



NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

**COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS IN
PREDICTING SOLAR POWER IN SOLAR PHOTOVOLTAIC SYSTEMS**

M.Sc. THESIS

Goodness Chinaza ORISH

Nicosia January

2024

**GOODNESS
CHINAZA ORISH**

**COMPARATIVE ANALYSIS OF
MACHINE LEARNING MODELS
IN PREDICTING SOLAR POWER
IN SOLAR PHOTOVOLTAIC**

MASTER THESIS

2024

**NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING**

**COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS IN
PREDICTING SOLAR POWER IN SOLAR PHOTOVOLTAIC SYSTEMS**

M.Sc. THESIS

Goodness Chinaza ORISH

Supervisor



Prof. Dr. Kamil DIMILILER

Nicosia January

2024

Approval

We certify that we have read the thesis submitted by Goodness Chinaza Orish titled “Comparative Analysis of Machine Learning Models in Predicting Solar Power in Photovoltaic Systems” and that in our combined opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Sciences.

Examining Committee	Name-Surname	Signature
Head of the Committee:	Assist. Prof. Dr. Cemal Kavalcioglu	
Committee Member:	Assist. Prof. Dr. Samuel Nii Tackie	
Supervisor:	Prof. Dr. Kamil Dimililer	

Approved by Head of the Department



06.15.2024

Prof. Dr. Bülent Bilgehan
Head of Department

Approved by the Institute of Graduate Studies



Prof. Dr. Kemal Hüsnü Can Baser
Head of the Institute

Declaration

I hereby declare that all information, documents, analysis, and results in this thesis have been collected and presented according to the academic rules and ethical guidelines of the Institute of Graduate Studies, Near East University. I also declare that as required by these rules and conduct, I have fully cited and referenced information and data that are not original to this study.

Goodness Chinaza Orish

...../...../.....

Acknowledgments

I am profoundly grateful to God for His grace throughout the completion of this thesis. His presence in my life has been my source of strength and inspiration, guiding me through every challenge. To my supervisor, Prof. Dr. Kamil Dimililer, I extend my heartfelt gratitude for their invaluable guidance, patience, and encouragement.

My deepest appreciation goes to my parents, Prof. Ebere Orish Orisakwe and Dr. Chinna Orish, whose unwavering support and encouragement have been my rock during the most trying times. At moments when I was on the verge of giving up, their belief in me never wavered, providing me with the strength and determination to persevere. Their sacrifices and unwavering love have been the driving force behind my success. I also want to thank my siblings for their understanding and support during this journey. They have always inspired me and given me confidence in my abilities.

To Dr. Oluwatayomi Rereloluwa Adegboye, your unwavering love, support, and encouragement have been instrumental in the completion of my thesis. Your constant push and belief in my capabilities propelled me forward, and I am deeply grateful for your unwavering support and encouragement.

Lastly, I extend my thanks to all those who have supported me along the way, whether through words of encouragement, assistance with research, or simply being there to lend an ear. Your contributions have made a significant impact on this journey and you have my sincere gratitude for being in my life.

Goodness Chinaza Orish

Abstract

Comparative Analysis of Machine Learning Models in Predicting Solar Power in Photovoltaic Systems.

Orish Goodness Chinaza

MSc, Department of Electrical and Electronics Engineering

January 2024, 90pages

Renewable energy has become a global focus in these recent times which also includes harnessing solar power using photovoltaic systems. Photovoltaic systems supply a renewable source of energy that is environmentally friendly by converting sunlight into electricity. Due to the variability of weather conditions, photovoltaic systems are not entirely reliable, to improve energy planning and consumption an accurate system for solar power prediction is needed. In this thesis, a careful comparison of six machine learning models which are Decision Tree (DT), Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and K-Nearest Neighbors (KNN), to examine the complex issue of solar power prediction. A large meteorological dataset is used to study the patterns and relationship between the variables which are majorly the weather conditions and the output which is the solar power generated. With the use of significant error metrics which are R-squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), the models undergo a thorough evaluation to observe how well each of them handles the complex nonlinear relationships of the meteorological data, the results of this study show that Random Forest outperforms the other five models consistently over the different ratios of data splits. It consistently has the lowest error metrics of RMSE of 0.1295, MSE of 0.0168, and MAE of 0.0605 in the training and RMSE of 0.2897, MSE of 0.0839, and MAE of 0.1393 in the testing and the highest R-squared scores of 0.9216, 0.8837, and 0.8879 on all the different ratios of data splits which were 10:90, 20:80, and 30:70 respectively, which shows a great capacity and ability to adapt and reduce over-fitting problems in prediction of solar power.

Keywords: Solar Power Prediction, Photovoltaic Systems, Machine Learning Models, Random Forest, Meteorological Dataset.

Özet

Fotovoltaik Sistemlerde Güneş Enerjisi Tahmininde Makine Öğrenmesi Modellerinin Karşılaştırmalı Analizi.

Goodness Chinaza Orish

Yüksek Lisans, Elektrik-Elektronik Mühendisliği Bölümü

Ocak, 2024, 90 sayfa

Yenilenebilir enerji, fotovoltaik sistemler kullanarak güneş enerjisinden yararlanmayı da içeren bu son zamanlarda küresel bir odak noktası haline gelmiştir. Fotovoltaik sistemler, güneş ışığını elektriğe dönüştürerek çevre dostu yenilenebilir bir enerji kaynağı sağlamaktadır. Hava koşullarının değişkenliği nedeniyle, fotovoltaik sistemler tamamen güvenilir değildir, enerji planlamasını ve tüketimini iyileştirmek için güneş enerjisi tahmini için doğru bir sisteme ihtiyaç vardır. Bu tezde, güneş enerjisi tahmini gibi karmaşık bir konuyu incelemek için Karar Ağacı (DT), Doğrusal Regresyon (LR), Destek Vektör Regresyonu (SVR), Rastgele Orman (RF) ve K-En Yakın Komşular (KNN) olmak üzere altı makine öğrenimi modeli dikkatli bir şekilde karşılaştırılmıştır. Büyük bir meteorolojik veri seti, büyük ölçüde hava koşulları olan değişkenler ile üretilen güneş enerjisi olan çıktı arasındaki kalıpları ve ilişkiyi incelemek için kullanılır. R-kare (R^2), Ortalama Karesel Hata (MSE), Ortalama Karesel Hatanın Kökü (RMSE) ve Ortalama Mutlak Hata (MAE) gibi önemli hata ölçütlerinin kullanılmasıyla, modeller, her birinin meteorolojik verilerin karmaşık doğrusal olmayan ilişkilerini ne kadar iyi ele aldığını gözlemlemek için kapsamlı bir değerlendirmeye tabi tutulur, bu çalışmanın sonuçları Rastgele Orman'ın farklı veri bölme oranlarında tutarlı bir şekilde diğer beş modelden daha iyi performans gösterdiğini göstermektedir. Eğitimde 0,1295 RMSE, 0,0168 MSE ve 0,0605 MAE ile en düşük hata metriklerine ve testte 0,2897 RMSE, 0,0839 MSE ve 0,1393 MAE ile

0,9216, 0,8837 ve 0,8878 ile en yüksek R-kare skorlarına sahiptir. Sırasıyla 10:90, 20:80 ve 30:70 olan tüm farklı veri bölme oranlarında 8837 ve 0.8879, bu da güneş enerjisinin tahmininde aşırı uyum sorunlarını uyarılma ve azaltma konusunda büyük bir kapasite ve yetenek göstermektedir.

Anahtar Kelimeler: Güneş Enerjisi Tahmini, Fotovoltaik Sistemler, Makine Öğrenmesi Modelleri, Rastgele Orman, Meteorolojik Veri Kümesi.

Table of Contents

Approval.....	1
Declaration	3
Acknowledgments.....	4
Abstract	5
Table of Content.....	9
List of Figures	12
List of Tables	14
List of Abbreviations	15

CHAPTER I

Introduction.....	16
Background of the study.....	16
Purpose of the Study.....	16
Research Question.....	17
Significance of the Study	18
Limitations.....	19
Structure of Research	20

CHAPTER II

Literature Review.....	21
Theoretical Framework on Renewable Energy Theories	21
Types of Photovoltaic Systems	24
Methods of Solar Power Prediction.....	28

Application of Solar Photovoltaic Power Prediction in Smart Grid	29
Machine Learning.....	32
Supervised Learning.....	33
Unsupervised Learning	34
Semi Supervised Learning	35
Reinforcement Learning.....	36
Related Research	36

CHAPTER III

Methodology	43
Simple Linear Regression	43
Support Vector Machine	44
Hard-Margin SVM Classification	46
KNN Classifier	48
Multi-Class K-Nearest Neighbors.....	50
KNN Regression	50
Decision Tree Regressor	52
Adaboost Regressor.....	54
Random Forest	57
Data Collection and Statistical Description of Dataset	59
Framework.....	60
Performance Metric	64

Mean Absolute Error (MAE):	64
Root Mean Square Error (RMSE):.....	64
Coefficient of Determination (R^2):.....	64
Mean Squared Error (MSE):	65

CHAPTER IV

Results and Discussion.....	66
Performance Analysis of Models using Error Evaluation Metrics:.....	67
Training Dataset	67
Testing Dataset.....	68
Performance Analysis of Models Using Accuracy Evaluation Metrics	70
Performance Analysis of Models on Observed Versus Predicted Data	73

CHAPTER V

Conclusion.....	77
Recommendation.....	78
Future Works.....	78
REFERENCES.....	80
APPENDICES.....	89

List of Figures

Figure 1. Various solar energy according to availability on the global market	21
Figure 2. Direct, Diffuse & Reflected Radiation	24
Figure 3. Grid-Connected Solar Photovoltaic Systems.....	25
Figure 4. Stand-Alone or Off-Grid PV Systems	26
Figure 5. Hybrid Solar PV Systems	28
Figure 6. The Architecture of Solar radiation predicting System.....	29
Figure 7. Minimizing RSS with Ordinary Least Squares (OLS) fit.....	44
Figure 8. Two-dimensional linear classifiers (hyperplane).....	45
Figure 9. Hyper-plane for SVR.....	46
Figure 10. A evaluation of KNN classification on two Gaussian-based data clouds for different neighborhood sizes ((a) $K = 1$ and (b) $K = 20$). While KNN ignores tiny agglomerations of patterns in favor of generalizing for bigger K , it over-fits and becomes local in in small neighborhoods	49
Figure 11. K nearest neighbor algorithm with $K=3$ and $K = 6$	50
Figure 12. Shows a uniform KNN regression for the values of (a) $K = 2$	52
Figure 13. A decision tree structure illustration.....	53
Figure 14. An example decision tree and a schematic design of Random Forest.....	59
Figure 15. The Location of 21 Photovoltaic systems in Germany.....	60
Figure 16. Machine Learning Models and Evaluation: Comparing SVR, LR, ADABOOST, RT, KNN, and DT	60
Figure 17. R2 Score results using 10% of the Dataset for Test	71
Figure 18. R2 Score results using 20% of the Dataset for Test	71
Figure 19. Figure 18: R2 Score results using 30% of the Dataset for Test.....	72

Figure 20. Predicted vs Observed using 10% of the Dataset for Test.....	74
Figure 21. Predicted vs Observed using 20% of the Dataset for Test.....	75
Figure 22. Predicted vs Observed using 30% of the Dataset for Test.....	76

List of Tables

Table 1. Error Metric Results of Training using 10% of the Dataset for Test	68
Table 2. Error Metric Results of Training using 20% of the Dataset for Test	68
Table 3. Error Metric Results of Training using 30% of the Dataset for Test	68
Table 4. Error Metric Results of Testing using 10% of the Dataset for Test	69
Table 5. Error Metric Results of Testing using 20% of the Dataset for Test	69
Table 6. Error Metric Results of Testing using 30% of the Dataset for Test	70
Table 7. R2 Score results using 10% of the Dataset.....	72
Table 8. R2 Score results using 20% of the Dataset.....	72
Table 9. R2 Score results using 30% of the Dataset.....	73

List of Abbreviations

ADABOOST: Adaptive Boosting

AI: Artificial Intelligence

DL: Deep Learning

KNN: K-Nearest Neighbor

LR: Linear Regression

ML: Machine Learning

PV: Photovoltaic

RF: Random Forest

SVR: Support Vector Regression

CHAPTER I

Introduction

Background of the study

In Germany, the proliferation of solar micro-grid systems is evident, with some integrated into the national grid and others contemplating future connections without prior studies. The dynamic nature of solar PV power generation significantly impacts system planning, operation, and economic analysis. Improved prediction methods for smart grid systems would be advantageous. Practical scenarios differ, influencing PV panel performance. Hence, it is crucial to explore the relationship between solar PV power output and external environmental factors such as solar irradiance, cell temperature, and wind speed. These uncertain factors contribute to power output uncertainty. A precise estimation of generated power is vital for investors incorporating such resources into the grid. This research aims to develop a solar PV data prediction model based on solar irradiance, emphasizing low complexity and acceptable modeling accuracy. Numerous studies in the literature have investigated PV characteristics for modeling and predicting PV power. With the rapid growth of PV system technology, a comprehensive understanding and research of PV system performance and accurate power output prediction is crucial (Ozerdem, Tackie, & Biricik, 2015).

Purpose of the Study

The purpose of this study is to research solar PV power prediction utilizing machine learning techniques (MLT) for the photo voltaic system. It aims to specifically analyze data, comparing solar PV power and irradiance and other parameters that affect the

generation of solar power. Including wind and temperature to determine how these variables relate to one another and affect solar power generation. Another goal is to create a solar PV power predicting model comparing different Machine learning models and evaluate them by using various evaluation metrics to compare the predicted and actual power. By comparing the results of these different machine learning models we will be able to suggest a very accurate model for predicting solar PV power. In achieving the aforementioned objectives, the study aims to instill confidence among stakeholders, attract investments, and catalyze the widespread adoption of solar microgrid systems. Thereby contributing to sustainable energy transitions. This contribution aligns with the broader goal of fostering a sustainable transition in the energy landscape.

Research Question

- How do external elements in the environment, such as solar irradiance, temperature, and wind speed impact the output of solar PV power?
- Can machine learning algorithms, specifically KNN, SVR, Random Forest, Decision Tree, Linear Regression and AdaBoost accurately predict solar PV power output?
- What is the relative performance of the chosen machine learning algorithms concerning precision, intricacy and applicability for practical deployment?
- How viable is the incorporation of the formulated prediction model into existing photovoltaic systems to enhance solar energy management?

Significance of the Study

- **Strategic Decision-Making for Stakeholders:** By enhancing the accuracy of solar PV power output predictions, this study empowers stakeholders, including investors, energy planners, and policymakers with reliable insights. The precise forecasting provided by the developed model serves as a strategic tool for informed decision-making allowing stakeholders to allocate resources efficiently and plan for the integration of solar resources into the grid.
- **Economic Viability and Investment Confidence:** As renewable energy investments gain momentum globally, the study's emphasis on optimizing model complexity ensures economic viability. The low-complexity prediction model proposed here not only facilitates seamless integration into existing photovoltaic systems but also instills confidence among investors. The ability to navigate complexities efficiently encourages sustained investments in solar microgrid systems.
- **Promoting Sustainable Energy Transition:** This research is important because it has the potential to hasten the shift to renewable energy sources. The research advances the goal of integrating renewable energy sources, especially solar power, into the global energy mix by boosting trust, offering precise prediction and streamlining integration procedures. In summation, the expanded significance of this study encompasses strategic decision-making, economic viability, advancements in machine learning applications, environmental impact mitigation, facilitation of photovoltaic systems integration and the overarching promotion of a sustainable energy transition on a global scale.

Limitations

While this study strives to provide valuable insights into solar power prediction and the incorporation of machine learning algorithms, it is imperative to recognize specific constraints that could impact the extent and applicability of the results (Fayyad, 2023)

- **Data Restrictions:** The precision and efficacy of the created prediction model heavily rely on the quality and representativeness of the accessible data. Constraints in the quantity or diversity of the dataset might influence the model's capacity to extrapolate to a broader spectrum of environmental conditions.
- **Presumptions in Machine Learning:** This study presupposes that the chosen machine learning algorithms (KNN, SVR, Random Forest, Decision Tree, Linear Regression and AdaBoost) are apt for the distinctive characteristics of the solar PV dataset. The efficacy of these algorithms hinges on optimal parameter adjustment, and less-than-optimal tuning could impact predictive capability.
- **Fluctuating Nature of Environmental Factors:** External environmental factors, such as solar irradiance, cell temperature and wind speed exhibit an inherently dynamic nature and are subject to swift alterations. The study may not encompass all potential variations, and real-time shifts in these factors might introduce uncertainties in the forecasts.
- **Single-Location Emphasis:** This study may concentrate on a particular geographic locale or a set of solar photovoltaic systems in Germany. While this provides insights at a local level, it constrains the generalizability of findings to diverse geographical and climatic settings.

- **Balancing Model Complexity and Accuracy:** Striking a balance between low model complexity and high accuracy poses a formidable challenge. The developed prediction model might prioritize simplicity, potentially compromising a degree of accuracy or vice versa.
- **External Factors Beyond the Scope:** The study focuses on the correlation between solar PV power output and specific external environmental factors. Other potential influencers, such as dust accumulation on solar panels or shading effects, are acknowledged but may not receive exhaustive treatment.
- **Variability in Photovoltaic Systems:** The assessment of feasibility for integrating the prediction model into photovoltaic systems assumes a certain level of standardization. Disparities in infrastructures across regions may impact the applicability of the proposed model.

Structure of Research

The research project comprises several elements, the first of which is the introduction that provides background information and context for the study topic. Solar power prediction using machine learning models and other relevant studies regarding the machine learning-based prediction are presented in the literature review section. Section on methodology providing an overview of the different machine learning model used in this study, along with an explanation of the data pre-processing steps, such as data cleaning details, as well as an explanation of the research dataset and performance evaluation metrics then the discussion section that carefully evaluates the results of the research.

CHAPTER II

Literature Review

Theoretical Framework on Renewable Energy Theories

Drawing upon relevant theories in renewable energy, machine learning, and photovoltaic systems, this framework offers a methodical approach to address the research questions and hypotheses. With the rise of the oil crisis in the 1970s solar energy has become on the increase in its use. Renewable energy has become a general source of energy which has made researchers and policy makers globally focus on the various manners to effectively utilise solar energy. There are two methods by which solar energy can be used in terrestrial regions: solar photovoltaic (SPV), as displayed in Figure 1 or solar Chimney, collectors, cookers, Air Conditioning System, and other sun thermal devices (Timilsina, Kurdgelashvili, & Narbel, 2012).

Photovoltaic energy conversion converts energy directly without an intermediary. A Solar photovoltaic (PV) system can generate power without an external inducer within the range of microwatts to megawatts. They are usually very simple to maintain and

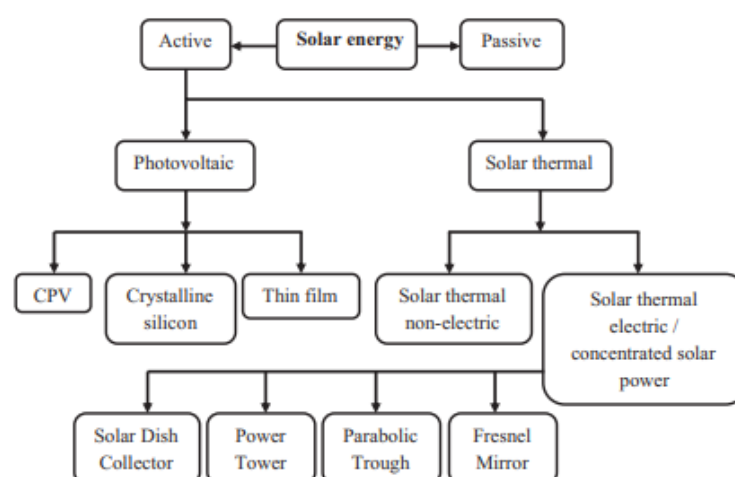


Figure 1: Various solar energy kinds according to availability on the global market(Timilsina et al., 2012)

adjustable in their design. Having an independent photovoltaic system has become increasingly important for rural areas in order to access electricity most especially in countries that are developing. The solar house lighting system is made of a charge controller, solar PV module and a battery that are all necessary to electrify a rural home (Elhadidy, 2002). Charge controller, solar PV module and batteries are also constituents of a solar house lighting system which are chosen based on the needs of the rural home. As a result, there is a greater need for solar PV across a range of fields. Future predictions indicate that the price of a watt of solar PV will decrease even further to 0.50 \$/Wp and 1.0 \$/Wp, from a value of 3.50 \$/Wp for first-generation and second-generation solar cells respectively (Díaz, Peña, Muñoz, Arias, & Sandoval, 2011). The popularity and installation of PV systems have been boosted by government feed-in tariff schemes put in place in many developing countries. The total installation capacity for the nations taking part in the IEAPVPS program was determined to have expanded from 103 MW in 1992 to 63,611 MW in 201 (Moosavian, Rahim, Selvaraj, & Solangi, 2013). The solar panel is like a sandwich made of silicon. It is usually made of a non-reflective coating made of tempered glass on the outside that serves as environmental protection. A conducting electrode on top of the solar panel serves as the cathode, or negative electrode, in most cases. Next is a thick layer of semi-conductors, either n-type or p-type. There are more free electrons than atoms are present in N-type solar panels. While the Solar panels of the P type has less free electrons. The opposite type of material is found on the other side of the depletion zone, which follows the top layer. The anode or positive electrode is the opposing electrode and makes up the bottom layer of the solar panel. Light in sun have energy which is transferred to them by the chemical processes taking place in the sun. As a result of the light's energy being transmitted to the solar cells' semi-conductor materials, a

constant flow of electrons is produced. This electron flow then produces electric current generating electricity which is used by the inverter to create AC energy, which can power your home or place of business. One thing to note is that, the sun is a huge star that emits energy continuously. The radiation from this energy is received by everything and eventually reaches the earth. This radiation is particularly crucial for solar installations since it affects peak power output. The impact of the different solar radiation components on photovoltaic system is covered in the section below and shown in Figure 2.

- **Direct Radiation:** This is all of the solar energy that makes it to Earth undisturbed, meaning that no buildings, trees, clouds or other obstructions are interfering with it. Going outside on a bright, sunny day and observing how light falls on the ground without obstruction is the simplest method to comprehend this. When solar PV systems are powered by direct sunlight, they produce the greatest electricity.
- **Diffused Radiation:** This type of radiation travels through more indirect routes to reach the planet. Usually, airborne particles, clouds, and water vapor disperse or obstruct the path that radiation takes to get to the surface of the earth. It can be experienced on a cloudy or rainy day, observe that light can still be seen but this is diffused light, not direct light. The environment is not entirely dark, solar panels can continue to generate electricity in the presence of clouds and overcast by utilizing the indirect radiation that is already present.
- **Reflected Radiation:** The quantity of solar energy reflected from a surface, determined by the albedo or solar reflectance of the substance on the surface, is known as reflected radiation. Either the planet's surface absorbs the solar

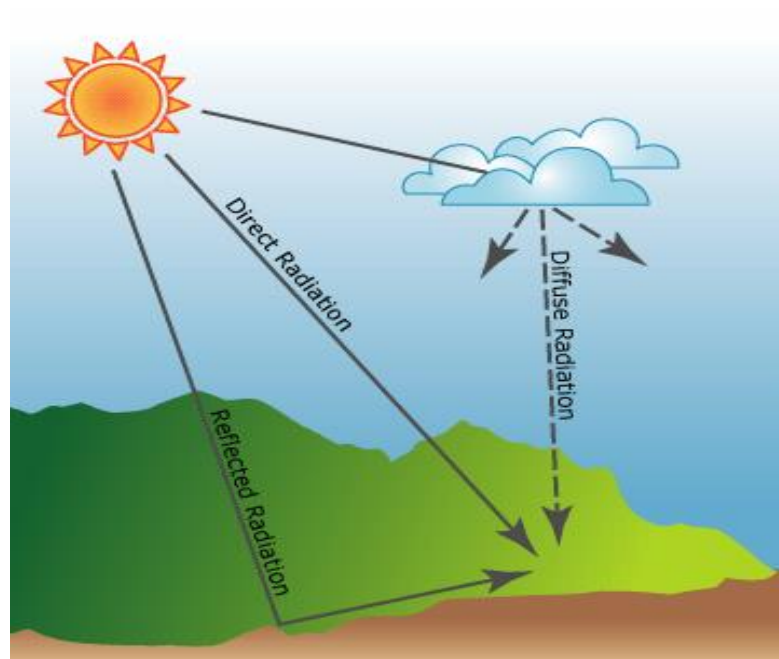


Figure 2: Direct, Diffuse & Reflected Radiation (Elhadidy, 2002)

energy that enters the atmosphere or snow, ice, and other surfaces reflect it back. Generally speaking, only a small percentage of radiation is reflected, with the exception of areas that are surrounded by highly reflecting surfaces like snow cover.

Types of Photovoltaic Systems

Photovoltaic systems vary in components, size, and application type with rural solar water pumping using slightly different components than residential rooftop systems, which use similar components for both types. The main types and component of a Photovoltaic systems which are shown in Figures 3, 4 & 5 are:

- **Grid-Connected Solar Photovoltaic Systems:** A DC-AC converter is used by a grid-connected solar photovoltaic (PV) system to transform solar energy into AC power. The inverter controls and regulates changes in the system by converting DC voltage from solar panels or the output voltage of a DC-DC converter into AC voltage. The inverter transforms the solar energy into AC power at a frequency compatible with the utility grid as this AC voltage is

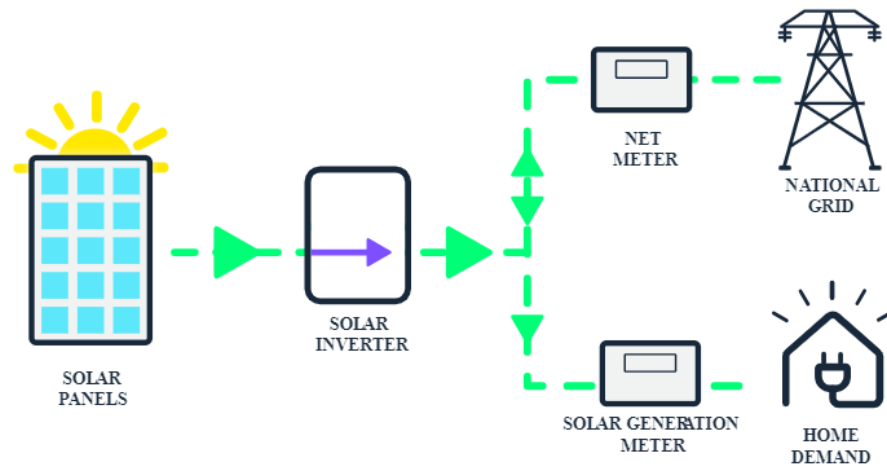


Figure 3: Grid-Connected Solar Photovoltaic Systems (Elhadidy, 2002)

included into the grid. The inverter AC output meets the grid's criteria for voltage and power quality. The solar PV system is typically deployed in conjunction with a metering system. In residential settings, utility grid power is not used when photovoltaic system power can cover the entire load. The remaining power is taken from the grid when photovoltaic power is limited. More photovoltaic power is put into the grid if it is generated. The solar photovoltaic system only generates electricity when the grid is operational. By connecting directly to the National Grid, the system lowers energy costs and its carbon footprint. It uses solar to power appliances instead of a battery storage system. Overage energy is exported back to the grid and reimbursed via SEG or Feed-in-Tariff. Because the National Grid consistently supports the energy supply, on-grid technologies provide security. Grid failure, however, can result power outages caused by grid-tied inverters. By including a battery in an existing hybrid system, solar power can be used even amid power outages.

- **Stand-Alone or Off-Grid PV Systems:** A PV system that is off-grid or stand-alone may run on DC or AC power. The photovoltaic system is not linked to the electric grid in either setup. The solar panels can provide the DC voltage if

DC loads remain attached to the solar PV structure, or a DC-DC converter can be utilized to convert the photovoltaic energy to higher DC levels. The PV voltage is increased by the DC-DC converter to a level appropriate for the DC loads. The number of solar panels in the PV system can be decreased by incorporating the DC-DC converter. Off-grid PV systems convert PV voltage to AC using inverters, reducing solar panel usage and storing voltage in batteries and can be integrated with DC-DC converters for stand-alone systems. For people who wish to be energy independent or who find it difficult to connect to the national grid, an off-grid solar system is a renewable energy source that doesn't depend on it. Demand for energy independence is growing as energy prices rise. Off-grid solutions include backup generators, solar energy generation equipment and renewable energy sources to guarantee battery charging throughout the year. They are energy independent since they can supply electricity even in isolated areas. Although off-grid systems can be modular, making them more flexible in fulfilling energy requirements, they are more expensive than regular grid-tied systems.

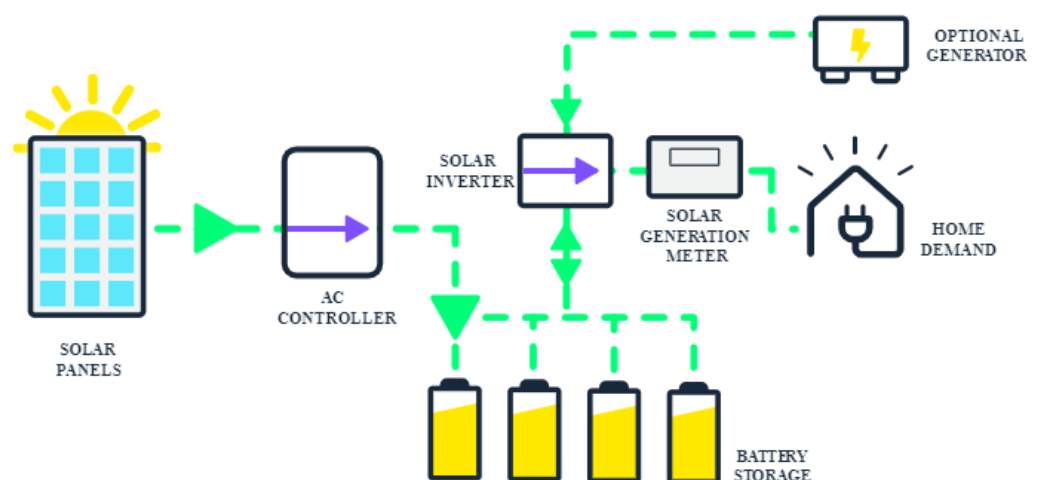


Figure 4: Stand-Alone or Off-Grid PV Systems (Elhadidy, 2002)

- **Hybrid Solar PV Systems:** In hybrid solar systems the technologies of solar panels and solar batteries are combined to generate green energy solutions that serves as a backup energy source. Even if the hybrid PV remains connected to the the national grid, solar enegy generated by it is initially held in a home battery then transferred to the grid. With hybrid solar system, excess solar energy is used to power homes at night and less enegy can be exported to the grid through battery storage. Additionally, in contrast to an on-grid system, the battery storage can be utilized for producing electricity in a case where the national grid experiences a disruption. This procedure known as "islanding" is particularly advantageous for owners living in areas with frequent blackouts. You can draw from the grid even in the event that your battery runs out of electricity when using hybrid solar panel structures, giving great deal of malleability. For this reason, a hybrid solar system is the best temporary solution. Although they are more expensive than on-grid systems, hybrid solar systems are a more affordable middle-of-the-road alternative to off-grid solutions. Two key advantages of a hybrid solar system are the capacity to expand your battery storage system at any time and the lower peak charging rates you receive from your continuous grid connection. However, because a hybrid solar system needs more components than a grid-tied system, it is less efficient. A hybrid PV system integrates solar PV with other power sources like diesel generators or wind, using converters to convert energy into DC or AC voltage. A maximum power point tracker (MPPT) is used for maximum power harnessing. While MPPT is not a requirement for solar PV systems, it

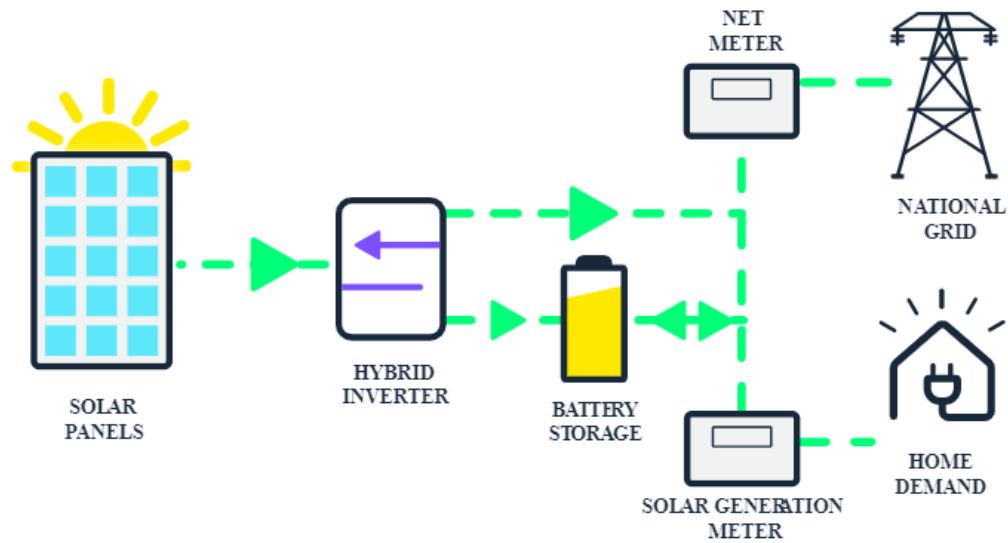


Figure 5: Hybrid Solar PV Systems (Elhadidy, 2002)

can assist raise overall system efficiency. The most promising answer to the worldwide energy problem that we are currently experiencing is solar photovoltaic systems. Any kind of solar PV system can be designed and simulated with the aid of Cadence's software

Methods of Solar Power Prediction

Four general categories can be used to group the methods used to forecast solar power generation:

- Meteorological models – These methods are usually indirect; they make use of satellite image processing and Numerical Weather Prediction (NWP) techniques. They initially predict how strong the sun will be, and then they translate that estimate into data for photovoltaic (PV) output.
- Statistical models - Statistical techniques like Exponential Smoothing (ES), Auto-Regressive Integrated Moving Average (ARIMA) and Auto-Regressive Moving Average (ARMA) are frequently utilized in these methods. Statistical models, in contrast to meteorological models, do not require the initial prediction of solar irradiance in order to directly predict PV power outputs.

- Machine learning models - In order to anticipate PV power output directly, machine learning algorithms like k-NN, Neural Networks (NN), Support Vector Regression (SVR), and Pattern Sequence-based Forecasting (PSF) are used. Creating a single forecasting model and assembling many forecasting models into an ensemble are the two main uses cases for machine learning techniques. (“Performance Evaluation and Viability Studies of Photovoltaic Power Plants in North Cyprus,” 2022)
- Hybrid models - These methods incorporate elements or models from the first three groups. In contrast to ensembles, which often comprise machine learning models, hybrid models frequently incorporate elements from statistical, machine learning, and meteorological methodologies.

Application of Solar Photovoltaic Power Prediction in Smart Grid

Large-scale integration of solar photovoltaic power into smart grids decreases system stability and dependability, particularly affecting smart grid energy management. This leads to issues with distribution network short-circuit current, power flow, grid losses, and voltage volatility. Electric power planning decision makers, system operators, and energy users may find useful assistance from solar PV power projection. Various forecasting models with varying prediction periods have been utilized for energy management in smart grids. PV outputs can fluctuate dramatically in a short amount of time based on the weather, including the passage of clouds. A precise very short-time solar PV power prediction model that can forecast for a few minutes to several hours could be useful to stabilize the PV outputs in order to prevent significant variations in the smart grid's frequency and voltage. Many techniques have been used

to smooth the PV outputs in order to reduce the ramp rate of PV generations. To absorb

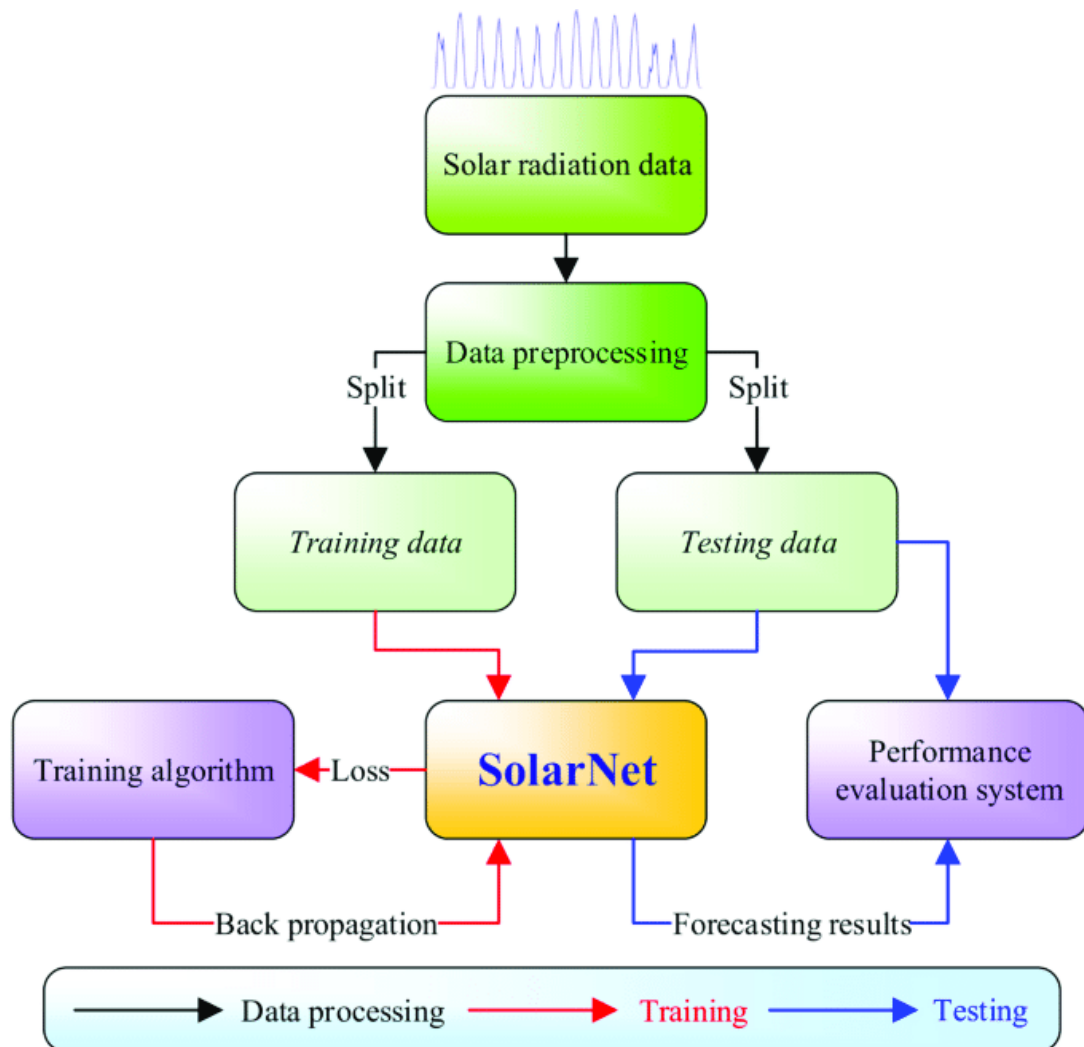


Figure 6. The Architecture of Solar radiation predicting System.(Kuo & Huang, 2018)

the sharp swings in solar PV power supply, fast ramping generators, battery storage systems, and electric double-layer capacitors are often used technologies. Many approaches are put forth for scheduling smart networks' intraday electric power consumption that integrate solar PV power generation. Intelligent energy management systems that are both grid-connected and island-connected are modeled, taking into account fluctuations in household load, storage capacity and charging rate, and

distribution network electricity pricing. The development of day-ahead energy management tools, such as demand response and storage units, for next-generation solar PV installations in the context of smart grids gives smart grid operators flexibility and uncertainty. It is suggested to implement a cost-based day-ahead energy management system with demand response and storage to mitigate swings caused by PV output uncertainties. Additionally, with thermal generators having sluggish ramp limitations, day-ahead power scheduling is becoming a crucial component of power systems. An assessment is conducted on the impact of prediction accuracy on large-scale aggregated solar power output. It is suggested to schedule PV generation in advance and combine it with battery storage for unit commitment issues. The bidding method used by PV businesses to compete in the day-ahead electricity markets is another way that the day-ahead prediction model is put to use (Amjady & Hemmati, 2009). Forecasting approaches for solar photovoltaic power enhance the quality of electricity supplied to the grid and reduce the extra expenses linked to weather dependence. The historical Solar PV Power data used in this paper were gathered by Belgium's power transmission system, Elia. The testing dataset is utilized to assess how well the suggested method predicts the future, while the training dataset is used to train the RNN network and uncover the nonlinear properties concealed in the PV power data.(Sharkawy, Ali, Mousa, Ali, & Abdel-Jaber, 2022). This study discusses four techniques to solar PV power forecasting: statistical, artificial, physical, and hybrid. It suggests that solar power forecasting is necessary to overcome certain technological and financial challenges. The statistical approach forecasts time series solar PV power data by data-driven formulation based on previously observed data (Behera, Majumder, & Nayak, 2018). A type of mathematical analysis known as statistics makes use of quantified models, representations for a particular group of

testing data, or empirical research. Statistics examines data collection, evaluation, analysis, and conclusion-drawing techniques. Artificial neural networks (ANN) are employed in artificial intelligence (AI) approaches to build solar builders, which are also included in the statistical approach category (Sfetsos & Coonick, 2000). It focuses on how machines that are programmed to process information and act like individuals mimic human intelligence. The phrase also refers to any machine that exhibits cognitive functions like problem-solving or learning skills that are comparable to those of a person. The optimal attribute of artificial intelligence is the ability to process information and act in a manner that optimizes the possibility of achieving a peculiar goal.

The physical model estimates solar irradiance and PV generation using satellite imagery or numerical weather prediction (NWP) (V. E. Larson, 2013). To forecast solar PV power, the hybrid model combines the three models mentioned above (Castillo-Rojas, Medina Quispe, & Hernández, 2023). A hybrid campaign could combine two performance-based models, or it could be a combination of impression-based (CPM) and performance-based (CPC or CPA). Sometimes, hybrid agreements are thought to be a means of further dividing the risk between marketers and publishers.

Machine Learning

Artificial intelligence (AI) allows computers to autonomously acquire knowledge from training and get better at something in the absence of explicit programming. One use of AI is machine learning. The primary objective of machine learning is the development of with the ability to acquire data and utilize it for independent learning. The training starts with training data that can be made of examples, first-hand experience or instructing, so that the learner can look for patterns in the data and apply

the examples given to improve their decision-making process. The primary aim is to empower computers to independently improve, devoid of human assistance or intervention, and update their operations correspondingly. Artificial intelligence and machine learning enable data engineers to assess outcomes with exceptional precision. Comparing deep learning methods for forecasting solar PV power. When faced with a dataset or scenario that has never been encountered before, machine learning algorithms are pre-trained to generate a particular result. But compared to humans, computers require more examples to learn. Intelligent decision-making may now be used in many fields and applications where significant algorithm development is required to get the desired results due to machine learning (Cohen, 2021).

There are various classifications for machine learning algorithms, which includes semi-supervised, supervised, unsupervised, and reinforcement learning. A quick description of these many algorithmic categories is provided below.

Supervised Learning

These techniques are able to forecast label values with the input consisting of not-labeled data once trained on a group of labeled data samples. Regression and classification are the two common issues with this kind of learning. The goal of the regression algorithm is to determine how the independent and dependent variables relate to one another. The procedure is used in classification to forecast the data's class label. Typical classification issues include of binary classification, wherein in contrast to conventional classification issues with mutually exclusive class labels, there are three types of classification: two class labels; multi-label classification, in which a single fragment of data is connected to multiple labels or classes; and multi-class classification, including above two class labels (Sarker, 2021)

A variety of techniques are employed in supervised learning, such as Logistic Regression (LR), Linear Regression (LR), Decision Trees, Naive Bayes, K-nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Convolutional Neural Networks (CNN), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost) for regression (Sarker, 2021). The DT, NB, and SVM algorithms are the most frequently used and used supervised learners in the literature (Alloghani, Al-Jumeily, Mustafina, Hussain, & Aljaaf, 2020) Text categorization and emotion prediction (for example, from Tweets or other posts on online); evaluating the sustainability of clothing products for the environment. (Satinet & Fouss, 2022) identifying, diagnosing, and treating mental illnesses (Jiang, Gradus, & Rosellini, 2020) and calculating peak energy demand are a few fascinating real-world uses.

Unsupervised Learning

In this kind of learning, unlabelled data is used. Instead of forecasting the right result in this instance, the algorithm investigates the unlabelled data to uncover hidden structures. Due to the uncertain potential values of the results, regression or classification problems do not instantly lend themselves to this type of learning.. Rather, it is frequently applied to dimensionality reduction, association, and grouping. Unlabelled data can be grouped using clustering according to their resemblances or dissimilarities. (El Bouchefry & De Souza, 2020). To find fresh and pertinent connections between the items in a set, association employs a variety of rules. Lastly, dimension reduction enables a dataset's features (or dimensions) to be lowered in order to remove elements that are unnecessary or of low importance, therefore lowering the model's complexity (Sarker, 2021). This feature reduction can be achieved in two ways: either by producing entirely new features (feature extraction) or by retaining a

portion of the original features (feature selection). The most often used clustering algorithm is likely K-means clustering, in which the k value indicates the dimension of the cluster. (Sarker, 2021) Apriori, Equivalency Class Transformation (ECLAT), and Frequent Pattern (F-P) Growth algorithms are examples of association algorithms. Last but not least, dimensionality reduction frequently makes use of the Pearson's correlation coefficient, Chi-squared test, Recursive Feature Elimination (RFE), Analysis of Variance (ANOVA) test, for feature selection, and Principal Components Analysis (PCA) for feature extraction. K-means, PCA, and hierarchical clustering are the most often utilized unsupervised learners, per (Alloghani et al., 2020) Numerous real-world uses for these unsupervised learners exist, including facial recognition, medical and consumer classification, data analysis and cyber-attack and intrusion detection in the field of astronomy (Y. Chen, Kong, & Kong, 2020).

Semi Supervised Learning

This technique, logically positioned between unsupervised and learning supervised, enables the utilization of the sizable unlabelled datasets that are occasionally paired with (typically smaller) quantities of labelled data (Van Engelen & Hoos, 2020). This creates intriguing opportunities because unlabelled data are more common than labelled data, and a semi supervised learner can produce predictions that are superior to those made with solely labelled data (Sarker, 2021). Applications that have a large amount of unlabelled samples and a limited amount of identified samples, or where the labelling effort is excessive, are considered candidates. In medical imaging, for instance, a modest quantity of data used for learning can result in an important increase in precision (Huynh, Nibali, & He, 2022). The main tasks that each type of training is utilized for (classification, regression versus clustering, association, and

dimensionality reduction) as well as the kind of input data that is utilized in every instance (labelled or un-labelled data) are highlighted in Table 4's contrast of unsupervised and supervised learning.

Reinforcement Learning

This method is reliant on the interaction between an agent engaged in an action and its surroundings, which offer constructive or critical criticism. The agent has to decide what to do in given situation to maximize reward. Monte Carlo, Q-learning, and Deep Q-learning are popular techniques (Sarker, 2021). Historically, prevalent uses have included strategy games like chess, self-driving cars, manufacturing and supply chain logistics, genetic algorithms, 5G mobility management, and customized healthcare (S. Liu et al., 2020).

Related Research

Building on the theoretical foundation, this section reviews existing literature relevant to solar power forecasting, machine learning applications in renewable energy, and the integration of solar microgrid systems into smart grids. It synthesizes findings from previous studies to identify gaps, trends, and methodologies employed in similar research domains.

In a study by (Mirjalili, Aslani, Zahedi, & Soleimani, 2023)The research uses Design Builder to model a neighborhood with solar panels and electric automobiles, predicting energy equilibrium for each building and micro grid by getting the final balance of the generation and use of energy for every construction and the entire neighborhood as a micro grid. The overall energy equilibrium is predicted using machine learning techniques such as K-Nearest Neighbor (KNN), Regression Support Vector (SVR),

Adaptive Boosting (AdaBoost), and Deep Neural Networks (DNN). Design Builder is used to model neighborhood structures. The effectiveness of the KNN, SVR, AdaBoost, and DNN algorithms was compared in order to ascertain which approach is best at predicting energy balance. To optimize the use of energy and minimize the effect on the environment, this research adopts a new method by creating a model that accounts for an integrated system of homes, solar cells, and electric usage for every construction in an area.

In another study by (Gaviria, Narváez, Guillen, Giraldo, & Bressan, 2022) modern machine learning approaches for photovoltaic systems are reviewed with a major emphasis on deep learning. How machine learning is used in control, PV system management, islanding identification, problem diagnosis and detection, irradiance and power production predictions, scaling, and region adaptability. The three major contributions by this study are reviews over hundred research articles that apply cutting-edge machine learning methods to PV systems; next, It assesses materials that offer researchers access to free datasets, source code, and experimental environments designed for evaluating machine learning algorithms. Third, to help academics seeking a deeper understanding of these subjects, encourage them to expose themselves to the applications of cutting-edge ML methods applied to PV systems. They also give a case study with open source code and data for each of the topics, gave guidelines, perspectives and opportunities for further advancement.

Although renewable energy sources, such as solar and wind power, are essential for satisfying global energy requirements, power system workers face difficulties due to their unpredictability and fluctuation. Grid stability depends on precise estimation of renewable energy production. Machine learning and deep learning (ML) and deep learning (DL) algorithms have found a potential use in forecasting renewable energy.

Classical ML models like linear regression are straightforward and facile to understand, but lack the ability to capture complex patterns. Random forest, SVMs, and XGBoost models perform better than linear regression in tackling non-linear relationships and intricate data. Inconsistently distributed data is best handled by dedicated time-series prediction methods like the autoregressive models, moving averages, and RNNs. Mixed models integrate ML and DL algorithms with conventional time-series analysis, but design is challenging due to the need for high-quality data and transparency (Benti, Chaka, & Semie, 2023)

San Diego, (Chow et al., 2011) presented a technique for intra-hour, sub-kilometer cloud irradiance forecasting utilizing a ground-based sky imager. Every thirty seconds, they captured sky photos, which were then analyzed using sunshine characteristics and a clear sky library to calculate the amount of sky cover. In order to predict cloud shadows at the surface, they created a two-dimensional cloud map using coordinate-transformed sky cover, which is then utilized to create forecasts. The projected horizon and cloud speed had the biggest effects on forecast accuracy. The findings demonstrated that the forecasting error in the 30s predictions was down to 50%–60% of the inaccuracy of the persistence models.

(Chu & Coimbra, 2017) suggested k-NN ensemble models to produce probability density function estimates for intra-hour Direct Normal Irradiance (DNI) utilizing lagged irradiance and image data. A variety of data sets (continental, coastal, and island) was used to assess the model by criteria like Prediction Interval Coverage Probability (PICP), Prediction Interval Normalized Averaged Width (PINAW), and other general error criteria. The model received measurements of diffuse irradiance and cloud cover information as exogenous feature inputs. They used a Gaussian probabilistic forecasting model and a persistent ensemble probabilistic forecasting

model as baselines. According to their findings, when the forecasting horizon was longer than five minutes, the suggested k-NN ensembles performed better than both reference models across all assessment criteria for every site. Chu et al., (2015) used a NN-optimized re-forecasting strategy to expand a k-NN model. The k-NN's performance over a period of 5, 10, and 15 minutes might be considerably enhanced by the reforecasting approach, according to the results of this model's evaluation, which used data from a 48 MW PV facility.

By using a k-NN model, Pedro & Coimbra (2012) demonstrated that it performed better than the comparative persistence model. A novel k-NN based technique for forecasting intra-hour GHI and DNI, together with the associated uncertainty estimation ranges. An optimization approach was used to establish the parameters, and the forecasting horizon varied from 5 to 30 minutes. The proposed method surpassed the persistence model by 10% to 25%, according to the results. Additionally, the scientists reported that incorporating sky photos into the optimization can result in a negligible improvement of roughly 5%.

Chen et al. presented a mechanism in C.-R. Chen & Kartini (2017) for hourly GSI value forecasting. To be more precise, they constructed a k-NN model to prepare the data before training a NN to predict the target PV station's GSI value one hour in advance. The k-NN model creates the inputs for the NN model that is utilized to produce the predictions, using meteorological data from eight nearby PV stations. The hybrid model produced an RMSE of 242 W/m² and a Mean Absolute Bias Error (MABE) of 42 W/m², according to the data.

The Pattern Sequence similarity Forecasting (PSF) approach was presented by Martinez Alvarez et al.(2011) as a means of forecasting time series connected to energy. PSF initially divides the previously recorded data into a number of sets, then

assigns a cluster label to each day. A pattern sequence of cluster labels is formed by the days leading up to the target day. Next, using the days just successive to the closest sequences to calculate the prediction for the upcoming days by averaging their values, PSF looks through the previously recorded data for these pattern sequences' closest neighbors. The outcomes demonstrated that PSF was an effective and successful forecasting technique.

In a study by (Chahboun & Maaroufi, 2021) one of the main objectives is to thoroughly compare three widely used machine learning methods for hourly power prediction from photovoltaic solar panels: multiple linear regression, support vector regression, and random forest. Residual analysis is done to visually test the regression models under investigation, With $R^2 = 96\%$ and $RMSE = 0.39$ KW, the outcomes demonstrated that random forest had the optimal prediction accuracy during the testing phase.

Performance indicators were used to measure the accuracy of the solar power forecast models that were developed in another study by Kuriakose et al.(2020). It is discovered that ANN outperforms support vector machines and linear regression in terms of results. While the accuracy of the SVM model is not as high as that of the ANN, it is nevertheless comparable. The linear regression model's accuracy is low. Datasets from the weather station and on-site pyranometer might be used to improve the precision of the power forecast.

After analyzing the meteorological station's weather report, (Mellit, Massi Pavan, & Lughfi (2014) categorized the days into groups such as sunny, foggy, cloudy, and wet before training an independent SVR model. It forecasts the PV power for the following day for each kind of day. They used the mean daily temperature estimation for the following day as well as the PV power production of the nearest day with an identical

label in the training data. On sunny days (RMSE = 1.57MW), the maximum accuracy was attained; on foggy days (RMSE = 2.52MW), the lowest.

A 2D-interval forecasting model utilizing SVR was presented by (Rana, Koprinska, & Agelidis (2015)). It estimates the 2D-interval PV power output directly based on previous meteorological data and solar power. Australian PV data was used to assess the model during a two-year period. Their findings demonstrated that, when compared to several baselines and various approaches for comparison, such as NN2D and two persistence models, SVR2D produced the most precise forecasts.

Ramli et al. (2015) utilized information from Saudi Arabia to contrast NN and SVM for solar irradiance projections. They assessed the models in terms of global solar radiation and direct diffuse on the horizontal surface as input data. Correlation coefficient, RMSE, MRE, and computing speed. The findings demonstrated that the SVM models offered improved computation robustness and accuracy, with MRE values of 0.33 and 0.51 for the two cities and a 2.15-second forecasting speed.

Seven SVM models with different inputs were proposed by J.-L. Chen et al.(2013) to forecast the daily sun irradiation levels. Five sunshine-based models (cubic, quadratic, linear, exponential and linear exponential,) that utilize data collected from three locations in China were contrasted to the suggested models. The potential of SVM models was demonstrated by the 10% reduced RMSE that the SVM models provided in comparison to the empirical models.

(Wolff, Kühnert, Lorenz, Kramer, & Heinemann, 2016) created SVR models to predict PV power data over horizons of five hours and fifteen minutes. The model was created as a substitute to neural network prediction models (NWP). According to the authors' findings, the cloud motion vector model was the most effective model among the NWP-based models, which offered superior period forecasts beginning at three hours

ahead. SVR, on the other hand, performed well for predictions made one hour ahead. The authors proposed that the accuracy could be increased even further by merging the output from various prediction models.

The Support Vector Machine Firefly Algorithm (SVM-FFA), which (Olatomiwa et al., 2015) devised, is used to estimate the mean horizontal global solar radiation levels. The length of the sun, the highest temperature, and the lowest temperature were their inputs. The outcomes of the comparison between the suggested model and the GP and NN models indicated that the suggested model produced the optimal RMSE, MAPE, r , and R^2 . Numerous research publications have conducted forecasting and modeling of solar PV output from PV installations in the American Southwest were published by Renewable Energy 91 in 2016. This publication by Larson et al., (2016) presents the research effort on predicting for hourly-averaged, day-ahead power outcomes from PV power plants based on least-square optimization of numerical weather prediction (NWP). We compare power output data from two nontracking facilities in California for the years 2011–2014 to three different predicting methods. The study confirms the suggested methodology's effectiveness in comparison to earlier research. Solar photovoltaic power is heavily reliant on the weather outside. They can also result in unanticipated variations in the voltages and PV power for the PV systems. Essentially, for the power system to operate safely and integrate economically, it is imperative to precisely estimate PV power generation (Wang et al., 2017).

CHAPTER III

Methodology

Simple Linear Regression

Basic linear regression is a simple way to predict a target variable's real value denoted as Y , by considering an input variable, denoted as X . It is thought that there is an approximately linear relationship between X and Y . A popular formal formulation for this connection is Y regressed onto X (Smys, Iliyasu, Bestak, & Shi, 2020).

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (1)$$

Here, β_1 is the mean variation in Y for an increment of a single unit of X , and β_0 denotes the predicted variable of Y if X is zero (the intercept). The coefficients β_0 and β_1 are unknown. The “term error, ϵ ” recognizes that. The variation in Y is not explained by this simple linear model. Put another way, since Y is probably impacted by other factors not included in the model, the possibility of a totally linear relationship between X and Y is quite low. After estimating the unknown model coefficients, β_0 and β_1 , we may use the following formula to estimate Y for a given value of X :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (2)$$

Where, Y is estimated for the i -th value of $X = x$ by \hat{y} . An estimated value for an unknown coefficient, parameter, or result is indicated by the hat $\hat{}$ symbol. The least squares approach, which was created at the start of the 19th century and it is used to solve astronomical issues and it is the oldest type of linear regression.

Ordinary least squares (OLS) is the most popular method for fitting a linear regression model, however, there are other alternatives as well. For $\hat{\beta}_0$ and $\hat{\beta}_1$, OLS chooses coefficients that minimize the residual sum of squares (RSS), which is defined by:

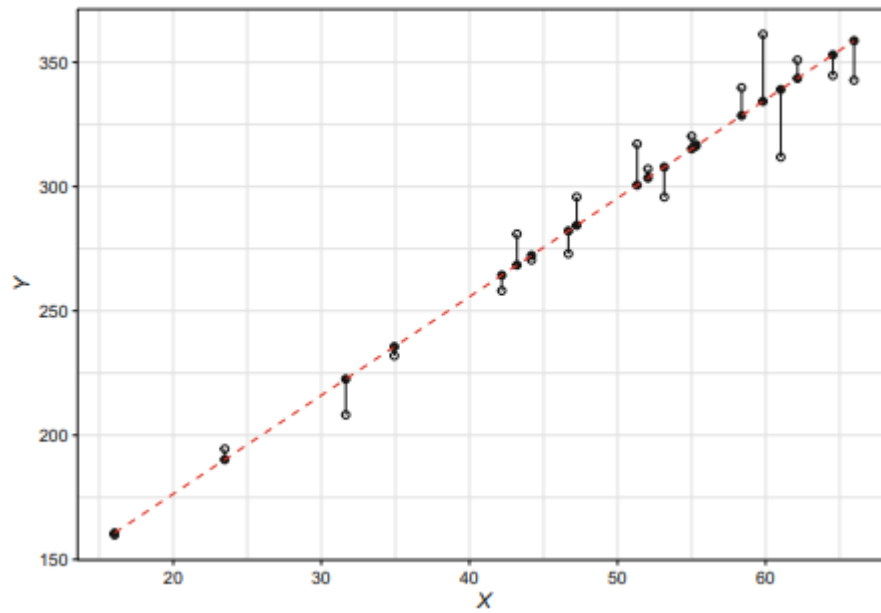


Figure 7: Minimizing RSS with Ordinary Least Squares (OLS) fit (Smys et al., 2020)

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \quad (3)$$

OLS builds a model for each value of X such that the squared difference between the actual (y_i) and anticipated ($\hat{\beta}_0 + \hat{\beta}_1 x_i$) values is as little as feasible. The outcome of employing OLS to minimize RSS is shown in Figure 6. The following defines the minimizers for the estimations of the least squares coefficients (Smys et al., 2020):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4)$$

Support Vector Machine

SVMs were first designed for classification and have been expanded for regression and classification learning. SVMs are first implemented as binary classifiers, where the learnt function returns a positive or negative value. By merging many binary classifiers using the pairwise coupling approach, a multiclass classification may be built (Jensen & Snodgrass, 2009). The two main characteristics of SVM, margin maximization and kernel technique are explained in this section along with how it was motivated and formalized as a binary classifier.

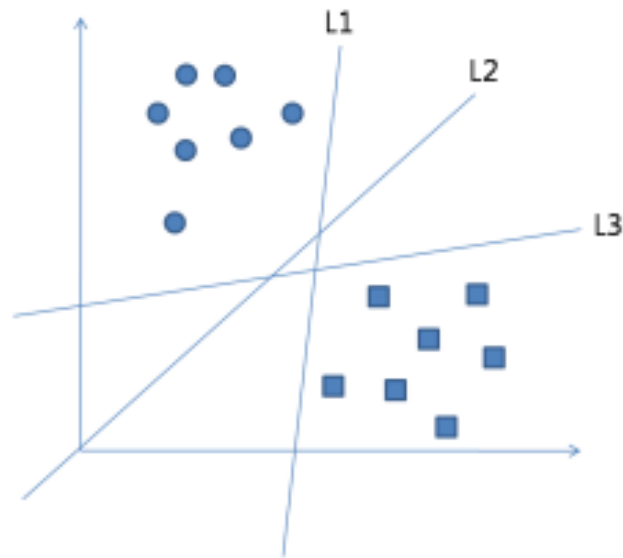


Figure 8: Two-dimensional linear classifiers (hyperplane). (Cortes & Vapnik, 1995)

Classifiers distinguishing between points of data from two groups are called binary SVMs. Every data is expressed by a vector with n dimensions. These particular data items are all part of just one of two classifications. A hyperplane is used to divide them using a linear classifier. For instance, two data sets and dividing hyperplanes—lines in a two-dimensional space—are displayed in Figure 7. Numerous linear classifiers accurately identify (or split) the two categories of data (Cortes & Vapnik, 1995).

SVM selects the largest margin hyperplane is needed to get the greatest prospective division between the two groups. The minimal distance between the splitting hyperplane and the nearest data point for each of the two classes adds up to the gap. As the hyperplane accurately classifies "unseen" or testing data points, it is more likely to generalize than other types of hyperplanes. (Ogidan, Dimililer, & Ever, 2018)

SVMs conducts the mappings from inputs space to features space to handle nonlinear classification issues. The kernel technique helps with this by permitting the mapping function to not be precisely defined, which may lead to the curse of dimensionality problem. This gives a linear categorization in the newer space comparable to nonlinear

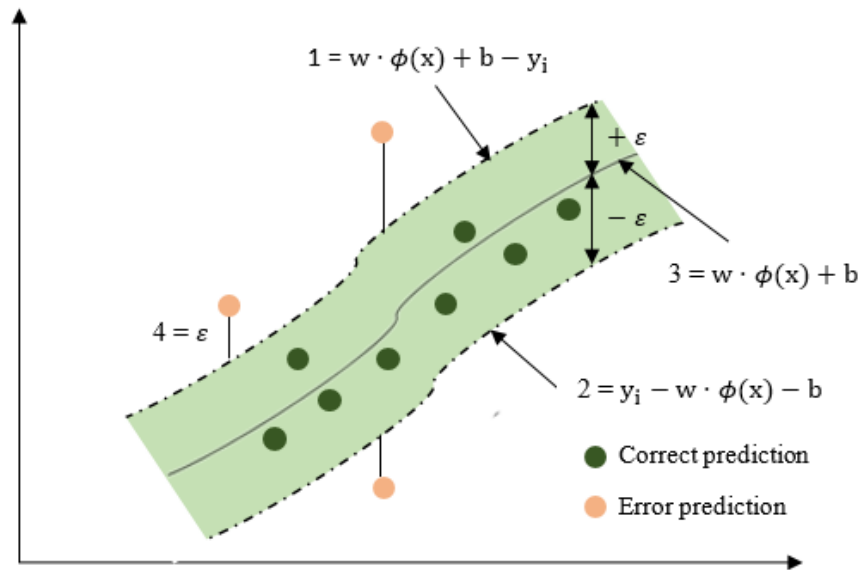


Figure 9: Hyper-plane for SVR (Smys et al., 2020).

categorization within the original space (or the data provided space). To do this, SVMs translate input vectors to a feature space which is a space with greater dimensions on which a maximum separating hyperplane is built (Smys et al., 2020).

Hard-Margin SVM Classification

This is a type of SVM that best separates various groups in a dataset without any points of data in the margin. Calculate the maximized marginal hyper-plane together with the support nonlinear categorization to grasp the working of SVM. Firstly, we discuss the hard-margin SVM when the data for training is devoid of noise and could be successfully categorized using a linear function. The training set, or points of data D , have the following mathematical expression.

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\} \quad (5)$$

If \mathbf{x}_i is an n-dimensional real vectors, y_i is either 1 or -1 representing the class in which the point \mathbf{x}_i belong. The form of the SVM classification function $F(\mathbf{x})$ is:

$$F(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b \quad (6)$$

During the training phase, SVM will compute the weight vector, denoted as \mathbf{w} , and the bias, represented as b . $F(\cdot)$ (or \mathbf{w} and b) must initially offer positive values for positive points of data and numbers that are negative otherwise to accurately categorize the training set. Put another way, for each point \mathbf{x}_i in D ,

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i - b &> 0 \text{ if } y_i = 1, \text{ and} \\ \mathbf{w} \cdot \mathbf{x}_i - b &< 0 \text{ if } y_i = -1 \end{aligned} \quad (7)$$

These requirements can be changed to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) > 0, \forall (\mathbf{x}_i, y_i) \in D \quad (8)$$

D is said to be linearly separable if there is a linear function F which properly classifies each point in D or fulfills Eq. (8). Second, the margin must be maximized by F , or the hyperplane. The gap between the nearest data points and the hyperplane is known as the margin. To do this, Eq. (8) is rewritten into the following Eq. (9).

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \forall (\mathbf{x}_i, y_i) \in D \quad (9)$$

Take note that the equality sign is included in Eq. (3.9), thus the value on the right end becomes 1 rather than 0.

There is such a F that meets Eq. (9) when D is linearly distinguishable or if each point in D meets Eq. (8). The reason for this is that, if such \mathbf{w} and b exist and fulfill Eq. (8), they can always be rescaled to meet Eq. (9). The distance between a vector \mathbf{x}_i and the

hyperplane is expressed as $\frac{|F(\mathbf{x}_i)|}{\|\mathbf{w}\|}$. The margin thus becomes

$$margin = \frac{1}{\|\mathbf{w}\|} \quad (12)$$

Thus, the margin becomes because, $F(\mathbf{x})$ will equal 1, if \mathbf{x} are the nearest vectors.

Support vectors are the nearest vectors that have an equality sign and meet Eq. (8).

Minimizing $\|\mathbf{w}\|$ results from maximizing the margin. As a result, the SVM training issue is transformed into the following restricted optimization problem.

$$\begin{aligned} \text{minimize: } Q(\mathbf{w}) &= \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subject to: } & y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \forall (\mathbf{x}_i, y_i) \in D \end{aligned} \quad (11)$$

In mathematics, the factor 1/2 is employed for convenience. Figure 8 is the graphical illustration for SVR. (Jensen & Snodgrass, 2009)

KNN Classifier

The nearest neighbor classification's underlying principle, sometimes referred to as K-nearest neighbors (KNN), is that patterns that most closely resemble the objective pattern, \mathbf{x}' , for which a label is needed, offer valuable label information. For the majority of K-nearest patterns, KNN represents the class label in data space. We have to be capable to develop a similarity metric in the data space. It is reasonable to apply the Minkowski metric. (p-norm) in \mathbb{R}^q (Subasi, Khateeb, Brahim, & Sarirete, 2020)

$$\|\mathbf{x}' - \mathbf{x}_j\|^p = \left(\sum_{i=1}^q |(x'_i) - (x_i)_j|^p \right)^{\frac{1}{p}} \quad (12)$$

This, for $p = 2$, is equivalent to the Euclidean distance. It is necessary to select appropriate distance functions in different data spaces, like the distance calculated by Hamming in \mathbb{B}^q . The label for the set $Y = \{1, -1\}$ is used in the binary classification scenario, and KNN is specified as with neighborhood dimension K and with the collection of indexes $\mathcal{N}_K(\mathbf{x}')$ of the K-nearest patterns. The KNN location is defined by selecting K.

$$f_{\text{KNN}}(\mathbf{x}') = \begin{cases} 1 & \text{if } \sum_{i \in \mathcal{N}_K(\mathbf{x}')} y_i \geq 0 \\ -1 & \text{if } \sum_{i \in \mathcal{N}_K(\mathbf{x}')} y_i < 0 \end{cases} \quad (13)$$

In locations where patterns from various classes are dispersed when $K = 1$ little neighborhoods appear. ($K = 20$, for example) Larger neighborhood sizes result in the

disregard of patterns whose labels are in the minority. The classification differences between KNN with $K = 1$ and $K = 20$ are shown in Figure 9 using a straightforward two overlapping data clouds comprise this 2-dimensional collection of data with 50 Gaussian-sampled blue and red points each. Red-classified places are depicted in white, whereas blue-classified data space locations are displayed in vivid blue. The predicted value is local for $K = 1$. For instance, the middle of the red cloud contains a point with a blue color that stands out within the blue class. The machine learning model ignores tiny agglomerations of patterns in favor of generalization for big K . In data space, KNN creates a Voronoi tessellation. KNN can already provide a fair estimate depending on the K -nearest neighbors in a screened subset, but it must look over the whole area to find the K -nearest patterns in cases of big data sets (Kramer, 2013). The issue of how to select K —that is, which neighborhood size produces the greatest categorization outcome. Model selection is another name for this issue and there are other methods, such as cross-validation, that may be used to pick the optimal model and parameters.

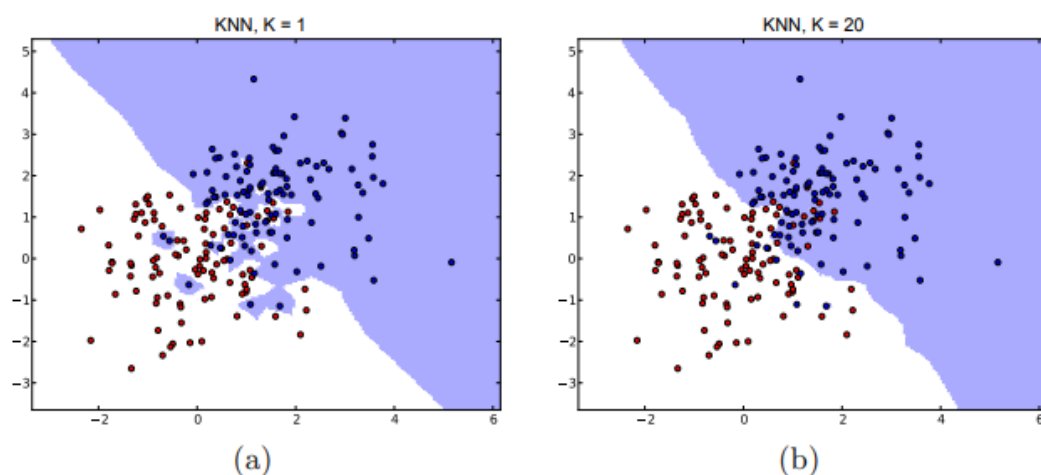


Figure 10: A evaluation of KNN classification on two Gaussian-based data clouds for different neighborhood sizes ((a) $K = 1$ and (b) $K = 20$). While KNN ignores tiny agglomerations of patterns in favor of generalizing for bigger K , it over-fits and becomes local in in small neighborhoods (Subasi, Khateeb, Brahim, & Sarirete, 2020)

Multi-Class K-Nearest Neighbors

Applications of KNN extend to multi-class classification issues. KNN for classification with multiple classes predicts the classification label of the majority of the data set's K -nearest patterns for an unknown pattern x (Kramer, 2013).

$$f_{\text{KNN}}(\mathbf{x}') = \arg \max_{y \in \mathcal{Y}} \sum_{i \in \mathcal{N}_K(\mathbf{x}')} \mathcal{J}(y_i = y) \quad (14)$$

Using the indicator function $\mathcal{J}(\cdot)$ which yields zero otherwise and one if its input is true.

KNN Regression

Regression is closely associated with categorization. Functional regression models relate patterns to continuous labels perhaps to a \mathbb{R}^d subspace. In contrast to classification issues, where the collection of labels is limited to a discrete collection of numbers, the distinction becomes evident when considering that in reality machine precision only permits a mapping to a very large set of labels.

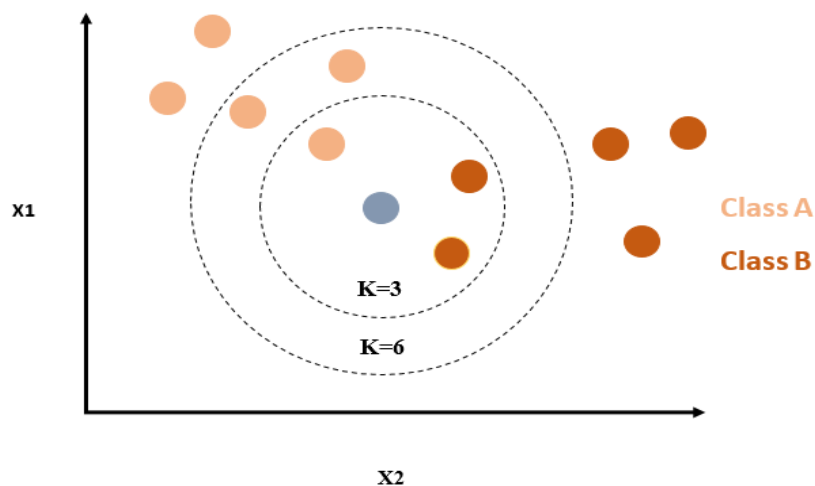


Figure 11: K nearest neighbor algorithm with $K=3$ and $K=6$. (Gabrieli D. Silva, Mariza Ferro, & Schulze, 2021)

A few components. Regression analysis's challenge is to forecast labels $\mathbf{y}' \in \mathbb{R}^d$ for novel patterns $\mathbf{x}' \in \mathbb{R}^q$ given a collection of N observations, or labeled patterns $\{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$. The objective is to become familiar with the regression function, $\mathbf{f}: \mathbb{R}^q \rightarrow \mathbb{R}^d$.

$$\mathbf{f}_{KNN}(\mathbf{x}') = \frac{1}{K} \sum_{i \in \mathcal{N}_K(\mathbf{x}')} \mathbf{y}_i \quad (15)$$

Using a set $\mathcal{N}_K(\mathbf{x}')$ that contains the indexes of the K -nearest neighbors of an unknown pattern \mathbf{x}' . KNN regression calculates the mean of the function's parameters values of its K -nearest neighbors. The localization of functions in data and label space is the underlying premise of the KNN average concept. Patterns \mathbf{x} are predicted to have continuous labels $f(\mathbf{x}')$ that are identical to \mathbf{y}_i in the immediate neighbors of \mathbf{x}' . Because of this, the label for an unknown \mathbf{x}' needs to resemble the labels of the patterns that are the closest, which are represented by the average of the K -nearest patterns' labels. KNN has been demonstrated well in numerous applications, e.g., with the discovery of quasars depending on spectroscopic data. The neighborhood size K of the KNN is a crucial parameter in regression as well (Dimililer, Kayali, & Tackie, 2023). While KNN regression averages across all patterns for $K = N$, it over fits to the label of \mathbf{x}' closest neighbor for $K = 1$. A comparison of the KNN regression model with the two neighborhood sizes, (a) $K = 2$ and (b) $K = 5$ is presented in Figure 11 and Figure 10 shows the K nearest neighbor algorithm with $K=3$ and $K = 6$. The regression of weighted KNNs causes plateaus.

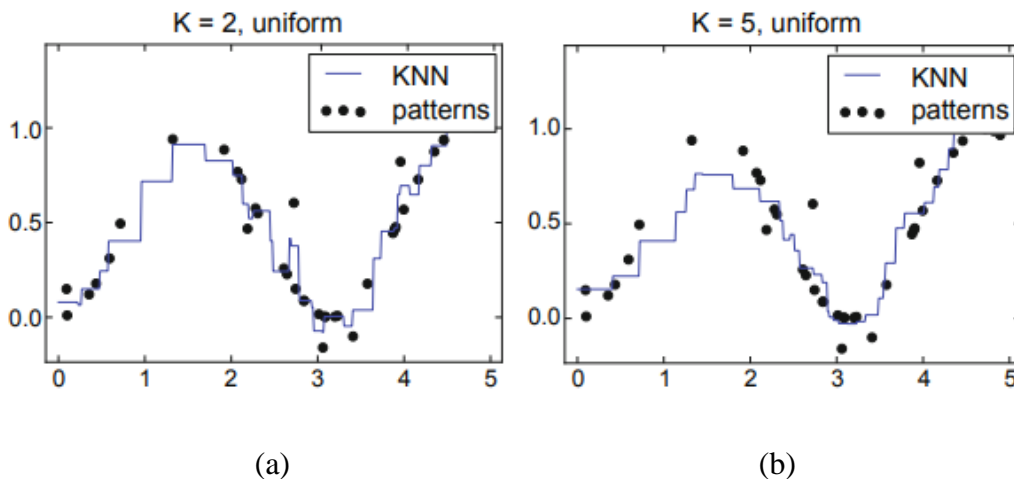


Figure 62: Shows a uniform KNN regression for the values of (a) $K = 2$ and (b) $K = 5$. (Gabrieli D. Silva et al., 2021)

Decision Tree Regressor

The learning function $f(\mathbf{X})$ is represented as a tree in DT algorithms. The branches of each node in the tree indicate a test using the values of the associated property, and each node itself symbolizes a test on an attribute. The y_i labels are represented by the leaves. Divide and conquer is how the algorithm constructs the categorization model. At the base of the tree is the characteristic that, in accordance with the term, "best" distinguishes the samples (Sekeroglu, Ever, Dimililer, & Al-Turjman, 2022). An attribute selection metric, such as knowledge gain or the Gini index, is used in this process. Every node (together with its accompanying properties) goes through this procedure repeatedly until every node covers the greatest amount of samples from a single class and, ideally, none from other classes. The structure of a decision tree (DT) is shown in Figure 12. A DT is made up of an initial element called the root, branches that have different decision nodes based on the best attribute that separates the data, and an end node, also called a leaf, that represents the final decision (the outcome of the classification or regression) (Gabrieli D. Silva et al., 2021).

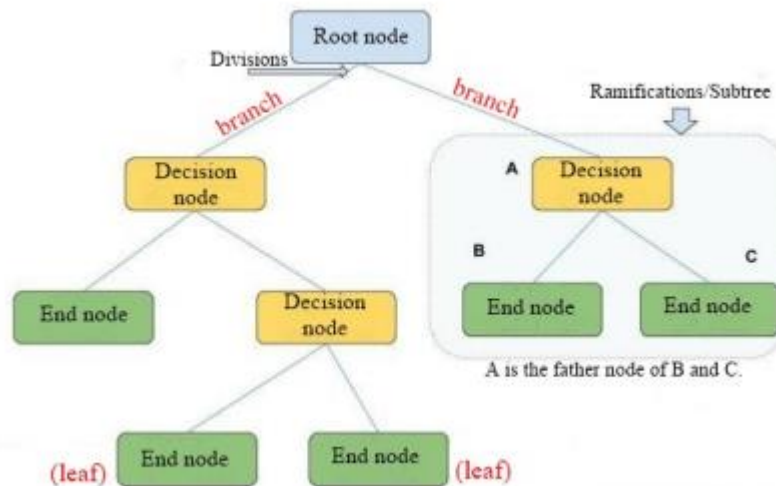


Figure13: A decision tree structure illustration. (Gabrieli D. Silva et al., 2021)

As indicated, despite their straightforwardness, DT are among the most extensively used ML algorithms and may be utilized for classifications or regression applications (KAGGLE, 2020). Classification is the objective of DT, same as it is in model C4.5 (QUINLAN, 1996), where the dependent variable (Y) denotes a category value. The goal of the regression tree is to forecast Y with a numerical value, just like in the CART (BREIMAN et al., 1984) method. Among the first Dictotomiser 3 (ID3) algorithms is the an iterative Dichotomiser 3 (QUINLAN, 1986). The technique builds a tree by identifying the category property that will result in the greatest information gain for each node. Numerical characteristics are not accepted by ID3, only category ones. Furthermore, it is incapable of handling missing values and lacks a post-pruning technique.

Equation 16 provides the information gain that ID3 uses, taking into account the T dataset displayed in Table 1, to determine which characteristic (attribute) to assign to each node. Entropy(T) is defined as the purity of data in the T set and X_j is a particular characteristic of the dataset T. T_j is the subset of rows in T for which the characteristic

column X_j has value v , $|T_v|$ is the amount of rows in T_v and similarly $|T|$ is the number of rows in T . N_{cl} is the total amount of classes in the Y characteristic, and p_i is the percentage of T that has class I , according to Equation 17.

$$Gain(T, X) = Entropy(T) - \sum_{v \in values(X_j)} \frac{|T_v|}{|T|} Entropy(T_v) \quad (16)$$

$$Entropy(T) = \sum_{i=1}^{N_{cl}} - p_i \log_2 p_i \quad (17)$$

Adaboost Regressor

By aggregating the forecasts produced by several learning algorithms/estimators, ensemble approaches aim to increase an estimator's resilience. Ensemble method can be divided into two categories

- **Averaging techniques:** The goal is to create and forecast utilizing a variety of estimators, then average these forecasts. According to some, the average method's combined estimators outperform single estimators on average because of a lower variance. A few common ensemble techniques are bagging and random forest trees, among others (Van Der Walt et al., 2014).
- **Boosting techniques:** These approaches aim to create an effective ensemble by merging numerous weak estimators, which is different from the average ensemble methods. By creating single estimators one after the other, these techniques aim to reduce the selection bias of the ensemble that is being built. Adaboost and gradient tree boosting are a couple of the boosting ensemble techniques. Boosting is the process of repeatedly executing a weak learning machine on various training sample distributions and aggregating the results. The machine's performance from the previous iteration determines the distributions of training samples for subsequent iteration. (Sekeroglu, Dimililer, & Tuncal, 2019). The process for determining the training example distribution

varies depending on the boosting technique used; for classification problems, the outputs of the many machines are aggregated using voting multiple classifiers, and for regression problems, they are merged using weighted average or median. (Bartlett, Freund, Lee, & Schapire, 1998) describes boosting via filtering, which is the original boosting strategy. The PAC (probably roughly correct) theory of learning (Haussler & Warmuth, 1993) served as its inspiration. In many real-world scenarios, it is not possible to obtain the huge number of training samples needed for boosting by filtering. AdaBoost (Cao, Xu, Liang, Zhang, & Li, (2010), a different boosting method, can be used in several versions to get around this restriction. In boosting by subsampling the, training instances and a fixed training size are employed, and during training, they are resampled in accordance with a specified probability distribution. During boosting by reweighting, the weak learning machine is trained using all of the training examples, with weights allocated to each example. This method is only useful when the weighted examples can be handled by the weak learning machine.

AdaBoost is one of the most well-known boosting techniques for classification, and it serves as the foundation for both ExpBoost and TrAdaBoost (Cao et al., (2010). AdaBoost assigns a weight w_i to each training instance, which is used to learn each hypothesis. This weight represents the relative relevance of each example and is used to calculate the inaccuracy of a hypothesis on the total amount of data. As in step 5 of Algorithm 1, instances are reweighted after each iteration, with bigger weights assigned to those that are incorrectly categorized by the final hypothesis. Learning therefore concentrates on the cases that are most challenging to categorize as the process proceeds. The basic idea behind the well-known boosting technique AdaBoost

is to combine weak estimators and apply them to better versions of the data. Next, each prediction's results are merged using hard voting or weighted majority voting. Regression issues are solved with the Adaboost Regressor.

Adaboost Regressor

1. Input

- A list of m samples $(x_1, y_1), \dots, (x_{0,1}, u_m)$ with an outcome of $y \in \mathbb{R}$ is the input.
- Inadequate learning algorithm Weak learner
- The number of iterations (machines) is indicated by the integer T .
- A threshold ϕ ($0 < \phi < 1$) to distinguish between accurate and inaccurate predictions

2. Initialize:

- Iteration or machine number: $t = 1$
- For every i , the distribution $D_t(i) = 1/m$
Error rate $\varepsilon_t = 0$

3. Continue until $t \leq T$.

- Call Weak Learner and supply it with the distribution, D_t
- $f_t(x) \rightarrow y$ is the regression model to build.
- Determine the absolute relative error for every training instance by

$$ARE_t(i) = \left| \frac{f_t(x_i) - y_i}{y_i} \right| \quad (18)$$

- Determine $f_t(x)$ error rate: $\sum_{i: ARE_t(i) > \phi} D_t(i)$
- Set $\beta_t = \varepsilon_t^n$, where n is the power coefficient (linear, square, or cubic, for example)

- Distribution of updates D_t as

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t & \text{if } ARE_t(i) \leq \phi \\ 1 & \text{otherwise} \end{cases} \quad (19)$$

Z_t is a normalization factor selected so that the distribution $D_{(t+1)}$ is obtained

- Assign $t = t + 1$.

4. Provide the end of the hypothesis:

$$f_{fin}(x) = \frac{\sum_t \left(\log \frac{1}{\beta_t} \right) f_t(x)}{\sum_t \left(\log \frac{1}{\beta_t} \right)} \quad (20)$$

Random Forest

An ensemble learning system called Random Forest is constructed using many decision trees. Using an RF classifier has several advantages, including:

- being extremely quick when working with large, high-dimensional datasets;
- being resilient to noise, multi-collinearity, and outliers;
- being unable to fail to fit on training datasets (which results in more accurate generalization compared to a single decision tree); and
- having a high degree of precision (Breiman, 2001)

RF gathers predictions from many decision trees that were trained using various random feature subsets and random sample subsets that were produced using bagging. The class with the greatest averaged probability score is then allocated when the forecasts are averaged (Sarna, Gutierrez, Mooney, & Zhu, 2022).

The CART algorithm maximizes the data gain at each node by utilizing the input attributes and potential thresholds. To group samples with comparable goal values, the decision tree divides the feature space recursively through each node. Let's assume for the purposes of our research that the data at node m of the decision tree is represented

by D_m , whose samples are n_m . Features k and its thresholds k_t are selected from our input features set (settlement reports from January 1995 to April 1997) for each possible split (k, k_t) to divide the data into two subsets: D_m^{left} with data n_m^l " and D_m^{right} " with samples n_m^r ". By reducing a function of loss determined by the mean squared error, the level of quality associated with this split is maximized as follows:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{D_m^{left}}{D_m} H(D_m^{left}(\theta)) + \frac{D_m^{right}}{D_m} H(D_m^{right}(\theta)) \quad (21)$$

In this in Equation (22) is the loss function, or mean squared error function, which is defined as follows:

$$H(D_m) = \frac{1}{n_m} \sum_{i=1}^{n_m} (y - \bar{y})^2 \quad (22)$$

Where, y represents the desired feature value and \bar{y} denotes the target feature value's mean inside the subgroup. Random Forest ensures independence of each decision tree by training separate subsets from the entire dataset using the bootstrap sampling approach. Figure 13 displays an illustration of the procedure. In this study, distinct Random Forest models for settlements and horizontal convergence, for example were trained for each goal feature.

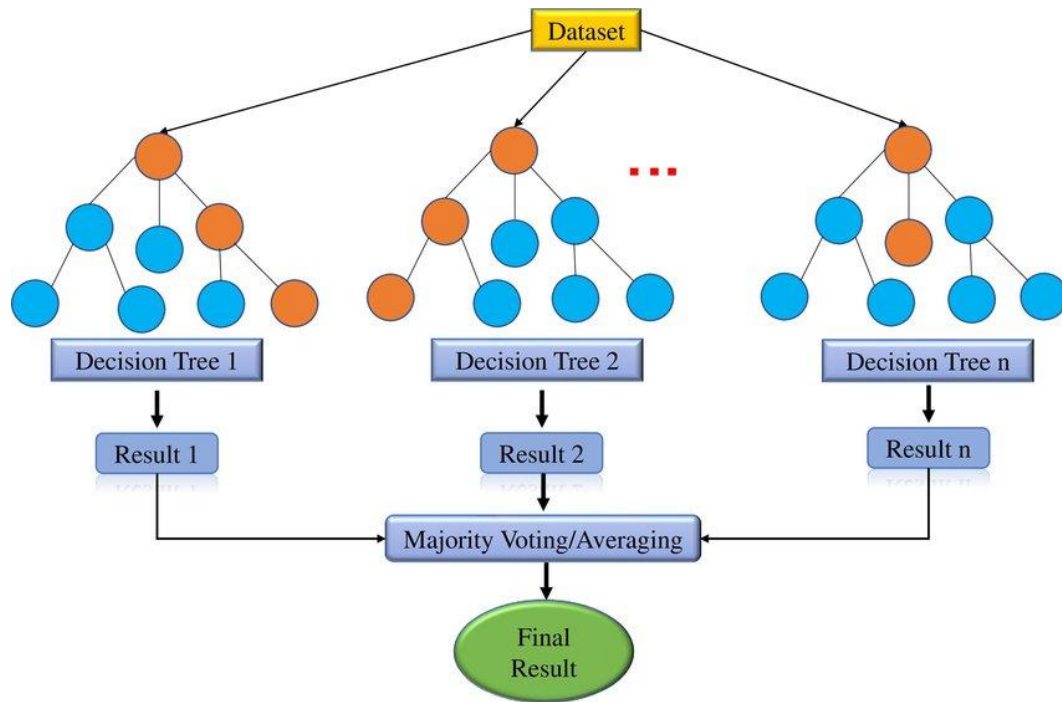


Figure 14: Random Forest Model(Pham & Tran, 2022)

Data Collection and Statistical Description of Dataset

The dataset is gotten from a Git-hub repository where a German PV system output and inputs are recorded and used to forecast solar power that will be produced. Germany has 21 distinct PV facilities established at various geographic regions for the model's training and testing. These establishments are situated in various locations, varying from rooftop to fully-functional solar farms. Each dataset includes historical power data for 990 days at a granularity of three hours together with NWP data. The PVs have nominal powers ranging from 100kW to 8500kW. The data was separated, and then it was normalized. All input values, with the exception of the output power, are normalized between 0 and 1. The target value or output power is normalized based on the power capacity of the corresponding PV installation. The figure 14 illustrates their distribution across Germany. The nominal output capacity of the relevant PV plant is used to normalize the target variable, the measured power output. Consequently,

permit the forecasting performance to be compared without accounting for the PV facilities' sizes.

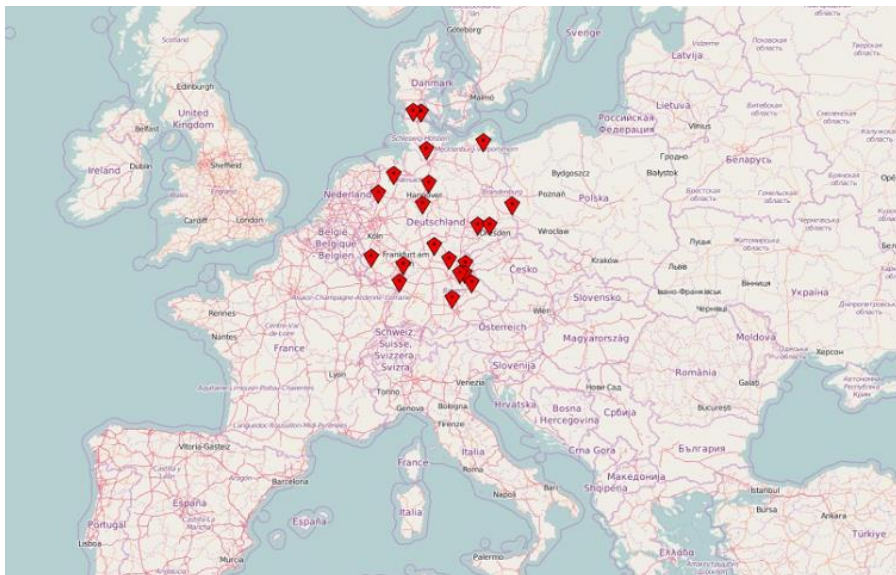


Figure 15: The Location of 21 Photovoltaic systems in Germany

Framework

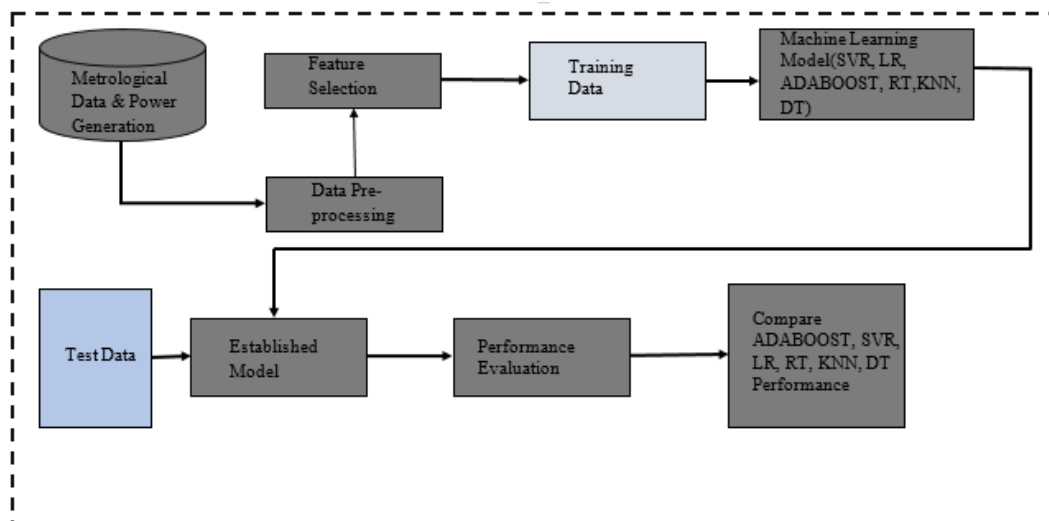


Figure 16: Machine Learning Models and Evaluation: Comparing SVR, LR, ADABOOST, RT, KNN, and DT

The framework of the research is discussed systematically in this section. Figure 15 shows the procedural framework used in this thesis, which makes use of machine learning methods which includes K-Nearest Neighbors, Random Forest, Support Vector Regression, Decision Tree, linear Regression and Adaboost. The first step involves collecting the metrological data and then data preprocessing which is the cleaning and converting the raw data into a format appropriate for machine learning. Cleaning the data and preprocessing are crucial phases in the creation of AI-based prediction techniques because the methodology create models for prediction from the information given. The prediction model becomes less accurate when raw data used is not in the appropriate format or data points are missing. Therefore no omitted data points and the meteorological data must be error-free. In this light the dataset on solar PV electricity underwent further procedures, including cleaning and analysis. This required locating and handling anomalies by making sure there is no missing data point. Another crucial stage in creating precise and useful predictive frameworks involves feature scaling. Scaling, normalization, and standardization which entails modifying the information to make it more suited for modeling this is an important components of feature engineering. By using these methods, one may guarantee that the information being analyzed is on a comparable scale, lessen the effect of unusual values, and enhance the accuracy of the model. The procedure of transforming the features in a data set to a comparable scale is known as feature scaling. The goal is to prevent variables with higher values from predominating and to make sure that every feature contributes similarly to the models. When working with data including elements that have varying ranges, degrees of magnitudes, or measurement units, scaling of features becomes important. Under such circumstances, the variance in the values of the features may result in skewed performance of models or issues with

learning. Therefore the type of feature scaling used in this thesis is normalization, which is a data preparation method used to standardize each feature in the dataset by converting them to a uniform scale. This procedure reduces the impact of different sizes on the models, improving the models ability to learn and modelling accuracy. Values are rescaled and altered throughout the normalization process to make them fall between 0 and 1. The normalization equation is expressed (23).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (23)$$

The maximum and minimum values of the features are denoted by X_max and X_min, respectively. Normalizing or standardizing the data. Furthermore in the next stage in order to further enhance the dataset, other procedures like choosing features was carried out. The study's objective is to create a predictive algorithm that can estimate solar PV power given a set of metrological input data. As a result, the generation of PV power was designated as the output variable, while meteorological information were employed as the input feature variables. The meteorological variables as shown in Appendix A, selected for the machine learning models are discussed below:

- Clear sky diffuse: This is the amount of solar energy that, on a clear sky day, reaches the Earth's surface after being scattered by gases and particles in the atmosphere in different directions.
- Clear sky direct: This describes solar energy that, on a clear sky day, reaches the Earth's surface straight from the sun without being impacted by air particle dispersion.
- Albedo: A surface's proportion of reflected sunlight is measured; higher numbers denote more reflectivity.
- WindComponentUat0 and WindComponentVat0: These represent the winds direction and strength in the east-west (U) and north-south (V) directions,

respectively, as well as the horizontal components of wind speed at ground level (0 meters above ground).

- `WindComponentUat100` and `WindComponentVat100`: At a height of 100 meters above sea level, they are comparable to the preceding.
- `TemperatureAt0`: This is the temperature of the air at zero meters above ground, or ground level.
- `RelativeHumidityAt1000`, `RelativeHumidityAt950`, and `RelativeHumidityAt0` : These represent the air's relative humidity at three distinct elevations: zero meters, 950 meters, and 1000 meters.
- `SolarRadiationGlobalAt0`: Indicates the entire amount of solar radiation that, at ground level, reaches the surface of the Earth.
- `SolarRadiationDirectAt0`: This is the amount of solar radiation that, at ground level, comes straight from the sun to the Earth's surface.
- `SolarRadiationDiffuseAt0`: This indicates the portion of solar radiation that reaches the surface of the Earth and scatters owing to ground-level air particles in different directions.
- `TotalCloudCoverAt0`: Indicates the portion of the sky that is cloud-covered at ground level.
- `Lower Wind Speed` and `Upper Wind Speed`: Defines the wind speed in the atmosphere at various vertical levels; "Lower" and "Upper" most likely relate to different ranges of altitude.

The next step is the training stage; two groups (Training and Testing) are formed from the chosen data points. The data is divided into the testing and training sets by 20:80 ratio, in thesis different ratios of testing and training data splitting is examined 30:70, 10:90. Afterwards each of the machine learning technique is trained using the training

dataset, subsequently in the trained models are tested using the testing dataset. Finally in the performance evaluation stage several machine learning evaluation metrics are employed, the following evaluation metrics are utilized stated in following section.

Performance Metric

Four different statistical assessment criteria were used to evaluate the suggested models' prediction performance. Which includes:

Mean Absolute Error (MAE):

The mean absolute deviation (MAE) of the variation between expected and actual values is quantified. By averaging the absolute differences between the values that were predicted and those that were seen, the MAE is calculated.

$$MAE = \frac{1}{N} \sum_{i=1}^N |o_i - p_i| \quad (24)$$

Root Mean Square Error (RMSE):

The standard deviation of the estimating errors is denoted by RMSE. The square root of the mean of the squared differences between the predicted and actual values is used to calculate the root mean square error, or RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (o_i - p_i)^2}{N}} \quad (25)$$

Coefficient of Determination (R²):

R² measures how strongly there is a linear connection between the model's predicted values and the real values. The percentage of the dependent variable's variation that can be predicted from the independent variables is how the R² is calculated.

$$R^2 = 1 - \frac{\sum_{i=1}^N (o_i - p_i)^2}{\sum_{i=1}^N (o_i - \bar{o})^2} \quad (26)$$

Mean Squared Error (MSE):

The average squared difference (MSE) between the actual and projected values is a statistic used to evaluate a predictive model's accuracy. It measures how well the model performs overall in terms of capturing the dataset's variability.

$$MSE = \frac{1}{N} \sum_{i=1}^N (o_i - p_i)^2 \quad (27)$$

Equations 5 through 8:

- The expected value for the i-th actual value o_i is denoted by p_i
- The number of samples is N.
- The average of the real data is shown by \bar{o}

CHAPTER IV

Results and Discussion

In this chapter, we evaluate models such as SVR, DT, AB, LR, RF, and KNN to determine the optimal model for power prediction in a PV system using the data collected. In the comparison, three metrics are used to understand the performance of the models in terms of how well they can keep prediction errors to the barest minimum. The error metrics used are MSE, MAE, and RMSE which measure the error value between the predicted and the actual value. Low values of RMSE, MSE, MAE shows a better performance of the model. Additionally, the accuracy of the models is examined using R² which quantifies the proportion of variance by the model, the value range from 0 to 1 and higher values of R² shows a better performance of the model. The ratio of data split is considered in the analysis because it has a significant impact on the performance and generalization of the machine learning model. In machine learning, the standard practice is to divide the whole dataset into two parts: training and testing. In this experiment, three ratios are tried: 80:20, 90:10, and 70:30. 80:20 is a ratio where 80% of the data is used for training each models while the remaining 20% is used for testing the performance of the model. In the same manner 90:10 is a ratio where 90% of the data is used for training each models while the remaining 10% is used for testing the performance of the model. This also applies to the 70:30 ratio.

Performance Analysis of Models using Error Evaluation Metrics:

Training Dataset

Tables 1, 2, and 3 show the well-detailed analysis of the machine learning models that were used for the different training of the dataset. It is seen that the decision tree's results are different from the other models with a really low error for the 10%, 20%, and as well as 30% datasets. This points to the problem of overfitting that the decision tree model has with this dataset. For example, at the 10% dataset, DT has values such as $5.1419E-15$ for RMSE, $2.6439E-29$ for MSE, and $2.8549E-15$ for MAE. Also, observing from the table, the Random Forest method gives the best results with the lowest RMSE, MSE, and MAE values of 0.1295, 0.0168, and 0.0605 respectively which shows a good ability to predict solar power. Observing the KNN and SVR, they tend to have similar results across the various datasets. The KNN has lower error metrics when compared to SVR. Linear Regression also gives more error values due to its less ideal fit for this dataset. Adaboost has the highest error metrics amongst these models making it the model with the poorest performance which shows there should be room for improving this model. One other thing to note from observing these results is, that as the data size increases from 10% to 30% the results of most of the models are better. The decision tree model's low error values are indicative of an overfitting problem on this dataset meaning it can learn extremely well on the training set but is unable to predict unseen datasets accurately. Further investigation will be done on the accuracy analysis. In conclusion, Random Forest has the best performance compared with the other models

Table 1: Error Metric Results of Training using 10% of the Dataset for Test

Model	RMSE	MSE	MAE
KNN	0.3044	0.0927	0.1486
SVR	0.3141	0.0987	0.1589
Random Forest	0.1295	0.0168	0.0605
Linear Regression	0.3626	0.1315	0.1993
Decision Tree Regressor	5.1419E-15	2.6439E-29	2.8549E-15
AdaBoost Regressor	0.4334	0.1878	0.3327

Table 2: Error Metric Results of Training using 20% of the Dataset for Test

Model	RMSE	MSE	MAE
KNN	0.3020	0.0912	0.1473
SVR	0.3106	0.0964	0.1576
Random Forest	0.1259	0.0159	0.0593
Linear Regression	0.3601	0.1297	0.1989
Decision Tree Regressor	4.7571E-15	2.2630E-29	2.7605E-15
AdaBoost Regressor	0.4255	0.1810	0.3217

Table 3: Error Metric Results of Training using 30% of the Dataset for Test

Model	RMSE	MSE	MAE
KNN	0.3017	0.0910	0.1474
SVR	0.3053	0.0932	0.1564
Random Forest	0.1283	0.0165	0.0607
Linear Regression	0.3590	0.1289	0.1991
Decision Tree Regressor	3.8470E-15	1.4799E-29	1.8254E-15
AdaBoost Regressor	0.4076	0.1662	0.3308

Testing Dataset

Below are Tables, 4, 5, and 6 of a comprehensive analysis of the different machine learning models on the testing dataset. Again, the Decision Tree is observed to have a higher error value when compared to the five other machine learning models, with the figures of RMSE of 0.4279 at the 10% dataset which shows a problem of model generalization. On the other hand, the Random Forest model still maintains a consistently good performance as it did in the training with low error metrics across all the datasets for testing. For example, in the 10% dataset, Random Forest has an

RMSE of 0.2897, MSE of 0.0839, and MAE of 0.1393, which shows its good ability to predict the solar output well. It can be observed that there is a slow degradation of the performance of models like Random Forest and KNN. Especially for the KNN model having an increased RMSE from 0.3348 at 10% to 0.3867 at 30%, which shows that this model is sensitive to how data varies. Also, LR and AB show a higher value of error as compared to other models which shows limitations in their prediction accuracy. For example, LR has an RMSE of 0.3215, MSE of 0.1033, and MAE of 0.1773 while testing the 10% dataset. It is seen that Random Forest has the best performance compared to other models.

Table 4: Error Metric Results of Testing using 10% of the Dataset for Test

Model	RMSE	MSE	MAE
KNN	0.3348	0.1121	0.1608
SVR	0.3006	0.0904	0.1555
Random Forest	0.2897	0.0839	0.1393
Linear Regression	0.3215	0.1033	0.1773
Decision Tree Regressor	0.4279	0.1831	0.1935
AdaBoost Regressor	0.4196	0.1761	0.3157

Table 5: Error Metric Results of Testing using 20% of the Dataset for Test

Model	RMSE	MSE	MAE
KNN	0.3648	0.1331	0.1780
SVR	0.3347	0.1120	0.1715
Random Forest	0.3414	0.1165	0.1620
Linear Regression	0.3534	0.1249	0.1921
Decision Tree Regressor	0.4695	0.2205	0.2169
AdaBoost Regressor	0.4436	0.1968	0.3274

Table 6: Error Metric Results of Testing using 30% of the Dataset for Test

Model	RMSE	MSE	MAE
KNN	0.3867	0.1496	0.1849
SVR	0.3507	0.1230	0.1815
Random Forest	0.3385	0.1146	0.1617
Linear Regression	0.3593	0.1291	0.1972
Decision Tree Regressor	0.4706	0.2215	0.2172
AdaBoost Regressor	0.4309	0.1857	0.3403

Performance Analysis of Models Using Accuracy Evaluation Metrics

In Figures 16, 17, 18 and also in Tables 7, 8 and 9. We conducted a comprehensive analysis of three testing and training ratio data splits, namely 10:90, 20:80, and 30:70, using the accuracy metric, namely the determinant coefficient (R²). The analysis will further provide insights into their predictive performance. As observed from the figures, Random Tree Forest consistently shows outstanding performance, leading to high R² scores for both the training and testing sets across all data split ratios. In Figures 17, 18, and 19, Random Forest obtained R² scores of 0.9216, 0.8837, and 0.8879, respectively. This showcases Random Forest's well-established predictive capabilities across different data split ratios. The effect of a model's ability to generalize can be better understood in models like KNN and SVR. The variation in the data split ratio had a slight impact on their prediction power, as their accuracy margin declined with the data split ratio varying from 10:90 to 20:80 and 30:70. On the other hand, Linear Regression and AdaBoost Regressor, in comparison to Random Forest, exhibited lower R² scores. This is indicative of their limited potential to learn the patterns that exist between the meteorological data and generated power. Finally, the Decision Tree Regressor achieved an extremely high accuracy in the training set but achieved the lowest accuracy in testing. This is indicative of overfitting meaning the decision tree regressor has a potential limitation on this type of dataset. It can learn the

training dataset well but is unable to predict unseen data accurately. In conclusion, the best-performing model is Random Forest, showcasing reliability and high prediction accuracy on both training and testing data across all data split ratios compared to other models.

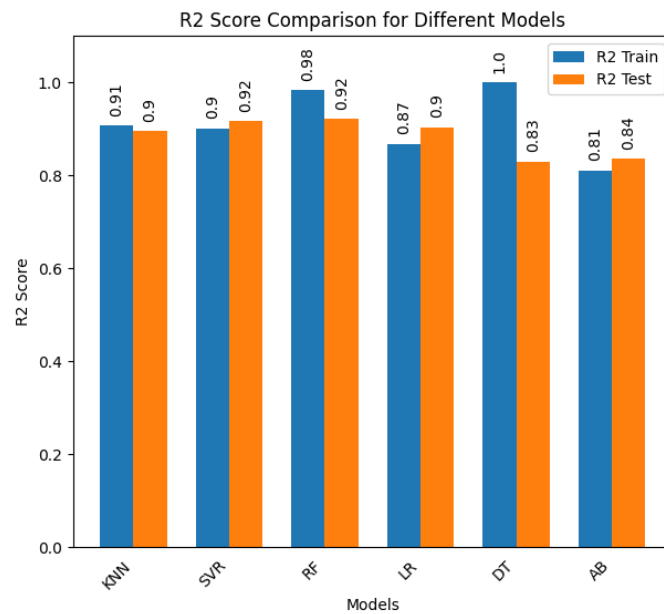


Figure 17: R2 Score results using 10% of the Dataset for Test

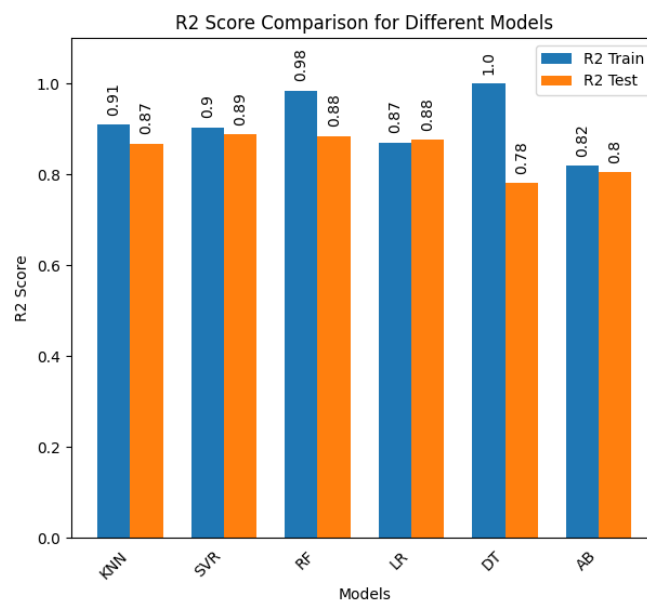


Figure 18: R2 Score results using 20% of the Dataset for Test

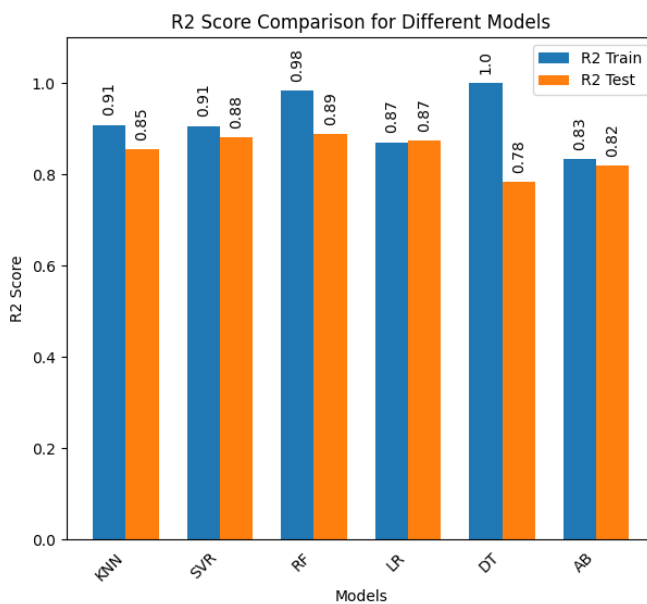


Figure 19: R2 Score results using 30% of the Dataset for Test

Table 7, 8 & 9 are the R2 Scores represented in a tabular form.

Table 7: R2 Score results using 10% of the Dataset

Model	Train	Test
KNN	0.91	0.9
SVR	0.9	0.92
Random Forest	0.98	0.92
Linear Regression	0.87	0.9
Decision Tree Regressor	1.0	0.83
AdaBoost Regressor	0.81	0.84

Table 8: R2 Score results using 20% of the Dataset

Model	Train	Test
KNN	0.91	0.87
SVR	0.9	0.89
Random Forest	0.98	0.88
Linear Regression	0.87	0.88
Decision Tree Regressor	1.0	0.78
AdaBoost Regressor	0.82	0.8

Table 9: R2 Score results using 30% of the Dataset

Model	Train	Test
KNN	0.91	0.85
SVR	0.91	0.88
Random Forest	0.98	0.89
Linear Regression	0.87	0.87
Decision Tree Regressor	1.0	0.78
AdaBoost Regressor	0.83	0.82

Performance Analysis of Models on Observed Versus Predicted Data

In the section, the actual vs predicted values of the model on various data split ratio is analyzed, Figures 19, 20, and 21 illustrate the scatter plot of these values, in the plots the x-axis represents the actual or observed data points while the y axis represents the predicted values of the model. The blue dots represent the individual data points of the predicted versus the actual data points while the red dashed line denotes the perfect prediction, this indicates where the blue dots should lie if the machine learning model predicted the actual data points perfectly meaning the predicted data points matches the actual datapoints exactly, if this is true the blue dot will be on or close to the red dashed line, otherwise the blue dotted line will be far away from the red dashed line. Worthy of note in practice, it's not uncommon for the predictions to deviate from the perfect prediction line depending on various factors e.g overfitting, underfitting, and noise, also the dataset or a machine learning model may not be suitable for the dataset due to inherent complexity in the dataset, here in Figures 19, 20 and 21, Random Forest has more blue dots closed to the red dashed line than other models, this validate the accuracy analysis established in the previous section.

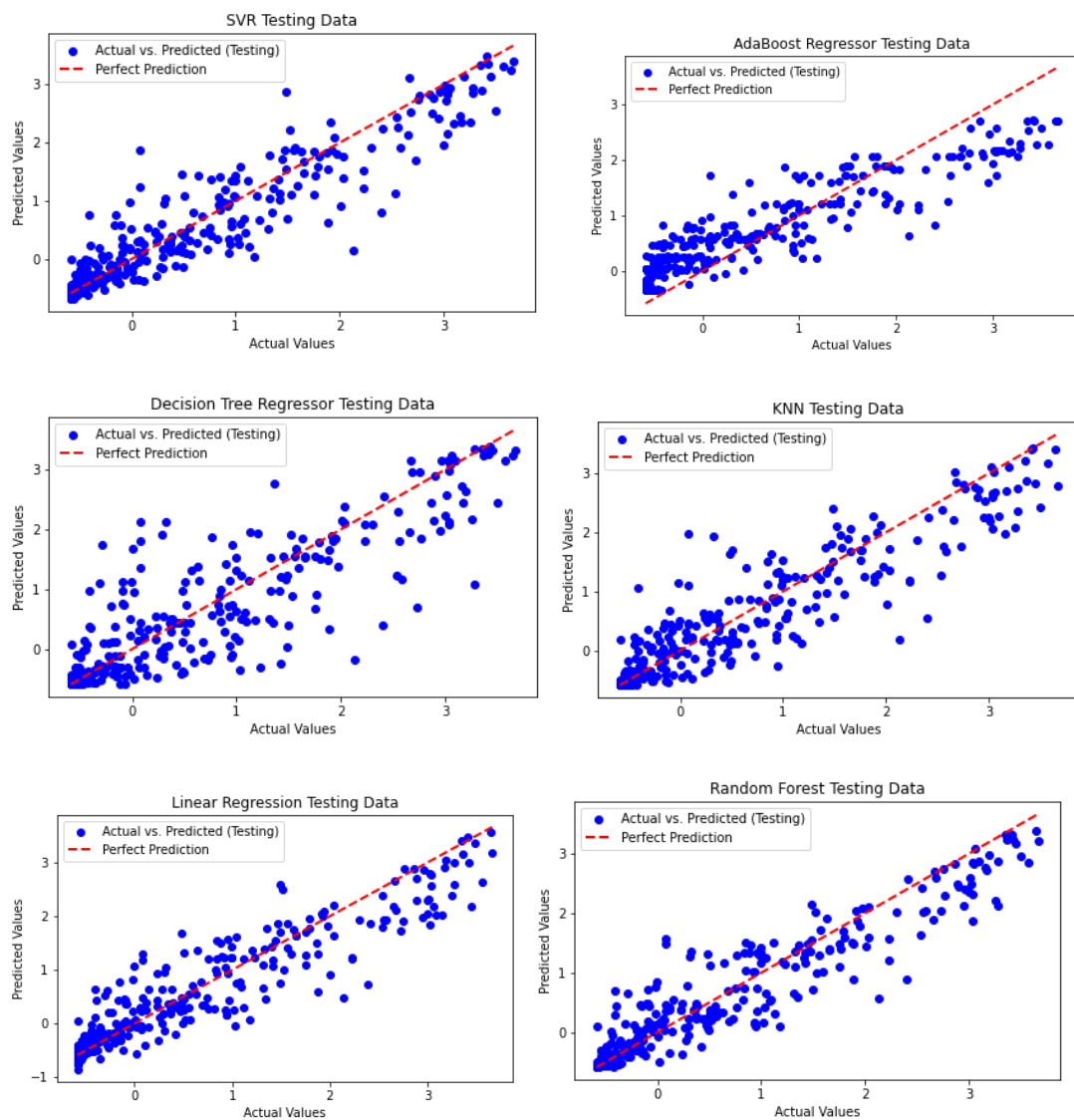


Figure 20: Predicted vs Observed using 10% of the Dataset for Test

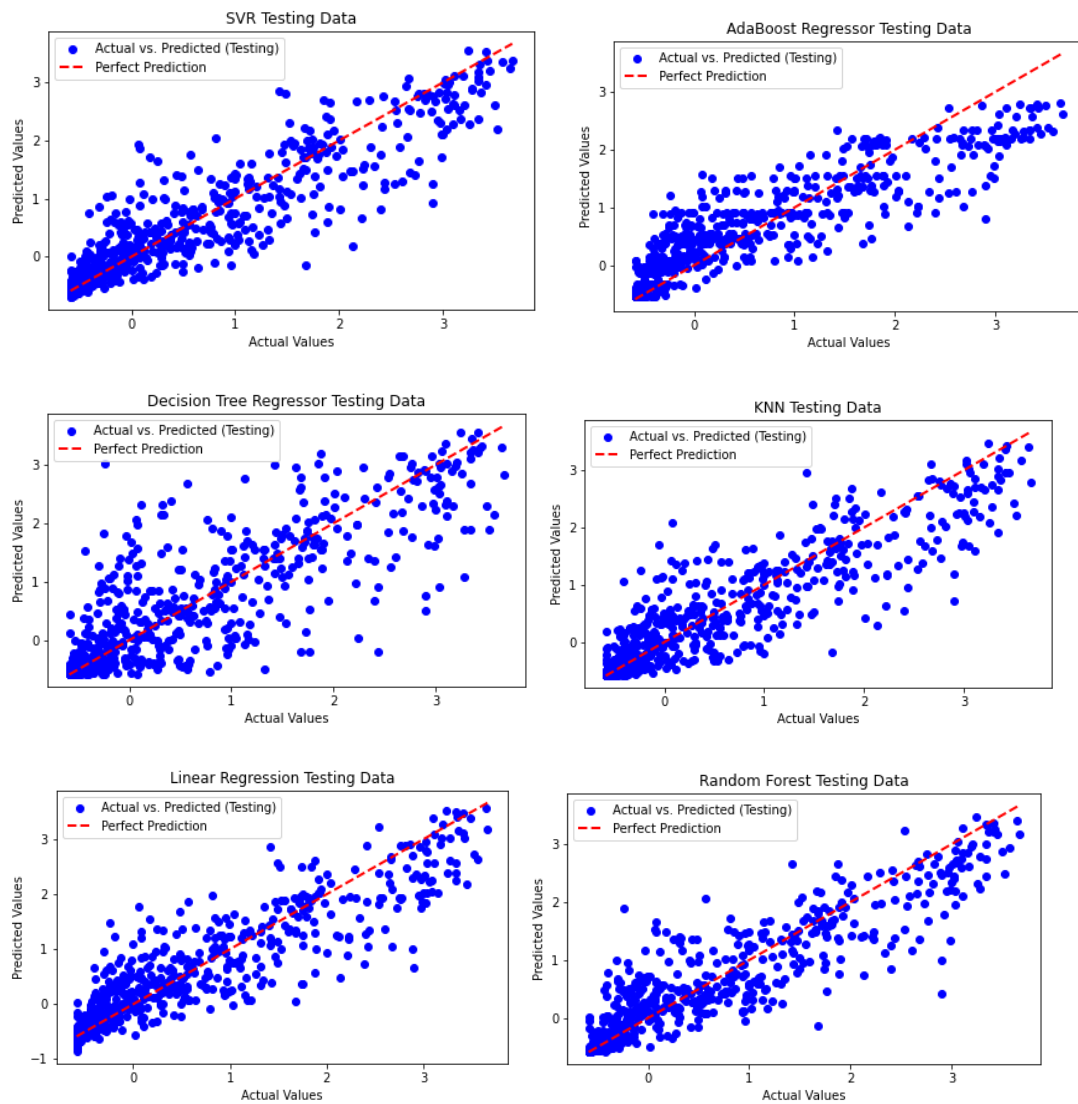


Figure 21: Predicted vs Observed using 20% of the Dataset for Test

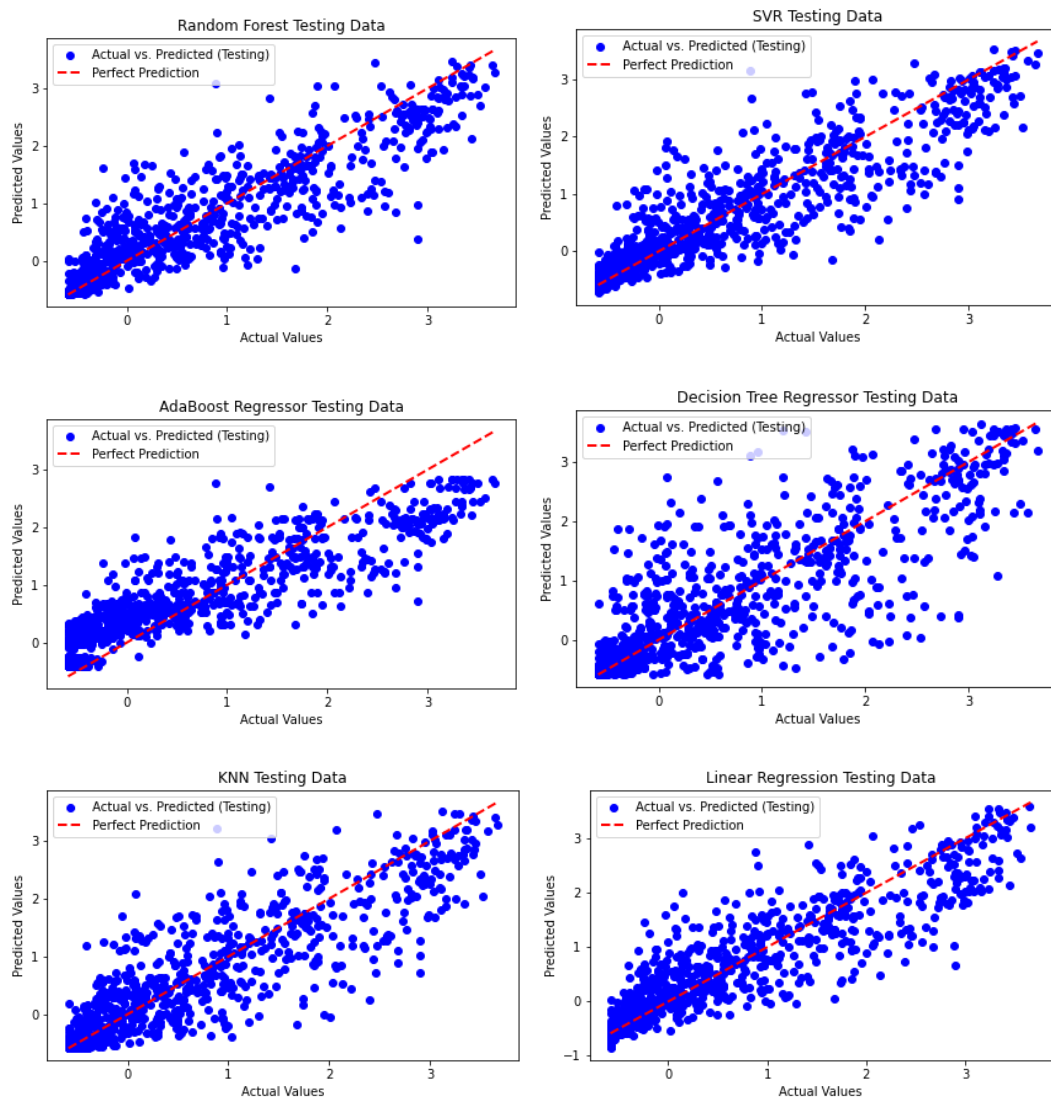


Figure 22: Predicted vs Observed using 30% of the Dataset for Test

Conclusion

This study is an extensive comparison of different machine learning models used in the prediction of solar power that is being generated in photovoltaic systems. Six unique models were employed which include the Decision Tree(DT), Linear Regression(LR), Support Vector Regression(SVR), Random Forest(RF), and the K-Nearest Neighbors(KNN). Applying these models over an already existing meteorological dataset with variables and records of the solar power generated. The models are evaluated carefully using error metrics such as Root Mean Squared Error, Mean Squared Error (MSE), R-squared (R²), and Mean Absolute Error (MAE) to measure the model's efficiency. The results of this study show that Random Forest is the best model relative to the five other models to give accurate predictions of the solar power generated across the training and testing sets of data. RF has consistently the lowest error metrics of 0.1295, 0.0168, and 0.0605 respectively in the training and RMSE of value of 0.2897, MSE of 0.0839, and MAE of 0.1393 in the testing which shows a good ability to predict solar power and the highest R-squared scores of 0.9216, 0.8837, and 0.8879 on all the different ratios of data splits which were 10:90, 20:80, and 30:70 respectively. This is connected to RF's ability to adapt effectively and handle complex nonlinear input and output relationships while handling over-fitting problems. The RMSE/MSE/MAE value are decreasing in the training results and seems to be increasing in the testing results and this suggests that the models might be overfitting to the training data which is when the model learns too well on the training data. The results of this study not only show a vivid recommendation for the use of Random Forests but also contribute to the wide use of machine learning in Renewable energy. The study highlights how the choice of a machine learning model can increase

the efficiency of the prediction which in turn, has a significant effect on reliability and efficiency in solar power predictions in photovoltaic systems.

Recommendations

The findings from this study are used to support the following suggestions whose objective is to improve solar power prediction in photovoltaic systems.

- The adoption of the Random forest method as the core model. From observing the outstanding performance of this model in handling the intricate nonlinear meteorological data it shows that RF will well fit into proffering solutions in solar power prediction.
- To improve the model's generalizability and accuracy adding/expanding the training dataset with more meteorological data is recommended
- Also, improving the prediction performance of the model integrating other machine learning models like the SVR or KNN into what is known as the ensemble technique is very well recommended.
- To keep the model consistent with changing seasons or weather conditions, it will be best to always update them with new meteorological data of those regions.

Future Works

This study shows that there is more research to be put in place concerning solar power prediction in photovoltaic systems, which includes but is not limited to:

- The application of deep learning models in solar power prediction using its ability to outperform the conventional machine learning models, working effectively with more complex nonlinear relationships.

- Having extended time horizons to have more effective energy management and planning, having meteorological data with the records of solar power generated over a longer time frame rather than the conventional daily predictions.
- To ensure the smooth integration of solar power into the energy grid system, a study on the real-time dataset can be included in the model for predicting solar power. This will advance the reliability and efficiency of photovoltaic systems and facilitate the wider use of renewable energy sources.

References

- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In M. W. Berry, A. Mohamed, & B. W. Yap (Eds.), *Supervised and Unsupervised Learning for Data Science* (pp. 3–21). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2_1
- Amjady, N., & Hemmati, M. (2009). Day-ahead price forecasting of electricity markets by a hybrid intelligent system. *European Transactions on Electrical Power*, *19*(1), 89–102. <https://doi.org/10.1002/etep.242>
- Bartlett, P., Freund, Y., Lee, W. S., & Schapire, R. E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, *26*(5). <https://doi.org/10.1214/aos/1024691352>
- Behera, M. K., Majumder, I., & Nayak, N. (2018). Solar photovoltaic power forecasting using optimized modified extreme learning machine technique. *Engineering Science and Technology, an International Journal*, *21*(3), 428–438. <https://doi.org/10.1016/j.jestch.2018.04.013>
- Benti, N. E., Chaka, M. D., & Semie, A. G. (2023). Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects. *Sustainability*, *15*(9), 7087. <https://doi.org/10.3390/su15097087>
- Breiman, L. (2001). Random Forest. *Machine Learning*, *45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

- Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., Zhang, L.-X., & Li, H.-D. (2010). The boosting: A new idea of building models. *Chemometrics and Intelligent Laboratory Systems*, *100*(1), 1–11. <https://doi.org/10.1016/j.chemolab.2009.09.002>
- Castillo-Rojas, W., Medina Quispe, F., & Hernández, C. (2023). Photovoltaic Energy Forecast Using Weather Data through a Hybrid Model of Recurrent and Shallow Neural Networks. *Energies*, *16*(13), 5093. <https://doi.org/10.3390/en16135093>
- Chahboun, S., & Maaroufi, M. (2021). Performance Comparison of Support Vector Regression, Random Forest and Multiple Linear Regression to Forecast the Power of Photovoltaic Panels. *2021 9th International Renewable and Sustainable Energy Conference (IRSEC)*, 1–4. Morocco: IEEE. <https://doi.org/10.1109/IRSEC53969.2021.9741154>
- Chen, C.-R., & Kartini, U. (2017). K-Nearest Neighbor Neural Network Models for Very Short-Term Global Solar Irradiance Forecasting Based on Meteorological Data. *Energies*, *10*(2), 186. <https://doi.org/10.3390/en10020186>
- Chen, J.-L., Li, G.-S., & Wu, S.-J. (2013). Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. *Energy Conversion and Management*, *75*, 311–318. <https://doi.org/10.1016/j.enconman.2013.06.034>
- Chen, Y., Kong, R., & Kong, L. (2020). Applications of artificial intelligence in astronomical big data. In *Big Data in Astronomy* (pp. 347–375). Elsevier. <https://doi.org/10.1016/B978-0-12-819084-5.00006-7>
- Chow, C. W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., & Washom, B. (2011). Intra-hour forecasting with a total sky imager at the UC

- San Diego solar energy testbed. *Solar Energy*, 85(11), 2881–2893.
<https://doi.org/10.1016/j.solener.2011.08.025>
- Chu, Y., & Coimbra, C. F. M. (2017). Short-term probabilistic forecasts for Direct Normal Irradiance. *Renewable Energy*, 101, 526–536.
<https://doi.org/10.1016/j.renene.2016.09.012>
- Chu, Y., Urquhart, B., Gohari, S. M. I., Pedro, H. T. C., Kleissl, J., & Coimbra, C. F. M. (2015). Short-term reforecasting of power output from a 48 MWe solar PV plant. *Solar Energy*, 112, 68–77.
<https://doi.org/10.1016/j.solener.2014.11.017>
- Cohen, S. (2021). The basics of machine learning: Strategies and techniques. In *Artificial Intelligence and Deep Learning in Pathology* (pp. 13–40). Elsevier.
<https://doi.org/10.1016/B978-0-323-67538-3.00002-6>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Díaz, P., Peña, R., Muñoz, J., Arias, C. A., & Sandoval, D. (2011). Field analysis of solar PV-based collective systems for rural electrification. *Energy*, 36(5), 2509–2516. <https://doi.org/10.1016/j.energy.2011.01.043>
- Dimililer, K., Kayali, D., & Tackie, S. N. (2023). Power Demand Prediction of North Cyprus using Machine Learning Regressor Models. *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 1–5. Sivas, Turkiye: IEEE. <https://doi.org/10.1109/ASYU58738.2023.10296786>
- El Bouchefry, K., & De Souza, R. S. (2020). Learning in Big Data: Introduction to Machine Learning. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (pp. 225–249). Elsevier. <https://doi.org/10.1016/B978-0-12-819154-5.00023-0>

- Elhadidy, M. A. (2002). Performance evaluation of hybrid (wind/solar/diesel) power systems. *Renewable Energy*, 26(3), 401–413. [https://doi.org/10.1016/S0960-1481\(01\)00139-2](https://doi.org/10.1016/S0960-1481(01)00139-2)
- Fayyad, M. (2023). Reconstructing lease-to-own contracts: A contemporary approach to Islamic banking standards. *Heliyon*, 9(9), e19319. <https://doi.org/10.1016/j.heliyon.2023.e19319>
- Gabrieli D. Silva, Mariza Ferro, & Schulze, B. (2021). *Performance and Energy efficiency Analysis of Machine Learning algorithms Towards Green AI: A case study of decision tree algorithms*. <https://doi.org/10.13140/RG.2.2.27740.31363>
- Gaviria, J. F., Narváez, G., Guillen, C., Giraldo, L. F., & Bressan, M. (2022). Machine learning in photovoltaic systems: A review. *Renewable Energy*, 196, 298–318. <https://doi.org/10.1016/j.renene.2022.06.105>
- Haussler, D., & Warmuth, M. (1993). The Probably Approximately Correct (PAC) and Other Learning Models. In A. L. Meyrowitz & S. Chipman (Eds.), *Foundations of Knowledge Acquisition* (pp. 291–312). Boston, MA: Springer US. https://doi.org/10.1007/978-0-585-27366-2_9
- Huynh, T., Nibali, A., & He, Z. (2022). Semi-supervised learning for medical image classification using imbalanced training data. *Computer Methods and Programs in Biomedicine*, 216, 106628. <https://doi.org/10.1016/j.cmpb.2022.106628>
- Jensen, C. S., & Snodgrass, R. T. (2009). Snapshot Equivalence. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 2659–2659). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-39940-9_1417

- Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy, 51*(5), 675–687.
<https://doi.org/10.1016/j.beth.2020.05.002>
- Kramer, O. (2013). *Dimensionality Reduction with Unsupervised Nearest Neighbors*. Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-38652-7>
- Kuo, P.-H., & Huang, C.-J. (2018). A Green Energy Application in Energy Management Systems by an Artificial Intelligence-Based Solar Radiation Forecasting Model. *Energies, 11*(4), 819. <https://doi.org/10.3390/en11040819>
- Kuriakose, A. M., Kariyalil, D. P., Augusthy, M., Sarath, S., Jacob, J., & Antony, N. R. (2020). Comparison of Artificial Neural Network, Linear Regression and Support Vector Machine for Prediction of Solar PV Power. *2020 IEEE Pune Section International Conference (PuneCon)*, 1–6. Pune, India: IEEE.
<https://doi.org/10.1109/PuneCon50868.2020.9362442>
- Larson, D. P., Nonnenmacher, L., & Coimbra, C. F. M. (2016). Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest. *Renewable Energy, 91*, 11–20.
<https://doi.org/10.1016/j.renene.2016.01.039>
- Larson, V. E. (2013). Forecasting Solar Irradiance with Numerical Weather Prediction Models. In *Solar Energy Forecasting and Resource Assessment* (pp. 299–318). Elsevier. <https://doi.org/10.1016/B978-0-12-397177-7.00012-7>
- Liu, S., See, K. C., Ngiam, K. Y., Celi, L. A., Sun, X., & Feng, M. (2020). Reinforcement Learning for Clinical Decision Support in Critical Care:

Comprehensive Review. *Journal of Medical Internet Research*, 22(7), e18477. <https://doi.org/10.2196/18477>

Martinez Alvarez, F., Troncoso, A., Riquelme, J. C., & Aguilar Ruiz, J. S. (2011).

Energy Time Series Forecasting Based on Pattern Sequence Similarity. *IEEE Transactions on Knowledge and Data Engineering*, 23(8), 1230–1243. <https://doi.org/10.1109/TKDE.2010.227>

Mellit, A., Massi Pavan, A., & Lughi, V. (2014). Short-term forecasting of power production in a large-scale photovoltaic plant. *Solar Energy*, 105, 401–413. <https://doi.org/10.1016/j.solener.2014.03.018>

Mirjalili, M. A., Aslani, A., Zahedi, R., & Soleimani, M. (2023). A comparative study of machine learning and deep learning methods for energy balance prediction in a hybrid building-renewable energy system. *Sustainable Energy Research*, 10(1), 8. <https://doi.org/10.1186/s40807-023-00078-9>

Moosavian, S. M., Rahim, N. A., Selvaraj, J., & Solangi, K. H. (2013). Energy policy to promote photovoltaic generation. *Renewable and Sustainable Energy Reviews*, 25, 44–58. <https://doi.org/10.1016/j.rser.2013.03.030>

Ogidan, E. T., Dimililer, K., & Ever, Y. K. (2018). Machine Learning for Expert Systems in Data Analysis. *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–5. Ankara: IEEE. <https://doi.org/10.1109/ISMSIT.2018.8567251>

Olatomiwa, L., Mekhilef, S., Shamshirband, S., Mohammadi, K., Petković, D., & Sudheer, C. (2015). A support vector machine–firefly algorithm-based model for global solar radiation prediction. *Solar Energy*, 115, 632–644. <https://doi.org/10.1016/j.solener.2015.03.015>

- Ozerdem, O. C., Tackie, S., & Biricik, S. (2015). Performance evaluation of Serhatkoy (1.2 MW) PV power plant. *2015 9th International Conference on Electrical and Electronics Engineering (ELECO)*, 398–402. Bursa: IEEE. <https://doi.org/10.1109/ELECO.2015.7394510>
- Pedro, H. T. C., & Coimbra, C. F. M. (2012). Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, *86*(7), 2017–2028. <https://doi.org/10.1016/j.solener.2012.04.004>
- Pham, T. A., & Tran, V. Q. (2022). Developing random forest hybridization models for estimating the axial bearing capacity of pile. *PLOS ONE*, *17*(3), e0265747. <https://doi.org/10.1371/journal.pone.0265747>
- Ramli, M. A. M., Twaha, S., & Al-Turki, Y. A. (2015). Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study. *Energy Conversion and Management*, *105*, 442–452. <https://doi.org/10.1016/j.enconman.2015.07.083>
- Rana, M., Koprinska, I., & Agelidis, V. G. (2015). 2D-interval forecasts for solar power production. *Solar Energy*, *122*, 191–203. <https://doi.org/10.1016/j.solener.2015.08.018>
- Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. <https://doi.org/10.1007/s42979-021-00592-x>
- Sarna, S., Gutierrez, M., Mooney, M., & Zhu, M. (2022). Predicting Upcoming Collapse Incidents During Tunneling in Rocks with Continuation Length Based on Influence Zone. *Rock Mechanics and Rock Engineering*, *55*(10), 5905–5931. <https://doi.org/10.1007/s00603-022-02971-z>

- Satinet, C., & Fouss, F. (2022). A Supervised Machine Learning Classification Framework for Clothing Products' Sustainability. *Sustainability*, *14*(3), 1334. <https://doi.org/10.3390/su14031334>
- Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student Performance Prediction and Classification Using Machine Learning Algorithms. *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, 7–11. Cambridge United Kingdom: ACM. <https://doi.org/10.1145/3318396.3318419>
- Sekeroglu, B., Ever, Y. K., Dimililer, K., & Al-Turjman, F. (2022). Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems. *Data Intelligence*, *4*(3), 620–652. https://doi.org/10.1162/dint_a_00155
- Sfetsos, A., & Coonick, A. H. (2000). Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Solar Energy*, *68*(2), 169–178. [https://doi.org/10.1016/S0038-092X\(99\)00064-X](https://doi.org/10.1016/S0038-092X(99)00064-X)
- Sharkawy, A.-N., Ali, M., Mousa, H., Ali, A., & Abdel-Jaber, G. (2022). Machine Learning Method for Solar PV Output Power Prediction. *SVU-International Journal of Engineering Sciences and Applications*, *3*(2), 123–130. <https://doi.org/10.21608/svusrc.2022.157039.1066>
- Smys, S., Iliyasu, A. M., Bestak, R., & Shi, F. (Eds.). (2020). *New Trends in Computational Vision and Bio-inspired Computing: Selected works presented at the ICCVBIC 2018, Coimbatore, India*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-030-41862-5>
- Subasi, A., Khateeb, K., Brahim, T., & Sarirete, A. (2020). Human activity recognition using machine learning methods in a smart healthcare

- environment. In *Innovation in Health Informatics* (pp. 123–144). Elsevier.
<https://doi.org/10.1016/B978-0-12-819043-2.00005-8>
- Timilsina, G. R., Kurdgelashvili, L., & Narbel, P. A. (2012). Solar energy: Markets, economics and policies. *Renewable and Sustainable Energy Reviews*, *16*(1), 449–465. <https://doi.org/10.1016/j.rser.2011.08.009>
- Van Der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... Yu, T. (2014). scikit-image: Image processing in Python. *PeerJ*, *2*, e453. <https://doi.org/10.7717/peerj.453>
- Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. <https://doi.org/10.1007/s10994-019-05855-6>
- Wang, H., Yi, H., Peng, J., Wang, G., Liu, Y., Jiang, H., & Liu, W. (2017). Deterministic and probabilistic forecasting of photovoltaic power based on deep convolutional neural network. *Energy Conversion and Management*, *153*, 409–422. <https://doi.org/10.1016/j.enconman.2017.10.008>
- Wolff, B., Kühnert, J., Lorenz, E., Kramer, O., & Heinemann, D. (2016). Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. *Solar Energy*, *135*, 197–208.
<https://doi.org/10.1016/j.solener.2016.05.051>

Appendices













Appendix A

Snippet of Excel Sheet of Dataset with Variables and Output

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	clearsky_clearsky_Albedo	WindCom	WindCom	WindCom	WindCom	WindCom	Temperat	RelativeH	RelativeH	RelativeH	SolarRadi	SolarRadi	SolarRadi	TotalCloud	LowerWir	UpperWir	UpperWir	UpperWir	power_normed	
2	0	0.038155	0.460675	0.758315	0.487756	0.763081	0.47329	0.907665	0.938805	0.951625	0	0	0	0.000521	1	0.458204	0.494272	0.56342	0	
3	0	0.031405	0.642463	0.454828	0.660619	0.435388	0.448161	0.881732	0.885329	0.927561	0	0	0	0.000521	1	0.432281	0.456086	0.73223	0	
4	0	0.031405	0.683629	0.52807	0.708264	0.509767	0.395531	0.829865	0.849678	0.893764	0	0	0	0.000521	1	0.513296	0.545543	0.707816	0	
5	0.450238	0.264241	0.031405	0.776618	0.602992	0.770723	0.579985	0.323382	0.867503	0.881761	0.071579	0.01771	0.134254	0.71875	0.676684	0.663653	0.693619	0.01456		
6	0.538696	0.405148	0.031405	0.770214	0.575414	0.75863	0.550991	0.41731	0.661299	0.7249	0.770379	0.254848	0.119574	0.366006	0.656101	0.634276	0.700475	0.152064		
7	0.280643	0.096953	0.031405	0.710567	0.655256	0.726366	0.635477	0.419674	0.570532	0.623889	0.703224	0.25795	0.194008	0.227581	0.613585	0.630121	0.671362	0.117406		
8	0	0.031405	0.668423	0.681782	0.697703	0.669127	0.39163	0.687232	0.736784	0.802464	0.011745	0.00213	0.023983	0.5625	0.577523	0.614073	0.656483	0.003783		
9	0	0.031405	0.589322	0.693925	0.621095	0.689544	0.380308	0.700199	0.742725	0.807672	0	0	0	0.000521	0.976563	0.495966	0.537326	0.631084	0	
10	0	0.031405	0.494542	0.752062	0.530813	0.762205	0.374249	0.680749	0.7249	0.79952	0	0	0	0.000521	1	0.470942	0.522476	0.58191	0	
11	0	0	0.0313	0.439302	0.939212	0.462968	0.935082	0.394182	0.823382	0.837794	0.890678	0	0	0.000521	1	0.663494	0.687923	0.53609	0	
12	0	0	0.0313	0.510748	0.993549	0.533232	0.986469	0.400303	0.894698	0.903154	0.935885	0	0	0.000521	1	0.754332	0.780047	0.594455	0	
13	0.45837	0.258296	0.0313	0.534658	0.95227	0.551307	0.933229	0.434506	0.881732	0.897212	0.925351	0.019945	0	0.047445	1	0.719122	0.728249	0.564971	0.004739	
14	0.535056	0.397918	0.0313	0.565906	0.938041	0.582559	0.918366	0.463075	0.823382	0.831852	0.886411	0.087313	0.009587	0.187174	0.992188	0.722246	0.732029	0.576056	0.051102	
15	0.275398	0.093168	0.0313	0.764483	0.824668	0.774361	0.800155	0.460652	0.855799	0.861561	0.908965	0.110139	0.034487	0.192127	0.945313	0.799754	0.80588	0.642499	0.048547	
16	0	0	0.031304	0.602578	0.245274	0.608964	0.207016	0.381661	0.836349	0.849678	0.890227	0.006759	0.000932	0.014599	0.945313	0.413917	0.440239	0.837845	0.002769	
17	0	0	0.031295	0.578074	0.404715	0.610409	0.366544	0.355602	0.849315	0.85562	0.909555	0	0	0.000521	0.734375	0.329882	0.37776	0.764356	0	
18	0	0	0.031295	0.571105	0.452125	0.602259	0.420525	0.335169	0.777999	0.784318	0.863438	0	0	0.000521	0.546875	0.32493	0.365165	0.735873	0	
19	0	0	0.031196	0.489196	0.542184	0.557209	0.532627	0.277629	0.855799	0.837794	0.930792	0	0	0.000521	0.03125	0.258364	0.340852	0.666894	0	
20	0	0	0.031196	0.448957	0.632661	0.500721	0.654729	0.269007	0.888215	0.891271	0.948289	0	0	0.000521	0.398438	0.309755	0.38364	0.592551	0	
21	0.441425	0.252445	0.031196	0.462179	0.76761	0.467953	0.733972	0.335161	0.803932	0.85562	0.8693	0.09651	0.048336	0.132951	1	0.469653	0.449139	0.55858	0.01199	
22	0.531421	0.390831	0.031196	0.480042	0.863673	0.488513	0.838587	0.349842	0.719649	0.817736	0.124765	0.010786	0.272941	1	0.589307	0.582815	0.533791	0.082348		
23	0.27022	0.089503	0.031196	0.45534	0.835392	0.475613	0.830284	0.358246	0.920632	0.938805	0.954949	0.036011	0.001465	0.082377	1	0.545479	0.566652	0.549563	0.056184	

Appendix B

Turnitin Similarity Report I

Goodness Orish	CHI 12032024	2318513083	12-Mar-2024	2%		
Goodness Orish	ALLTH 12032024	2318513089	12-Mar-2024	10%		
Goodness Orish	CH2 12032024	2318513091	12-Mar-2024	8%		
Goodness Orish	CH4 12032024	2318513092	12-Mar-2024	5%		
Goodness Orish	CNC 12032024	2318513096	12-Mar-2024	0%		
Goodness Orish	CH3 12032024	2318513098	12-Mar-2024	13%		
Goodness Orish	ABS 12032024	2318513079	12-Mar-2024	0%	