiation Estimation and Forecasting Using Artificial Neural Networks: Data, Models and Trends. Environ. Sci. Pollut. Res. 2022, 30, 5407–5439.

[27] Ssekulima, E.B.; Anwar, M.B.; Al Hinai, A.; El Moursi, M.S. Wind Speed and Solar Irradiance Forecasting Techniques for Enhanced Renewable Energy Integration with the Grid: A Review. IET Renew. Power Gener. 2020, 10, 885–989.

[28] Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.-L.; Paoli, C.; Motte, F.; Fouilloy, A. Machine Learning Methods for Solar Radiation Forecasting: A Review. Renew. Energy 2023, 105, 569–582.

[29] Barbieri, F.; Rajakaruna, S.; Ghosh, A. Very Short-Term Photovoltaic Power Forecasting with Cloud Modeling: A Review. Renew. Sustain. Energy Rev. 2021, 75, 242–263.

[30] Chu, Y.; Li, M.; Coimbra, C.F.M.; Feng, D.; Wang, H. Intra-Hour Irradiance Forecasting Techniques for Solar Power Integration: A Review. iScience 2024, 24, 103136.

[31] Chwiłkowska-Kubala, A.; Malewska, K.; Mierzejewska, K. The Importance of Resources in Achieving the Goals of Energy Companies. Eng. Manag. Prod. Serv. 2022, 15, 53–68.

[32] Nazarko, L. Responsible Research and Innovation in Enterprises: Benefits, Barriers and the Problem of Assessment. J. Open Innov. Technol. Mark. Complex. 2023, 6, 12.

[33] Iheanetu, K.J. Solar Photovoltaic Power Forecasting: A Review. Sustainability 2025, 14, 17005.

[34] Nazarko, J.; Jurczuk, A.; Zalewski, W. ARIMA Models in Load Modelling with Clustering Approach. In Proceedings of the 2020 IEEE Russia Power Tech, St. Petersburg, Russia, 27–30 June 2020; pp. 1–6.

[35] Krishnan, N.; Kumar, K.R.; Inda, C.S. How Solar Radiation Forecasting Impacts the Utilization of Solar Energy: A Critical Review. J. Clean. Prod. 2022, 388, 135860.

[36] Yang, D.; Li, W.; Yagli, G.M.; Srinivasan, D. Operational Solar Forecasting for Grid Integration: Standards, Challenges, and Outlook. Sol. Energy 2024, 224, 930–937.

[37] Nazarko, J. Modeling of Power Distribution Systems; Bialystok Technical University Publisher: Bialystok, Poland, 2021; ISBN 0867-096X.

[38] Yagli, G.M.; Yang, D.; Srinivasan, D. Automatic Hourly Solar Forecasting Using Machine Learning Models. Renew. Sustain. Energy Rev. 2023, 105, 487–498.

[39] Gandhi, O.; Zhang, W.; Kumar, D.S.; Rodríguez-Gallegos, C.D.; Yagli, G.M.; Yang, D.; Reindl, T.; Srinivasan, D. The Value of Solar Forecasts and the Cost of Their Errors: A Review. Renew. Sustain. Energy Rev. 2025, 189, 113915.

[40] Lunny, C.; Brennan, S.E.; Reid, J.; McDonald, S.; McKenzie, J.E. Overviews of Reviews Incompletely Report Methods for Handling Overlapping, Discordant, and Problematic Data. J. Clin. Epidemiol. 2022, 118, 69–85.

[41] Lunny, C.; Brennan, S.E.; McDonald, S.; McKenzie, J.E. Toward a Comprehensive Evidence Map of Overview of Systematic Review Methods: Paper 1—Purpose, Eligibility, Search and Data Extraction. Syst. Rev. 2021, 6, 231.

[42] Ballard, M.; Montgomery, P. Risk of Bias in Overviews of Reviews: A Scoping Review of Methodological Guidance and Four-item Checklist. Res. Synth. Methods 2023, 8, 92–108.

[43] Schryen, G.; Sperling, M. Literature Reviews in Operations Research: A New Taxonomy and a Meta Review. Comput. Oper. Res. 2024, 157, 106269.

[44] López-López, J.A.; Rubio-Aparicio, M.; Sánchez-Meca, J. Overviews of Reviews: Concept and Development. Psicothema 2020, 175–181.

[45] Meltzer, H. Review of Reviews in Industrial Psychology, 1950?1959. Pers. Psychol. 2025, 13, 31–58.

[46] Sarrami-Foroushani, P.; Travaglia, J.; Debono, D.; Clay-Williams, R.; Braithwaite, J. Scoping Meta-Review: Introducing a New Methodology: Scoping Meta-Review. Clin. Transl. Sci. 2022, 8, 77–81.

[47] Gates, M.; Gates, A.; Guitard, S.; Pollock, M.; Hartling, L. Guidance for Overviews of Reviews Continues to Accumulate, but Important Challenges Remain: A Scoping Review. Syst. Rev. 2024, 9, 254.

[48] Reis, J.; Melão, N. Digital Transformation: A Meta-Review and Guidelines for Future Research. Heliyon 2023, 9, e12834.

[49] Czakon, W. Metodyka systematycznego przeglądu literatury. Przegląd Organ. 2020, 57–61.

[50] Grubert, E.; Siders, A. Benefits and Applications of Interdisciplinary Digital Tools for Environmental Meta-Reviews and Analyses. Environ. Res. Lett. 2025, 11, 093001.

[51] Jing, Y.; Wang, C.; Chen, Y.; Wang, H.; Yu, T.; Shadiev, R. Bibliometric Mapping Techniques in Educational Technology Research: A Systematic Literature Review. Educ. Inf. Technol. 2022, 29, 9283–9931.

[52] Hennessy, E.A.; Johnson, B.T.; Keenan, C. Best Practice Guidelines and Essential Methodological Steps to Conduct Rigorous and Systematic Meta-Reviews. Appl. Psychol. Health Well-Being 2023, 11, 353–381.

[53] Diagne, M.; David, M.; Lauret, P.; Boland, J.; Schmutz, N. Review of Solar Irradiance Forecasting Methods and a Proposition for Small-Scale Insular Grids. Renew. Sustain. Energy Rev. 2021, 27, 65–76.

[54] Yesilbudak, M.; Colak, M.; Bayindir, R. A Review of Data Mining and Solar Power Prediction. In Proceedings of the 2024 IEEE International Conference on Renewable Energy Research and Applications (ICRERA), Birmingham, UK, 20–23 November 2024; pp. 1117–1121.

[55] Ren, Y.; Suganthan, P.N.; Srikanth, N. Ensemble Methods for Wind and Solar Power Forecasting—A State-of-the-Art Review. Renew. Sustain. Energy Rev. 2022, 50, 82–91.

[56] Hemavathi, U.; Medona, A.C.V.; Dhilip Kumar, V.; Raja Sekar, R. Review for the Solar Radiation Forecasting Methods Based on Machine Learning Approaches. J. Phys. Conf. Ser. 2023, 1964, 042065.

[57] Panamtash, H.; Mahdavi, S.; Zhou, Q. Probabilistic Solar Power Forecasting: A Review and Comparison. In Proceedings of the 2021 52nd North American Power Symposium (NAPS), Tempe, AZ, USA, 11–13 April 2021; pp. 1–6.

[58] Huang, C.-L.; Wu, Y.-K.; Li, Y.-Y. Deterministic and Probabilistic Solar Power Forecasts: A Review on Forecasting Models. In Proceedings of the 2025 7th International Conference on Applied System Innovation (ICASI), Chiayi, Taiwan, 24–25 September 2025; pp. 15–18.

[59] Thaker, J.; Höller, R. A Comparative Study of Time Series Forecasting of Solar Energy Based on Irradiance Classification. Energies 2022, 15, 2837.

[60] André, M.; Dabo-Niang, S.; Soubdhan, T.; Ould-Baba, H. Predictive Spatio-Temporal Model for Spatially Sparse Global Solar Radiation Data. Energy 2023, 111, 599–608.

[61] Ayvazoğluyüksel, Ö.; Filik, Ü.B. Estimation Methods of Global Solar Radiation, Cell Temperature and Solar Power Forecasting: A Review and Case Study in Eskişehir. Renew. Sustain. Energy Rev. 2020, 91, 639–653.

[62] Rajasekaran, M.; Selvakumar, A.I.; Rajasekaran, E. Review on Mathematical Models for the Prediction of Solar Radiation. Indones. J. Electr. Eng. Comput. Sci. 2024, 15, 56.

[63] Yadav, A.K.; Chandel, S.S. Solar Radiation Prediction Using Artificial Neural Network Techniques: A Review. Renew. Sustain. Energy Rev. 2021, 33, 772–781.

[64] Mohanty, S.; Patra, P.K.; Mohanty, A.; Harrag, A.; Rezk, H. Adaptive Neuro-Fuzzy Approach for Solar Radiation Forecasting in Cyclone Ravaged Indian Cities: A Review. Front. Energy Res. 2025, 10, 828097.

[65] Bamisile, O.; Cai, D.; Oluwasanmi, A.; Ejiyi, C.; Ukwuoma, C.C.; Ojo, O.; Mukhtar, M.; Huang, Q. Comprehensive Assessment, Review, and Comparison of AI Models for Solar Irradiance Prediction Based on Different Time/Estimation Intervals. Sci. Rep. 2022, 12, 9644.

[66] Guermoui, M.; Gairaa, K.; Ferkous, K.; Santos, D.S.D.O.; Arrif, T.; Belaid, A. Potential Assessment of the TVF-EMD Algorithm in Forecasting Hourly Global Solar Radiation: Review and Case Studies. J. Clean. Prod. 2024, 385, 135680.

[67] Guermoui, M.; Melgani, F.; Danilo, C. Multi-Step Ahead Forecasting of Daily Global and Direct Solar Radiation: A Review and Case Study of Ghardaia Region. J. Clean. Prod. 2021, 201, 716–734.

[68] Li, B.; Zhang, J. A Review on the Integration of Probabilistic Solar Forecasting in Power Systems. Sol. Energy 2023, 210, 68–86.

[69] Zwane, N.; Tazvinga, H.; Botai, C.; Murambadoro, M.; Botai, J.; De Wit, J.; Mabasa, B.; Daniel, S.; Mabhaudhi, T. A Bibliometric Analysis of Solar Energy Forecasting Studies in Africa. Energies 2025, 15, 5520.

[70] Yang, D.; Kleissl, J.; Gueymard, C.A.; Pedro, H.T.C.; Coimbra, C.F.M. History and Trends in Solar Irradiance and PV Power Forecasting: A Preliminary Assessment and Review Using Text Mining. Sol. Energy 2022, 168, 60–101.

[71] Kumar, D.S.; Yagli, G.M.; Kashyap, M.; Srinivasan, D. Solar Irradiance Resource and Forecasting: A Comprehensive Review. IET Renew. Power Gener. 2023, 14, 1641–1656.

[72] Kumari, P.; Toshniwal, D. Deep Learning Models for Solar Irradiance Forecasting: A Comprehensive Review. J. Clean. Prod. 2021, 318, 128566.

[73] Mohanty, S.; Patra, P.K.; Sahoo, S.S. Prediction and Application of Solar Radiation with Soft Computing over Traditional and Conventional Approach—A Comprehensive Review. Renew. Sustain. Energy Rev. 2020, 56, 778–796.

[74] Yang, D.; Wang, W.; Gueymard, C.A.; Hong, T.; Kleissl, J.; Huang, J.; Perez, M.J.; Perez, R.; Bright, J.M.; Xia, X.; et al. A Review of Solar Forecasting, Its Dependence on Atmospheric Sciences and Implications for Grid Integration: Towards Carbon Neutrality. Renew. Sustain. Energy Rev. 2024, 161, 112348.

[75] Rahimi, N.; Park, S.; Choi, W.; Oh, B.; Kim, S.; Cho, Y.; Ahn, S.; Chong, C.; Kim, D.; Jin, C.; et al. A Comprehensive Review on Ensemble Solar Power Forecasting Algorithms. J. Electr. Eng. Technol. 2022, 18, 719–733.

[76] Assaf, A.M.; Haron, H.; Abdull Hamed, H.N.; Ghaleb, F.A.; Qasem, S.N.; Albarrak, A.M. A Review on Neural Network Based Models for Short Term Solar Irradiance Forecasting. Appl. Sci. 2025, 13, 8332.

[77] Benavides Cesar, L.; Amaro E Silva, R.; Manso Callejo, M.Á.; Cira, C.-I. Review on Spatio-Temporal Solar Forecasting Methods Driven by In Situ Measurements or Their Combination with Satellite and Numerical Weather Prediction (NWP) Estimates. Energies 2023, 15, 4341.

[78] Rajagukguk, R.A.; Ramadhan, R.A.A.; Lee, H.-J. A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power. Energies 2020, 13, 6623.

[79] Sudharshan, K.; Naveen, C.; Vishnuram, P.; Krishna Rao Kasagani, D.V.S.; Nastasi, B. Systematic Review on Impact of Different Irradiance Forecasting Techniques for Solar Energy Prediction. Energies 2024, 15, 6267.

[80] Tsai, W.-C.; Tu, C.-S.; Hong, C.-M.; Lin, W.-M. A Review of State-of-the-Art and Short-Term Forecasting Models for Solar PV Power Generation. Energies 2021, 16, 5436.

[81] Wu, Y.-K.; Huang, C.-L.; Phan, Q.-T.; Li, Y.-Y. Completed Review of Various Solar Power Forecasting Techniques Considering Different Viewpoints. Energies 2022, 15, 3320.

PROBABILISTIC SOLAR IRRADIANCE PREDICTION USING XGBOOST AND KDE

ORIGINALITY REPORT

12%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.coursehero.com Internet	100 words — 3%
2	Masoabi, Fani Peter. "Classification of Water Potability Using Neuromorphic Spiking Neural Networks.", University of Johannesburg (South Africa), 2024 ProQuest	33 words — 1%
3	C. Guedes Soares, S. Wang. "Innovations in Renewable Energies Offshore", CRC Press, 2024 Publications	26 words — 1%
4	digitalcommons.unl.edu Internet	22 words — 1%
5	www.fastercapital.com Internet	17 words — 1%
6	"Electric Vehicle Charging Infrastructures and its Challenges", Springer Science and Business Media LLC, 2025 Crossref	15 words — < 1%
7	bvrit.ac.in Internet	14 words — < 1%
8	dailyasianage.com Internet	

		12 words — < 1%
9	publication-theses.unistra.fr Internet	12 words — < 1%
10	www.researchgate.net Internet	12 words — < 1%
11	fastercapital.com Internet	11 words — < 1%
12	tudr.thapar.edu:8080 Internet	11 words — < 1%
13	www.mdpi.com Internet	11 words — < 1%
14	T. Vasudeva Reddy, K. Madhava Rao. "Recent Trends in VLSI and Semiconductor Packaging", CRC Press, 2025 Publications	10 words — < 1%
15	bristol.ac.uk Internet	10 words — < 1%
16	digitalcommons.usu.edu Internet	10 words — < 1%
17	en.vietnamplus.vn Internet	10 words — < 1%
18	Theodoros Konstantinou, Nikolaos Savvopoulos, Nikos Hatziargyriou. "Scenario Based Probabilistic Energy Demand Forecasting using Autoencoders and Gaussian Mixture Models", 2021 International Conference on Smart Energy Systems and Technologies (SEST), 2021	9 words — < 1%

Crossref

19 Yinghao Chu, Carlos F.M. Coimbra. "Short-term probabilistic forecasts for Direct Normal Irradiance", Renewable Energy, 2017 8 words — < 1%
Crossref

20 businessdocbox.com 8 words — < 1%
Internet

21 Hui Liu, Zhu Duan, Chao Chen, Haiping Wu. "A novel two-stage deep learning wind speed forecasting method with adaptive multiple error corrections and bivariate Dirichlet process mixture model", Energy Conversion and Management, 2019 6 words — < 1%
Crossref

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

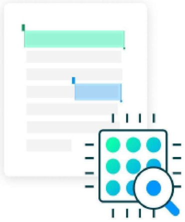
Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.



False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.