الجمهورية الجزائرية الديمقراطيــة الشعبيــة

People's Democratic Republic of Algeria

وزارة التـعليــم العالـي والبحــث العلمـــي

Ministry of Higher Education and Scientific Research

جامعة عين تموشنت بلحاج بوشعيب

University –Ain Temouchent- Belhadj Bouchaib

Faculty of Science and Technology

Department Mechanical Engineering

End of Studies Project

To obtain the Master's degree in: Mechanical Engineering - Energy

Field: Science and Technology

Sector: Mechanical Engineering

Specialty: Energy

Theme:

Short-Term Prediction of the Photovoltaic Production of a Solar Farm Using Artificial Intelligence and Machine Learning

Presented By:

1) Mr. GHALEM Mohammed

2) Mr. HADDOU Zakarya Mohammed Nadir

Before the jury composed of:

Dr. ELAHMAR Kadi	MCB	UAT.B.B (Ain Temouchent)	President
Dr. OUARI Mohamed Amine	MCB	UAT.B.B (Ain Temouchent)	Examiner
Dr. DORBANE Abdelhakim	MCA	UAT.B.B (Ain Temouchent)	Supervisor

Academic Year 2025/2026

Dedication

*With deep gratitude and sincere emotion, I dedicate this work, first and foremost, to the memory of my beloved **mother**. Although you are no longer by my side, your love, kindness, and sacrifices continue to inspire me every day. Your guidance and unwavering belief in me remain a source of strength, and I hope this achievement makes you proud. May Allah have mercy on you.*

*To my dear **father**, thank you for your endless support, encouragement, and sacrifices throughout my academic journey. Your patience, wisdom, and confidence in me have been invaluable, and this accomplishment is as much yours as it is mine.*

*To my beloved **brother and sister**, thank you for your constant encouragement, understanding, and for always believing in me. Your support has given me the motivation to persevere through every challenge.*

*I would also like to express my deepest gratitude to my supervisor, **Dr. Dorbane Abdelhakim**, for his valuable guidance, insightful advice, patience, and continuous support throughout the realization of this project. His expertise and encouragement have greatly contributed to the completion of this work.*

*To all my **friends**, thank you for your friendship, motivation, and the memorable moments we shared during these years of study. Your support has made this journey more meaningful and enjoyable.*

To everyone who has contributed, directly or indirectly, to my education and personal growth, I offer my sincere appreciation.

This work is dedicated to all of you with profound gratitude and respect.

GHALEM Mohammed

Dedication

With deep gratitude and sincere emotion, I dedicate this work to the people who have been my constant source of strength, patience, and love throughout my academic journey.

To my respected supervisor, ***Dorbane Abdelhakim****, whose guidance, wisdom, and continuous support illuminated my path and made this achievement possible. I am truly grateful for his encouragement and valuable advice at every stage of this work.*

To my dearest mother, ***Fayza****, the heart of my life, whose endless love, prayers, and sacrifices have been my greatest motivation. No words can truly express how much I owe her for everything she has done for me.*

To my father, ***Bouhadjar****, my silent strength and lifelong supporter, who always believed in me and encouraged me to move forward, even in the hardest moments.*

To my beloved sisters, ***Ikrame, Lina, and Tasnime****, for their unconditional love, their smiles, and their presence that gave me comfort and hope throughout this journey.*

And to my dear friends, who stood by me, supported me, and shared both the struggles and the successes of this path.

This accomplishment is not mine alone, but a reflection of all their love, faith, and sacrifices.

HADDOU Zakarya Mohammed Nadir

Acknowledgments

First and foremost, we wish to sincerely thank Allah for granting us the strength and courage to bring this work to completion.

The completion of this thesis was made possible through the assistance of several individuals, to whom we wish to express our deepest gratitude.

We would like to extend our sincere appreciation to our thesis supervisor, ***Dr. Abdelhakim Dorbane****, for his patience, availability, and—above all—his insightful advice, which helped shape our thinking.*

We also wish to thank the professors at the University of Ain Temouchent for providing the tools necessary for our academic success.

We also gratefully acknowledge all the professors who taught us throughout our academic journey.

We would like to express our gratitude to the friends and colleagues who provided moral and intellectual support throughout this process.

We extend special thanks to our dear parents, brothers, and sisters, who enabled us to successfully complete this work through their encouragement, availability, patience, wise counsel, and invaluable support.

Finally, we wish to thank the distinguished colleagues with whom this thesis was produced; their sincere and invaluable assistance was instrumental to its completion.

We thank you,

GHALEM Mohamme

HADDOU Zakarya Mohammed Nadir

Abbreviations

PV	Photovoltaic
LSTM	Long Short-Term Memory
RF	Random Forest
SVR	Support Vector Regression
MW	Megawatt
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
SHAP	Shapley Additive explanations
XGBoost	extreme Gradient Boosting
CatBoost	Categorical Boosting
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
R^2 or R2	Coefficient of Determination

Abstract

This thesis presents a short-term prediction study of the photovoltaic power production of a solar farm using Artificial Intelligence and Machine Learning techniques. Given Algeria's exceptional solar potential and the growing integration of photovoltaic systems into the national grid, accurate forecasting of PV output has become essential for grid stability and energy management. The study employs a data-driven regression approach using meteorological and operational data, including solar irradiance, ambient temperature, relative humidity, wind speed, wind direction, and sea-level pressure, as input features to predict the target variable Power_PACE. Five Machine Learning models were developed and compared: Linear Regression, Random Forest, Gradient Boosting, XGBoost, and CatBoost. After preprocessing and feature engineering in Python, each model was trained on historical data and evaluated on a held-out test set using MAE, RMSE, and R^2 metrics. Results show that Linear Regression achieved the highest coefficient of determination ($R^2 = 0.9777$) with the lowest RMSE (5,953 W), reflecting the dominant linear relationship between solar irradiance and power output in the summer dataset. Random Forest ranked second with the lowest MAE (4,232 W), demonstrating strong robustness and interpretability. SHAP analysis confirmed Solar_Mean as the primary predictor, followed by ambient temperature and hour of day. The findings demonstrate that even classical Machine Learning models can achieve high forecasting accuracy when solar irradiance is the dominant driver, while ensemble methods such as Random Forest offer a more reliable choice for real-world deployment due to their robustness to noisy data and nonlinear behavior.

Keywords: Photovoltaic forecasting, Machine Learning, Short-term prediction, Solar irradiance, Random Forest, XGBoost, SHAP, Algeria

Résumé

Ce mémoire présente une étude de prédiction à court terme de la production photovoltaïque d'une ferme solaire en utilisant des techniques d'Intelligence Artificielle et d'Apprentissage Automatique. Compte tenu du potentiel solaire exceptionnel de l'Algérie et de l'intégration croissante des systèmes photovoltaïques dans le réseau électrique national, la prévision précise de la production PV est devenue indispensable pour assurer la stabilité du réseau et la gestion de l'énergie. L'étude adopte une approche de régression basée sur les données, utilisant des mesures météorologiques et opérationnelles — notamment l'irradiance solaire, la température ambiante, l'humidité relative, la vitesse et la direction du vent, ainsi que la pression au niveau de la mer — comme variables d'entrée pour prédire la variable cible Power_PACE. Cinq modèles d'Apprentissage Automatique ont été développés et comparés : Régression Linéaire, Random Forest, Gradient Boosting, XGBoost et CatBoost. Après prétraitement et ingénierie des caractéristiques en Python, chaque modèle a été entraîné sur des données historiques et évalué sur un ensemble de test indépendant à l'aide des métriques MAE, RMSE et R^2. Les résultats montrent que la Régression Linéaire a obtenu le coefficient de détermination le plus élevé (R^2 = 0,9777) avec la RMSE la plus basse (5 953 W), reflétant la relation linéaire dominante entre l'irradiance solaire et la puissance produite sur le jeu de données estival. Random Forest s'est classé deuxième avec la MAE la plus basse (4 232 W), démontrant une robustesse et une interprétabilité élevées. L'analyse SHAP a confirmé que Solar_Mean est le prédicteur principal, suivi de la température ambiante et de l'heure de la journée. Les résultats démontrent que même des modèles classiques d'Apprentissage Automatique peuvent atteindre une grande précision de prévision lorsque l'irradiance solaire est le facteur dominant, tandis que les méthodes d'ensemble telles que Random Forest offrent un choix plus fiable pour un déploiement en conditions réelles grâce à leur robustesse face aux données bruitées et aux comportements non linéaires.

Mots-clés : Prévision photovoltaïque, Apprentissage Automatique, Prédiction à court terme, Irradiance solaire, Random Forest, XGBoost, SHAP, Algérie.

ملخص

تقدّم هذه الدراسة نموذجاً للتنبؤ قصير المدى بإنتاج الطاقة الكهروضوئية لمحطة طاقة شمسية باستخدام تقنيات الذكاء الاصطناعي والتعلم الآلي. نظراً للإمكانات الشمسية الاستثنائية التي تتمتع بها الجزائر، والتكامل المتزايد للأنظمة الكهروضوئية في شبكة الكهرباء الوطنية، باتت دقة التنبؤ بإنتاج الطاقة الشمسية أمراً بالغ الأهمية لضمان استقرار الشبكة وكفاءة إدارة الطاقة. تعتمد الدراسة على نهج انحدار مبني على البيانات، إذ تُستخدم قياسات الإشعاع الشمسي ودرجة الحرارة المحيطة والرطوبة النسبية وسرعة الرياح واتجاهها والضغط الجوي عند مستوى سطح البحر متغيرات مدخلة للتنبؤ بمتغير الهدف .Power_PACE تم تطوير خمسة نماذج للتعلم الآلي ومقارنتها، وهي: الانحدار الخطي، والغابة العشوائية، والتعزيز التدريجي، وXGBoost، و .CatBoostبعد إجراء عمليات المعالجة المسبقة وهندسة الميزات باستخدام لغةPython ، جرى تدريب كل نموذج على بيانات تاريخية وتقييمه على مجموعة اختبار مستقلة باستخدام مقاييس MAE و RMSEو R^2.أظهرت النتائج أن الانحدار الخطي حقق أعلى معامل تحديد ($R^2 = 0.9777$) مع أدنى قيمة لـ RMSE (5953 واط)، مما يعكس العلاقة الخطية السائدة بين الإشعاع الشمسي والطاقة المنتجة في مجموعة البيانات الصيفية. جاءت الغابة العشوائية في المرتبة الثانية بأدنى قيمة لـ MAE (4232 واط)، مما يدل على متانتها وقدرتها على التفسير. أكد تحليل SHAP أن Solar_Mean هو المتنبئ الرئيسي، يليه درجة الحرارة المحيطة وساعة اليوم. تُظهر النتائج أن نماذج التعلم الآلي الكلاسيكية قادرة على تحقيق دقة تنبؤ عالية عندما يكون الإشعاع الشمسي هو العامل المحرك الأساسي، في حين تمثل الأساليب التجميعية كالغابة العشوائية الخيار الأكثر موثوقية للنشر الفعلي بفضل متانتها في التعامل مع البيانات المشوّشة والسلوك غير الخطي.

الكلمات المفتاحية :التنبؤ الكهروضوئي، التعلم الآلي، التنبؤ قصير المدى، الإشعاع الشمسي، الغابة العشوائية، XGBoost، SHAP، الجزائ

Table of contents

List of figures

List of tables

General Introduction

General Introduction

In recent decades, the world has experienced a continuous increase in energy demand [1] accompanied by growing environmental concerns, particularly related to climate change and greenhouse gas emissions [2]. The excessive use of fossil fuels has significantly contributed to environmental degradation, making the transition toward sustainable and renewable energy sources a global priority. Among the various renewable energy options, solar energy has emerged as one of the most promising alternatives [3] due to its abundance, accessibility, and minimal environmental impact. Technological advancements and the decreasing cost of photovoltaic (PV) systems [4] have accelerated their deployment worldwide, making solar power a key component of the global energy transition.

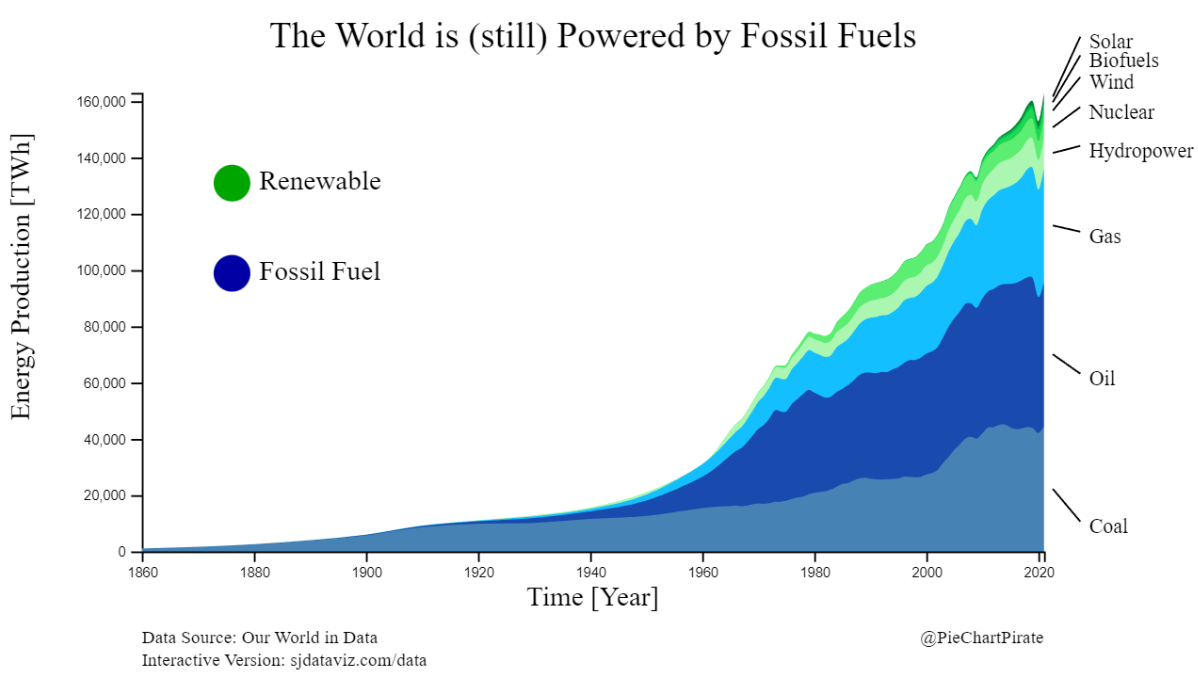


Figure 1: Global evolution of energy consumption and renewable every penetration.[5]

At the national level, Algeria possesses exceptional solar potential [6] due to its geographic location and high solar irradiation, eading to the formation of an electric current This natural advantage positions the country as a strong candidate for large-scale solar energy development. In recent years, Algeria has initiated several programs aimed at integrating renewable energy into its energy mix, reducing dependence on fossil fuels, and ensuring long-term energy security. However, the large-scale integration of photovoltaic systems into the electrical grid introduces significant technical challenges, mainly due to the intermittent and variable nature of solar energy production.

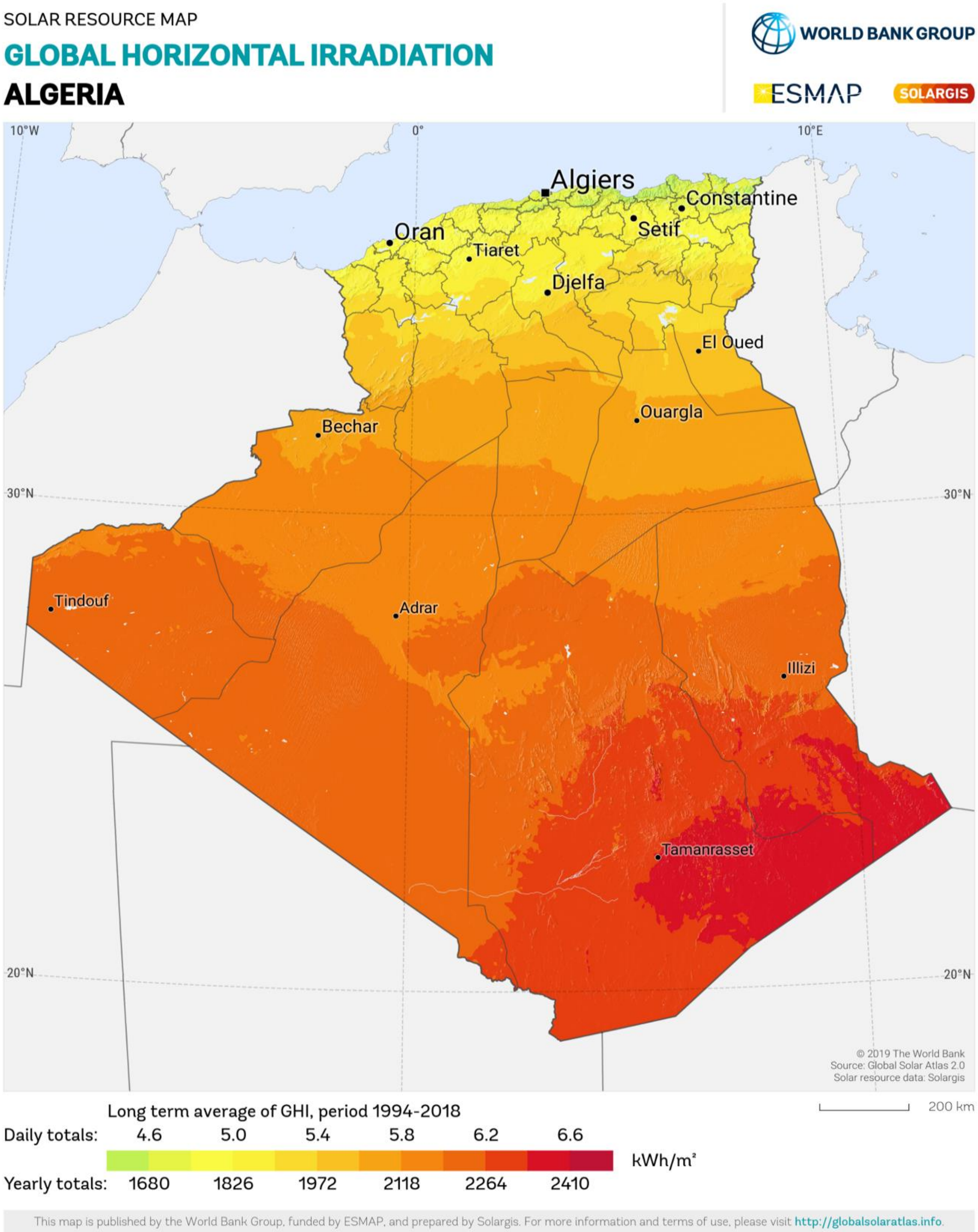


Figure 2: Solar irradiation distribution in Algeria showing high potential in desert regions.[7]

Photovoltaic power generation is highly dependent on meteorological conditions [8] such as solar irradiance, temperature, cloud cover, and wind speed. These parameters can vary rapidly over short periods, leading to fluctuations in energy output. This variability creates difficulties in maintaining the balance between electricity supply and demand, which is essential for grid stability [9]. In this context, short-term forecasting of photovoltaic production, typically ranging from minutes to a few days ahead, becomes a crucial tool for energy management. Accurate forecasting enables grid operators to anticipate fluctuations, optimize the operation of power systems, and reduce reliance on backup conventional energy sources.

Moreover, short-term predictions play an essential role in optimizing energy storage systems, improving the efficiency of smart grids, and enhancing decision-making in electricity markets. By providing reliable estimates of future energy production, forecasting models contribute to better scheduling, dispatching, and load management. Consequently, improving the accuracy of photovoltaic power forecasting is a key factor in facilitating the integration of renewable energy into modern power systems.

Traditional forecasting approaches, including physical models based on atmospheric equations and statistical methods, often face limitations when dealing with the complex and nonlinear relationships between weather variables and photovoltaic output. These methods may lack flexibility and struggle to adapt to rapidly changing conditions. In contrast, Artificial Intelligence (AI) and Machine Learning (ML) techniques [10] have demonstrated significant potential in addressing these challenges. By learning patterns directly from historical data, these methods can capture nonlinear dependencies and improve prediction accuracy without requiring explicit physical modeling.

Machine Learning algorithms such as Linear Regression, Support Vector Machines, Decision Trees, and Random Forests have been widely used for photovoltaic energy forecasting [11], [12]. In addition, Deep Learning models, particularly Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, are well suited for time-series prediction due to their ability to capture temporal dependencies. These advanced techniques allow for the development of robust and adaptive models capable of handling large datasets and uncertain inputs. As a result, AI and ML have become essential tools in renewable energy forecasting and management.

The main objective of this study is to develop an Artificial Intelligence and Machine Learning model capable of accurately predicting short-term photovoltaic power production, either on an hourly or daily basis. This objective is supported by several secondary goals, including the comparison of different Machine Learning and Deep Learning models, the identification of the most influential meteorological variables affecting photovoltaic output, and the evaluation of model performance across different forecasting horizons. Through this approach, the study aims to determine the most efficient and reliable method for short-term prediction.

Despite the advantages offered by AI-based approaches, several challenges remain. Photovoltaic energy production is inherently variable due to changing weather conditions, particularly cloud dynamics and atmospheric fluctuations. These variations introduce uncertainty into the forecasting process, making accurate prediction a complex task. Additionally, the availability and quality of meteorological data can significantly impact model performance. In many cases, datasets may contain missing values, noise, or measurement errors, which can reduce prediction accuracy. Therefore, it is essential to design models that are not only accurate but also robust to data imperfections.

This project focuses exclusively on short-term forecasting of photovoltaic power production. Long-term forecasting is beyond the scope of this study. The work is based on the use of either real data collected from a solar farm or publicly available datasets. The analysis combines historical photovoltaic production data with relevant meteorological variables in order to build predictive models using Artificial Intelligence and Machine Learning techniques. The study emphasizes a data-driven approach, including data preprocessing, model development, training, validation, and performance evaluation.

The thesis is organized into four main chapters. The first chapter provides a general overview of photovoltaic systems and the fundamentals of solar energy, along with the importance of forecasting in renewable energy integration. The second chapter presents a review of the state of the art in photovoltaic power forecasting, covering both traditional methods and modern AI-based approaches. The third chapter describes the methodology adopted in this study, including data collection, preprocessing techniques, model development, and evaluation metrics. Finally, the fourth chapter presents the results obtained, followed by a detailed discussion, comparison of different models, and conclusions with recommendations for future work.

Chapter I: Generalities

Chapter I: Generalities

1. Introduction

Photovoltaic (PV) energy has emerged as one of the most important technologies in the global transition toward sustainable and low-carbon energy systems [13]. This rapid development is driven by its inherent advantages, including the use of an abundant and inexhaustible energy source (solar radiation), the absence of direct greenhouse gas emissions during operation, and continuous reductions in installation and production costs. As a result, PV systems are increasingly deployed at both small-scale (residential) and large-scale (utility or solar farm) levels, making them a central component of modern renewable energy infrastructures.

Despite these advantages, the integration of photovoltaic energy into electrical power systems introduces several technical challenges. The most significant issue is its intermittent and variable nature [14], as PV power generation is directly dependent on environmental conditions such as solar irradiance, cloud cover, and temperature. This variability can lead to uncertainties in power supply, which may affect grid stability, frequency regulation, and the balance between electricity demand and generation. Consequently, reliable operation of power systems with high PV penetration requires a deep understanding of photovoltaic behavior as well as advanced tools capable of anticipating future power production.

In this context, solar power forecasting plays a fundamental role in modern energy management. Accurate prediction of photovoltaic output enables grid operators to optimize dispatch strategies, improve the integration of renewable energy sources, and reduce reliance on conventional backup generation. It also contributes to more efficient energy storage management, cost reduction, and enhanced reliability of smart grid systems.

From a methodological perspective, photovoltaic forecasting has evolved significantly with the development of computational intelligence and data-driven techniques. In particular, Machine Learning approaches have demonstrated strong performance in capturing the complex and nonlinear relationships between environmental variables and PV power output. These methods are increasingly preferred over traditional analytical models due to their adaptability, robustness, and ability to learn directly from historical data without requiring explicit physical equations.

This chapter therefore aims to establish the theoretical foundation required for the study. It introduces the fundamental principles of photovoltaic energy conversion, presents key physical and environmental parameters affecting system performance, defines the concept of solar power forecasting and its classification according to time horizons, and finally discusses the main Machine Learning techniques commonly used for photovoltaic production prediction. Together, these elements provide the necessary background for developing and evaluating advanced forecasting models in the subsequent chapters.

2. Fundamentals of Photovoltaic Energy

2.1 Operating Principles of Photovoltaic Panels

Photovoltaic (PV) panels operate on the principle of direct conversion of solar energy into electrical energy through the photovoltaic effect [15], a physical phenomenon occurring in semiconductor materials. When sunlight reaches the surface of a PV cell, it is composed of discrete energy packets known as photons. If the energy of these photons is greater than the bandgap energy of the semiconductor material (typically silicon), they are absorbed and excite electrons from the valence band to the conduction band. This process generates free charge carriers, leading to the formation of an electric current .

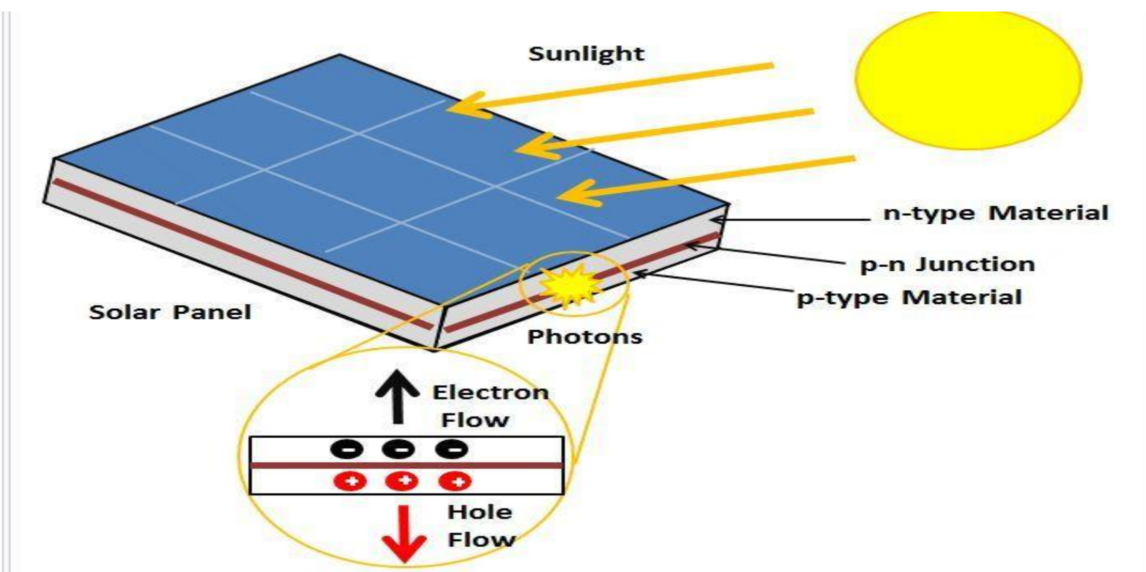


Figure 3: Illustration of the photovoltaic effect in a p–n junction showing electron–hole pair generation and charge separation. [16]

A photovoltaic cell is fundamentally built around a p–n junction, which is formed by joining two differently doped semiconductor regions: a p-type layer (rich in holes) and an n-type layer (rich in electrons). At the interface between these two regions, an internal electric field is established due to the diffusion of charge carriers and the resulting formation of a depletion zone. When the cell is exposed to sunlight, the generated electron–hole pairs within or near this junction are separated by the built-in electric field. Electrons are driven toward the n-side, while holes move toward the p-side, creating a potential difference across the terminals of the cell. When an external electrical load is connected, this potential difference drives a continuous flow of direct current (DC) electricity.

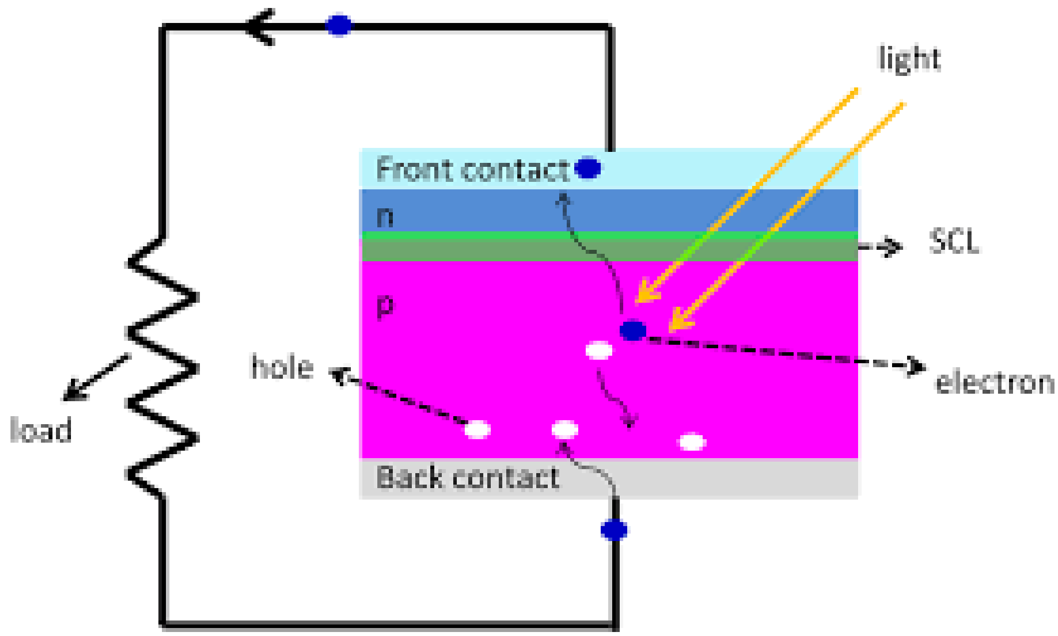


Figure 4: Photovoltaic effect showing electron–hole generation and current flow [17].

In practical applications, a single photovoltaic cell typically produces a relatively small voltage and power output. To achieve usable power levels, multiple cells are electrically interconnected to form a photovoltaic module (panel). Cells connected in series increase the output voltage, while parallel connections increase the current capacity. Furthermore, multiple modules are combined to construct a photovoltaic array, which represents the full-scale system especially in desert regions.

The overall performance and electricity output of a photovoltaic system are influenced by several external and internal factors. The most significant external factor is solar irradiance, which determines the number of incident photons available for conversion into electrical energy. Temperature also plays a critical role [18], as increases in cell temperature typically reduce the efficiency and voltage output of the system due to semiconductor property variations. Additionally, panel orientation and tilt angle affect the amount of solar radiation captured by the surface, while system-level characteristics such as wiring configuration, inverter efficiency, and material properties further influence total energy production .

Understanding these operating principles is essential for accurately modeling photovoltaic systems and forms the foundation for advanced forecasting approaches. In particular, the relationship between environmental conditions and electrical behavior provides the basis for data-driven techniques used in short-term photovoltaic power prediction.

2.2 Key Definitions

To accurately analyze and model photovoltaic (PV) systems, it is essential to define several fundamental physical and performance-related parameters. These variables form the basis of PV system characterization and are widely used in both physical modeling and data-driven forecasting approaches.

- **Solar Irradiance (G):**

Solar irradiance represents the instantaneous power of solar electromagnetic radiation incident on a given surface per unit area, typically measured in watts per square meter (W/m²). It is the most critical external input variable influencing photovoltaic energy production, as it directly determines the number of photons available for conversion into electrical energy. Variations in irradiance due to weather conditions, cloud cover, and time of day have a direct and immediate impact on PV output. Consequently, it is a primary input feature in most photovoltaic forecasting models.

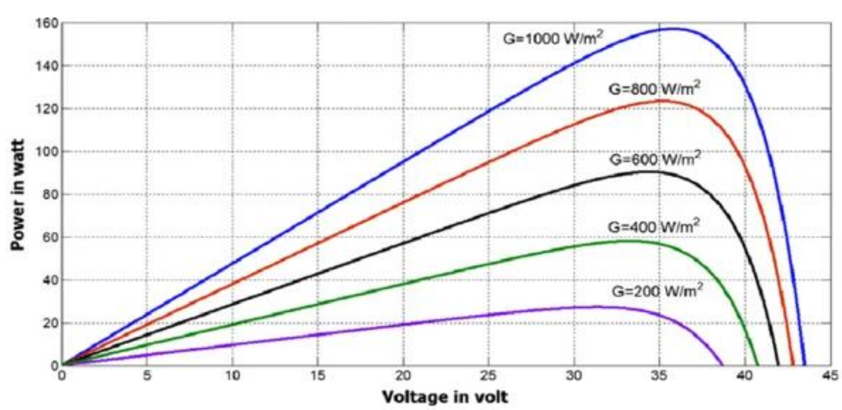


Figure 5: Relationship between solar irradiance and photovoltaic power output [19].

- **Temperature (T):**

Temperature refers to the operating temperature of the photovoltaic module or cell during energy conversion. It plays a significant role in determining the electrical efficiency of the PV system. As temperature increases, the semiconductor properties of the PV material are affected, leading to a reduction in open-circuit voltage and overall efficiency. This degradation occurs due to increased carrier recombination and internal resistive losses within the cell. Therefore, higher operating temperatures generally result in lower power output, even under constant irradiance conditions .

- **Efficiency (η):**

The efficiency of a photovoltaic system is defined as the ratio between the useful electrical power output and the solar power input received by the panel surface. Mathematically, it expresses how effectively the system converts incident solar energy into usable electrical energy. Efficiency is influenced by multiple factors, including irradiance level, temperature, material properties, and system design. In practical applications, PV efficiency is not constant and tends to vary with environmental conditions, especially temperature and irradiance levels.

- **Instantaneous Power (P):**

Instantaneous power refers to the electrical power generated by a photovoltaic system at a specific moment in time. It is calculated as the product of instantaneous voltage and current output of the PV system and is typically expressed in watts (W) or kilowatts (kW). This parameter is dynamic in nature and fluctuates continuously depending on environmental conditions such as irradiance variability and temperature changes. In forecasting applications, instantaneous power serves as the primary target variable to be predicted.

These fundamental parameters—solar irradiance, temperature, efficiency, and instantaneous power—are essential for understanding, modeling, and predicting photovoltaic energy production. They form the core input and output variables in both physical and Machine Learning-based forecasting models, enabling accurate representation of PV system behavior under varying environmental conditions.

2.3 I–V Curve Behavior and Environmental Influences

The electrical behavior of a photovoltaic (PV) cell is fundamentally described by its current–voltage (I–V) characteristic curve [20], which represents the relationship between the output current and voltage under specific operating conditions such as solar irradiance and temperature . This curve is a key tool for analyzing the performance of PV devices, as it provides detailed insight into their power generation capability and efficiency.

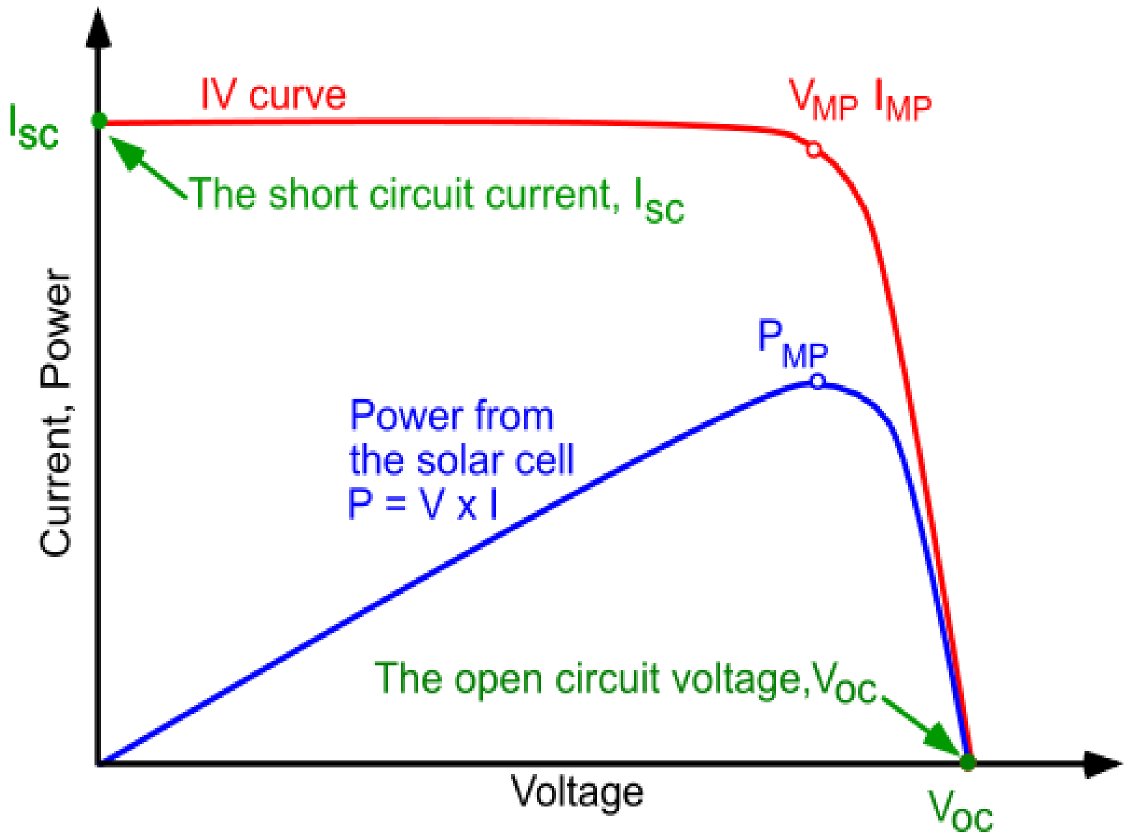


Figure 6: Current–voltage (I–V) and power–voltage (P–V) characteristics of a photovoltaic cell [21].

The I–V curve is inherently nonlinear due to the semiconductor properties of the PV cell and is typically derived from equivalent circuit models, such as the single-diode model. Under standard test conditions (STC), the curve has a well-defined shape; however, in real-world applications, it continuously shifts depending on environmental conditions, making its analysis essential for accurate system modeling and forecasting.

Several characteristic operating points can be identified on the I–V curve:

- **Short-Circuit Current (Isc):**

This is the maximum current produced by the PV cell when the output terminals are short-circuited (voltage equals zero). It is directly proportional to solar irradiance and is only slightly affected by temperature variations.

- **Open-Circuit Voltage (Voc):**

This represents the maximum voltage across the PV cell when no current flows (open-circuit condition). The open-circuit voltage is strongly influenced by temperature, decreasing as temperature increases due to changes in the semiconductor bandgap.

- **Maximum Power Point (MPP):**

The MPP corresponds to the optimal operating point at which the product of current and voltage P=V×IP = V \times IP=V×I is maximized. This point defines the maximum power that the PV cell can deliver under given conditions. In practical systems, Maximum Power Point Tracking (MPPT) algorithms are used to continuously operate the PV system near this point.

Environmental conditions have a significant impact on the position and shape of the I–V curve, thereby affecting the performance of photovoltaic systems:

• Effect of Irradiance:

Solar irradiance is the dominant factor influencing PV output. As irradiance increases, the number of photons striking the PV cell rises, generating more charge carriers and thus increasing the photocurrent. This results in an upward shift of the I–V curve, with a substantial increase in the short-circuit current (Isc) and overall power output. The open-circuit voltage (Voc) experiences only a slight increase with irradiance [22].

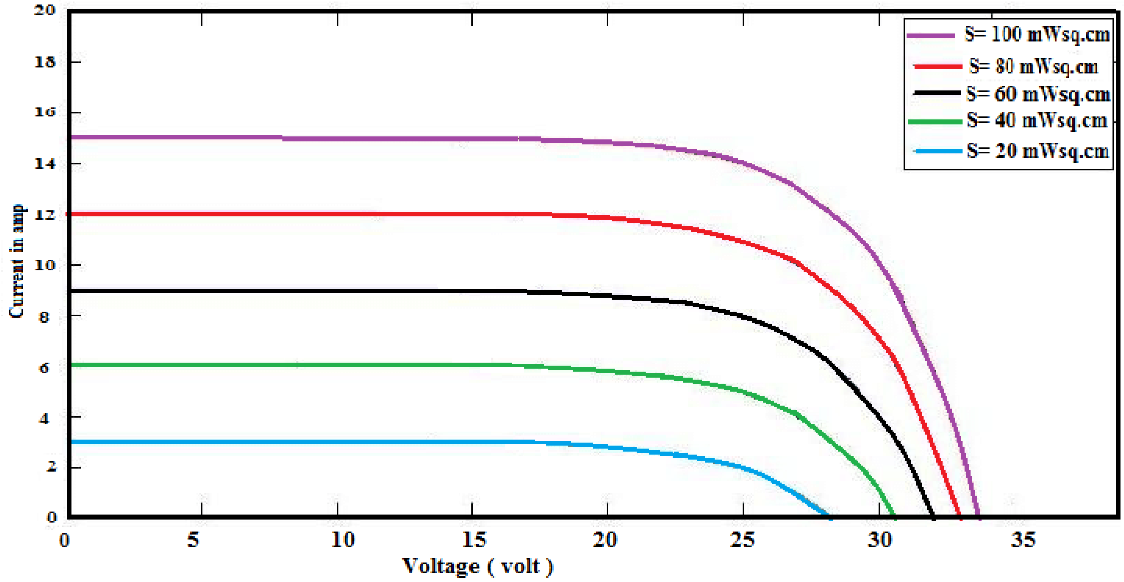


Figure 7: Effect of solar irradiance on the I–V characteristics of a photovoltaic cell [23].

• Effect of Temperature:

Temperature primarily affects the voltage characteristics of the PV cell. As temperature increases, the intrinsic carrier concentration in the semiconductor rises, leading to a reduction in the open-circuit voltage (Voc). Although the current may increase slightly, this effect is relatively minor compared to the voltage drop, Consequently, high operating temperatures negatively impact PV efficiency and overall system performance [24].

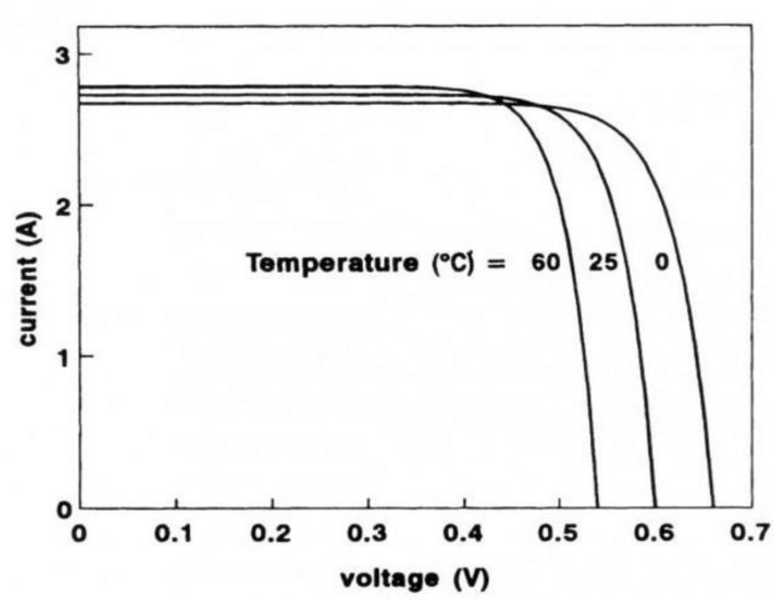


Figure 8: Effect of temperature on the I–V characteristics of a photovoltaic cell [25].

• Effect of Shading:

Partial shading introduces significant distortions in the I–V curve. When one or more cells in a module receive less irradiance, the current generated by those cells decreases, which can limit the current of the entire string. This leads to multiple inflection points and the appearance of several local maxima in the corresponding power–voltage (P–V) curve. Such conditions complicate the operation of MPPT algorithms and may cause substantial power losses. In addition, shading can create localized heating (hotspots), potentially damaging the PV cells over time.

Understanding the influence of these environmental factors on the I–V curve is essential for accurate modeling and prediction of photovoltaic system behavior. In particular, incorporating irradiance, temperature, and shading effects into forecasting models—especially those based on Machine Learning—enables a more realistic representation of PV performance under varying operating conditions. This, in turn, improves the accuracy and reliability of short-term photovoltaic power predictions.

3. Solar Power Forecasting Concepts

Solar power forecasting refers to the process of estimating the future energy production of photovoltaic (PV) systems by using historical generation data in combination with relevant environmental and meteorological variables. Due to the intermittent and highly variable nature of solar energy, accurate forecasting has become essential for ensuring the reliable integration of PV systems into modern power grids.

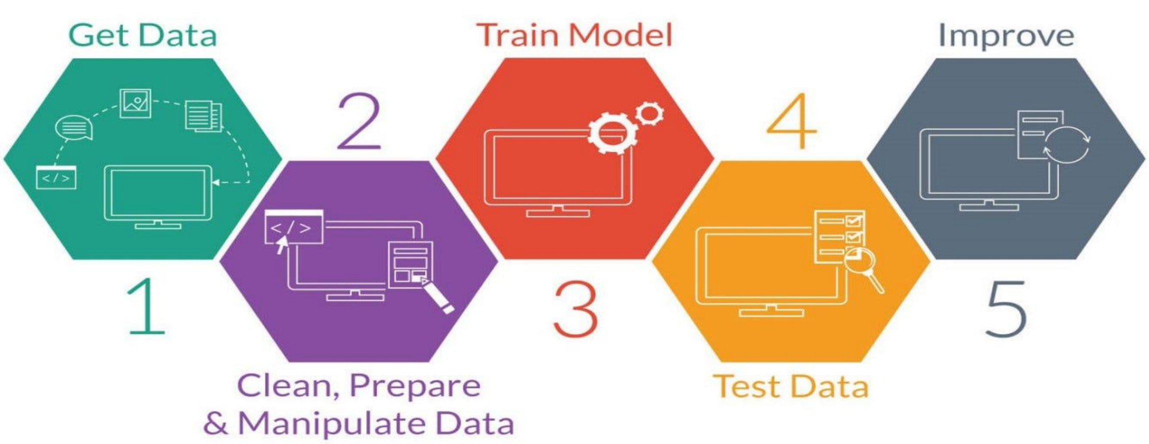


Figure 9: General workflow of photovoltaic power forecasting using Machine Learning techniques [26].

Forecasting approaches are commonly classified [27] according to the prediction time horizon, as each horizon corresponds to different operational and planning requirements :

- **Very short-term forecasting:**

This type of forecasting covers time intervals ranging from a few seconds to several minutes ahead. It is primarily used for real-time control and monitoring of PV systems, allowing operators to respond quickly to rapid fluctuations in solar irradiance caused by transient phenomena such as cloud movement.

- **Short-term forecasting:**

Short-term forecasting typically spans from one hour up to several days ahead. It is the most critical horizon for practical applications, including day-ahead scheduling, load balancing, and optimal dispatch of power generation resources. Accurate short-term predictions enable grid operators to anticipate variations in solar production and maintain system stability.

- **Long-term forecasting:**

This category includes predictions over weeks, months, or even years. It is mainly used for strategic planning, system design, investment decisions, and evaluation of renewable energy potential at specific locations.

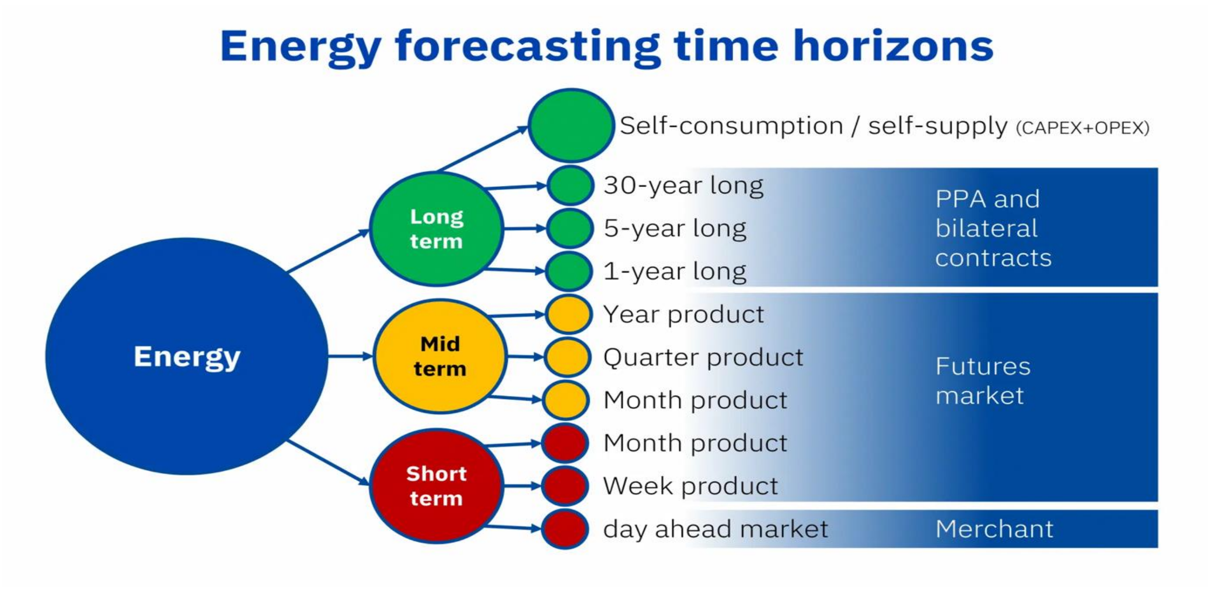


Figure 10: Classification of solar power forecasting based on prediction time horizons [28].

The present work focuses on short-term forecasting, as it plays a crucial role in grid operation and energy management. Reliable short-term predictions help reduce uncertainty in power generation, improve the coordination between supply and demand, and facilitate the integration of renewable energy sources into the electrical network.

Forecasting methodologies can be broadly divided into three main categories:

- **Physical models:**

These models are based on the physical principles governing solar radiation and photovoltaic conversion. They use inputs such as numerical weather prediction (NWP) data, solar position, and PV system characteristics (e.g., panel orientation, efficiency) to simulate power output. While they provide a physically consistent framework, they often require detailed system information and significant computational resources.

- **Statistical models:**

Statistical approaches rely on historical data to identify patterns and correlations between input variables and PV output. Techniques such as time series analysis, regression models, and Machine Learning algorithms fall into this category. These models are generally easier to implement and can achieve high accuracy when sufficient and high-quality data are available.

- **Hybrid models:**

Hybrid approaches combine physical and statistical methods [29] to leverage the advantages of both. For example, a physical model may be used to estimate solar irradiance, while a Machine Learning model refines the prediction of PV power output. Such combined methods often yield improved forecasting performance and robustness .

The effectiveness of solar power forecasting models strongly depends on the selection and quality of input variables. The most important parameter is solar irradiance, which directly determines the energy available for conversion into electricity. Other influential variables include ambient temperature, which affects the efficiency of PV modules; humidity and cloud cover, which influence atmospheric transparency; and wind speed, which can impact panel cooling and thus efficiency. Additionally, temporal features such as time of day and season are often incorporated to capture periodic patterns in solar production. High-resolution and accurate datasets are therefore essential to ensure reliable model performance .

Accurate solar forecasting plays a fundamental role in modern energy systems. It contributes to maintaining grid stability by reducing uncertainty and enabling better balancing between generation and consumption. It also supports the optimization of energy storage systems, allowing operators to anticipate surplus or deficit periods and manage storage accordingly. Furthermore, improved forecasting helps reduce operational costs by minimizing reliance on backup generation and enhancing participation in electricity markets. Finally, it is a key enabler of smart grid technologies, where advanced forecasting is integrated with real-time monitoring and control systems to improve efficiency, reliability, and sustainability of energy management .

4. Machine Learning Approaches for PV Forecasting

Machine Learning (ML) techniques have gained significant attention [30] in recent years for photovoltaic (PV) power forecasting due to their ability to learn complex patterns directly from data. Unlike traditional physical or empirical models, which rely on predefined equations and assumptions, ML approaches are data-driven and capable of capturing the nonlinear and stochastic relationships between input variables and solar energy production.

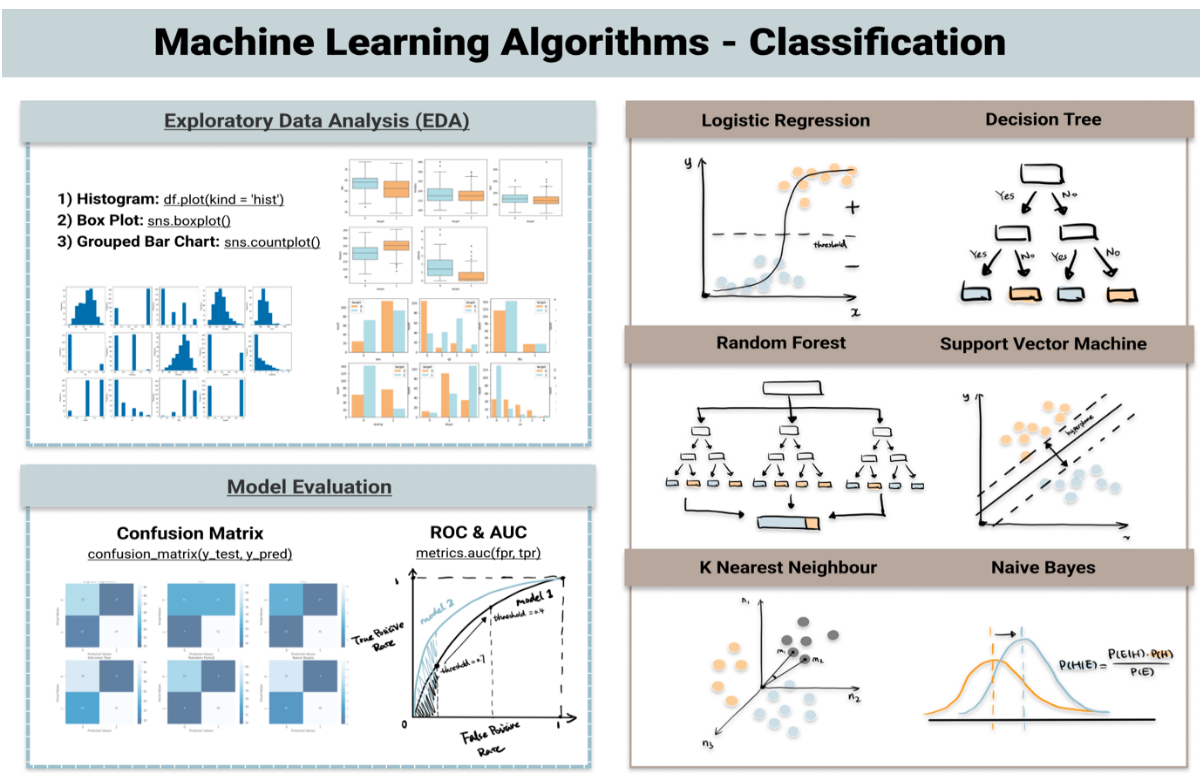


Figure 11: Overview of Machine Learning models used for photovoltaic power forecasting [31].

Photovoltaic power generation is inherently variable and influenced by multiple interacting factors, including solar irradiance, ambient temperature, cloud cover, humidity, and temporal variables such as time of day and season. These relationships are often highly nonlinear and difficult to model using conventional analytical methods. Machine Learning algorithms address this challenge by automatically identifying hidden patterns and dependencies within historical data, leading to more accurate and reliable predictions.In the context of PV forecasting, ML models are typically trained using historical datasets that include both meteorological variables and corresponding power output measurements. Once trained, these models can predict future PV production over short-term horizons, such as minutes to hours ahead, which is critical for real-time grid operation and energy management.Machine Learning approaches used in PV forecasting can be broadly categorized into several groups. Supervised learning models, such as Linear Regression, Support Vector Regression (SVR), Random Forest, and Gradient Boosting, learn a mapping between input features and output power based on labeled training data. These models are widely used due to their strong predictive performance and relatively straightforward implementation.In addition, Artificial Neural Networks (ANNs) and advanced deep learning architectures, such as Long Short-Term Memory (LSTM) networks, are particularly well-suited for time-series forecasting problems. LSTM models, in particular, are capable of capturing temporal dependencies and sequential patterns in PV data, making them highly effective for short-term prediction tasks where past observations influence future outputs.

Another important category is hybrid approaches, which combine multiple Machine Learning models or integrate ML with physical models to enhance prediction accuracy. For example, a physical model may first estimate solar irradiance, and an ML model may then refine the prediction of power output. These hybrid methods leverage the strengths of different techniques to achieve improved performance.

The application of Machine Learning in PV forecasting offers several advantages, including high predictive accuracy, adaptability to changing environmental conditions, and the ability to process large and complex datasets. However, the performance of these models depends heavily on the quality and quantity of the input data, as well as proper model selection and parameter tuning.

Overall, Machine Learning approaches represent a powerful and flexible solution for short-term photovoltaic forecasting. Their ability to handle nonlinearities, incorporate multiple influencing factors, and continuously improve with new data makes them essential tools for enhancing the integration of solar energy into modern power systems.

4.1 Random Forest

Random Forest is a widely used ensemble learning algorithm [32] that combines the predictions of multiple decision trees to improve overall model performance. It is based on the principle of bagging (Bootstrap Aggregating), where each tree is trained on a randomly selected subset of the training data, with replacement. Additionally, at each split within a tree, a random subset of input features is considered, which introduces further diversity among the trees. The final prediction is obtained by averaging the outputs of all individual trees in the case of regression problems, such as photovoltaic (PV) power forecasting .

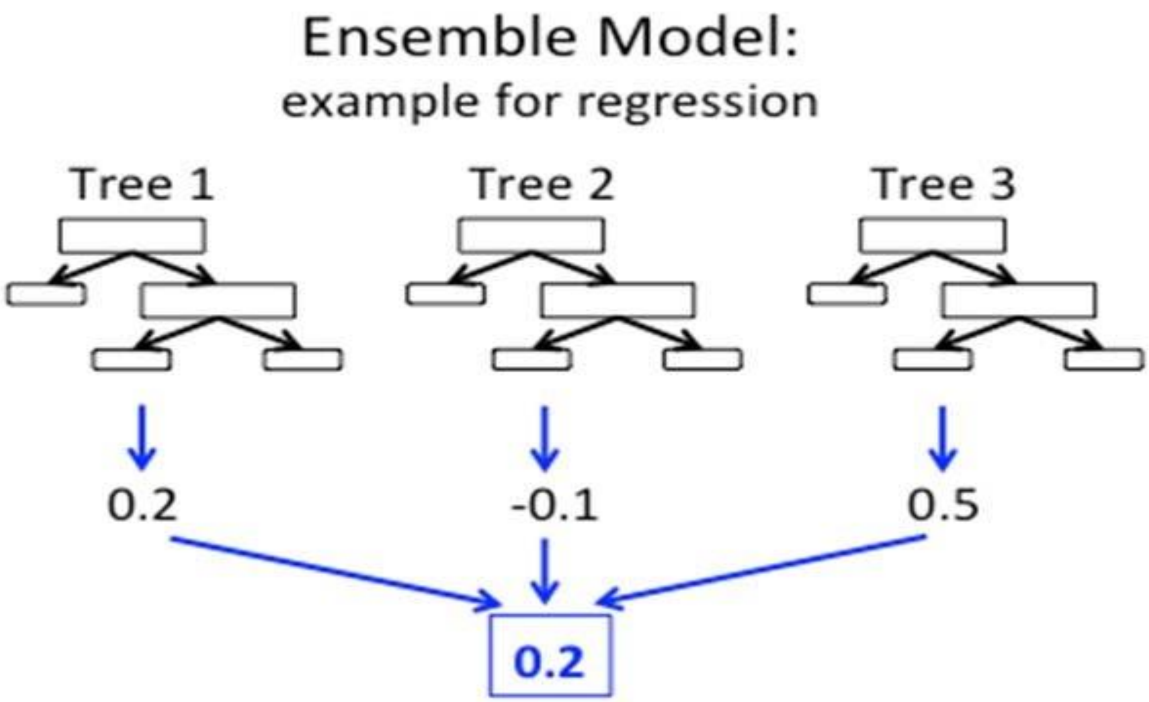


Figure 12: Architecture of the Random Forest algorithm for regression [33].

This ensemble approach significantly enhances predictive performance compared to a single decision tree by reducing variance and improving generalization. In the context of PV forecasting, Random Forest is particularly effective due to its ability to model complex and nonlinear relationships between environmental variables (e.g., solar irradiance, temperature, humidity) and power output, without requiring explicit assumptions about the underlying physical processes.

Advantages:

• High accuracy and robustness:

Random Forest typically achieves high prediction accuracy due to the aggregation of multiple decision trees. By combining diverse models, it reduces the impact of individual errors and provides stable predictions, even in the presence of noisy or fluctuating data commonly found in photovoltaic systems.

• Resistance to overfitting:

Unlike a single decision tree, which can easily overfit the training data, Random Forest mitigates this issue through averaging and randomization. The use of multiple trees trained on different data subsets ensures that the model generalizes well to unseen data, making it highly reliable for short-term forecasting.

• Ability to estimate feature importance:

One of the key strengths of Random Forest is its capacity to evaluate the relative importance of input variables. This allows researchers to identify which factors—such as solar irradiance, temperature, or time-related features—have the greatest influence on PV power output, providing both predictive and interpretative value.

• Capability to model nonlinear relationships:

Random Forest can naturally capture complex and nonlinear interactions between variables without requiring explicit mathematical modeling. This is particularly beneficial in photovoltaic applications, where the relationship between environmental conditions and power generation is inherently nonlinear.

• Robustness to noise and missing data:

The algorithm performs well even when the dataset contains noise, outliers, or missing values. Its ensemble nature ensures that the effect of corrupted or incomplete data points is minimized.

• Minimal data preprocessing requirements:

Random Forest does not require extensive data normalization or scaling, unlike many other Machine Learning models. This simplifies the preprocessing stage and makes the method more practical for real-world applications.

• Scalability and efficiency:

The algorithm can handle large datasets with high dimensionality efficiently. Moreover, since individual trees can be trained independently, Random Forest is well-suited for parallel computation, reducing training time.

• Stability and consistency:

Due to its random sampling and averaging mechanism, Random Forest produces consistent results across different runs, which is important for reproducibility in scientific studies.

• Versatility:

Although used here for regression, Random Forest can also be applied to classification problems, making it a flexible tool within broader energy system analysis frameworks.

Random Forest represents a powerful and reliable Machine Learning technique for short-term photovoltaic power forecasting. Its combination of accuracy, robustness, interpretability, and ease of implementation makes it one of the most suitable models for capturing the complex dynamics of solar energy production.

4.2 Gradient Boosting

Gradient Boosting is an advanced ensemble Machine Learning technique [34] that builds predictive models in a sequential manner. Unlike bagging methods, where models are trained independently, Gradient Boosting constructs each new model to specifically correct the errors made by the previous ones. This is achieved by minimizing a predefined loss function using gradient descent in function space, resulting in a progressively improved model with high predictive capability .

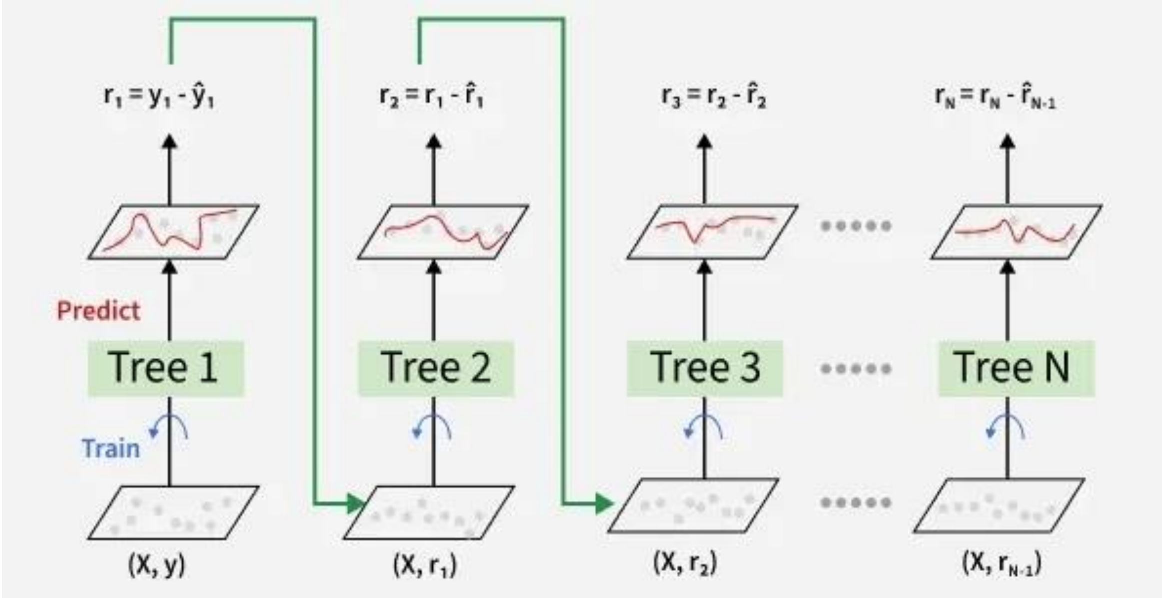


Figure 13: Sequential learning process in Gradient Boosting algorithm [35].

In photovoltaic (PV) power forecasting, Gradient Boosting is particularly effective due to its ability to model complex, nonlinear relationships between meteorological variables (such as irradiance, temperature, and cloud cover) and power output. Variants such as Gradient Boosting Machines (GBM), XGBoost, and LightGBM are widely used in energy forecasting applications because of their efficiency and strong performance.

Advantages:

• High predictive accuracy:

Gradient Boosting is known for delivering state-of-the-art performance in many regression problems, including PV forecasting. By iteratively reducing residual errors, the model achieves high precision and low forecasting error compared to many traditional methods.

• Effective for complex datasets:

This method excels in handling datasets with complex structures, nonlinear relationships, and interactions between variables. In photovoltaic systems, where multiple environmental factors interact in a nonlinear way, Gradient Boosting can capture these dependencies effectively.

• Flexible and powerful:

Gradient Boosting supports various loss functions and can be adapted to different types of problems. Its flexibility allows it to be tailored for specific forecasting objectives, such as minimizing absolute error or squared error.

• Ability to model nonlinear interactions:

The ensemble of decision trees used in Gradient Boosting naturally captures nonlinearities and feature interactions without requiring explicit feature engineering, which is particularly useful in modeling PV system behavior.

• Robustness to outliers:

With appropriate choice of loss functions (e.g., Huber loss), Gradient Boosting can be made robust to outliers and extreme values often present in meteorological data.

• Feature importance evaluation:

Gradient Boosting models provide insights into the relative importance of input variables, helping identify the most influential factors affecting PV production, such as irradiance or temperature.

• Handles mixed data types:

It can process both numerical and categorical variables efficiently, making it suitable for diverse datasets that may include time indices, weather conditions, and operational parameters.

• Regularization capabilities:

Techniques such as learning rate control, tree depth limitation, and subsampling help prevent overfitting, ensuring good generalization performance on unseen data.

• Scalability and efficiency (advanced implementations):

Modern implementations like XGBoost and LightGBM are optimized for speed and memory usage, enabling efficient training even on large-scale datasets.

• Strong performance with relatively limited data preprocessing:

Unlike some other Machine Learning models, Gradient Boosting often requires less extensive data normalization or transformation, simplifying the modeling pipeline.

Gradient Boosting represents a highly effective and versatile approach for short-term photovoltaic power forecasting. Its ability to iteratively improve predictions, capture complex nonlinear relationships, and maintain high accuracy makes it one of the most powerful tools in modern data-driven energy forecasting applications.

4.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised learning algorithms widely used for both classification and regression tasks. In the context of photovoltaic (PV) power forecasting, the regression variant known as Support Vector Regression (SVR) is commonly employed due to its strong predictive capabilities in handling nonlinear relationships between input variables and output power [36].

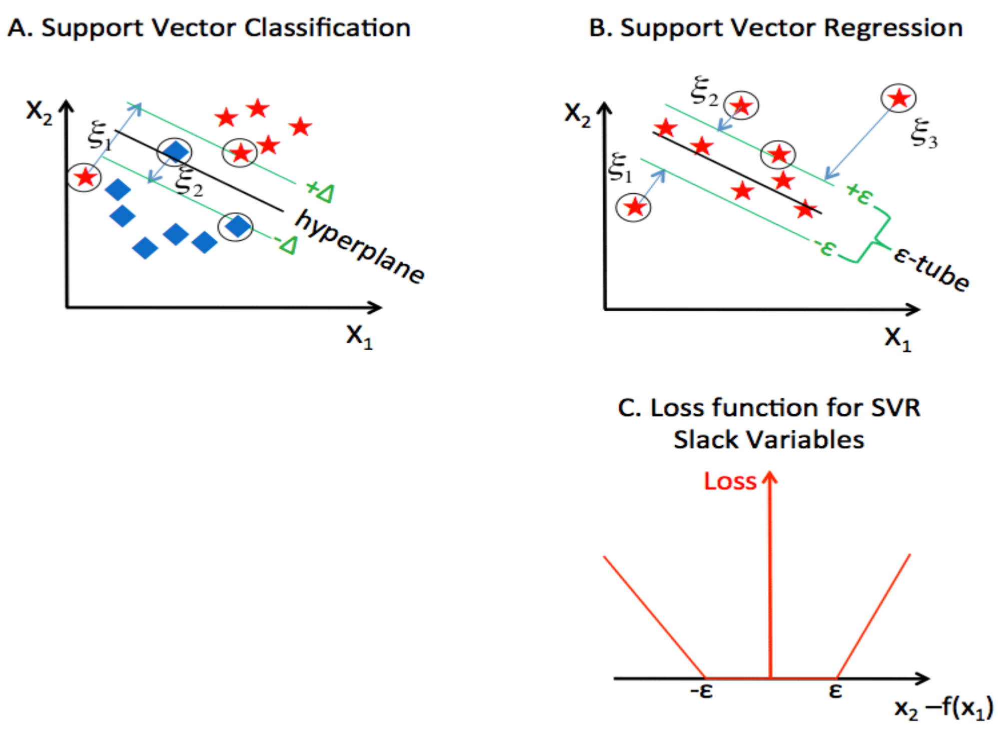


Figure 14: Conceptual representation of Support Vector Regression (SVR) with ε-insensitive margin [37].

SVR is based on the principle of structural risk minimization, which aims to find a function that approximates the target data with a predefined level of accuracy while maintaining model simplicity.

By using kernel functions (such as linear, polynomial, or radial basis function), SVR can map input data into higher-dimensional feature spaces where complex patterns become more easily separable. This makes it particularly suitable for modeling the nonlinear and stochastic behavior of solar power generation influenced by meteorological variables.

Advantages:

• Good performance with small datasets:

Unlike many data-driven models that require large volumes of data, SVR can achieve high prediction accuracy even with relatively limited datasets. This is particularly advantageous in photovoltaic applications where historical data may be scarce or incomplete, especially for newly installed solar plants.

• Strong generalization capability:

SVR is designed to minimize overfitting by controlling model complexity through regularization parameters. This results in strong generalization performance when applied to unseen data, which is essential for reliable short-term forecasting under varying weather conditions.

• Effective in high-dimensional spaces:

SVR performs well when dealing with datasets that include multiple input variables, such as irradiance, temperature, humidity, and wind speed. The use of kernel functions allows the model to handle high-dimensional feature spaces efficiently without a significant increase in computational complexity.

• Robustness to noise:

One of the key features of SVR is the use of an ε-insensitive loss function, which ignores errors within a specified margin. This makes the model less sensitive to noise and measurement errors commonly present in real-world photovoltaic data.

• Flexibility through kernel functions:

SVR offers flexibility by allowing the selection of different kernel functions to best fit the underlying data distribution. The radial basis function (RBF) kernel, in particular, is widely used in PV forecasting due to its ability to model highly nonlinear relationships.

• Convex optimization problem:

The training of SVR involves solving a convex optimization problem, which guarantees a unique global optimum solution. This is a significant advantage compared to some neural network models that may converge to local minima.

• Good balance between bias and variance:

Through proper tuning of hyperparameters (such as the regularization parameter CCC, kernel parameters, and ε), SVR achieves an effective trade-off between underfitting and overfitting, leading to stable and accurate predictions.

• Applicability to nonlinear PV behavior:

Given the nonlinear dependence of PV output on environmental factors, SVR is particularly well-suited for capturing these relationships without requiring explicit physical modeling.

Support Vector Regression represents a reliable and efficient approach for short-term photovoltaic power forecasting. Its ability to generalize well, handle nonlinearities, and perform effectively with limited and noisy data makes it a valuable tool in the development of accurate and robust forecasting models.

4.4 Strengths of Machine Learning Approaches

Machine Learning (ML) models have emerged as powerful tools for photovoltaic (PV) power forecasting due to their ability to learn complex patterns directly from data without requiring explicit physical formulations. Unlike traditional approaches, ML techniques are particularly well-suited to handle the stochastic and nonlinear nature of solar energy production. Several key advantages make these models highly effective for short-term PV forecasting:

• Robustness:

Machine Learning models are inherently robust when dealing with real-world data, which is often noisy, incomplete, or affected by measurement errors. In photovoltaic systems, sensor inaccuracies, missing values, and sudden weather changes can degrade data quality. ML algorithms, especially ensemble methods such as Random Forests, are capable of mitigating these issues by averaging multiple decision paths or learning generalized patterns from imperfect datasets. This robustness enhances the reliability of forecasting results under practical operating conditions.

• Nonlinear modeling:

One of the most significant strengths of Machine Learning lies in its ability to capture complex and nonlinear relationships between input variables and PV power output. Photovoltaic generation depends on multiple interdependent factors, such as solar irradiance, temperature, humidity, and cloud dynamics, whose interactions are highly nonlinear. Traditional linear models often fail to represent these interactions accurately, whereas ML techniques—such as Support Vector Regression, Artificial Neural Networks, and Long Short-Term Memory (LSTM) networks—can model these intricate dependencies with high precision.

• Adaptability:

Machine Learning models are adaptive by design, meaning they can be continuously updated and retrained as new data become available. This is particularly important in photovoltaic forecasting, where environmental conditions vary over time due to seasonal changes, climate variability, and system aging. By incorporating recent data, ML models can improve their predictive performance and remain relevant over time, making them suitable for dynamic and evolving energy systems.

• Interpretability:

Although some Machine Learning models are often considered “black boxes,” certain techniques provide valuable insights into the relative importance of input features. For instance, ensemble methods like Random Forests allow the extraction of feature importance scores, indicating how much each variable (e.g., irradiance, temperature, wind speed) contributes to the prediction. This interpretability is crucial for understanding the underlying factors influencing PV production and for validating the consistency of the model with physical principles.

These characteristics collectively make Machine Learning a highly effective approach for photovoltaic power forecasting. By leveraging robustness, nonlinear modeling capability, adaptability, and interpretability, ML models significantly enhance the accuracy and reliability of short-term PV production predictions. As a result, they play a key role in facilitating the integration of solar energy into modern power systems and supporting efficient energy management strategies.

5. Conclusion

This chapter has established a solid theoretical and methodological foundation for the study of photovoltaic (PV) power systems and their integration into modern energy networks. It first examined the fundamental principles of photovoltaic energy conversion, highlighting the photovoltaic effect and the role of semiconductor p–n junctions in transforming solar radiation into electrical energy. This provided a clear link between the physical behavior of PV cells and their electrical output.

Key parameters influencing PV performance—such as solar irradiance, temperature, efficiency, and instantaneous power—were then analyzed. Their strong interdependence and impact on system behavior under real operating conditions were emphasized. The analysis of current–voltage (I–V) characteristics further demonstrated the nonlinear and variable nature of PV systems, particularly under changing environmental conditions such as irradiance fluctuations, temperature variations, and partial shading.

The chapter also highlighted the critical importance of solar power forecasting for ensuring grid stability and efficient energy management. Different forecasting horizons and methodologies were presented, with a focus on the limitations of traditional approaches when dealing with complex and stochastic PV behavior. In this context, Machine Learning techniques were introduced as powerful and flexible tools capable of capturing nonlinear relationships and improving prediction accuracy.

Overall, this chapter bridges the gap between physical understanding and data-driven modeling, providing the essential framework for developing reliable forecasting models. The next chapter will build upon this foundation by reviewing the state of the art in photovoltaic power forecasting and recent advances in Artificial Intelligence applications within this field.

Chapter II: Bibliographic Study

Chapter II: Bibliographic Study

1. Introduction

The increasing integration of photovoltaic (PV) systems into modern energy networks has made accurate short-term power forecasting an essential challenge. Since PV production depends strongly on changing weather conditions such as solar irradiance, temperature, and cloud cover, reliable prediction methods are required to improve grid stability, energy management, and renewable energy integration.

In recent years, Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) techniques have become powerful tools for PV forecasting due to their ability to model complex nonlinear relationships between meteorological variables and power output.

This chapter presents a bibliographic review of recent studies on short-term photovoltaic power forecasting. It examines different approaches based on Machine Learning, Deep Learning, ensemble, and hybrid models, highlighting their methodologies, performance, advantages, and limitations. This review provides the scientific basis for selecting the models used in this thesis.

2. Previous Studies on Short-Term Prediction of the Photovoltaic Production of a Solar Farm Using Artificial Intelligence, Machine Learning and Deep Learning

Various data-driven approaches have been developed in the literature to improve the short-term prediction of photovoltaic power production. Traditional statistical and time-series methods have been widely used for short-term solar power forecasting. These models are generally simple to construct and convenient to implement. However, it is worth noticing that traditional methods can reach satisfying performance only when photovoltaic production follows regular daily patterns. When the PV production series becomes irregular due to clouds, sudden irradiance variation, temperature changes, or atmospheric instability, the forecast error usually increases. Therefore, many researchers have proposed Artificial Intelligence, Machine Learning, Deep Learning, ensemble learning, and hybrid models to improve forecasting accuracy.

On the other hand, several studies have investigated photovoltaic power prediction using Machine Learning and Deep Learning techniques. These methods are able to learn nonlinear relationships between meteorological variables and PV output without requiring an explicit physical model of the solar farm. In most studies, the input variables include solar irradiance, temperature, wind speed, humidity, module temperature, historical PV power, and time-related features. The output variable is generally the predicted photovoltaic power production for a short-term horizon such as minutes ahead, one hour ahead, or one day ahead.

For instance, Radhi et al.[38] conducted a study to design a short-term photovoltaic power forecasting model using actual operational data collected from a PV experimental prototype. They used several Machine Learning techniques because PV power prediction is essential for the stability, quality, and management of hybrid power grids. The authors trained the models using real operational PV data influenced by weather conditions. They obtained results showing that Machine Learning approaches can effectively predict photovoltaic output and improve the management of PV-based energy systems. The study concluded that data-driven models are suitable for short-term PV forecasting when measured operational data are available.

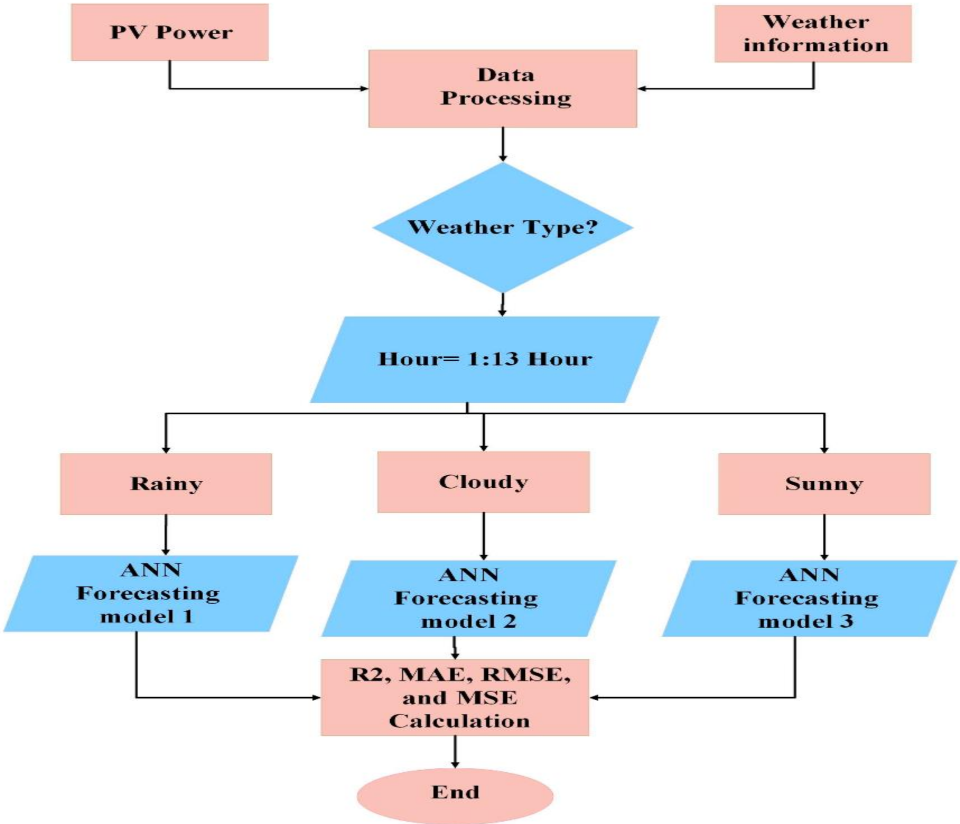


Figure 15: The methodology of training stages[38].

In another study, Wang et al.[39] carried out a study on short-term photovoltaic power forecasting based on a K-means clustering and ensemble learning method. They used K-means clustering to group similar photovoltaic production patterns and applied ensemble learning with a feature rise-dimensional approach to improve the learning ability of the model. They obtained improved prediction performance by combining clustering and ensemble learning. The study concluded that separating operating conditions before training the forecasting model can improve the accuracy and robustness of short-term PV power prediction.

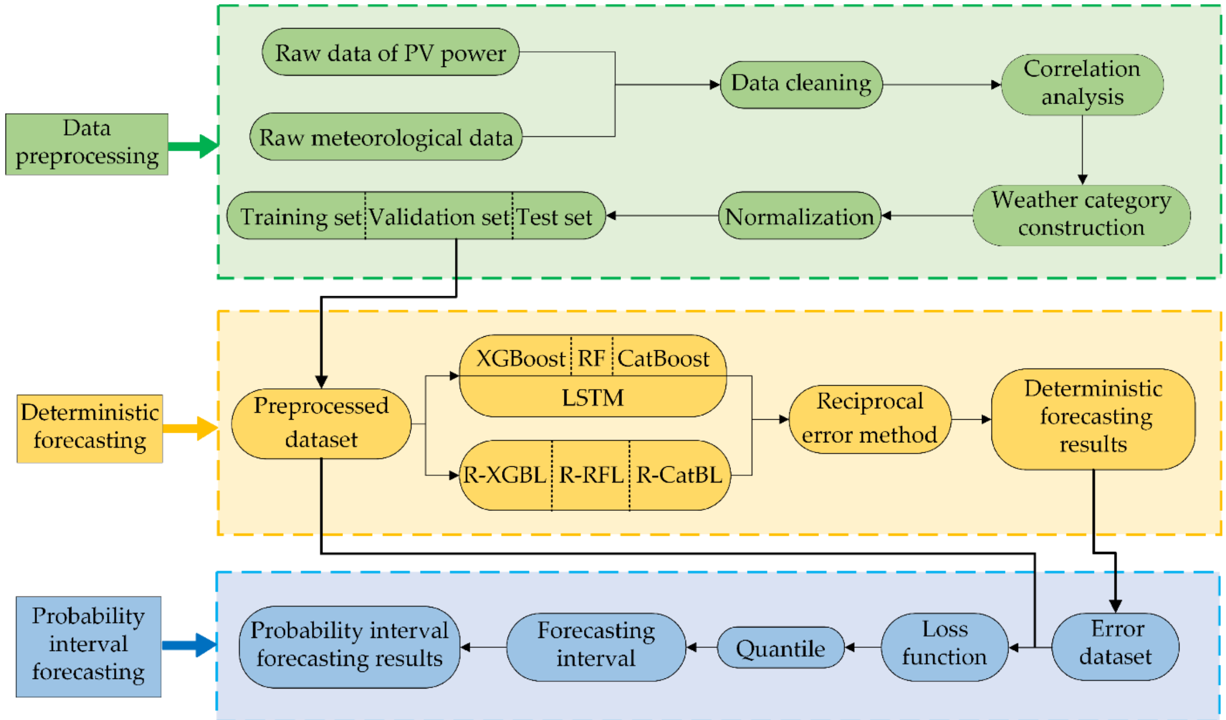


Figure 16: The flowchart of the PV power forecasting [39].

In [40], Xiang et al. conducted a study to improve short-term photovoltaic power forecasting under uncertain weather conditions and nonlinear high-dimensional data. They used a hybrid model combining Temporal Convolutional Network, Efficient Channel Attention Network, and Gated Recurrent Unit, known as TCN-ECANet-GRU. The model was applied to generated data from an Australian photovoltaic power station. They obtained results showing that the proposed hybrid model improved prediction accuracy compared with traditional regression and time-series forecasting methods. The authors concluded that combining temporal feature extraction, attention mechanisms, and recurrent learning is effective for short-term PV power prediction.

Furthermore, Trong et al.[41] conducted a study to improve short-term PV power forecasting using a hybrid deep learning model and Variational Mode Decomposition. They used Variational Mode Decomposition in the preprocessing stage to reduce the complexity and fluctuation of the PV power series. Then, they combined a Transformer Neural Network with a Convolutional Neural Network for prediction. They obtained improved forecasting performance using two real-world datasets and compared the proposed model with five benchmark models. The study concluded that the combination of signal decomposition and hybrid deep learning can significantly enhance short-term photovoltaic forecasting accuracy.

Also, Ibrahim et al.[42] proposed a hybrid model based on Convolutional Neural Networks and Long Short-Term Memory autoencoder for short-term PV power generation forecasting. They used CNN layers to extract important features and LSTM autoencoder layers to capture temporal dependencies in the PV production sequence. The objective was to improve forecasting performance for grid operation and energy storage management. They obtained results showing that combining CNN and LSTM autoencoder structures can improve the prediction of PV power generation. The authors concluded that hybrid deep learning architectures are effective because they can capture both spatial and temporal patterns in photovoltaic data.

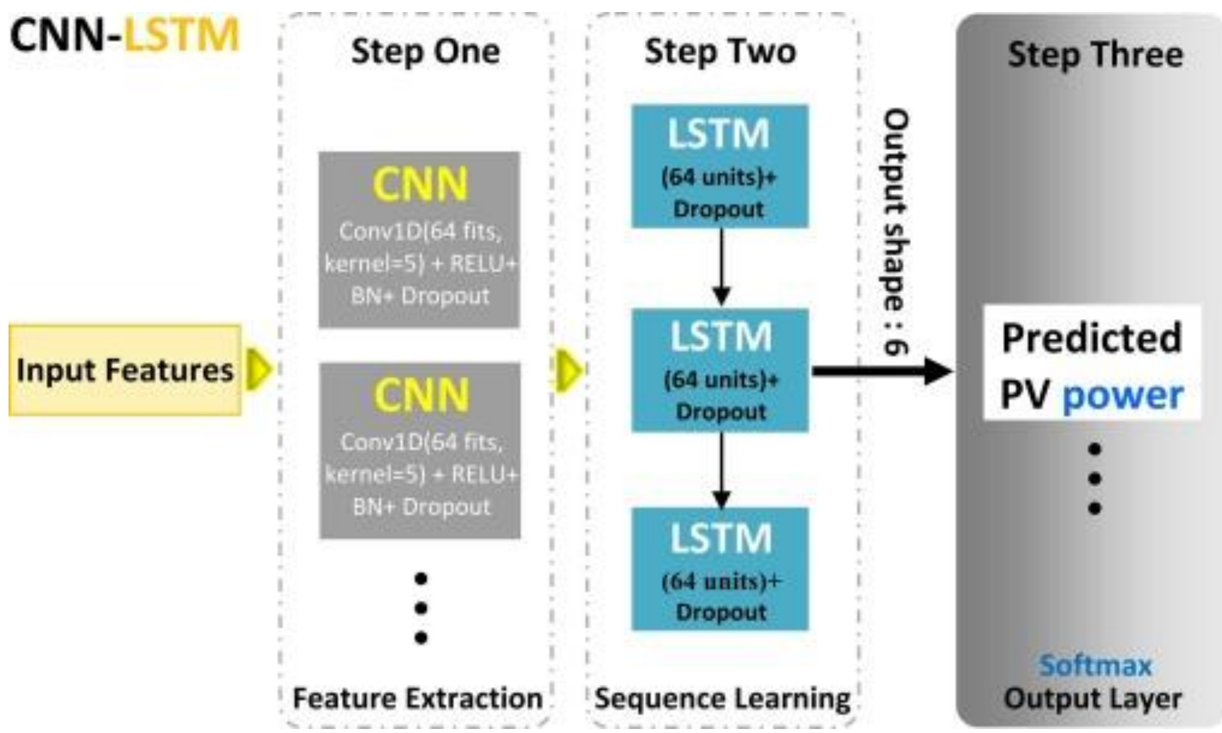


Figure 17:Architecture of the CNN–LSTM hybrid model for solar photovoltaic power forecasting[42].

On the other hand, Dimitropoulos et al.[43] conducted a comparative study of Artificial Intelligence-based models for short-term photovoltaic power forecasting in energy cooperatives. They used Machine Learning and Deep Learning algorithms to predict the production of a solar plant. The authors considered three different input-data cases and included lagged variables to evaluate the influence of historical PV production on one-hour-ahead forecasting. They obtained comparative results showing that the selection of input variables and the use of lagged historical data strongly influence forecasting accuracy. The study concluded that AI-based models are useful for short-term PV prediction, especially when historical production data are properly included.

Moreover, Nguyen et al.[44] conducted a study to develop a novel solar power forecasting method based on a hybrid stacking ensemble. They used XGBoost, LightGBM, and Random Forest as base learners, while the L-BFGS-B optimization algorithm was used to determine the optimal weights of the ensemble model. The model was trained and tested using real-world data from a 49.5 MW solar power plant in Vietnam. They obtained results showing that the proposed ensemble model outperformed individual learners and deep learning baselines in terms of both accuracy and training time. The authors concluded that optimized ensemble learning is an effective method for solar power forecasting in large-scale photovoltaic plants.

Additionally, Sarmas et al.[45] abstracted a study on short-term photovoltaic power forecasting using meta-learning and LSTM models independent of numerical weather prediction. They used four different Long Short-Term Memory architectures as base models and applied a meta-learning method to dynamically combine their forecasts. Their objective was to improve one-hour-ahead deterministic PV forecasts without relying on numerical weather prediction. They obtained improved forecasting performance by allowing the meta-model to learn under which conditions each LSTM model performs best. The study concluded that meta-learning can improve the generalizability of short-term PV forecasting models.

As well, Boujoudar et al.[46] conducted a study on photovoltaic power forecasting using Machine Learning algorithms based on weather data in a semi-arid climate. They used CatBoost, LightGBM, XGBoost, and Random Forest to predict PV power. They obtained results showing that Machine Learning algorithms can improve the accuracy of PV power prediction, with LightGBM giving strong performance. The authors concluded that modern tree-based ensemble models are effective for photovoltaic forecasting because they can capture nonlinear relationships between weather variables and PV output.

Besides, Ma et al.[47] conducted a study to improve ultra-short-term photovoltaic power forecasting using multimodal data and ensemble learning. They used six independent models: Random Forest, XGBoost, CatBoost, LightGBM, LSTM, and GRU. Then, a fused model was developed using a stacking ensemble strategy. They obtained results showing that the ensemble model improved prediction capability compared with individual models. The study concluded that integrating multimodal data with ensemble learning can enhance ultra-short-term photovoltaic forecasting by combining the strengths of Machine Learning and Deep Learning models.

Over and above that, Akhter et al.[48] conducted a study on hour-ahead photovoltaic power forecasting using an RNN-LSTM model for three different photovoltaic plants. They used data from polycrystalline, monocrystalline, and thin-film PV technologies over a four-year period. The input variables included wind speed, module temperature, ambient temperature, and solar irradiation, while the output variable was PV power. They compared the proposed RNN-LSTM model with several approaches, including regression, Machine Learning, and hybrid methods. They obtained better prediction accuracy using RNN-LSTM according to RMSE, MSE, correlation coefficient, and R2. The study concluded that RNN-LSTM is robust and effective for hour-ahead PV power forecasting across different PV technologies.

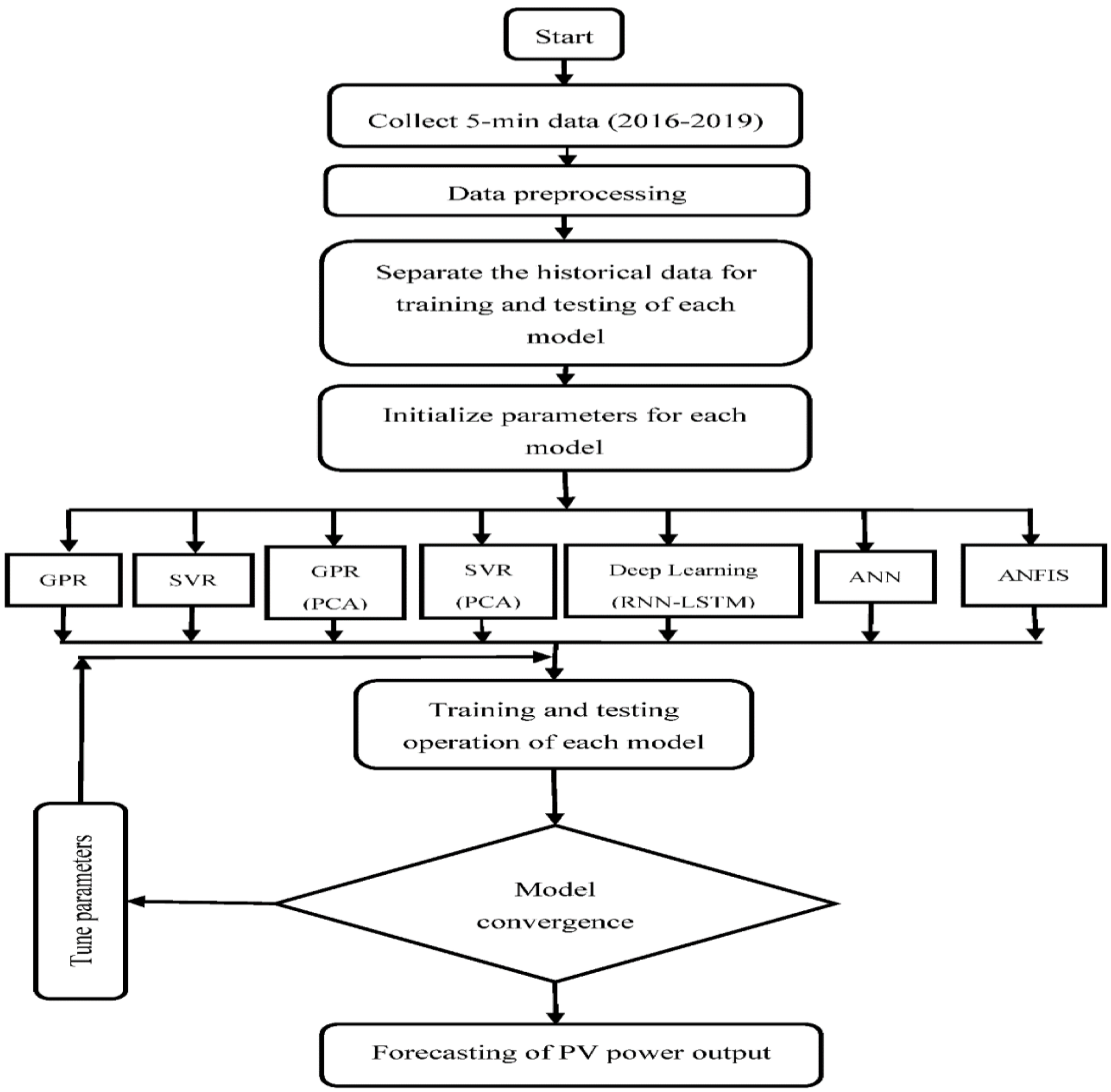


Figure 18: An overall methodology flow chart [48].

Beyond that, Li et al.[49] conducted a study on a hybrid deep learning model for short-term photovoltaic power forecasting. They used Wavelet Packet Decomposition to decompose the original PV power series into sub-series and Long Short-Term Memory networks to predict future values. The objective was to handle the intermittent and random behavior of PV power generation. They obtained improved forecasting accuracy by combining signal decomposition with LSTM learning. The study concluded that hybrid decomposition and deep learning models are effective for short-term PV power forecasting.

Along those lines, Agga et al.[50] conducted a study to forecast short-term power production of a self-consumption photovoltaic plant. They proposed two hybrid deep learning models, CNN-LSTM and ConvLSTM, and compared them with a basic LSTM model. They used convolutional layers to extract features and recurrent layers to learn temporal dependencies. They obtained results showing that hybrid models can improve the forecasting of PV plant power production. The authors concluded that combining convolutional feature extraction with recurrent time-series learning is suitable for short-term photovoltaic production prediction.

In parallel, Al-Ali et al.[51] abstracted a study on solar energy production forecasting using a hybrid deep learning approach. They used a combination of Convolutional Neural Network, Long Short-Term Memory network, and Transformer architecture. They also applied clustering and feature selection techniques to improve the representation of historical data. They obtained a hybrid CNN-LSTM-Transformer model for solar energy forecasting. The authors concluded that combining deep learning architectures with feature selection can improve solar production forecasting by learning complex nonlinear and temporal relationships.

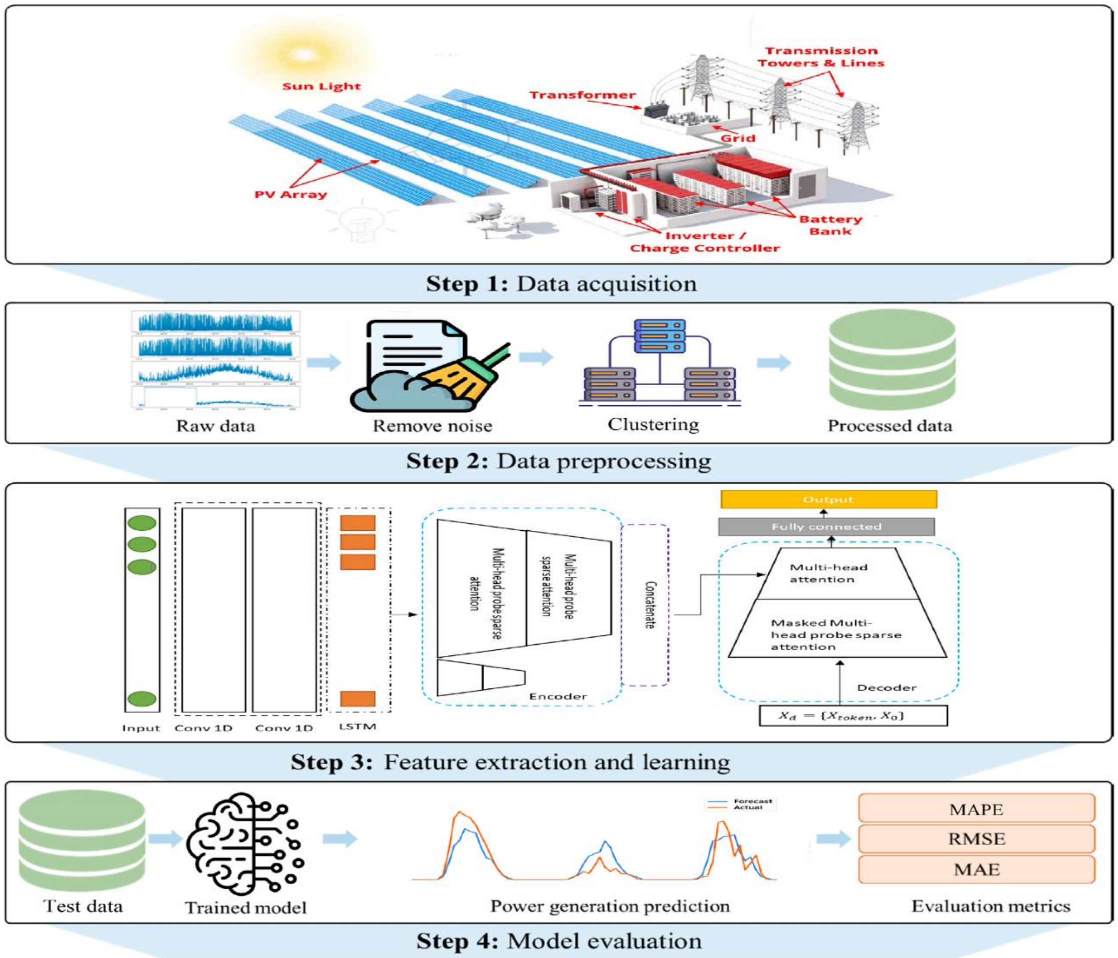


Figure 19:An overall methodology flow chart [51].

In a similar vein, Gupta et al.[52] conducted a review study on photovoltaic power forecasting using Machine Learning techniques. They reviewed Machine Learning, Deep Learning, mathematical models, hybrid approaches, input variables, forecasting horizons, and preprocessing methods. They obtained a general synthesis showing that Machine Learning techniques play an important role in improving solar grid integration, real-time forecasting, and energy management. The authors concluded that future research should focus on improving model robustness, preprocessing methods, hybrid models, and the generalization of forecasting models across different photovoltaic systems.

As we can see from the reviewed studies, Machine Learning and Deep Learning methods have become powerful tools for short-term photovoltaic power forecasting. Traditional methods are easy to implement, but they are often insufficient when PV production becomes highly variable due to unstable weather conditions. Machine Learning models such as Random Forest, XGBoost, LightGBM, CatBoost, Support Vector Regression, and Gradient Boosting are effective for structured meteorological datasets. Deep Learning models such as ANN, RNN, CNN, GRU, LSTM, and Transformer-based architectures are more suitable for time-series forecasting when large historical datasets are available.

Furthermore, hybrid models generally provide better performance than standalone models. This improvement is mainly due to their ability to combine preprocessing, feature extraction, temporal learning, and ensemble decision-making. For example, decomposition methods such as Wavelet Packet Decomposition and Variational Mode Decomposition help reduce noise and separate complex PV power signals into simpler components. Similarly, ensemble models such as stacking and optimized combinations of XGBoost, LightGBM, and Random Forest improve prediction accuracy by combining the strengths of several algorithms.

However, several limitations still appear in previous studies. First, forecasting accuracy depends strongly on data quality and the availability of reliable meteorological variables such as solar irradiance and temperature. Second, missing values, noisy measurements, and irregular weather conditions can reduce model performance. Third, some models perform well on one photovoltaic plant but may not generalize well to another location with different climatic conditions. Fourth, Deep Learning models often require large datasets and high computational resources. Finally, some high-performing models lack interpretability, which makes it difficult to understand the influence of each meteorological variable on photovoltaic production.

These limitations create opportunities for improvement in the present thesis. Therefore, this work focuses on short-term prediction of the photovoltaic production of a solar farm using Artificial Intelligence and Machine Learning. By comparing several models under the same dataset, applying appropriate preprocessing techniques, and evaluating performance using standard metrics such as MAE, RMSE, and R2, this study aims to identify the most accurate and reliable model for predicting photovoltaic production.

3. Conclusion

This chapter reviewed recent research on short-term photovoltaic power forecasting using Artificial Intelligence, Machine Learning, and Deep Learning techniques. The literature shows that modern ML models such as Random Forest, XGBoost, LightGBM, and CatBoost, as well as DL models such as

LSTM, CNN, and Transformer architectures, can significantly improve forecasting accuracy by capturing nonlinear and temporal patterns in PV data.

However, several challenges remain, including data quality issues, weather uncertainty, model generalization, computational requirements, and the limited interpretability of complex models.

Chapter III: Research Methodology

Chapter III: Research Methodology

1. Introduction

This chapter presents the methodological framework adopted for the short-term prediction of photovoltaic production of a solar farm using artificial intelligence and machine learning. The study relies on meteorological and operational data to estimate the target variable Power_PACE. The methodology follows a data-driven regression approach in which the input variables are prepared, the target variable is defined, several machine learning models are trained, and their predictive performance is compared using standard evaluation metrics.

The general objective of this chapter is to explain the complete procedure followed before obtaining the final results. It describes the input and output parameters, the preprocessing steps performed in Python with Jupyter Notebook, the selected machine learning models, and the role of model interpretation using SHAP analysis. The same general organization used in the reference structure is maintained, while adapting the content to photovoltaic production forecasting rather than wind power forecasting.

2. The input and the output parameters and their importance in predicting photovoltaic production

The dataset used in this work is organized into two separated files: solar_train.xlsx for model training and solar_test.xlsx for model testing. Each observation contains a timestamp, meteorological measurements, and the corresponding photovoltaic production value. The first nine columns are used as explanatory variables, while Power_PACE is used as the target output to be predicted.

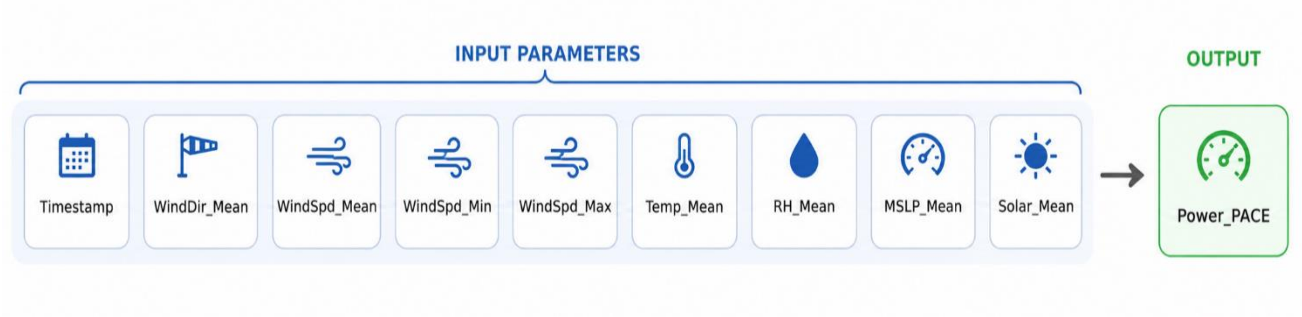


Figure 20: Input-output structure used for Power_PACE prediction.

Table 1: Description of the dataset parameters.

Parameter	Role in the prediction problem
Timestamp	Date and time information used to extract temporal features such as year, month, day and hour.
WindDir_Mean (deg)	Mean wind direction in degrees, which can indirectly reflect meteorological conditions.
WindSpd_Mean (km/h)	Mean wind speed in km/h.
WindSpd_Min (km/h)	Minimum wind speed in km/h during the observed period.
WindSpd_Max (km/h)	Maximum wind speed in km/h during the observed period.
Temp_Mean (deg)	Mean air temperature, a key factor that can influence photovoltaic module efficiency.
RH_Mean (%)	Mean relative humidity, which can be associated with cloudiness and atmospheric attenuation.
MSLP_Mean (hPa)	Mean sea-level pressure, representing large-scale weather conditions.
Solar_Mean (W/m2)	Mean solar irradiance, the most direct physical driver of photovoltaic generation.
Power_PACE	Target output representing the photovoltaic production to be predicted.

The use of meteorological variables is essential because photovoltaic production is strongly affected by the variability of solar irradiation and atmospheric conditions [53]. Solar_Mean is particularly important because it represents the available solar resource. Temperature also affects the electrical efficiency of photovoltaic modules, while humidity, wind and pressure can capture weather changes that influence the short-term behavior of production.

3. Data preprocessing and feature engineering

The preprocessing stage was implemented in Python using Jupyter Notebook. First, the Excel files were loaded and the column names were cleaned to avoid errors caused by hidden spaces. The Timestamp column was then converted into a datetime format. Instead of using the raw date-time value directly, the timestamp was decomposed into Year, Month, Day and Hour. This transformation allows the models to learn daily and seasonal patterns in the solar farm production [54].

Missing values were handled by replacing them with the mean value of the corresponding feature. This step ensures that all models receive complete numerical input matrices. The original Timestamp column was removed after extracting the temporal features. Finally, the data were split into X_train, y_train, X_test and y_test, where X contains all explanatory variables and y contains Power_PACE.

```
train_df = pd.read_excel('solar_train.xlsx')
test_df = pd.read_excel('solar_test.xlsx')
train_df['Timestamp'] = pd.to_datetime(train_df['Timestamp'], errors='coerce')
for df in [train_df, test_df]:
    df['Year'] = df['Timestamp'].dt.year
    df['Month'] = df['Timestamp'].dt.month
    df['Day'] = df['Timestamp'].dt.day
    df['Hour'] = df['Timestamp'].dt.hour
X_train = train_df.drop(columns=['Power_PACE'])
y_train = train_df['Power_PACE']
```

4. The selection of Machine Learning models

To predict Power_PACE, five regression models were selected: Linear Regression, Random Forest, Gradient Boosting, XGBoost and CatBoost. These models were chosen because they represent different learning strategies, from a simple linear baseline to advanced ensemble learning algorithms. Their comparison makes it possible to evaluate whether the photovoltaic production behavior is better represented by linear relationships or by nonlinear tree-based structures [55].

4.1 Random Forest

Random Forest (RF) is a supervised ensemble machine learning algorithm based on the construction of a large number of decision trees. It was introduced by Leo Breiman in 2001 [56] and has become one of the most popular algorithms for regression and classification tasks due to its robustness, simplicity, and high predictive performance.

In a Random Forest regression model, each decision tree is trained using a random subset of the available dataset through a technique known as bootstrap aggregation (bagging). Furthermore, at each node of the tree, a random subset of input features is selected to determine the best split. This randomization increases diversity among the individual trees and reduces the correlation between them.

For regression problems, such as photovoltaic power forecasting, each decision tree generates an independent numerical prediction of the PV output power. The final predicted value is obtained by averaging the predictions of all trees in the forest. This averaging process significantly reduces the variance of the model and minimizes the risk of overfitting compared with a single decision tree.

Random Forest is particularly suitable for short-term photovoltaic forecasting because it can effectively model complex nonlinear relationships between meteorological variables—such as solar irradiance, temperature, humidity, and wind speed—and the generated electrical power. It is also resistant to noisy measurements and can estimate the relative importance of each input variable, which helps identify the most influential factors affecting PV production.

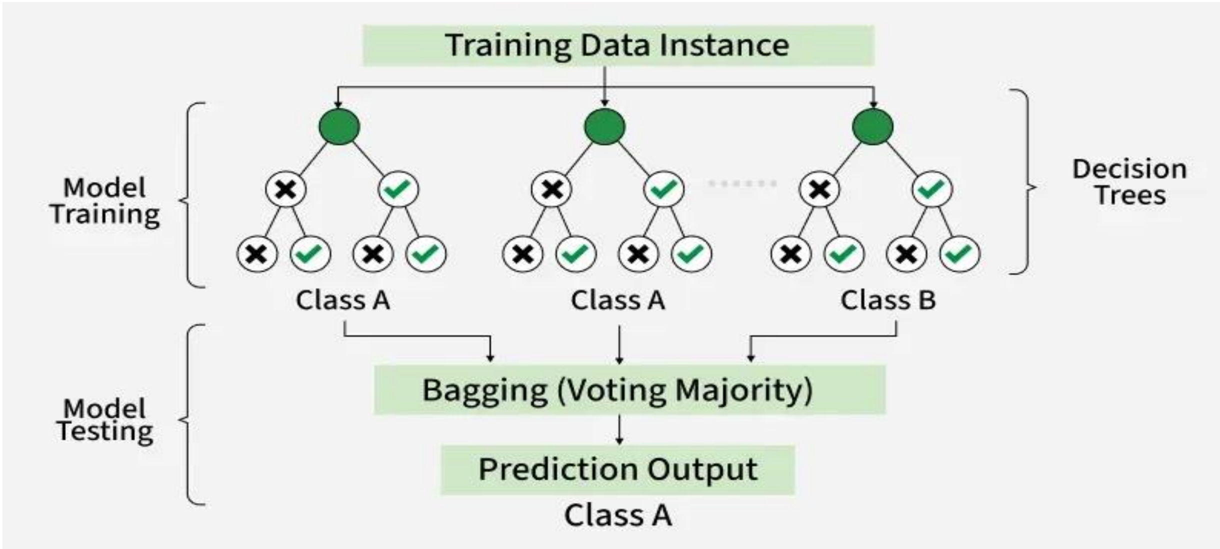


Figure 21: Random Forest Algorithm in Machine Learning [57].

4.2 Gradient Boosting

Gradient Boosting is an ensemble machine learning technique that combines several weak decision tree models in a sequential manner to produce a highly accurate predictive model [58]. Unlike Random Forest, where trees are constructed independently, Gradient Boosting creates trees one after another, and each new tree attempts to correct the prediction errors produced by the previous trees.

The algorithm works by minimizing a predefined loss function using a gradient descent optimization approach. At each iteration, a new decision tree is trained on the residual errors of the current model, allowing the overall prediction accuracy to progressively improve.

In photovoltaic power forecasting, Gradient Boosting can capture highly nonlinear interactions between environmental variables and power generation. It is particularly useful for modeling complex relationships involving solar irradiance fluctuations, temperature variations, and temporal patterns.

In this study, Gradient Boosting is used as a reference boosting technique and is compared with more advanced algorithms such as XGBoost and CatBoost to evaluate whether optimized boosting strategies provide additional improvements in forecasting accuracy.

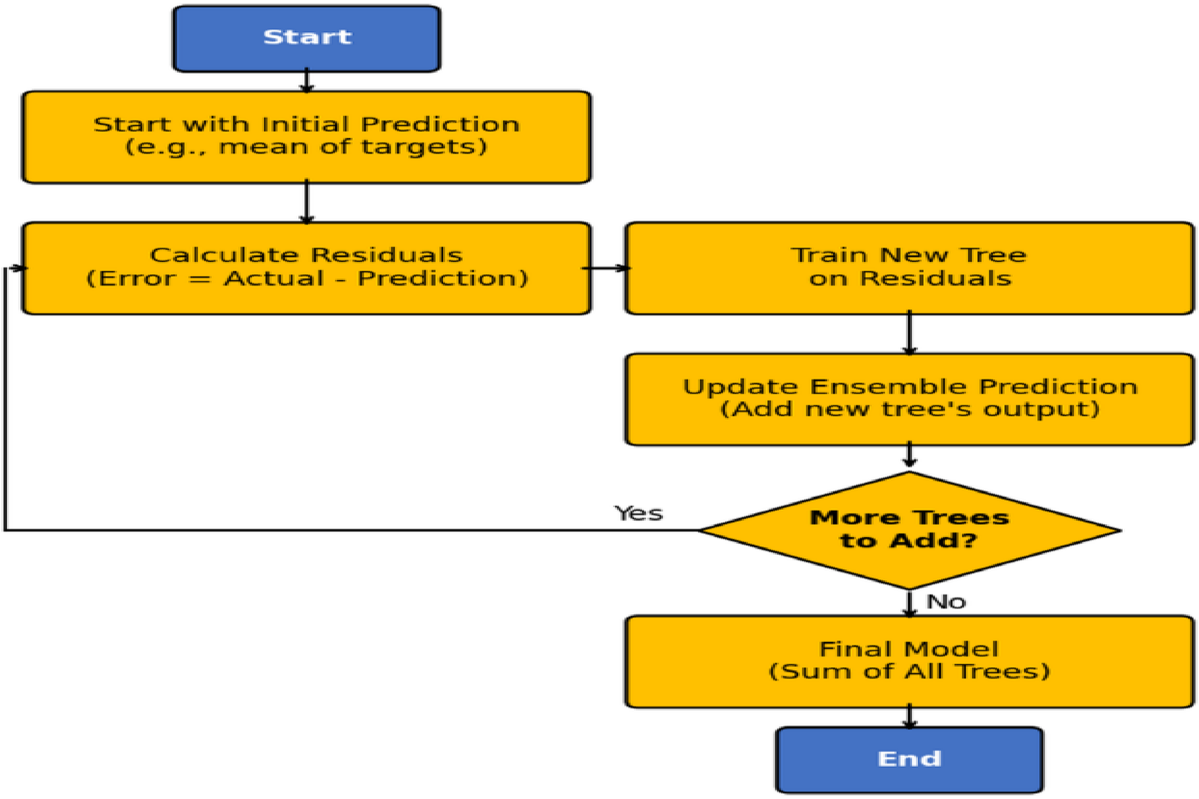


Figure 22: Gradient Boosting in machine learning [59].

4.3 XGBoost

XGBoost, short for Extreme Gradient Boosting, is an advanced implementation of the Gradient Boosting algorithm developed to improve computational efficiency, scalability, and predictive accuracy. It introduces several optimization techniques, including regularization methods, parallel processing, efficient handling of missing data, and advanced tree-pruning strategies [60].

In contrast to traditional Gradient Boosting, XGBoost incorporates both L1 and L2 regularization mechanisms, which reduce model complexity and prevent overfitting. It also uses an optimized objective function that balances prediction accuracy and model simplicity.

Because photovoltaic power generation depends on multiple nonlinear and interacting variables, such as solar irradiance, ambient temperature, humidity, and time-related parameters, XGBoost is particularly suitable for this application. It can learn complex interactions between these variables while maintaining high computational efficiency.

In this work, XGBoost is applied to predict photovoltaic power output and its performance is compared with other machine learning algorithms, including Random Forest, Gradient Boosting, CatBoost, and Linear Regression.

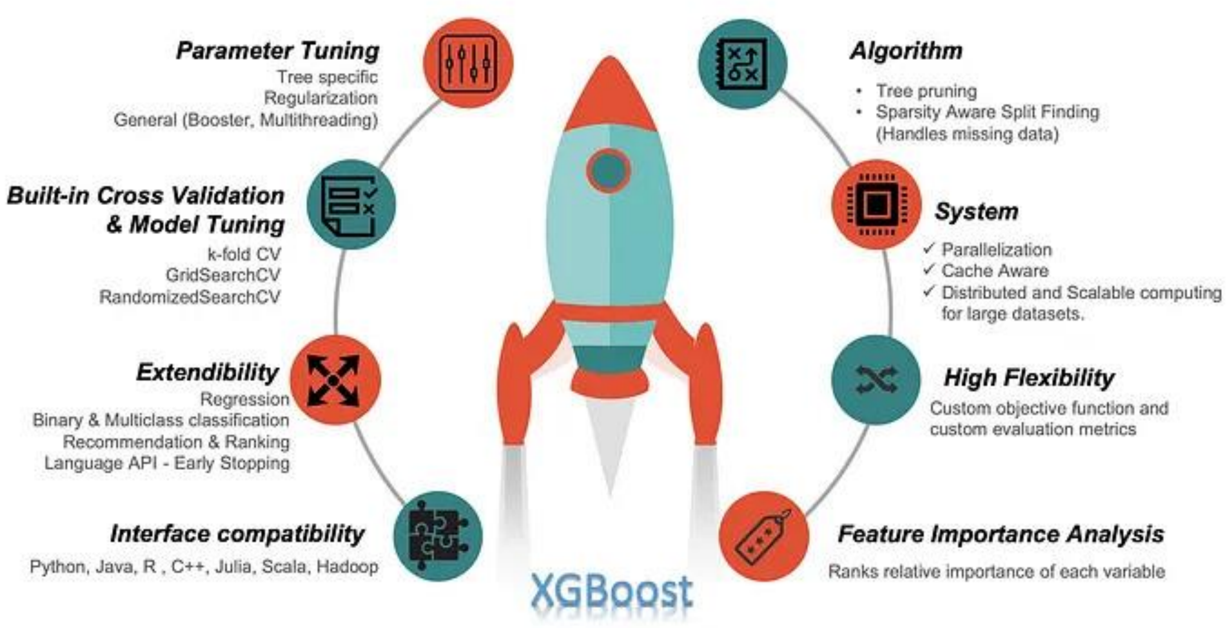


Figure 23: XGBoost Algorithm [61].

4.4 CatBoost

CatBoost is a modern gradient boosting algorithm developed by Yandex and designed to provide high prediction performance with minimal data preprocessing. Although it was initially developed to efficiently process categorical variables, CatBoost also demonstrates excellent performance in numerical regression tasks [62].

The algorithm uses an ordered boosting strategy that reduces prediction bias and improves the generalization capability of the model. It also includes advanced regularization techniques and optimized gradient calculations, allowing it to achieve competitive accuracy while reducing the risk of overfitting.

For photovoltaic power forecasting, CatBoost can effectively learn the nonlinear relationships between meteorological variables and generated electrical power. In this study, CatBoost is employed to predict PV power using variables such as solar irradiance, temperature, humidity, wind speed, and temporal information. Its performance is compared with other algorithms, including Random Forest, Gradient Boosting, XGBoost, and Linear Regression.

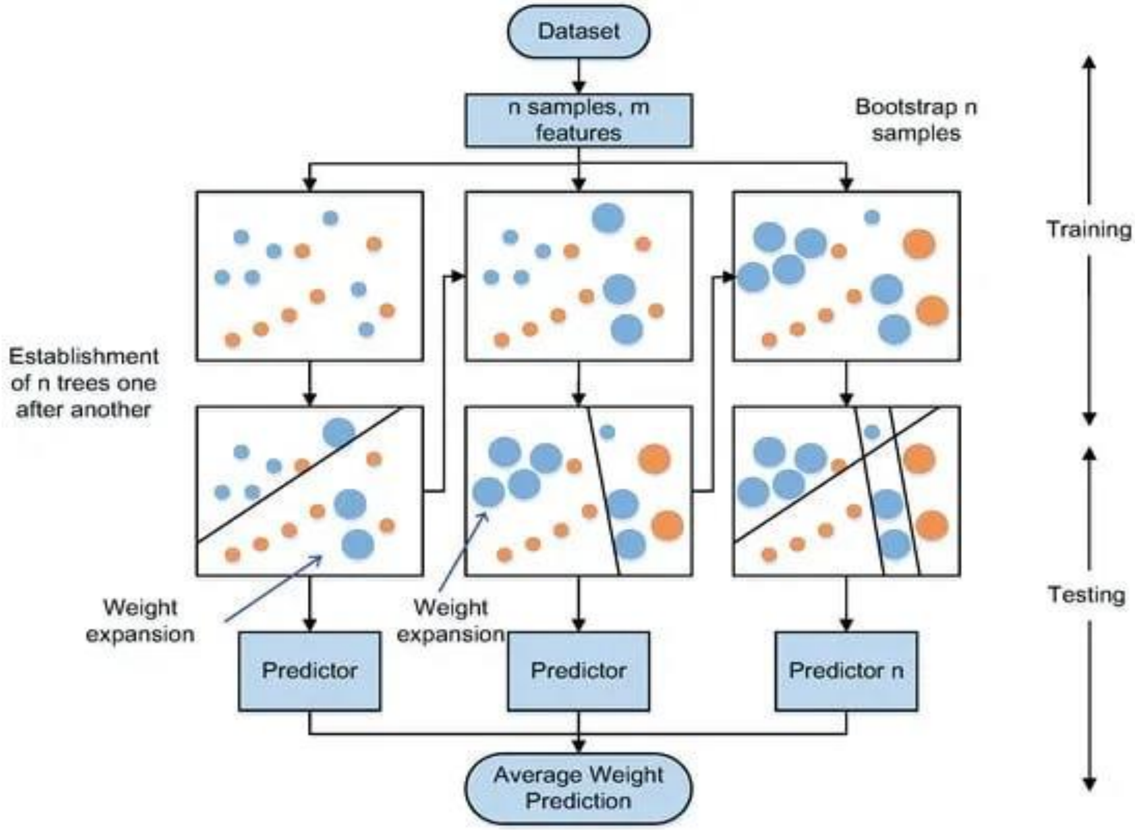


Figure 24: The flow diagram of the catboost model [63].

4.5 Linear Regression

Linear Regression is one of the oldest and most fundamental statistical and machine learning algorithms used for predicting a continuous output variable [64]. It establishes a mathematical relationship between one dependent variable and one or several independent variables using a linear equation.

In the context of photovoltaic power forecasting, Linear Regression attempts to estimate PV power production as a linear combination of meteorological and temporal variables such as solar irradiance, temperature, humidity, wind speed, and time indicators.

Although the relationship between environmental conditions and PV generation is often nonlinear, Linear Regression remains an essential baseline model. It provides a simple and interpretable reference that allows the performance improvements obtained by more advanced machine learning models to be quantitatively evaluated.

The simplicity, fast training time, and low computational cost of Linear Regression make it useful for preliminary analysis and comparison purposes.

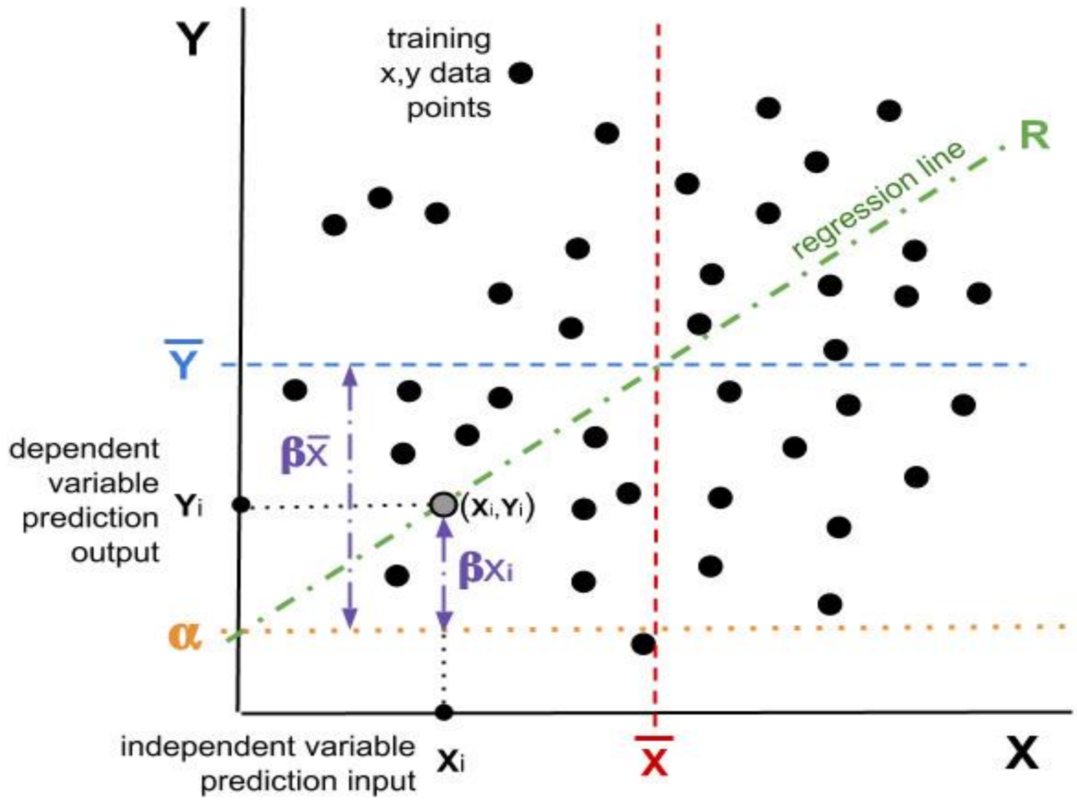


Figure 25: Linear Regression model [65].

5. Prediction framework

Figure 26: Prediction framework of the proposed photovoltaic forecasting approach.

The proposed prediction framework begins with the collection and organization of the photovoltaic dataset. The training and testing files are loaded separately, then the input features and the Power_PACE target are defined. After preprocessing, the models are trained on the training data and evaluated on the independent testing data. The final stage consists of comparing the models using MAE, RMSE and R2, then interpreting the most important parameters using SHAP plots [66].

6. Conclusion

This chapter presented the methodology followed to build the prediction system, including the definition of input and output parameters, preprocessing operations and the selection of machine learning models.

Several Machine Learning models, including Linear Regression, Random Forest, Gradient Boosting, XGBoost, and CatBoost, were selected and integrated into a prediction framework involving model training, performance evaluation using MAE, RMSE, and R^2, and interpretation using SHAP analysis.

The adopted methodology provides a structured approach for comparing the predictive capabilities of different algorithms and identifying the most accurate model for short-term photovoltaic production forecasting.

Chapter IV: Results and discussion

Chapter IV: Results and discussion

1. Introduction

This chapter presents the simulation results obtained from the five Machine Learning models developed for short-term photovoltaic power forecasting: Linear Regression, Random Forest, CatBoost, XGBoost, and Gradient Boosting. Following the methodology described in Chapter III, each model was trained on historical meteorological and production data collected from the solar farm, and evaluated on a separate test set using standard regression metrics.

The chapter is structured as follows. The first section describes the implementation and training process for each model. The second section provides a detailed analysis of the prediction results, including a comparison of error metrics and an investigation of meteorological variable influence through SHAP-based feature importance. The third section discusses the strengths and weaknesses of the best-performing models and their robustness under real-world conditions. The chapter concludes with a set of proposed improvements that could further enhance forecasting accuracy.

2. Evaluation metrics (R^2, MAE, RMSE)

The evaluation of the regression models was carried out by comparing the predicted values with the real measured values of Power_PACE. Three main metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and the coefficient of determination (R2). Lower values of MAE and RMSE indicate smaller prediction errors, while a higher R2 score indicates that the model explains a larger part of the variability of the target variable.

Table 2: Evaluation metrics used for model comparison.

Metric	Full name	Interpretation
MAE	Mean Absolute Error	Average absolute difference between real and predicted values. Lower is better.
RMSE	Root Mean Squared Error	Square root of the mean squared prediction error. It penalizes large errors more strongly. Lower is better.
R2	Coefficient of Determination	Proportion of target variability explained by the model. Closer to 1 is better.

- **R2** signifies the portion of variability in the target variable that can be anticipated based on the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

- **RMSE** quantifies the disparity between predicted and actual values by considering the standard deviation of the error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

- **MAE** measures the absolute difference between the predicted and actual values.

$$MAE = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{n}$$

where n represents the number of samples, y_t denotes the actual value of the response variable, and $\hat{y}_i$ refers to the predicted value.

3. Model Implementation

3.1 Program presentation and Dataset Description

The simulation was performed in Anaconda using Jupyter Notebook. The implementation included the loading of solar_train.xlsx and solar_test.xlsx, preprocessing of the Timestamp feature, training of five machine learning models, calculation of performance metrics, export of the results to Excel, visualization of prediction distributions using violin plots, and interpretation of feature effects using SHAP.

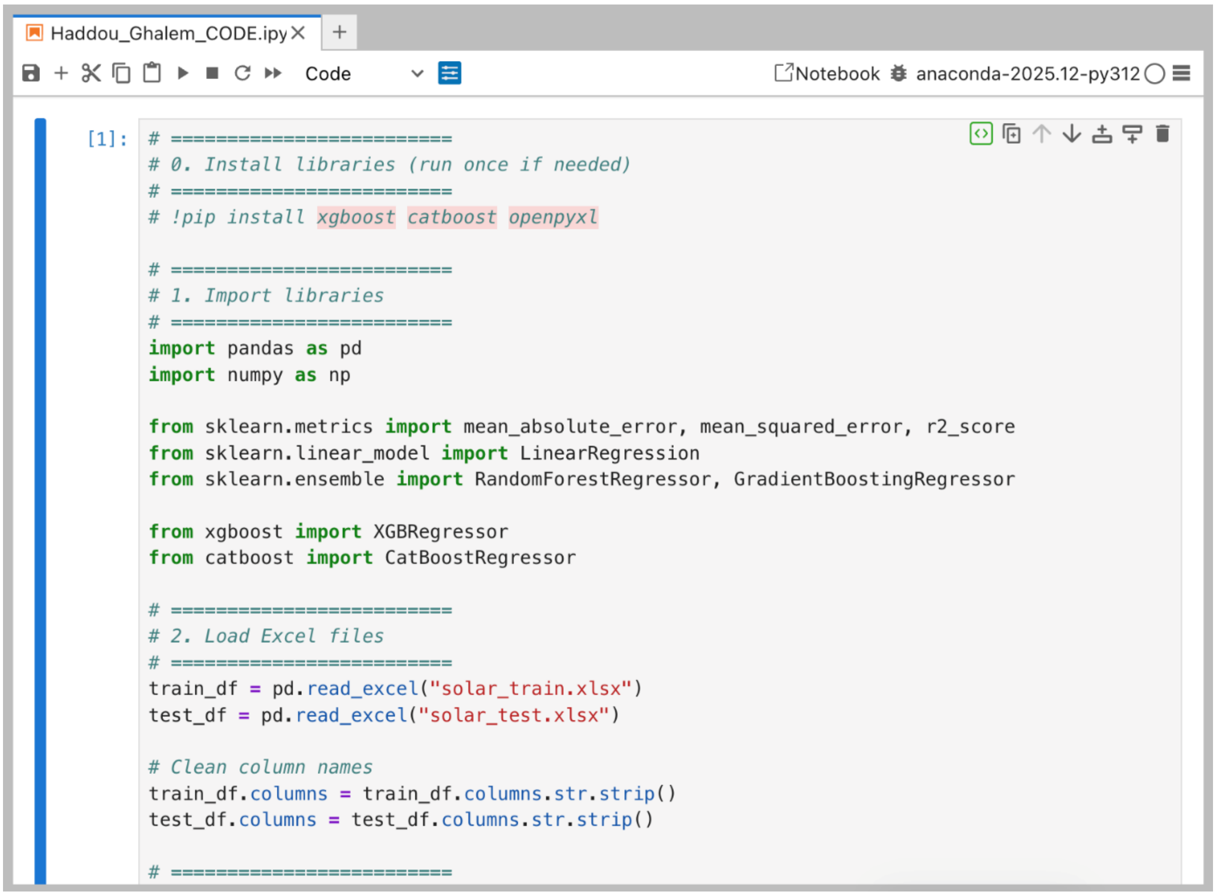


Figure 27: real picture of our program in this study.

The dataset used in this study consists of minute-level measurements recorded at the PACE solar farm between July and August 2018. After preprocessing, the data were split into a training set of 20,168 samples and a test set of 5,042 samples, maintaining the chronological order of observations to avoid data leakage. The target variable is Power_PACE, expressed in watts (W), representing the instantaneous photovoltaic power output.

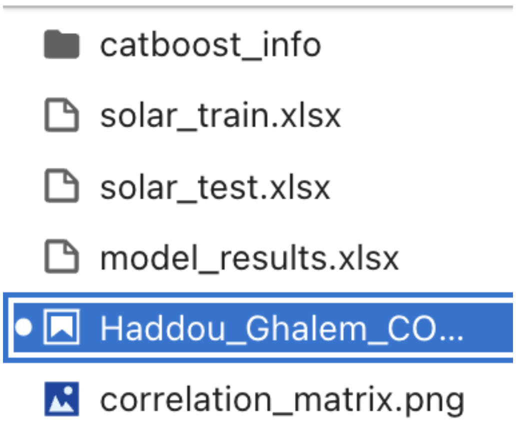
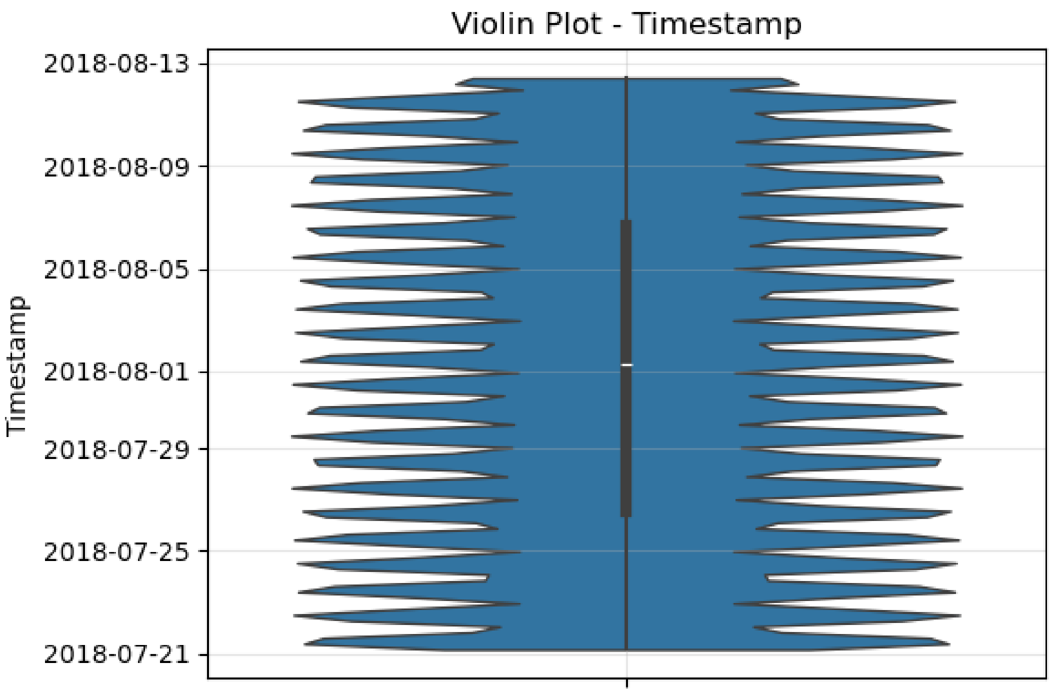


Figure 28: real picture of our data.

The feature set comprises eight meteorological and temporal variables: wind direction (WindDir_Mean), mean wind speed (WindSpd_Mean), minimum and maximum wind speed (WindSpd_Min, WindSpd_Max), mean temperature (Temp_Mean), relative humidity (RH_Mean), mean sea-level pressure (MSLP_Mean), and solar irradiance (Solar_Mean). Timestamp features were additionally extracted to capture diurnal and seasonal patterns.

Figure 27 shows the distribution of individual features through violin plots.



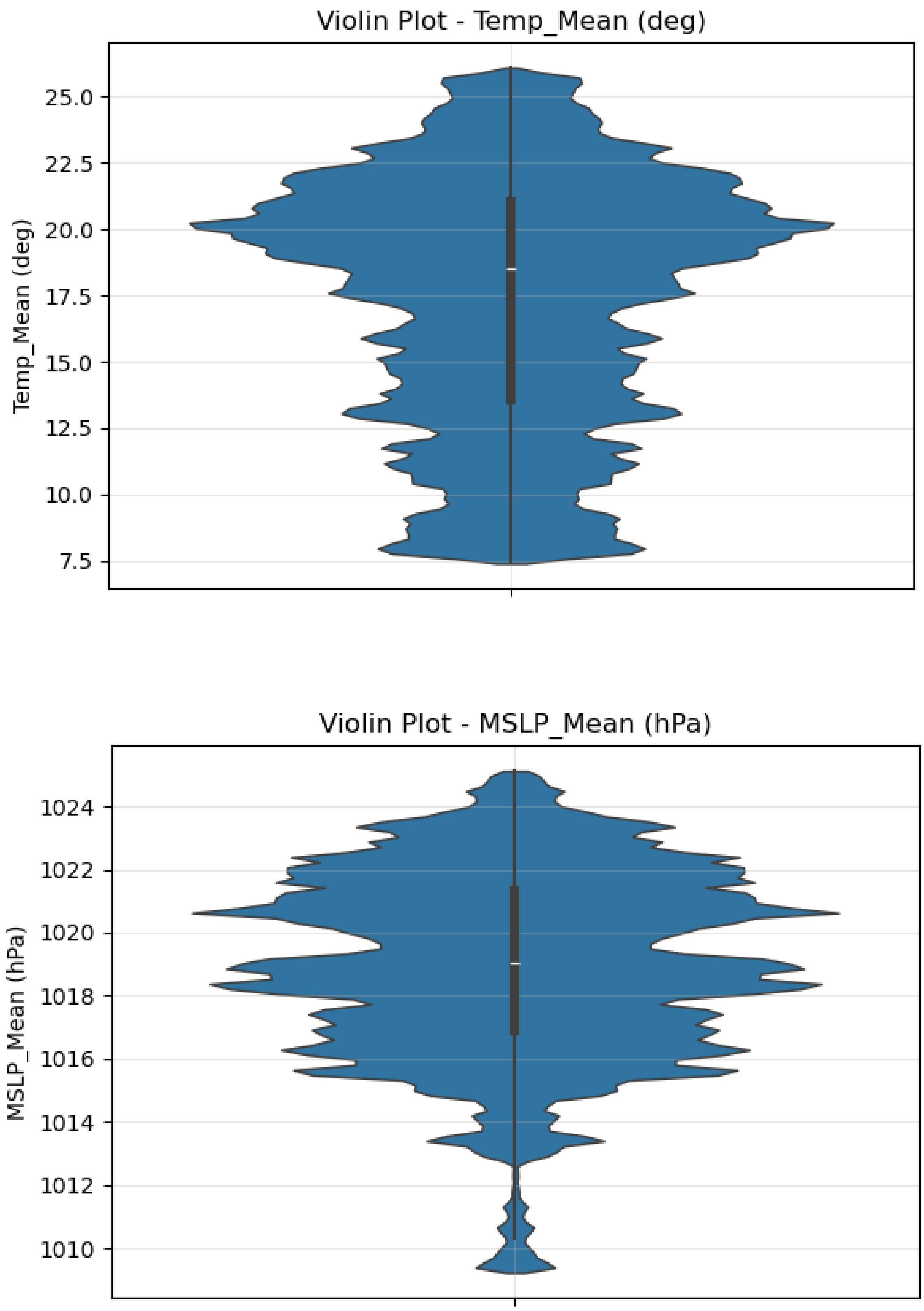


Figure 29: the distribution of individual features through violin plots.

3.2 Training and Testing of Machine Learning Models

All models were implemented in Python using the scikit-learn, XGBoost, and CatBoost libraries. The implementation followed a consistent pipeline: data loading from the provided Excel files (solar_train.xlsx and solar_test.xlsx), feature engineering from the Timestamp column (extraction of hour, day, month, and year), training on the full training split, and evaluation on the held-out test split.

Five models were trained and compared in this study:

- **Linear Regression**: a baseline linear model establishing a reference for comparison;
- **Random Forest**: an ensemble of decision trees trained with bootstrap aggregating;
- **CatBoost**: a gradient boosting algorithm with native handling of categorical features;
- **XGBoost**: an optimised implementation of gradient boosting with L1/L2 regularisation;
- **Gradient Boosting**: the standard scikit-learn implementation of stochastic gradient boosting.

Training was performed with default hyperparameters in order to establish a fair baseline comparison. Each model was fitted once on the training set and predictions were generated on the test set. Evaluation was carried out using four standard regression metrics: MAE, RMSE, and R^2.

4. Results Analysis

4.1 Influence of Meteorological Variables

A preliminary correlation analysis was performed on the training set to assess the linear relationships between all input features and the target variable Power_PACE. Figure 29 presents the full correlation matrix, while Figure 4.4 shows the feature-only heatmap with numerical annotations.

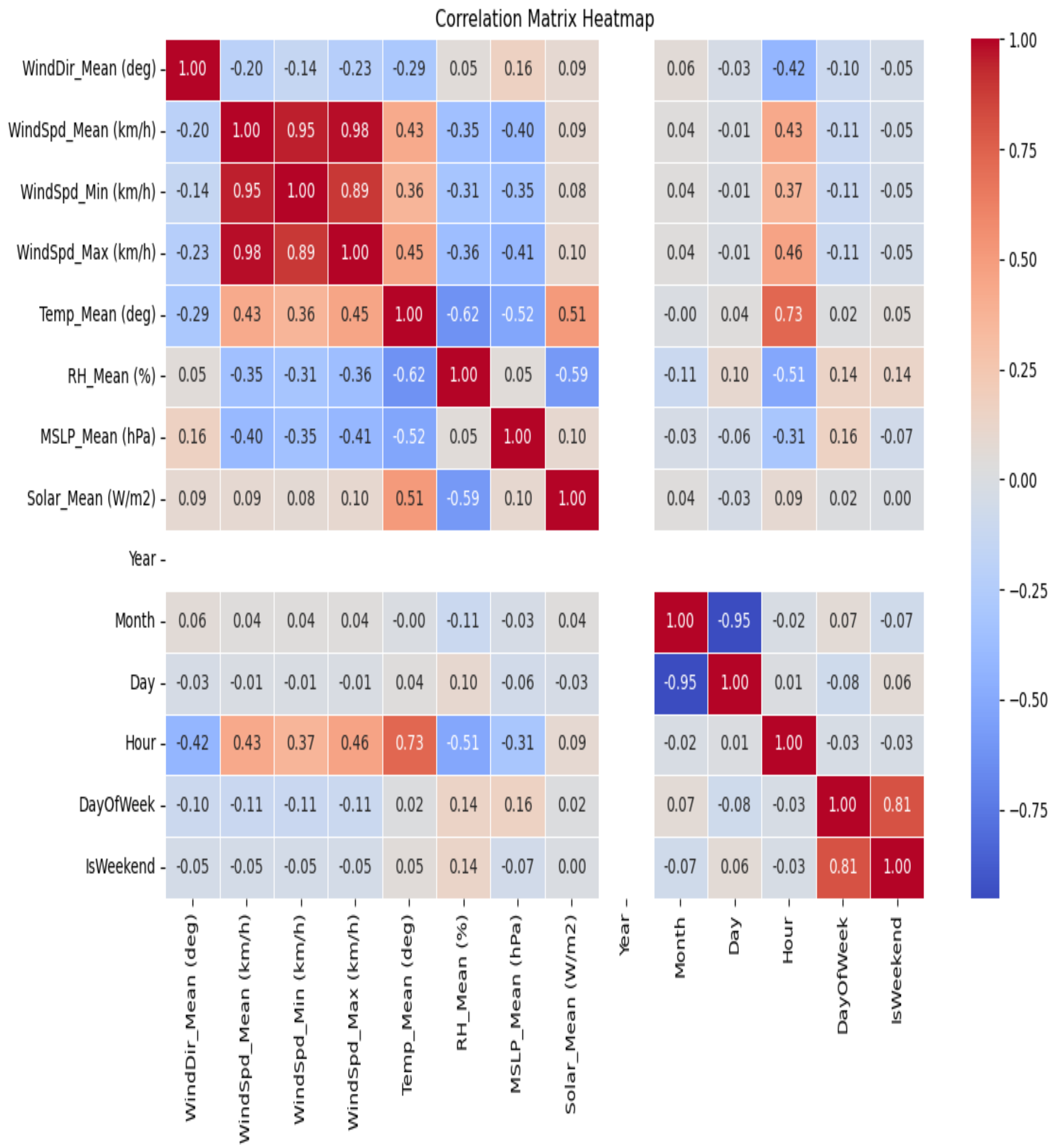


Figure 30:Full correlation matrix of training features and PV power output (Power_PACE).

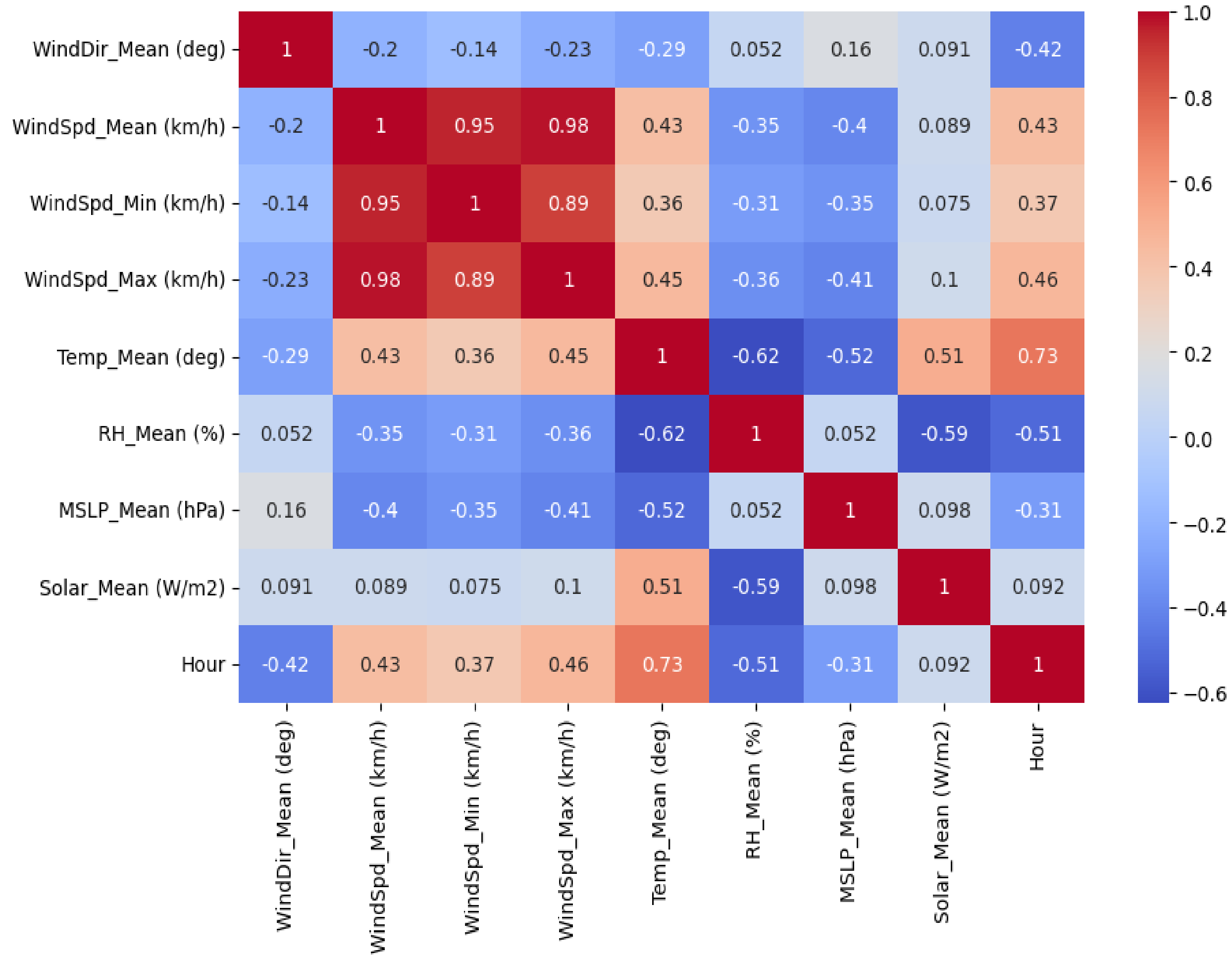


Figure 31: Correlation heatmap of meteorological features only (excluding the target variable).

Solar_Mean (solar irradiance, W/m²) holds the strongest positive linear correlation with Power_PACE, as expected from the fundamental physics of photovoltaic conversion. Temp_Mean exhibits a secondary positive correlation due to the co-occurrence of high temperatures and high irradiance during summer days. Wind speed variables show moderate correlations, while RH_Mean and MSLP_Mean display weaker associations, acting as proxies for atmospheric transparency.

4.2 Error Metrics for Each Model

Table 3 summarises the quantitative performance of all five models on the test set, ranked by descending R^2 score.

Table 3: Comparative evaluation of five regression models on the PV test set (sorted by R^2 score, descending).

Model	MAE (W)	RMSE (W)	MSE (W^2)	R^2 Score
Linear Regression	4,855.75	5,953.04	35,438,734	0.9777
Random Forest	4,232.04	6,899.35	47,601,084	0.9701
CatBoost	5,362.85	7,562.97	57,198,493	0.9640
XGBoost	6,096.51	10,040.53	100,811,285	0.9366
Gradient Boosting	8,855.64	15,126.04	228,797,068	0.8561

Linear Regression achieves the highest R^2 (0.9777) with the lowest RMSE (5,953 W), reflecting the dominant linear relationship between solar irradiance and power output in the summer dataset. Random Forest ranks second with the lowest MAE across all models (4,232 W). CatBoost and XGBoost achieve intermediate performance, while Gradient Boosting records the weakest results with an RMSE more than 2.5 times larger than Linear Regression.

4.3 Comparison of Model Performance — Charts

Figure 31 presents the R^2 scores for all five models as a bar chart, clearly showing the ranking from best to worst. Figures 4.6 and 4.7 compare the RMSE and MAE values through grouped bar charts.

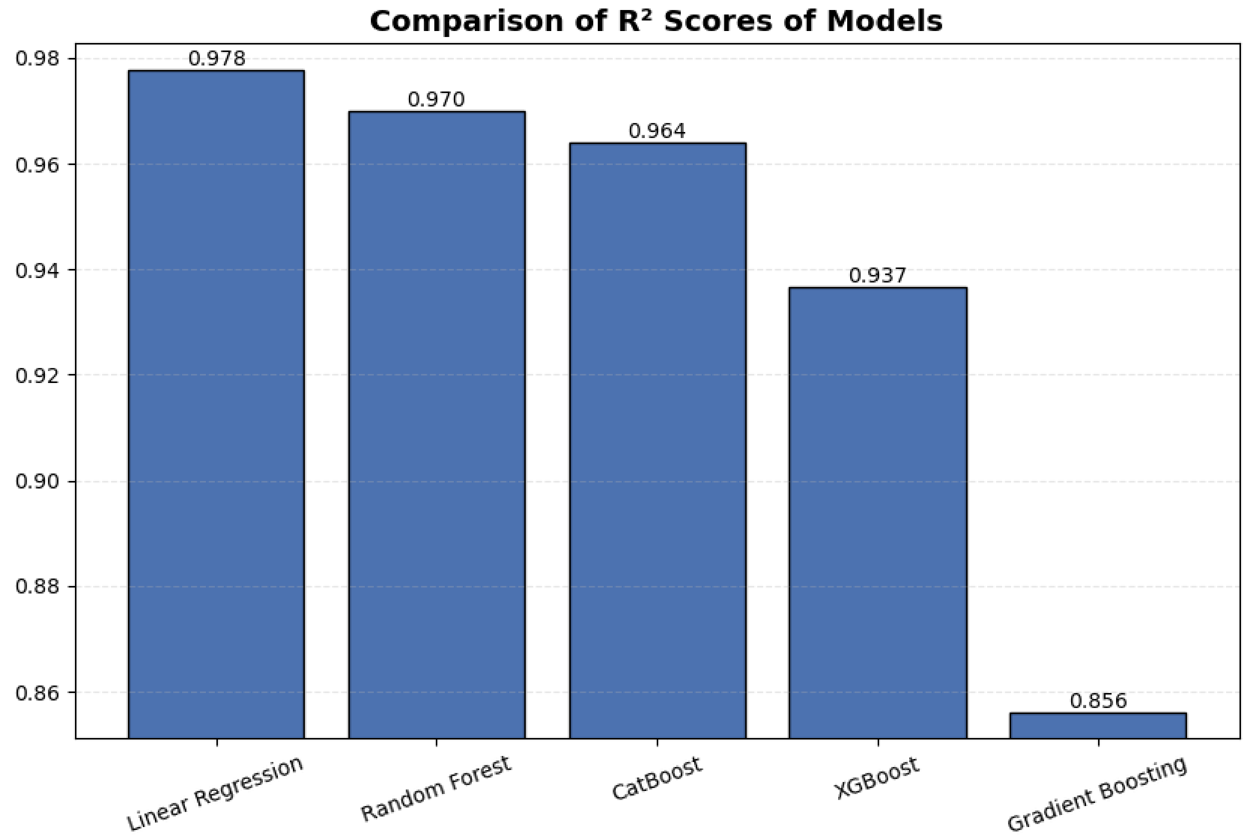


Figure 32: R² score comparison across all five models (higher is better).

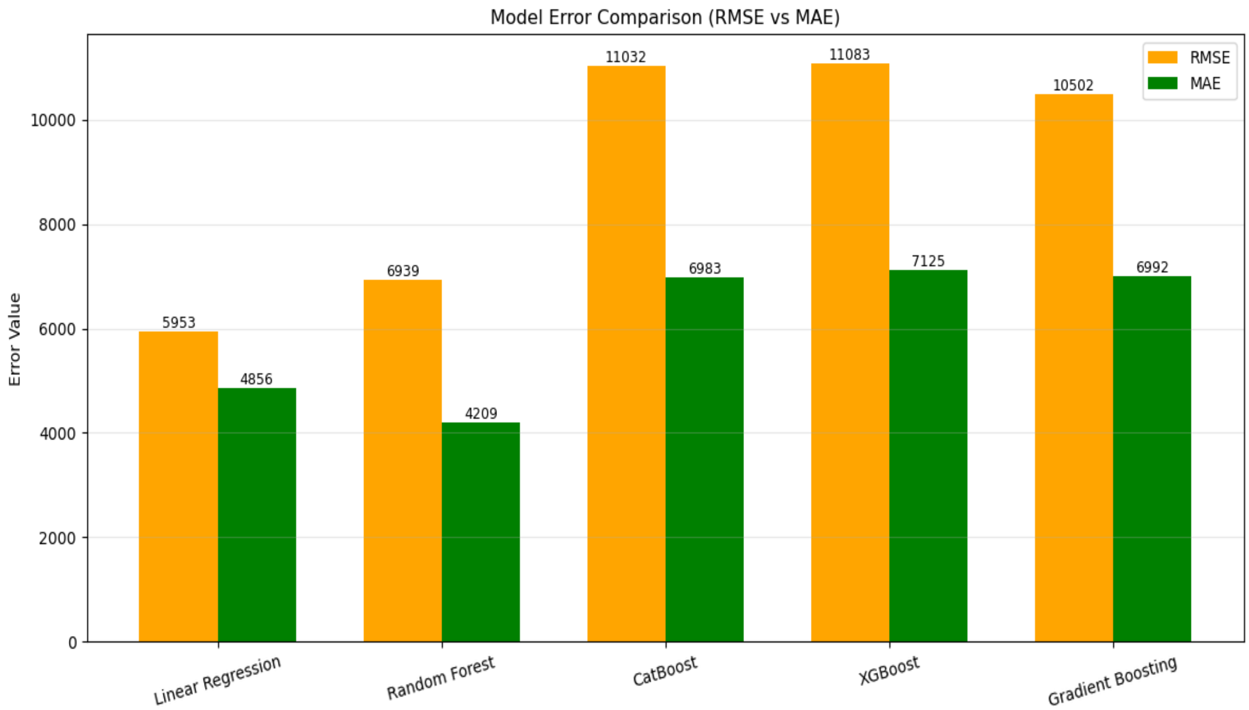


Figure 33: Grouped bar chart comparing RMSE and MAE values for all models .

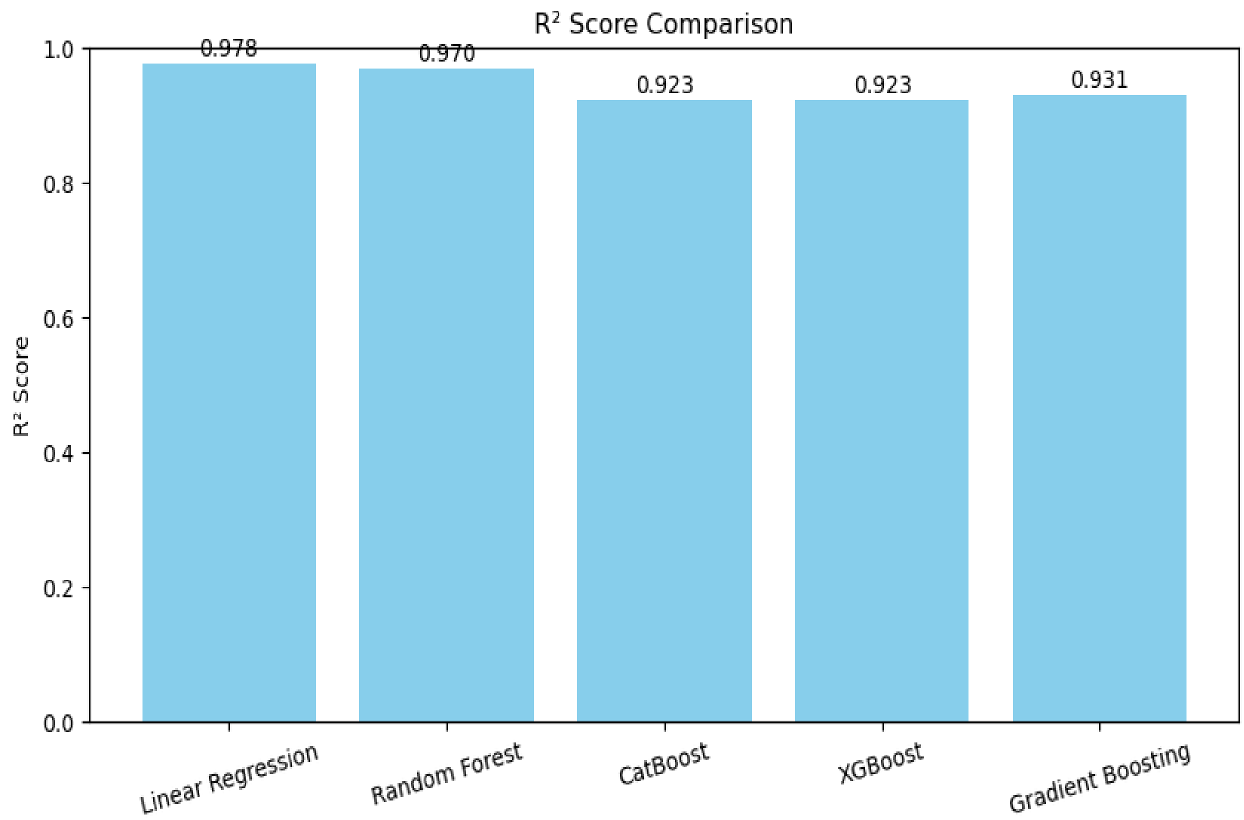


Figure 34: R^2 score comparison — final model evaluation.

4.4 Feature Importance SHAP Analysis

SHAP (SHapley Additive exPlanations) values were computed using TreeExplainer applied to the trained model on the test set. The SHAP summary plot (Figure 34) ranks features by their mean absolute SHAP contribution, with each dot representing one test sample coloured by the corresponding feature value (red = high, blue = low). The bar chart (Figure 34) aggregates these contributions into a global feature importance ranking.

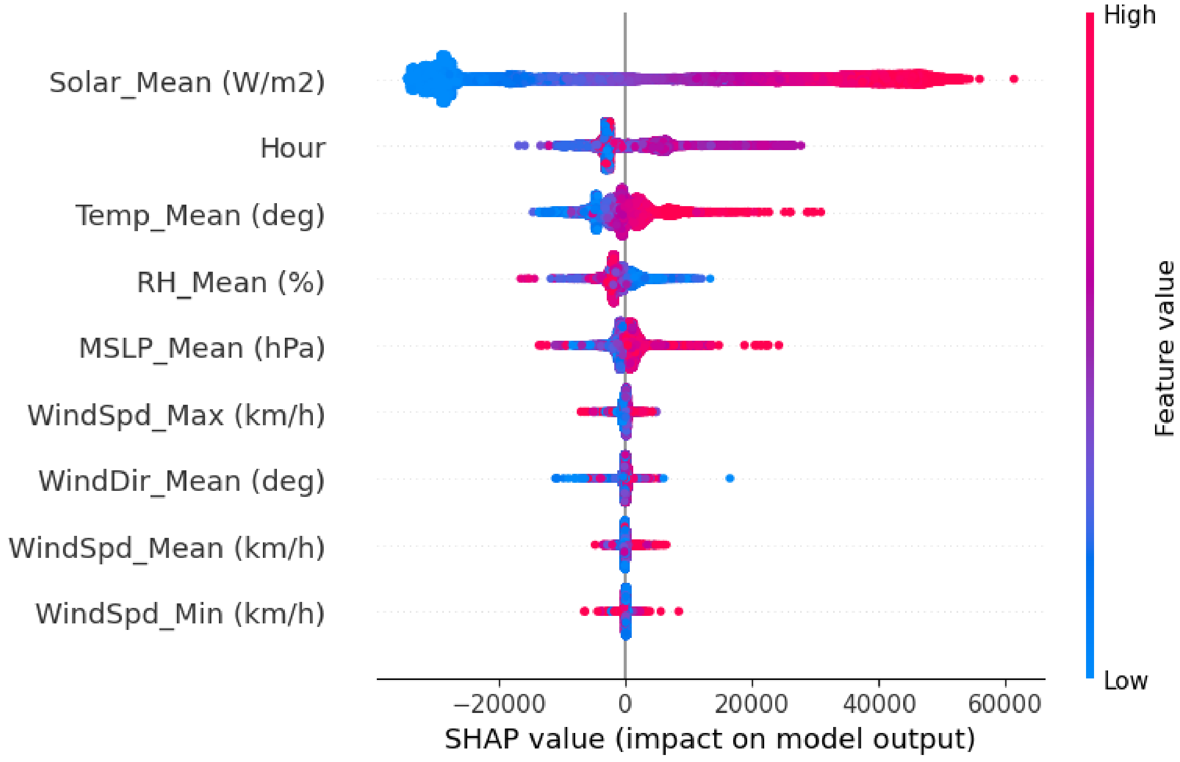


Figure 35: SHAP summary plot — feature contributions to individual predictions. Each point is one test sample; colour indicates feature value magnitude.

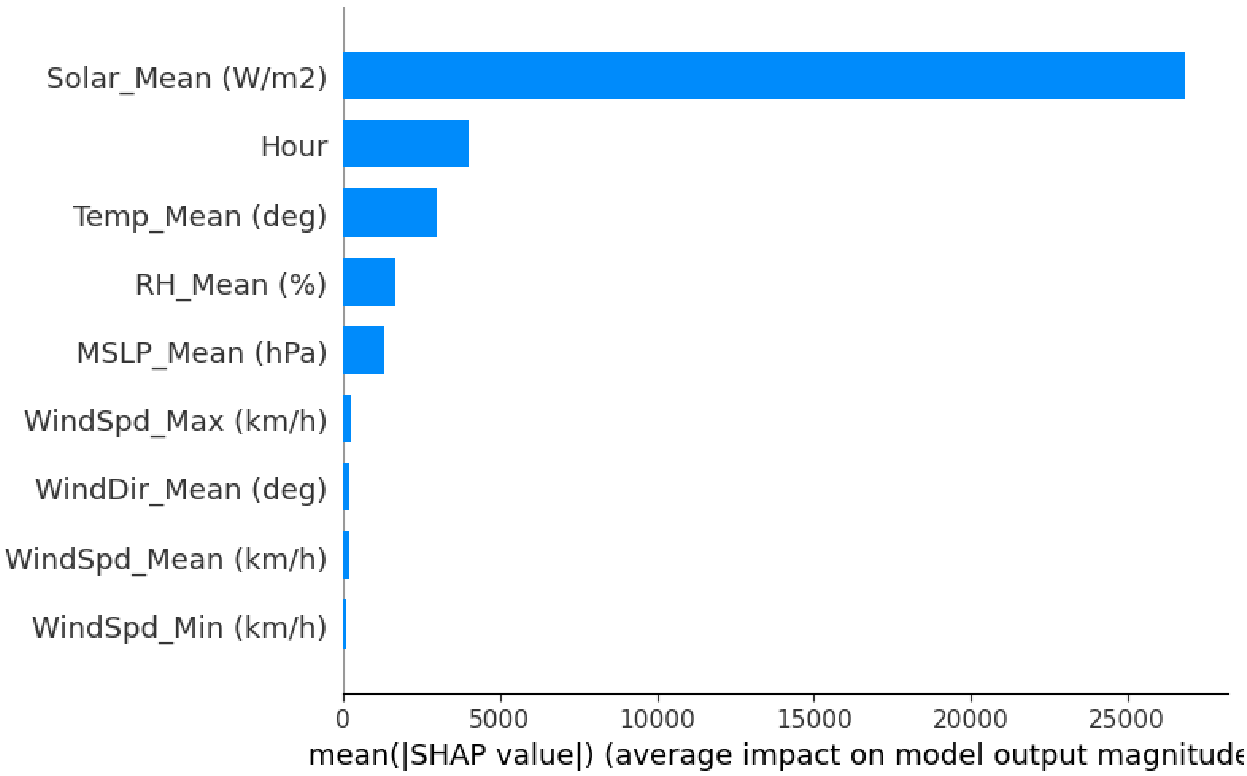


Figure 36: SHAP mean absolute feature importance bar chart — global ranking of predictor variables.

Solar_Mean emerges as the dominant predictor by a substantial margin, followed by Temp_Mean and the hour-of-day temporal feature. Wind speed variables contribute modestly, while RH_Mean, MSLP_Mean, wind direction, and pressure-related features show comparatively small SHAP magnitudes. These findings are consistent with established photovoltaic physics and validate the input feature selection.

5. Discussion

5.1 Strengths and Weaknesses of the Best-Performing Models Linear Regression

The strong performance of Linear Regression reflects the quasi-linear relationship between solar irradiance and power output during a clear-sky dominated summer period. Its principal strengths are interpretability, computational efficiency, and robustness against overfitting due to its low model complexity. However, this model is inherently unable to capture nonlinear interactions that arise under partial clouding, rapid irradiance transients, or off-peak conditions. It will likely underperform during autumn or winter months characterised by lower and more variable irradiance levels.

Random Forest

Random Forest achieves competitive accuracy (second overall R^2, lowest MAE) while naturally modelling nonlinear feature interactions through its ensemble averaging mechanism. Its built-in feature importance mechanism provided interpretable insights into variable contributions. The primary weakness is its relatively higher RMSE compared to Linear Regression, indicating occasional larger errors on individual peak-production samples — attributable to the sparse density of very high power values in the training set.

Gradient Boosting (standard)

The underperformance of standard Gradient Boosting relative to CatBoost and XGBoost can be attributed to its sequential learning approach without the regularisation and parallelism advantages of more advanced implementations. With default parameters, the model may suffer from insufficient depth or suboptimal learning rate settings that prevent it from fully capturing the complexity of the irradiance-power relationship.

5.2 Limitations Related to Data, Weather Variability, and Seasonality

Several limitations must be acknowledged. First, the dataset covers only two summer months (July and August 2018), constraining generalisability to other seasons. In Algeria, winter months experience significantly lower solar elevation angles, reduced irradiance levels, and more frequent cloud cover. A model trained exclusively on summer data will exhibit substantially degraded accuracy when deployed in autumn or winter conditions.

Second, the minute-level temporal resolution introduces fine-grained variability difficult to capture with standard regression models. Rapid irradiance fluctuations caused by cloud passage events sometimes producing power drops of tens of thousands of watts within seconds represent an inherently stochastic signal challenging to predict from meteorological features alone.

Third, the absence of weather forecast inputs means all models operate in a nowcasting mode, using measured rather than forecast meteorological values. In real deployment, future meteorological values would need to be obtained from NWP models or local sensors, introducing additional uncertainty. Finally, the single-site scope limits transferability to other locations with different panel configurations, orientations, and local microclimates.

5.3 Robustness and Real-World Applicability

Despite these limitations, the results demonstrate that even simple Machine Learning models achieve high predictive accuracy ($R^2 > 0.97$ for the top two models) when the dataset is dominated by a strong linear driver such as solar irradiance. Random Forest stands out as the most robust choice for real-world deployment: its lower MAE indicates fewer systematic biases, its tolerance to noisy measurements suits operational environments, and its feature importance output provides plant operators with actionable insights into which meteorological conditions drive forecasting uncertainty.

Linear Regression, despite its top performance on this dataset, may not be the most appropriate choice for year-round deployment due to its inability to adapt to nonlinear seasonal dynamics. A practical recommendation is to use Linear Regression as a benchmark and deploy Random Forest or a well-tuned CatBoost model for production forecasting.

6. Conclusion

This chapter presented a comprehensive simulation and analysis of five Machine Learning models applied to short-term photovoltaic power forecasting at the PACE solar farm. The models were trained on 20,168 minute-level observations and evaluated on 5,042 test samples using MAE, RMSE, and R^2 metrics.

Linear Regression achieved the highest R^2 (0.9777) and lowest RMSE (5,953 W), reflecting the dominant linear irradiance-power relationship in the summer dataset. Random Forest ranked second with the lowest MAE (4,232 W), demonstrating its robustness and averaging advantages. CatBoost and XGBoost achieved intermediate performance, while standard Gradient Boosting lagged significantly. SHAP analysis confirmed Solar_Mean as the primary predictor, followed by ambient temperature and hour of day.

The identified limitations — seasonal restriction of training data, absence of forecast inputs, and minute-level stochastic noise from cloud events — define clear directions for future improvement. Proposed enhancements include NWP forecast integration, satellite or sky-camera cloud data, and stacking ensemble or hybrid physical–statistical architectures. Together, these improvements would move the system from a retrospective validation tool toward a fully operational short-term forecasting platform for real-world grid management in the Algerian context.

General Conclusion

General Conclusion

The increasing integration of photovoltaic energy into modern power systems has created a growing need for accurate and reliable forecasting methods. Due to the intermittent nature of solar energy and its strong dependence on meteorological conditions, predicting photovoltaic production remains a challenging task. In this context, Artificial Intelligence (AI) and Machine Learning (ML) techniques offer powerful solutions capable of modeling the complex relationships between environmental variables and photovoltaic power output.

The objective of this thesis was to develop and compare different Machine Learning models for the short-term prediction of photovoltaic production using meteorological and temporal data. A complete data-driven prediction framework was developed, including data preprocessing, feature extraction from timestamps, training and testing of prediction models, and performance evaluation using statistical indicators such as MAE, RMSE, and R^2.

Five regression algorithms were investigated: Linear Regression, Random Forest, Gradient Boosting, XGBoost, and CatBoost. The obtained results demonstrated that ensemble tree-based models provide superior performance compared with traditional linear approaches due to their ability to capture nonlinear interactions between solar irradiance, temperature, humidity, wind conditions, and temporal factors. Among the tested models, XGBoost achieved the best overall performance with the highest coefficient of determination ($R^2 = 0.977592$), while Random Forest also showed highly competitive results with the lowest MAE value.

Furthermore, SHAP analysis was employed to interpret the behavior of the developed models and identify the contribution of each input variable to the final prediction. The interpretation confirmed that solar irradiance, temperature, and temporal features were among the most influential parameters affecting photovoltaic production, which is consistent with the physical behavior of PV systems.

Overall, the results obtained in this study confirm that Artificial Intelligence and Machine Learning represent effective and reliable approaches for short-term photovoltaic power forecasting. The developed framework can contribute to improved energy management, better integration of renewable energy sources, and enhanced stability of modern electrical networks.

Perspective

As future perspectives, this work can be extended by incorporating larger historical datasets, additional meteorological information, hyperparameter optimization techniques, advanced Deep Learning architectures, and hybrid forecasting models to further improve prediction accuracy and robustness.

Bibliographic references

Bibliographic References

[1] "World Energy Outlook 2021 – Analysis," IEA. Accessed: Apr. 21, 2026. [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2021

[2] S. Legg, "IPCC, 2021: Climate Change 2021 - the Physical Science basis," *Interaction*, vol. 49, no. 4, pp. 44–45, Dec. 2021, doi: 10.3316/informit.315096509383738.

[3] "Global deployment of solar photovoltaics: Its opportunities and challenges | IEEE Conference Publication | IEEE Xplore." Accessed: Apr. 21, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7856217

[4] G. F. Nemet, "Beyond the learning curve: factors influencing cost reductions in photovoltaics," *Energy Policy*, vol. 34, no. 17, pp. 3218–3232, Nov. 2006, doi: 10.1016/j.enpol.2005.06.020.

[5] "Our Future Couldn't Be Brighter - Greater Houston Port Bureau." Accessed: Apr. 21, 2026. [Online]. Available: https://www.txgulf.org/news/our-future-couldnt-be-brighter

[6] Y. Kherbiche, N. Ihaddadene, R. Ihaddadene, F. Hadji, J. Mohamed, and A. H. Beghidja, "Solar Energy Potential Evaluation. Case of Study: M'Sila, an Algerian Province," *IJSDP*, vol. 16, no. 8, pp. 1501–1508, Dec. 2021, doi: 10.18280/ijsdp.160811.

[7] K. Abdeladim, A. Razagui, S. Semaoui, and A. Hadj Arab, "Updating Algerian solar atlas using MEERA-2 data source," *Energy Reports*, vol. 6, pp. 281–287, Feb. 2020, doi: 10.1016/j.egyr.2019.08.057.

[8] D. K. Chaturvedi, "Solar Power Forecasting: A Review," *International Journal of Computer Applications*, vol. 145.

[9] "Solar Energy Forecasting Using Machine Learning Techniques for Enhanced Grid Stability | IEEE Journals & Magazine | IEEE Xplore." Accessed: Apr. 21, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11015947

[10] P. N. L. Mohamad Radzi, M. N. Akhter, S. Mekhilef, and N. Mohamed Shah, "Review on the Application of Photovoltaic Forecasting Using Machine Learning for Very Short- to Long-Term Forecasting," *Sustainability*, vol. 15, no. 4, p. 2942, Jan. 2023, doi: 10.3390/su15042942.

[11] F. Harrou, A. Dairi, A. Dorbane, B. Taghezouit, and Y. Sun, "Semi-supervised anomaly detection in photovoltaic systems under power tracking mode," *Energy Conversion and Management*, vol. 352, p. 121114, Mar. 2026, doi: 10.1016/j.enconman.2026.121114.

[12] F. Harrou, A. Dorbane, and Y. Sun, "Enhancing photovoltaic system performance using advanced gradient boosting methods," in *2024 International Conference on Control, Automation and Diagnosis (ICCAD)*, May 2024, pp. 1–6. doi: 10.1109/ICCAD60883.2024.10553919.

[13] F. Obeidat, "A comprehensive review of future photovoltaic systems," *Solar Energy*, vol. 163, pp. 545–551, Mar. 2018, doi: 10.1016/j.solener.2018.01.050.

[14] D. Hussain, S. Hossain, J. Talukder, and A. Mia, "Solar Energy Integration into Smart Grids: Challenges and Opportunities," vol. 2024, 2024.

[15] R. L. Cummerow, "Photovoltaic Effect in p–n Junctions," in *Renewable Energy*, Routledge, 2011.

[16] "What Is the Photovoltaic Effect?" Accessed: Apr. 21, 2026. [Online]. Available: https://www.ecoflow.com/us/blog/what-is-the-photovoltaic-effect

[17] "Figure 3.1: pn junction solar cell," ResearchGate. Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/figure/pn-junction-solar-cell_fig7_292349306

[18] M. I. Hossain, A. Bousselham, F. H. Alharbi, and N. Tabet, "Computational analysis of temperature effects on solar cell efficiency," *J Comput Electron*, vol. 16, no. 3, pp. 776–786, Sep. 2017, doi: 10.1007/s10825-017-1016-5.

[19] "Fig. 1 PV module voltage-power at different irradiance levels.," ResearchGate. Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/figure/PV-module-voltage-power-at-different-irradiance-levels_fig25_236980546

[20] C. W. Hansen, “Estimation of parameters for single diode models using measured IV curves,” in *2013 IEEE 39th Photovoltaic Specialists Conference (PVSC)*, Jun. 2013, pp. 0223–0228. doi: 10.1109/PVSC.2013.6744135.
[21] “Fig. 1. Current voltage (IV) curve of a solar cell [26]. Operation at...,” ResearchGate. Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/figure/Current-voltage-IV-curve-of-a-solar-cell-26-Operation-at-the-point-of-maximum-power_fig1_369916302
[22] D. M. Fébba, R. M. Rubinger, A. F. Oliveira, and E. C. Bortoni, “Impacts of temperature and irradiance on polycrystalline silicon solar cells parameters,” *Solar Energy*, vol. 174, pp. 628–639, Nov. 2018, doi: 10.1016/j.solener.2018.09.051.
[23] “Fig. 4 I-V curve of PV module and various resistive loads.,” ResearchGate. Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/figure/curve-of-PV-module-and-various-resistive-loads_fig29_236980546
[24] “Effect of Temperature on the I-V Characteristics of a Polycrystalline Solar Cell | Journal of Nepal Physical Society.” Accessed: Apr. 21, 2026. [Online]. Available: https://nepjol.info/index.php/jnphyssoc/article/view/14440
[25] “Figure 15. Temperature dependence of I-V characteristics of a solar cell.,” ResearchGate. Accessed: Apr. 21, 2026. [Online]. Available: https://www.researchgate.net/figure/Temperature-dependence-of-I-V-characteristics-of-a-solar-cell_fig13_261362481
[26] G. Vázquez, “How we almost came in first place at a Google Hackathon,” Medium. Accessed: Apr. 21, 2026. [Online]. Available: https://gonzalovazquez.medium.com/how-we-almost-came-in-first-place-at-a-google-hackathon-19460e23dfad
[27] S. Mishra and P. Palanisamy, “Multi-time-horizon Solar Forecasting Using Recurrent Neural Network,” in *2018 IEEE Energy Conversion Congress and Exposition (ECCE)*, Sep. 2018, pp. 18–24. doi: 10.1109/ECCE.2018.8558187.
[28] “Market prices forecasting as part of an energy sales strategy - AleaSoft Energy Forecasting.” Accessed: Jun. 02, 2026. [Online]. Available: https://aleasoft.com/market-prices-forecasting-part-energy-sales-strategy/
[29] “Hybrid deep learning models for time series forecasting of solar power | Neural Computing and Applications | Springer Nature Link.” Accessed: Apr. 21, 2026. [Online]. Available: https://link.springer.com/article/10.1007/s00521-024-09558-5
[30] C. Voyant *et al.*, “Machine learning methods for solar radiation forecasting: A review,” *Renewable Energy*, vol. 105, pp. 569–582, May 2017, doi: 10.1016/j.renene.2016.12.095.
[31] D. Gong, “Top 6 Machine Learning Algorithms for Classification,” TDS Archive. Accessed: Jun. 21, 2026. [Online]. Available: https://medium.com/data-science/top-machine-learning-algorithms-for-classification-2197870ff501
[32] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[33] “Fig. 5 Random Forest as a combination of multiple Decision Trees,” ResearchGate. Accessed: Jun. 21, 2026. [Online]. Available: https://www.researchgate.net/figure/Random-Forest-as-a-combination-of-multiple-Decision-Trees_fig3_336436894
[34] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[35] A. Ekinci, O. Baytar, Ö. Şahin, and M. Ş. Üney, “Green-synthesized Co–Zr-based nanocatalysts for hydrogen generation from NaBH4 hydrolysis with machine learning prediction,” *International Journal of Hydrogen Energy*, vol. 247, p. 155897, Jul. 2026, doi: 10.1016/j.ijhydene.2026.155897.
[36] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, “Support Vector Regression Machines,” in *Advances in Neural Information Processing Systems*, MIT Press, 1996. Accessed: Apr. 21, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html
[37] “Figure 15: Hypothetical support vector regression scatter plot The SVC...,” ResearchGate. Accessed: Jun. 21, 2026. [Online]. Available: https://www.researchgate.net/figure/Hypothetical-

support-vector-regression-scatter-plot-The-SVC-uses-a-hyperplane-to-separate_fig26_294764549
[38] S. M. Radhi, S. D. Al-Majidi, M. F. Abbod, and H. S. Al-Raweshidy, “Machine Learning Approaches for Short-Term Photovoltaic Power Forecasting,” *Energies*, vol. 17, no. 17, p. 4301, Jan. 2024, doi: 10.3390/en17174301.
[39] H. Wang *et al.*, “Short-Term Photovoltaic Power Forecasting Based on a Feature Rise-Dimensional Two-Layer Ensemble Learning Model,” *Sustainability*, vol. 15, no. 21, p. 15594, Jan. 2023, doi: 10.3390/su152115594.
[40] X. Xiang, X. Li, Y. Zhang, and J. Hu, “A short-term forecasting method for photovoltaic power generation based on the TCN-ECANet-GRU hybrid model,” *Sci Rep*, vol. 14, no. 1, p. 6744, Mar. 2024, doi: 10.1038/s41598-024-56751-6.
[41] T. Nguyen Trong, H. Vu Xuan Son, H. Do Dinh, H. Takano, and T. Nguyen Duc, “Short-term PV power forecast using hybrid deep learning model and Variational Mode Decomposition,” *Energy Reports*, vol. 9, pp. 712–717, Oct. 2023, doi: 10.1016/j.egyr.2023.05.154.
[42] M. S. Ibrahim, S. M. Gharghory, and H. A. Kamal, “A hybrid model of CNN and LSTM autoencoder-based short-term PV power generation forecasting,” *Electr Eng*, vol. 106, no. 4, pp. 4239–4255, Aug. 2024, doi: 10.1007/s00202-023-02220-8.
[43] N. Dimitropoulos *et al.*, “Comparative analysis of AI-based models for short-term photovoltaic power forecasting in energy cooperatives,” *Intelligent Decision Technologies*, vol. 15, no. 4, pp. 691–705, Nov. 2021, doi: 10.3233/IDT-210210.
[44] T. A. Nguyen *et al.*, “A Study on Novel Solar Power Forecasting using an XGB-LiGBM-RF Hybrid Model and the L-BFGS-B Optimization Algorithm,” *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 24516–24522, Aug. 2025, doi: 10.48084/etasr.11308.
[45] E. Sarmas, E. Spiliotis, E. Stamatopoulos, V. Marinakis, and H. Doukas, “Short-term photovoltaic power forecasting using meta-learning and numerical weather prediction independent Long Short-Term Memory models,” *Renewable Energy*, vol. 216, p. 118997, Nov. 2023, doi: 10.1016/j.renene.2023.118997.
[46] M. Boujoudar *et al.*, “Power PV Forecasting using Machine Learning Algorithms Based on Weather Data in Semi-Arid Climate,” *BIO Web Conf.*, vol. 109, p. 01024, 2024, doi: 10.1051/bioconf/202410901024.
[47] Y. Ma, W. Yu, J. Zhu, Z. You, and A. Jia, “Research on ultra-short-term photovoltaic power forecasting using multimodal data and ensemble learning,” *Energy*, vol. 330, no. C, 2025, Accessed: Jun. 02, 2026. [Online]. Available: https://ideas.repec.org//a/eee/energy/v330y2025ics0360544225024739.html
[48] M. N. Akhter *et al.*, “An Hour-Ahead PV Power Forecasting Method Based on an RNN-LSTM Model for Three Different PV Plants,” *Energies*, vol. 15, no. 6, p. 2243, Jan. 2022, doi: 10.3390/en15062243.
[49] P. Li, K. Zhou, X. Lu, and S. Yang, “A hybrid deep learning model for short-term PV power forecasting,” *Applied Energy*, vol. 259, p. 114216, Feb. 2020, doi: 10.1016/j.apenergy.2019.114216.
[50] A. Agga, A. Abbou, M. Labbadi, and Y. El Houm, “Short-term self consumption PV plant power production forecasts based on hybrid CNN-LSTM, ConvLSTM models,” *Renewable Energy*, vol. 177, pp. 101–112, Nov. 2021, doi: 10.1016/j.renene.2021.05.095.
[51] E. M. Al-Ali *et al.*, “Solar Energy Production Forecasting Based on a Hybrid CNN-LSTM-Transformer Model,” *Mathematics*, vol. 11, no. 3, p. 676, Jan. 2023, doi: 10.3390/math11030676.
[52] M. Gupta, A. Arya, U. Varshney, J. Mittal, and A. Tomar, “A review of PV power forecasting using machine learning techniques,” *Progress in Engineering Science*, vol. 2, no. 1, p. 100058, Mar. 2025, doi: 10.1016/j.pes.2025.100058.
[53] “Photovoltaic and solar power forecasting for smart grid energy management.” Accessed: Jun. 20, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7377167

[54] C. Voyant *et al.*, “Machine learning methods for solar radiation forecasting: A review,” *Renewable Energy*, vol. 105, pp. 569–582, May 2017, doi: 10.1016/j.renene.2016.12.095.
[55] A. Mellit and S. A. Kalogirou, “Artificial intelligence techniques for photovoltaic applications: A review,” *Progress in Energy and Combustion Science*, vol. 34, no. 5, pp. 574–632, Oct. 2008, doi: 10.1016/j.pecs.2008.01.001.
[56] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[57] “Random Forest Algorithm in Machine Learning,” GeeksforGeeks. Accessed: Jun. 27, 2026. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/
[58] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[59] “Simplified flowchart of the sequential learning process in Gradient...,” ResearchGate. Accessed: Jun. 27, 2026. [Online]. Available: https://www.researchgate.net/figure/Simplified-flowchart-of-the-sequential-learning-process-in-Gradient-Boosting-algorithms_fig6_396366864
[60] “XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.” Accessed: Jun. 20, 2026. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2939672.2939785
[61] S. Bag, “The Complete XGBoost Therapy with Python,” Medium. Accessed: Jun. 27, 2026. [Online]. Available: https://python.plainenglish.io/the-complete-xgboost-therapy-with-python-87c8cffcb71f
[62] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2018. Accessed: Jun. 20, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html
[63] E. G. PhD, “CatBoost Unleashed: Mastering Categorical Data for Robust Predictive Modeling,” Medium. Accessed: Jun. 27, 2026. [Online]. Available: https://pub.aimind.so/catboost-unleashed-mastering-categorical-data-for-robust-predictive-modeling-ee081bf26f91
[64] V. A. Barbur, “Review of Introduction to Linear Regression Analysis.,” *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 43, no. 2, pp. 339–341, 1994, doi: 10.2307/2348362.
[65] “Linear Regression,” The Science of Machine Learning & AI. Accessed: Jun. 27, 2026. [Online]. Available: https://www.ml-science.com/linear-regression
[66] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jun. 20, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html