

University of Bahrain

College of Science

**Department of
Postgraduate Programs**



Short-term Forecasting for Photovoltaic Power Output Using Machine Learning Algorithms

Thesis submitted in partial fulfilment of the requirements for the
degree of M.Sc. In Big Data Science and Analytics

Prepared by:

Hajar Al-Bahri

20134920

Supervised by:

Dr. Hesham Al-Ammal

Assistant Professor

University of Bahrain

Kingdom of Bahrain

May 2022

Defence Committee's Approval

University of Bahrain
Deanship of Graduate Studies
And scientific Research



جامعة البحرين
عمادة الدراسات العليا والبحث العلمي

Thesis Discussion Committee Decision

Student Name: HAJAR ZOUHAIR KHALIFA ALBAHRI	Student ID: 20134920
Department: Choose an item.	College: Science
Program: MSc in Big Data Science and Analytics	
Thesis Title: Short-Term Forecasting for Photovoltaic Power Output Using Machine Learning Algorithms	
Supervisor Name: Hesham Al-Ammal	Student GPA: 3.50

The Discussion Committee met on 11th May 2022 at 11:00 AM. In Session 1 to enter text.. Upon discussing the student whose details are stated above in public, the discussion committee recommends the following:

The student is awarded a grade Pass in the thesis subject.

☐ Without making any changes to the thesis.

☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

Name		Signature
1. Hesham Al-Ammal	(Advisor)	
2. Nabil M. Hewahi	(Internal Examiner)	
3. Ahmed Fawzi Ootom	(External Examiner)	

Abstract

As the demand for energy rises significantly in all countries, Photovoltaic power generation has thus increased significantly in popularity and demand as it offers clean and accessible energy. Nonetheless, the challenge of solar power is in its instability as it depends significantly on weather variables. Thus, forecasting day-ahead PV power generation is essential for smart grid systems scheduling and operation and solar energy providers. This thesis aims to establish the relationship between the meteorological variables that mostly affect solar power generation in the kingdom of Bahrain. Secondly, investigating other ML models that have not been deployed in past literature to forecast the PV power output, which are GRU, Bi-LSTM, and Bi-GRU. In this thesis, the Spearman Correlation Coefficient was used to discover the strength of 15 meteorological attributes with the PV power output, which were seven variables. Using a simple vanilla structure and a hybrid structure containing one layer of CNN, the models were tested and compared against the LSTM. To further test the model's ability to accurately predict on less historical data, the five-month dataset was split into three and two months that capture the winter and summer seasons, respectively. The tests were run using data collected from the renewable energy lab at UoB. After analyzing the results, it was concluded that LSTM gave the worst performance across all datasets using both structures, the simple and hybrid. The hybrid form enhanced the performance most significantly in the LSTM than the other models, nonetheless still producing the highest RMSE and MAE. The CNN-GRU and CNN-Bi-LSTM gave low RMSE and MAE when trained on the summer data in contrast to CNN-Bi-GRU, which worsened radically, proving robust to provide reliable and accurate predictions on less historical data. Although overall results were comparable between GRU, Bi-LSTM, and Bi-GRU, both in hybrid and vanilla structures, the best model that gave the highest accuracy for the five-month dataset was CNN-Bi-GRU with RMSE 20.74 and MAE 9.64. As for summer and winter data, the CNN-Bi-LSTM gave the best performance for both, with RMSE of 8.17, 21.72, and MAE of 4.23 and 10.32, respectively.

Keywords: Photovoltaic Power Output, Renewable Energy, Solar Power Generation, Machine Learning Algorithms, Time Series Forecast, GRU, Bi-GRU, Bi-LSTM, LSTM

Dedication

To my mother

Sonia Hadrich

My strongest pillar of support, who guided me through all difficulties in life. Everything I am and will be is because of you.

Your unconditional love has kept me going.

Acknowledgements

First and foremost, all thanks and praises for Allah the almighty, who has showered me with countless blessings so that I have reached where I am today and finally accomplished this thesis. Without his grace, this would not have been a reality.

The success of this thesis comes from the encouragement and the support of so many people. I would like to extend my appreciation and sincere thanks to:

Dr. Hesham Al-Ammal, my supervisor, for giving me continuous support and invaluable insights that helped me immensely in writing this thesis. It was an honor to work under your supervision.

Dr. Hanan Buflasa, Assistant Professor, for providing valuable datasets and the many meetings to assist me with the necessary information.

The thesis committee: Dr. Nabil Hewahi, Internal Examiner, and Prof. Ahmed Otoom, External Examiner, for their encouraging words and insightful comments.

Dr. Mohammed Redha Qadar, Dean of Graduate Studies & Scientific Research, for believing in me and making this journey happen.

Dr. Sawsan Hilal, program coordinator, who assisted me from the beginning of the master's journey until the very end and challenged me to bring out the best in me. Your remarkable effort and patience are highly appreciated.

Dr. Ahmed Zeki, Assistant Professor, for his thoughtful wisdom and sincerity in giving valuable advice during the course.

Dr. Abdullah Alqadoumi, Assistant Professor, for his kind and caring words and the knowledge you shared in teaching.

Ahmed Khedr, the student representative of the second cohort, for always being there and all the immeasurable help and support you gave.

All my colleagues whom I had the privilege of meeting and who have made this journey memorable.

Table of Contents

Defence Committee's Approval	A
Abstract	B
Dedication	C
Acknowledgements	D
Table of Contents	F
List of Tables	I
List of Figures	J
List of Abbreviations	L
1 Chapter One Introduction to the Study	2
1.1 Introduction	2
1.2 Research Problem	3
1.3 Research Objectives	5
1.4 Research Questions	5
1.5 Research Contribution	6
1.6 Research Limitations	6
1.7 Research Outline	7
2 Chapter Two Literature Review	10
2.1 Background	10
2.1.1 Solar Power	10
2.1.2 Renewable Energies in the Gulf	11
2.2 Literature Review	13
2.2.1 Process Approach and Model Input	14

2.2.2	Photovoltaic Power Forecasting Techniques	22
2.2.3	Research In Bahrain	27
2.2.4	Conclusions and Insights	29
2.3	Theoretical Framework and Mathematical Background	34
2.3.1	Time Series Forecasting	34
2.3.2	Machine Learning in Time Series Forecasting	35
2.3.3	Machine Learning Algorithms	37
3	Chapter Three Methods and Procedures	44
3.1	Raw Data.....	44
3.1.1	Data Collection & Description	44
3.1.2	Solar Data	45
3.1.3	Poly Data	47
3.1.4	Meteorological Data	48
3.2	Data Pre-processing	50
3.2.1	Feature Analysis	53
3.2.2	Handling Outliers	58
3.3	Performance Metrics.....	59
3.3.1	Root Mean Squared Error.....	59
3.3.2	Mean Absolute Error	59
3.4	Forecasting Models.....	60
4	Chapter Four Results and Discussion	64
4.1	Results.....	64
4.1.1	Case 1- Selection of data time.....	64
4.1.2	Case 2 – PV Power Forecasting on Full Dataset.....	65

4.1.3 Case 3 – PV Power Forecasting Based on Weather Seasons	66
4.1.4. Results Summary	68
4.2. Discussion.....	70
4.2.1 Machine Learning Techniques	70
4.2.2 Data.....	74
5 Chapter Five Conclusion and Recommendations	80
5.1 Conclusions and Recommendations	80
References.....	83
Appendices.....	96
الملخص	100

List of Tables

Table (1): Solar installations across the three locations.	13
Table (2): Summary of all research in literature which concerns PV prediction	30
Table (3): Summary of all research in literature which concerns PV correlation with meteorological attributes	32
Table (4): Solar dataset showing the start and end date for logged data and the number of instances.....	46
Table (5): Poly dataset showing the start and end date for logged data and the number of instances.....	48
Table (6): Meteorological variables and their respective description.	49
Table (7): Average solar irradiance readings for the 5 months in 2019 and 2022	51
Table (8): Spearman correlation coefficient results for the meteorological al variables.	56
Table (9): Interpretation of Spearman's correlation coefficients.	57
Table (10): Hyperparameter values for the vanilla and hybrid network	62
Table (11): Result for the full day cycle and half day cycle.	65
Table (12): RMSE and MAE values for models trained on 5 months data.	65
Table (13): RMSE and MAE values for models trained on summer data.	66
Table (14): RMSE and MAE values for models trained on winter data.	67
Table (15): Time recorded in minutes for the models to train, both vanilla and hybrid network	81

List of Figures

Fig. 1: The SLTM network structure with its general information flowchart (Rahman et al., 2019).	38
Fig. 2: Bi-LSTM network architecture shown unfolded in time for three time-steps (Cui et al., 2018).	39
Fig. 3: Structure of a GRU network (Huang et al., 2019).	40
Fig. 4: Architecture of a Bi-GRU network structure (Meng et al., 2020).	41
Fig. 5: Location of the campus from which the data is extracted, and the solar-grid modules deployed at UoB- part of 5mw BAPCO pilot project.....	45
Fig. 6: Hourly reading of Global Solar Irradiance from Jan 2019 to Jan 2020	46
Fig. 7: Daily reading of current and voltage captured from the solar panels.	47
Fig. 8: Boxplot of the PV power output for the month of September, with mean values calculated for some of the days.	52
Fig. 9: Average PV power Output for each month per Watts	53
Fig. 10: Distribution and Q-Q plot of the data, where the Shapiro test gave 0.767, this means rejecting hypothesis that data is normally distributed, the data is not gaussian	56
Fig. 11: Box-plot of hourly reading of Relative Humidity for the whole 5 months.	58
Fig. 12: Structure of the proposed vanilla network for all models, X are the variables as input and Y is the output which is the PV power generated	60

Fig. 13: Structure of the proposed hybrid network for all models, X are the variables as input and Y is the output which is the PV power generated.....	61
Fig. 14: RMSE and MAE values for 5 months data.....	69
Fig. 15: Top to down: CNN-LSTM and CNN-Bi-LSTM using 128 neurons for the 5 months data.....	72
Fig. 16: Clockwise: 5 months, winter and summer data loss graph using the CNN-LSTM 128 neurons model.....	73
Fig. 17: First row shows two days in the month of October, second row two days from July.	75
Fig. 18: First two figures are the Relative Humidity and Dry Temperature during July August, summer data, and the last two during September to November, winter data.	77

List of Abbreviations

ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
BAPCO	Bahrain Petroleum Company
Bi	Bidirectional Recurrent Neural Networks
CNN	Convolutional Neural Networks
CO ₂	carbon dioxide
DBN	Deep Belief Networks
DL	Deep Learning
DRM	Dynamic Regression Models
GCC	Gulf Cooperation Council
GPR	Gaussian Process Regression
GRU	Gated Recurrent Units
km ²	Kilometre squared
kW	kilowatt
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
NARX	Nonlinear Autoregressive Network with Exogenous Inputs
NN	Neural Network
NWP	Numerical Weather Predictions
PV	photovoltaic
Random Forest	RF
R ²	R Squared
RMSE	Root Mean Squared Error
SVR	Support Vector Regression
UN	United Nations

WT

Wavelet Transform

Chapter One

Introduction to the Study

1.1. Introduction

1.2. Research Problem

1.3. Research Objectives

1.4. Research Questions

1.5. Research Contribution

1.6. Research Limitations

1.7. Research Outline

1 Chapter One

Introduction to the Study

1.1 Introduction

Energy is an essential pillar in today's modern world (Ahmad et al., 2020). The consumption of energy grows exponentially due to industrialization and population growth. To meet this demand, the amount of pollution generated from traditional energy sources causes severe effects on our climate, such as greenhouse gas emissions, and the resources are depleting (Torino, 2020). Thus, electric power generation is shifting towards renewable resources, making them crucial and of great importance (Shboul et al., 2021).

In light of this, the United Nations has adopted 17 Sustainable Development Goals (SDG) in 2015 that tackle many issues, one of which is the earth's environment and climate, where renewable energy is of great focus (Bexell & Jönsson, 2017). Accordingly, countries compete to align their government's programs to the SDGs to achieve these goals (Alsammarræe & Alshareeda, 2019). Bahrain has put tremendous action into reaching these goals as guided by Bahrain's Economic Vision 2030 and succeeded with great strides in achieving many of the SDG's requirements. In recognition of these efforts, Bahrain has received several awards from the UN.

Renewable energy has various sources from which it can be generated (Sharma et al., 2021). Most commonly are Solar energy, Wind energy, Hydroelectric energy, Tidal energy, and Geothermal energy (Herzog et al., 2001). Nonetheless, not all sources are safe, such as Nuclear Power, which

carries many risks and concerns, making it not a very suitable energy supply in the long term (Wolff et al., 2017). In contrast, Solar and Wind energy are considered the best and most promising sources as they are easily accessible. Further, the significant improvements in their technology and efficiency have made them further popular and high in demand. For instance, the production of photovoltaic (PV) power modules used in solar energy has been made affordable for the public, where many PV panels appear on house rooftops in urban and rural places (Stanojevic, 2021). Forecasts indicate that in the next four years, there will be a rapid increase, up to three times, in the capacity of the PV power systems installed, reaching 540GW (Z. Wang, 2019). Thus, Solar power will contribute significantly to the future global energy supply.

However, being a non-linear intermittence energy source is a downside to all the benefits of solar power (Son & Jung, 2021). The unstable nature is due to the fact that solar energy is highly dependent on weather conditions and day/night cycles. Consequently, this makes it very difficult to plan the future electricity supply, negatively affecting the grid balance (Casaca de Rocha Vaz, 2014). Accordingly, accurate and reliable prediction of electricity produced by solar systems is essential to maintain the stability of integrating the produced energy into the electrical grids.

1.2 Research Problem

Producing accurate forecasting needs sufficient data that has a specific pattern for some time (KAVAKÇI, 2021). However, when there is insufficient historical data for the model to learn or the data does not show any trend or pattern, this causes the models' performance to deteriorate. These trends are essential to understanding solar energy's characteristics by analyzing the relationship with the meteorological factors. This

behavior would vary from region to region; thus, forecasting models cannot produce a consistent result for all countries. That is to say, some countries which experience all four weather seasons will have a different solar energy behaviours and patterns and thus production than the Gulf countries, for instance, as they only experience two seasons, mild winter and hot summer (Radhi, 2009). Accordingly, it is essential to establish first the relationship of the meteorological variables with the solar energy production, specifically in Bahrain, from which accurate prediction can be made from the model.

The time horizon on which the PV power output would be forecasted plays a vital role for different market participants (Gensler et al., 2017). Each type of forecast is tailored to a specific application to which they are targeted (Nguyen et al., 2021). For instance, the short-term forecasting models are helpful for accurate time control of electrical systems operations. They are applied to automate the control of solar power generation sources for load balancing and plant management of the power grid production and delivery. This time horizon is of great interest to many actors in the field, specifically because electricity is primarily traded in the day-ahead market; therefore, a day-ahead forecast will emphasize this type of forecast in this research.

According to recent research, Machine Learning (ML) techniques predict more accurately than traditional time series techniques in forecasting solar power generation (Carrera & Kim, 2020). In contrast to statistical tools, these algorithms are much superior because they capture the non-linear relationship in solar energy. ML techniques have many models that can be utilized, each with strengths and shortcomings. However, specifically for time series, it is observed from the literature that few models have not been tested yet in past research and hence room to

test other models for accurate predictions. Thus, the thesis will focus only on the ML techniques, specifically for time series.

1.3 Research Objectives

The objective of this thesis is twofold. The first is to establish the relationship of the meteorological variables that affect Photovoltaic (PV) solar energy, specifically in the kingdom of Bahrain. The second is to exploit the robustness of other ML algorithms that has not been investigated yet in literature and compare the performance with different standard machine learning techniques to forecast PV power output for the short-term, one day ahead, based on historical PV power and meteorological data.

1.4 Research Questions

Based on the introductory section discussed above, the problem investigated in the thesis can be summarized with the following research question:

- What are the main meteorological attributes that affect the Solar power output?
- Which machine learning algorithm, specifically for time series, would perform best in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for a short-term (1 day ahead) forecasting of solar power based on historical power and meteorological data.
- How is the accuracy of the models affected when using less historical data?

1.5 Research Contribution

The main contributions of the thesis are as follows:

- (1) Forecast the PV power output rather than the global solar irradiance, which is more efficient than forecasting the latter since the balanced grid needs actual power generation from solar PV systems, which has not been investigated in Bahrain.
- (2) Considering the non-linearity relationship of PV power output with the meteorological variables when conducting the correlation analysis, using the Spearman Correlation test, which has not been implemented in previous literature.
- (3) Considering the meteorological variables with PV power output, we propose multiple ML models that have not been investigated in past literature to forecast short-term (one-day-ahead) solar power.
- (4) Working with high-resolution datasets, which record data at a 1-minute interval.

1.6 Research Limitations

The limitations of this thesis are as follows. First, there are assumptions made and had to be followed in order to predict the solar power output, which are the azimuth angle, inclination and tilt of PV panels are assumed to be kept constant and are not changed throughout the research.

Dust accumulation on solar panels is responsible for the poor performance of the PV systems in the MENA region and thus needs to be maintained regularly (Naser Waheeb Alnaser & Alnaser, 2019). Bahrain is also known for having many dust storms, particularly significant in the month of August and September. For instance, one of the heaviest dust

storm Bahrain experienced was on 2nd of April 2015, where the PV panels at the UoB Site project were covered by about 2 mm of Dust. This leads to the second assumption that the solar panels are being cleaned and maintained regularly.

Though using machine learning approaches can give adequate results, one drawback to these techniques is that they require historical data that span more than one year which is sometimes unavailable (Casaca de Rocha Vaz, 2014). Moreover, there are limitations on the algorithm itself since they have specific assumptions and resulting in limitations (Malik, 2020). Since this research will focus only on ML algorithms to forecast a day ahead of solar power output, it is crucial to analyse these limitations in this specific area/domain and hence propose new methods to address these limitations and improve the accuracy.

1.7 Research Outline

The remaining parts of the thesis will be structure as follows:

- **Chapter 2: Background and Literature Review**

Provide overview of previous research achievements and findings on time series data using Machine Learning techniques. Theory behind the algorithms to be investigated in this research and the evaluation measures are also presented with their mathematical formula.

- **Chapter 3: Methods and Data**

Data collection, description and the pre-processing parts including the correlation analysis will be presented in this chapter, along with the methodology for testing the models is discussed.

- **Chapter 4: Results and Discussion**

The results of this thesis will be presented and the analysis will be discussed.

- **Chapter 5: Conclusion and Recommendations**

The limitations of the study and the potential improvements that can be addressed in future research are outlined.

Chapter Two

Literature Review

2.1. Background

2.2. Literature Review

2.3. Theoretical Framework and Mathematical Background

2 Chapter Two

Literature Review

2.1 Background

2.1.1 Solar Power

Light is energy, otherwise known as solar radiation, emitted from the sun (Rhodes, 2010). Every place on earth receives sunlight. However, the amount of solar radiation differs from one place to another. Solar technologies capture solar radiation and turn it into a useful form of energy (Zabihian, 2021). The use of solar energy is divided into two main types: solar thermal and solar electricity. The first uses the sun as a direct source of heat energy and is most commonly used for supplying hot water to houses and swimming pools. The latter seeks to convert sunlight directly into electricity through a process known as photovoltaic (Singh, 2013).

Photovoltaics (PV) is a Greek word that was introduced in 1890, where Photo means light and Volt refers to electricity energy (Ravaee et al., 2012). As the name suggests, Photovoltaics is the direct conversion of light into electric power using semiconducting materials such as silicon. This electric power is then generated as direct electric current (DC), which is then converted to alternate electric current (AC) using inverters and other components in order to be distributed onto the power grid network (Zeman, 2014).

The output electrical power of solar PV is given by equation:

$$P = I \times V$$

Accordingly, the maximum electrical power of solar PV is given by equation:

$$P_{max} = I_{mpp} \times V_{mmp}$$

Although solar power is more adequate to forecast when it comes to balancing grid rather than solar irradiance, there comes with it a challenge and the accurate conversion of solar irradiance into actual power output (Reindl et al., 2017).

2.1.2 Renewable Energies in the Gulf

The Gulf holds a third of global crude oil reserves and nearly a fifth of the world's gas reserves (Al-Badi & AlMubarak, 2019). The region's rich fossil fuels made them heavily reliant on these conventional energy resources to generate electricity, as well as the primary source of income in the region (Alharbi & Csala, 2020). While the population of this region is relatively small, the increase of urban population rate, which was up to 83% between 2005 and 2008, increased with it the use of energy, specifically the air conditioning during summer (Al-Maamary et al., 2017). It was recorded that oil and gas consumption in the Gulf is greater than that of Indonesia and Japan. As a result of the excessive use of fossil fuels, the CO₂ emission levels have increased from 2000 to 2010 up to 80%, with an average increase of 8% per year and dominating 79% of the other emissions (Alharbi & Csala, 2020).

As much as the Gulf countries are rich with fossil fuels, they are in renewable energy resources (Frroukhi et al., 2016). The abundant and

robust sunshine and significant wind resources make the Gulf countries a great place to harness both energy sources. Along with climate change and global warming affecting all sectors in the GCC region (Al-Maamary et al., 2017), it has motivated the use and utilization of solar and wind energy sources to overcome the environmental issues and eliminate their reliance on fossil fuels. Consequently, the majority of the GCC countries, including Bahrain, have committed to the agreement of the United Nations Framework Convention on Climate Change 2015 (Elgabiri et al., 2021). Reducing greenhouse gases and CO₂ emissions is one of the targets to fulfill the goal. Bahrain has set an obligation to achieve by 2025, and that is to generate 5% of capacity from renewable sources and 10% by 2035. This is further detailed in the National Renewable Energy Action Plan (NREAP 2019). As of now, all GCC countries have conducted relatively large solar and wind energy projects, most notably in Kuwait, where they currently have renewable energy of about 70 MW with a plan to have a 200 MW by 2030, and the UAE has currently a 300 MW with the plan to have a renewable energy target of 2500 MW. Saudi Arabia has one of the most ambitious plans: to develop the world's largest solar power plant worth \$200 billion, which would be up and running fully by 2030 with 200 GW of solar power (N.W. Alnaser & Alnaser, 2019).

Bahrain, just like other neighboring GCC countries, enjoys some of the highest solar radiation levels in the world, giving this region an advantage in utilizing the solar energy in particular among other renewable energy sources and, in turn, would gain the most from it (Waheeb Essa Alnaser & Alnaser, 2019). According to Waheeb Essa Alnaser & Alnaser (2019), Bahrain has the potential to generate 33 terawatt-hours (TWh) per year from solar power only, which exceeds twice the power demand in 2013.

Bahrain initiated its first 5 MW pilot PV solar electricity grid-connected project that was installed, owned, and operated by BAPCO in 2014 (SEU TEAM, 2017). The project was deployed in three locations covering a total area of 34 km²: Campus of UoB, BAPCO refinery car parking lot, and Awali town (Waheeb Essa Alnaser & Alnaser, 2019). Table 3 summarizes the installation size and PV system capacity for all locations.

Table (1): Solar installations across the three locations.

Site Name	Installation Type	Installation Site
Awali	Carport	1.51 MW
	Solar Trees	2.40 kW
	Streetlights Poles	48 kW
UoB	Ground Mount	501 kW
BAPCO Refinery	Car Ports	2.7 MW
total		5.0 MW

It is worth noting that although this project is considered a pilot project, the solar system established on a 10,000 m site at the campus of UoB makes it the world's most extensive PV system installed within a university campus. This launch marks the first crucial step in achieving Bahrain's goals of reducing dependency on conventional energy resources and fulfilling its commitment to producing 5% of its total electricity output from renewable sources. In addition, this project was installed as a pilot project to test the technical viability of the project (SEU TEAM, 2017).

2.2 Literature Review

The urgent necessity to harness solar energy from one side, and its complex nature from the other, has motivated many scholars in recent years

to research in this specific domain and investigate different approaches with varying degrees of success to forecast PV power output. As a result, many works of literature implemented various modeling approaches to forecast the PV power output, such as using statistical techniques and simple machine-learning models. In addition, in recent years, scholars have developed a hybrid approach, which incorporates the first two techniques together. However, each of these approaches has its unique strengths and shortcomings.

Some scholars aimed at predicting the PV power generated, while others the solar irradiance. The choice of target predictor is maybe the scholar's aim, or due to the fact there was unavailable data of PV power and hence use Solar Irradiance. However, both predictors will be included in this section as they are much the same way linked in modeling approaches; thus, valuable insights might be retrieved.

The following literature is categorized into three main sections: Process Approach and Model Input, photovoltaic power forecasting techniques, and Research in Bahrain. The first two sections reflect the main focus of research laid on forecasting the PV power or solar irradiance, whether it was the input parameters, the applied algorithm, the type of neural network, or even the time horizon for prediction, and the latter concerns all Researchers conducted in Bahrain.

2.2.1 Process Approach and Model Input

The quality and accuracy of model depends significantly on the input data (Ahmed et al., 2020). The pre-processing can include feature selection, treatment of outliers, and other measures in regards to data, or a specific framework and method to predicting the solar power output (Raza et al., 2016).

2.2.1.1 Model Input

Monteiro et al., (2013), Researchers aimed at forecasting the hourly electrical production of a PV plant by initially predicting the NWP using SVR, where the model provides hourly meteorological predictions of irradiance as input for the analytical PV power forecasting model, APVF model, and predictions of irradiance and temperature and other meteorological variables as inputs for the multiplayer perceptron PV forecasting model, the MPVF model. Both models have also utilized past hourly recorded values for PV electric power production corresponding to one year. The APFV model is an analytical approach based on irradiation data to adjust the modeling of the clear sky by irradiation and PV power attenuation index. In contrast, the MPVF model uses a more extensive set of meteorological inputs and the PV power generated. Research has also used GA to optimize their MPVF model. Both models achieved very similar results, obtaining RMS errors varying from 11.95% to 12.10% for both models. They concluded that both approaches are suitable for PV power output prediction.

In a similar approach to Monteiro et al., (2013), Scholars in (Nguyen et al., 2021) proposed forecasting the power generation capacity of a large-scale solar power plant using the LSTM algorithm by considering two cases, actual historical weather, data, and forecasted weather data as inputs to the model. Although there is an inevitable error that comes with forecasting weather data, not to mention the significant increase in error as it magnifies with forecasting, researchers believed that this error might be improved through hidden rules the LSTM will be able to learn and capture, thus improving it. Using data collected from Thanh Thanh Cong 1 solar power plant, which records 13 months with increments of 5 minutes, researchers first conducted feature analysis using the Pearson test correlation on the meteorological data. They treated outliers by removing

them from the dataset. Upon testing different LSTM structures for both cases, the best LSTM structure is of 4 layers with 100 nodes. Finally, to test the model's superiority and approach, scholars compared the model with Multilayer Perceptron (MLP), Autoregressive Integrated Moving Average (ARIMA), Persistence, and Dynamic Regression Models (DRM). Results indicated that the LSTM gave higher performance using the second case, reducing RMSE from 10.857 to 9.811.

2.2.1.2 Normalization and Wavelet

Among the data pre-processing techniques, Wavelet Transform (WT) and Normalization are widely used as they can convert large input data to a smaller range which improves computational complexity (Alhakeem et al., 2015). Normalization compresses and transforms large data into a smaller degree, confined between the ranges 0 and 1. This is aimed to limit the regression errors and maintain correlation among the datasets. The WT transforms the signals into time-frequency domains and decomposes the signals into one approximate set and several detailed sets, thus, overcoming the problem with non-stationary datasets.

Researchers integrated the wavelet transformation to decompose the solar power time series into a set of input data with better features for prediction (Nazaripouya et al., 2016). They proposed a univariate model that captures linear and non-linear patterns, specifically incorporating RNN, with an ARMA model after the wavelet method is employed to predict the solar power output for one minute ahead. Due to the nature of the forecast time horizon, which is a super short-term prediction, only the solar power output is used as input and does not account for the metrological factors, which simplifies the forecast. Firstly, they used ARMA as the linear predictor, that is, to model the linear features of the data, after which RNN is utilized as the non-linear pattern recognition tool

to capture the non-linear patterns in the residual part to estimate and compensate in the wavelet-ARMA model prediction. Researchers applied the algorithm on two days (sunny and cloudy) from the dataset. The model was then compared with other algorithms, where results validated the accuracy and efficiency of the proposed hybrid method.

Mohammadi et al., (2015) also proposed a new hybrid approach incorporating with WT algorithm to forecast horizontal global radiation for an Iranian coastal city. Three sets of features were fed as input to model, which consists of different combination of six parameters, hence establishing three models. Results showed relative sunshine duration, difference between air temperatures, relative humidity, average temperature and extra-terrestrial solar radiation as inputs delivered higher performance than other models. Further, the proposed model was compared against ARMA, NN and Genetic Programming (GP) models. The proposed model outperformed all other methods.

Others also used two years of hourly PV power output data from solar plant installed at Applied Science Private University (ASU) in Jordan, and real-time records of weather data that was measured at the university campus by AUS weather station (Alomari et al., 2018). Data filtering and association was done for weather and PV power datasets, where any PV power record with missing weather data, or vice versa, was removed from the dataset. Moreover, negative solar radiation can often have negative values in early and late hours of the day, due to solar radiation sensors offset and inverter failures. Accordingly, Researchers have set the values for solar radiation and PV power output to zero. Data normalization was done on the radiation and PV output, where values were confined between 0 and 1. This enhanced the model's performance according to researchers. Lastly, the Levenberg-Marquardt (LM) and Bayesian Regularization (BR) algorithms applied on the ANN algorithm, were both models were able to

capture the correlation of temperature, solar irradiance and timing with PV power generated for real-time day ahead predictions. After testing different training configuration, the BR technique suggested 28 hidden layers and all-weather inputs while LM dictated 23 layers for the data, with the BR-ANN.

2.2.1.3 PV power forecasting based on forecast horizon

The future period for output forecasting or the time duration between the actual and practical prediction time is known as the forecast horizon (Das et al., 2018). The accuracy of the PV power depends significantly on the forecast horizon, and in fact, the model's performance is heavily affected by the time horizon.

A research conducted by Reikard (2009) who analyzed the performance of several algorithms based on different time resolutions. Scholars conducted the study on six different datasets, three of which covered a time span of 3 years with hourly samples. The other three datasets covered a time period of one year, with minute sample resolution. As a result, the last three datasets had more observation than the first three. The hourly data was averaged to create time series at 1, 2, 3 and 4 hours. Similarly, the minute data set was averaged to make the following resolution: 5, 15, 30, and 60 min. The predictor in all 6 datasets was the solar irradiance. As the data exhibited non-linear variability, the seasonality was eliminated from all datasets using statistical methods. To compare the performance of the forecast at high resolutions, Researchers used the ARIMA, NN and hybrid approached to predict the solar irradiance. Best results obtained in nearly all test was using the ARIMA in logs, with time-varying coefficients. Exceptions were at the high resolutions, transfer function using the cloud cover variable is found to improve over ARIMA. Other cases, neural networks found to improve the

forecast accuracy at very high resolutions, on the order of 5 min. The success of the ARIMA is attributable mainly to its ability to capture the diurnal cycle more effectively than other methods

2.2.1.4 Feature Selection

Imprudent input selection can lower the model's accuracy not to mention increase time delay in modeling and computational complexity (Yujing Sun et al., 2017). PV power is most correlated with solar radiation, while the rest of the meteorological factors are dependent on the geographical location, from which the strength varies from country to country.

AlSkaif et al. (2020) presented a systematic approach to conduct a feature analysis selection on the relationship between the PV power output and 9 different meteorological variables. Three years of meteorological and PV power output data were used to establish the proposed approach in two case studies, one in the U.S. and one in the Netherlands. The analysis covers correlation test using PCA among the weather attributes to better understand the interdependence between them. After which several ML regression methods were implemented, Multiple Linear Regression, Lasso LR, SVM regression and Random Forrest (RF) to understand the performance of these models considering a lower-dimension subspace of meteorological variables. Results showed that using a lower-dimension meteorological variable gave similar accuracy to using all attributes.

The impact of the meteorological attributes on the model's performance was investigated by Ziane et al. (2021), who performed correlation studies using the PCA. Scholars then proposed the Random Forrest model to forecast the target variable instead of neural networks due to their nature of adequate predictability and interpretation in contrast. Results proved Random Forest succeeded in capturing the nonlinear

dependence of variables and the noise data found in the meteorological variables while giving high performance.

Weather classification has become a necessary pre-processing on the meteorological variables step to enhance the model's performance, as many researchers concluded in their study such as in (F. Wang et al., 2017). Although this method is able to increase the accuracy of the model, it carries with it many considerations such as the type of classifiers and the quality of training data.

Researchers proposed making a 24 hour ahead forecast of the power output of PV system using its historical data along with the online meteorological data (Chen et al., 2011). Using a one-year data with hourly time stamp, scholars first classified the variables of the NWP using a self-organized map (SOM) into the following categories: cloudy, rainy and sunny days. For each type of day, NN model was used and several simulations was carried out to identify the best parameters for each category, using the 10 – cross validation. Moreover, scholars utilized the RBFNN in the NN model to further enhance the accuracy, by optimizing number of neurons in each hidden layer during the learning step of the network. It was found that the number of neurons for hidden layers differed for each category, while the first and last layer had the same number of neurons for all categories: 6 and 24 respectfully. They concluded from the results obtained that the use of weather type classification improved the forecast accuracy (Chen et al., 2011).

Weather classification has proven to be a practical approach in improving the accuracy of PV power prediction under various weather statuses. However, this approach becomes difficult to classify once there are missing records in the historical data. Hence, F. Wang et al. (2015) aimed at identifying these missing weather records, through a solar irradiance feather extraction and SVM based weather statuses pattern

recognition (WSPR) technique was designed. The input vector of the SVM model is composed of features extracted from solar irradiance data only. Through a case study using data of a 5MWp grid-connected PV plant is carried out, and the effectiveness of the model is evaluated. Scholars concluded that the missing weather records in the time series data can be effectively recovered by the proposed model (F. Wang et al., 2015).

In contrast, F. Wang et al. (2019) reclassified 33 meteorological categories into ten groups by gathering several single weather types to create a single new weather type. Further, GAN was utilized to generate more data in the training set in each weather category, where the data was imbalanced, and extreme weathers conditions were a minority class. Then the Convolutional Neural Networks (CNN) model was trained using both the original and generated data to predict one day ahead PV power. The results showed that CNN classification gave better performance than the traditional machine learning models. Moreover, all models were further improved using the GAN, however to different extents, where GAN-CNN gave the best; however, it shows that GAN could generate high-quality samples that capture the intrinsic features of the original data. In addition, the authors noted that classifying the weather contributed significantly to the model's performance and gave precise day-ahead PV power output with high efficiency.

Clustering is an unsupervised machine learning techniques that groups similar objects based on a similar characteristics of a given set of inputs (Zanero & Savaresi, 2004). They are often used as a pre-processing step for supervised problems, by creating new dataset by grouping similar observations (Agarwal, 2014).

In (Najibi et al., 2021) scholars proposed a pipeline of processes, which would enhance the Gaussian Process Regression (GPR). The process consisted of feature selection and using clustering before building

the model. The data consisted of power output and meteorological data, where the features most correlated to the power output were identified. Using k-means clustering, the data were categorized into a total of 4 groups based on day-time, as solar output relies on solar irradiance. Researchers found two clusters that appeared to represent early morning and night times, while the remaining two clusters appear to represent seasonal variations. The model is then trained for each cluster using the GPR model, using the Mat'ern 5/2 as the kernel function. To confirm the model's accuracy, Researchers used a 5-fold cross validation; and hold-out validation, holding out 30 random days as test data. This proposed framework was applied to different datasets from different sites: Denver, New York, Dallas, San Francisco, and St. Lucia. The results noted for all datasets in different locations were nearly of the same range, indicating that the model may be used in any site if the data has the same attributes. In addition, Researchers noted that the use of clustering and the kernel selected increased the accuracy of the model when compared with existing methods in past literature.

2.2.2 Photovoltaic Power Forecasting Techniques

Several modeling approaches have been utilized to forecast PV power output: physical, statistical, and ML techniques which includes DL (Lai et al., 2020). Most common techniques used in past literature were simple statistical tools, ML model and hybrid.

2.2.2.1 Traditional Statistical Methods

Statistical time series models were tested against each other, such as the moving average and exponential smoothing and a decomposition model, where the data was decomposed into seasonal patterns, summer and winter where each of them predicted separately (Prema & Uma Rao, 2015). The models were tested on multiple durations of data for further evaluation

to identify best possible set of data: 1 month, 2 month and 4 months, which is the total data available from February through May 2013. Results showed that the prediction with decomposition model for 2 months data gave best results with MAPE less than 9.28%. Since researchers concluded through analysis that there are sudden changes in weather in India that happen every two months, where seasonal variations are very prominent as a result worsens accuracy more when more historical data is used. The MAPE would increase on cloudy days for the winter season, however far less than what was reported in past literature. Researchers highlighted for future studies that the performance can be further improved when the data is viewed as signals. As a result, the high frequency components may be captured by using the wavelets.

2.2.2.2 Machine Learning Techniques

Simple Network

Isaksson & Conde (2018) aimed to predict the solar power across 5 different sites in Sweden, by using ML techniques and comparing the performance with statistical models. Researchers eliminated the nighttime data of the solar power and only included daytime readings, as Researchers justified that predicting the nighttime data is more straightforward than daytime and with this the model's performance would not be fair as the accuracy would be boosted from the easy prediction of nighttime data. It was found that employing time series was a much-complicated process due to non-stationary time series in contrast to ML models which was easy to implement and gave accurate results, particularly ANN and Gradient Boosting Regression Trees performed best.

Similarly, Poudel & Jang (2017) compared the performance of an LSTM neural network, which consists of one input node two hidden layers and one output node, and the moving average to model a day solar power

data. Using only PV power data, the LSTM model was used 500 epochs to train the model, choosing the linear regression function as the final output activation function. The results obtained from the comparison indicate that the LSTM neural network approach is more reasonable for short term solar power prediction than the moving average, where LSTM gave RMSE value of 8.679, and the moving average 10.677.

Rogier & Mohamudally (2019) accounted for the non-linearity characteristics in the PV power output with the meteorological data by using the Nonlinear Autoregressive Network with Exogenous Inputs (NARX), which is an RNN variant with dynamic architecture system with feedback connections enclosing multilayers of the network. Authors aimed at conducting two experiment using data that has been collected for 3 years. The first experiment was designed to forecast the future values for ambient temperature and Solar radiation, from which the results would then be fed to a Simulink model to predict the PV power. The second experiment was set to forecast future PV power output from previous PV values. Authors compared the results obtained from ML algorithm with a persistence model, which results concluded that ML model achieved better accuracy. Further, the RMSE produced by the model trained in the second experiment was much lower than the multi-step, that is in the first experiment.

Likewise, Louzazni et al., (2020) proposed to forecast PV power output plant installed in Heliopolis University, Egypt, using a NARX which was trained with the non-linear optimization, Levenberg-Marquardt algorithm, on a one-month data that is sampled in hourly intervals. Scholars noted that the proposed model has an advantage over the static neural networks, which is the simplicity and rapid training time because of the favorable dynamics and resistance to interference present in its

architecture. Further, results indicated that it outperformed the standard NN.

Hybrid and Sequential Deep Learning Techniques

Gensler et al., (2017) introduced a more sophisticated approach, where Researchers aimed to forecast solar power using DL algorithms such as an AutoEncoder with an LSTM (Auto-LSTM), and DBN. These algorithms are compared with standard MLP and LSTM. For the basic algorithms, MLP and LSTM, Researchers selected the best correlation to the target variable and then inputted the data to the model. However, for the Auto-LSTM and DBN, all variables were fed as input since these two methods are able to extract features within data. All of the models had 4 layers in their structure using the tanh activation except for the last output layer, ReLU was used as the activation function. Results showed that the Auto-LSTM performed the best, and DBN gave similar results with slightly more RMSE. Nevertheless, these two proposed methods outperformed the basic MLP and LSTM. Scholar noted that the combination of DBN with LSTM might further improve the forecasting accuracy in future work, as well as the use of CNN with LSTM algorithm.

In most recent studies, Researchers in (Zheng et al., 2020) aimed to predict the sum of the solar power output for power system with solar power plants in multiple regions, a gap found in literature that received minimal attention in previous research. Further, Researchers used the intelligent algorithm to optimize the parameters of the LSTM model to improve prediction, which included the learning rate, time step, batch size and number of neural units, this too was not found in literature. Upon obtaining the optimal parameters, different LSTM structures are compared, and the suitable structure was chosen according to the training loss. Moreover, researchers compared the LSTM to other algorithms: artificial neural network, and extreme gradient boosting. Researchers concluded

based on prediction accuracy that the PSO-LSTM method outperformed the other algorithms including the basic LSTM.

Researchers in (Ying Sun, 2020) were motivated to have an accurate forecast of the power PV systems in short term. Scholars used a one year reading of data collected from Collect the SCADA data from the PV system since January 2018 till December 2018 collected from a 9 MWp grid-connected plant. Pearson correlation test was carried out that showed a high correlation between solar irradiance and power production a low correlation between power and wind velocity. Data were split into training and testing datasets and were fed to a one layer of LSTM, with number of hidden units 200 and epoch value of 100. The model showed adequate efficiency of the LSTM model to forecast PV power production even under a cloudy day, with an $R^2 = 0.98$.

Researchers proposed to forecast a day-ahead PV power based on a two-stage attention mechanism over LSTM model, where the performance of the attention mechanism is analyzed in respect with specific attributes such as the solar radiation, temperature and snowfall (Aslam et al., 2021). In addition, Researchers used Bayesian optimization algorithm to obtain the optimal combination of the hyperparameters for the proposed deep-learning model. This model was compared against many machine learning algorithms and single attention mechanism to test its performance: LSTM-Attention, CNN-LSTM, SNAIL, Raffel, and Hierarchical attention etc. Researchers concluded from the results obtained that the proposed model outperformed the traditional methods, which is consistent to what authors in (Gensler et al., 2017) have proposed for future work. After splitting the data, the data has been normalized. Except the output power, all input data are normalized between 0 and 1. The output power or the target value is normalized according to the power capacity of the respective PV facility

2.2.3 Research In Bahrain

No research has been found that concerns the solar power output directly, rather all past literature were with regard to predicting the global solar radiation. None the less it has been included since the PV power and irradiance are very much correlated and behave in much the same way.

Solar irradiance is the radiant energy that is emitted from the sun in the form of electromagnetic radiation, where it reaches the surface of earth as either direct radiation or diffused radiation (Ahmed et al., 2020). It has the highest/strongest correlation with the PV power output, in contrast to other weather factors, and therefore of great importance that can be harnessed.

Researchers aimed at estimating the total and diffuse irradiation in Bahrain using meteorological data recorded from 1984 to 1987 (W. E. Alnaser, 1989). The total radiation was obtained by calculation and afterwards an empirical equation was used to estimate the monthly means and annual solar radiation on a horizontal surface, where the diffuse solar radiation can be obtained from two parameters, the clearness index and the fraction of sunshine.

Given the unpredicted nature of solar irradiance, which exhibits pattern and is non-stationary, Researchers in (Al-Sadah et al., 1990) removed these trends and seasonality by using statistical means to make time series stationary, high order polynomial models, and then modelled the resulting stationary series to estimate the hourly solar radiation in Bahrain. Results noted from model compared well with the experimental data where it gave a good fit to the daily GHI trend, below is their empirical formula developed:

$$r_t = b_1 + b_2t + b_3t^2$$

Where b_1 , b_2t and b_3 are the regression parameters.

Likewise, (Al-Sadah & Ragab, 1991) attempted at estimating the monthly daily average global solar radiation of Bahrain, using the Ångström equation, a method which defines the relationship between solar energy and the sunshine duration. The data used covered the period from 1983-1987, where the constants have been derived from the monthly-specific approach.

Similarly, a model was proposed by (W. E. Alnaser, 1993) to estimate the monthly mean solar global irradiance, using linear regression equations starting with one parameter first, and then the parameters are increased, reaching six parameters, which is the total number of attributes used in this study, they are: declination angle, humidity, ratio of sunshine duration, mean temperature, soil temperature and the extra-terrestrial solar radiation. The data used covered a period of 27 years, collected from Qatar weather station. After constructing 63 models, from which the equation containing all six variables gave the highest correlation coefficient.

In a similar approach, Researchers aimed at investigating the correlation of global solar radiation with meteorological factors in Bahrain to propose a numerical formula, these variables are: daily global radiation, sunshine duration hours, minimum and maximum temperatures and relative humidity (Arabia & Town, 1993). The data covered a time period of 19 years, from 1968 to 1987 that was collected from Bahrain National Airport. Scholars used the Least Squares Techniques to conduct their study, starting with one parameter, then two parameters and so on. This procedure was repeated twice, the first time under the assumption of linearity and the second time with non-linear case. To assess the performance of each formula, the R^2 and RMSE were both considered for evaluation. Researchers concluded that the best model was when using all parameters considering the non-linear case. Moreover, the non-linear case

for all parameters gave better results than the linear case. In addition, scholars concluded that the approximation would be better with more parameters to be included.

In most recent study in this field, (Bin Shams et al., 2016) conducted the study to accurately forecast four meteorological parameters: wind speed, solar irradiance, ambient temperature and PV module temperature using the BoxeJenkins, otherwise known as the ARIMA model, framework. The data under study was collected was from the city of Awali, and covered the period from May 2010 till April 2011 where Researchers noted of missing data issue in all 4 attributes. Researchers omitted these missing observations from the list using Linear Interpolation. The data set was collected every 2 min daily average data in their study. Results show that the framework provided adequate results to identify and estimate the models for meteorological time series using ARIMA (1,0,0), ARIMA (0,1,2) models were found reasonable in capturing the autocorrelative structures of the 4 attributes.

2.2.4 Conclusions and Insights

2.2.4.1 Summary

The challenge of building a PV power forecasting is incorporated in its instability, non-linearity, and complexity. In addition, the PV is strongly dependent on solar radiation and is affected by various weather conditions, which differ from country to country.

The statistical methods offer simple linear models that are inadequate for non-linear PV and solar radiation. On the other hand, the NN, specifically the CNN and its hybrid forms, held the best forecasting accuracy, especially for the short-term forecast horizon. Hence, these

methods have the most promise of accurate predictions for PV power generation.

The use of optimizers for the input data contributed significantly to providing reliable and accurate forecasts, however, at the expense of computational complexity. Some scholars have concluded that reliable historical data is essential for accurate and consistent forecasting for all techniques.

In Bahrain, the prediction was concerned with global solar irradiance forecasting. Some scholars have conducted a correlation test of the solar irradiance with some meteorological variables, with a maximum of 4 attributes. Indeed, having more variables to perform the correlation test would give an adequate understanding of the solar irradiance behavior or the solar power. Tables 1 and 2 summarize all mentioned researchers in chapter 2.

Table (2): Summary of all research in literature which concerns PV prediction

Author	Year	Data	Forecasting Model	Approach	performance
Alnaser	1989	3 years	Empirical equation	Use meteorological data to predict solar irradiance	Clearness index and fraction of sunshine are the main parameters to calculate the solar irradiance
Al-Sadah et al	1990	Not mentioned	Empirical formula	Removed trends and seasonality from data	-
Al-Sadah & Ragab	1991	4 years	Ångström equation		
W.E Alnaser	1993	27 years	Linear regression	Using 6 weather attributes to construct the regression model	Model included all model
Arabia & Town	1993	19 years	Least Squares Technique	Using 4 attributes to construct the model in two cases, first under the assumption of linearity and second non-linear	Best model using all attributes under the assumption of non-linearity
Reikard	2009	3 years and 1 year	ARIMA, ANN and Hybrid	Tested different time horizons on different datasets with different time resolution.	ARIMA

Chen et al	2011	1 year	RBFNN	Classified days as rainy, sunny and cloudy using weather classification using SOM	Use of weather classification improved accuracy of model
Monteiro et al	2013	1 year	MPVF and APVF	Compared Univariate and Multivariate model	Both models achieved similar results
F. Wang et al	2015	Not mentioned	SVM	Identify missing records in weather data.	Time series data was effectively recovered.
Mohammedi et al	2015	Not-mentioned	Genetic Programming model	Compared 3 models with different input of weather variables. Compared with other models	Proposed model outperformed the other models.
Prema & Uma Rao	2015	1 month, 2 month and 4 months	Statistical models	-	2 months of data gave best results
Bin Shams et al	2016	1 year	Box-Jenkins	Forecast 4 meteorological attributes including solar irradiance	Model gave good performance
Nazaripouya et al	2016	Not-mentioned	Hybrid RNN-ARIMA, Univariate model	Implemented Wavelet.	Results validated accuracy of model
Gensler et al	2017	Not mentioned	Auto-LSTM	Compared proposed model with basic ML model	Proposed model outperformed standard ML models
Poudel & Jang	2017	Not mentioned	LSTM	Compared LSTM against statistical models	LSTM gave best results
Alomari et al	2018	2 years	Applied Levenberg-Marquart and Bayesian Regularization on ANN model	Data normalization, filtering missing data and outliers. Tested different parameter setting.	Bayesian suggested deeper network than Levenberg-Marquardt
Isaksson & Conde	2018	Not mentioned	ML models	Compare ML with statistical tools, researchers eliminated night-time data	ANN, Boosting Regression Trees performed best
F. Wang et al	2019	Not mentioned	ML algorithms	Classify types and using GAN to generate more data in training of model	CNN gave best performance when implementing GAN.
Rogier & Mohamudally	2019	3 years	NARX	Compared univariate and multivariate model of PV	Univariate NARX performed best

Louazani et al	2020	1 month	NARX	Trained using Levenberg-Marquardt	Proposed model performed better standard Models
Zheng et al	2020	Not mentioned	PSO-LSTM	Compared against standard ML model	PSO optimizer enhanced LSTM considerably including basic LSTM
Aslam et al	2021	Not mentioned	Two-stage mechanism CNN-LSTM	Compared against ML algorithms and single attention mechanism	Two-stage CNN-LSTM gave best performance
Najibi et al	2021	Not mentioned	GPR model	Cluster data based on time-of-day for PV and meteorological data.	Proposed model outperformed other models and clustering increased the accuracy
Nguyen et al	2021	1 year	LSTM, MLP, ARIMA, Persistence, and DRM	PCA feature selection. Used historic weather data and predicted weather data	LSTM gave best results

Table (3): Summary of all research in literature which concerns PV correlation with meteorological attributes

Author	Year	Data	Forecasting Model	Approach	Accuracy
AlSkaif	2020	3 years	ML regression models	PCA feature selection	Lower dimensions gave better performance for all models
Ying Sun	2020	1 year	LSTM	Constructed PCA feature analysis	Validation accuracy confirmed good performance
Ziane et al	2021	Not mentioned	Random Forrest	PCA feature selection	Identified most variables correlated with PV power generated

2.2.4.2 Conclusion

All of the studies reviewed above have an essential role in developing this thesis. Although the data and parameters differed in each study, imperative details helped construct the framework for this thesis. After a thorough review of past literature, it is concluded that all models have their advantages and shortcomings. This depends significantly on the availability and quality of data and the forecasting horizon; the model selection varies. Hence, there is still room to compare some of the standard ML algorithms in short forecasting horizons, specifically those which have not yet been investigated, which are other variants of the Recurrent Neural Networks (RNN).

All scholars that deployed ML techniques have investigated primary LSTM by varying the parameters, and very few tested other techniques such as SVM and RF. The ML techniques extend more than the LSTM model, specifically for time series. Other variants of the RNN, which are very powerful, have not been tested yet in the context of PV power generation prediction. These models are Bidirectional Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and the Bidirectional GRU. Thus, this thesis will address the gap found in the literature that is, to the best of our knowledge, neither GRU nor the Bidirectional LSTM and GRU have been studied in the context of PV power prediction.

2.3 Theoretical Framework and Mathematical Background

2.3.1 Time Series Forecasting

Time-series is a sequence of patterns or observations on a parameter measured at successive points in time (Hewamalage et al., 2021). The instances or rows in the time series data is ordered according to time, thus each data has a definite relationship with the time obtained (Prema & Uma Rao, 2015). These instances can be taken as per minute, hour, month year or any regular interval. Hence, the nature of time series data includes large size data and high dimensionality and necessity to update continuously (Fu, 2011). Time series remains a complex problem yet.

Time series analysis is estimating the next unknown observation at a certain date (Hewamalage et al., 2021). This forecasting can be done on a univariate or a multi-variate approach. Univariate is a single scalar observations, whereas multivariate time series consists of multiple variables as inputs, whose values are changing throughout the given time horizon (Karaca et al., 2022). The multivariate is hence much more costly than univariate time series data in terms of memory and time due to complex relations among variables level in time intervals.

For a time series analysis to make a prediction, data pattern is an important factor which can help in making an accurate forecast (Karaca et al., 2022). The characteristics and behavior of the time series is then understood from this pattern. There are three types of time series patterns which are (Cang & Seetaram, 2012):

- 1- Trend: can be linear, or exponential that change direction during time.

- 2- Seasonal: Occurs when the data is influenced by seasonal factors, such as day of the week, month or year. Seasonality exists of a fixed known period.
- 3- Cyclic: Occurs when the data follows a pattern of a rise and fall. This does not depend on a fixed time, and the duration of these fluctuations is usually at least every 2 years.

Time series forecasting has always been a key area of interest to academic research in many fields, such as climate modelling, biological sciences, medicine as well as finance and many more (Torres et al., 2021). While the traditional methods focused on parametric models such as Autoregressive, exponential smoothing and many more, ML algorithms have also provided the means to learn dynamics and patterns in time series data. With the increase in data availability as well as its complexity, and computing power in recent years, the ML has become a vital part of the next generation of time series forecasting models, which in many Researchers, have proven its superiority over the traditional methods.

2.3.2 Machine Learning in Time Series Forecasting

In recent decades, ML algorithms have been widely used in various fields (Lai et al., 2020). The improvements of computational power combined with the availability of large datasets has led to large interest in the topic of machine learning from Researchers (Schmidt et al., 2019). ML techniques are successfully employed for classification, regression and clustering tasks of big datasets, proving to have the human abilities in numerous tasks.

Data processing is an essential step in building a robust model, that can either improve or worsen the performance of a model. A poor representation of data can even reduce the performance of the most

advanced and complex machine algorithm (Najafabadi et al., 2015). Likewise, for a relatively simpler algorithm, it can give a robust forecast due to the good representation of data. Consequently, feature engineering, which focuses on extracting features from raw data, is an important step in the machine learning process. After which the model may be trained and adequate results may be acquired to aid decision makers.

With the new development of software and information technology, DL approaches have emerged vigorously, which is a subset of ML, and is thriving in the research field and is being used by many scholars in many fields (Najafabadi et al., 2015). DL has the characteristics of extracting non-linear relationships between attributes and high-level invariant data configuration, that intelligently analyze these complex big data. Thus, this technique managed to obtain satisfying results in various fields such as computer vision, natural language processing and recently embraced for solving time series forecasting (Wen et al., 2021). In addition to DL methods, SVM have also become increasingly popular in the field of machine learning (Lai et al., 2020), and specifically in the solar power forecasting which will be reviewed in depth in this section.

In this research, different ML algorithms, specifically for timeseries data, will be applied to forecast the PV power outputs in the Kingdom of Bahrain. The statistical models will be excluded since they only capture the linear relationships, unlike ML algorithms that can learn a far more complex relationship in the dataset (Wolff et al., 2017). To forecast the PV power, there are multiple parametric models that need to be used to generate adequate forecasts. The parametric models use details regarding the PV systems used: its material, the location installed, orientation and tilt of the panel, temperature and other meteorological variables to determine the PV power output. These empirical models will not be suitable for a

specific PV system. That is to say, a model developed in America will not work properly in Bahrain as it was trained under certain conditions that is not similar geographically. At least, the model would need to be readjusted for it to function for a specific location or place.

2.3.3 Machine Learning Algorithms

In this section, the mathematical background of the ML models adopted in this study will be discussed, along with their differences, advantages and shortcomings.

2.3.3.1 LSTM

LSTM is an augmented type of RNN model, where it learns the sequential information with long term dependencies and capable of handling the problem of vanishing gradient which RNN suffers from (Santra & Lin, 2019). This problem arises when the number of layers increase using the same activation function the gradients of the loss function approaches zero making it hard to train the network through the backpropagation of errors. The LSTM prevents this from happening through the memory cells, where each of these cells maintain a cell state and a hidden cell state uses three gates, the input, output and forget gate (Sundermeyer et al., 2012). These gates control the amount of information of the past that flows in the cell. Fig. 1 is a representation of the architecture of the LSTM network.

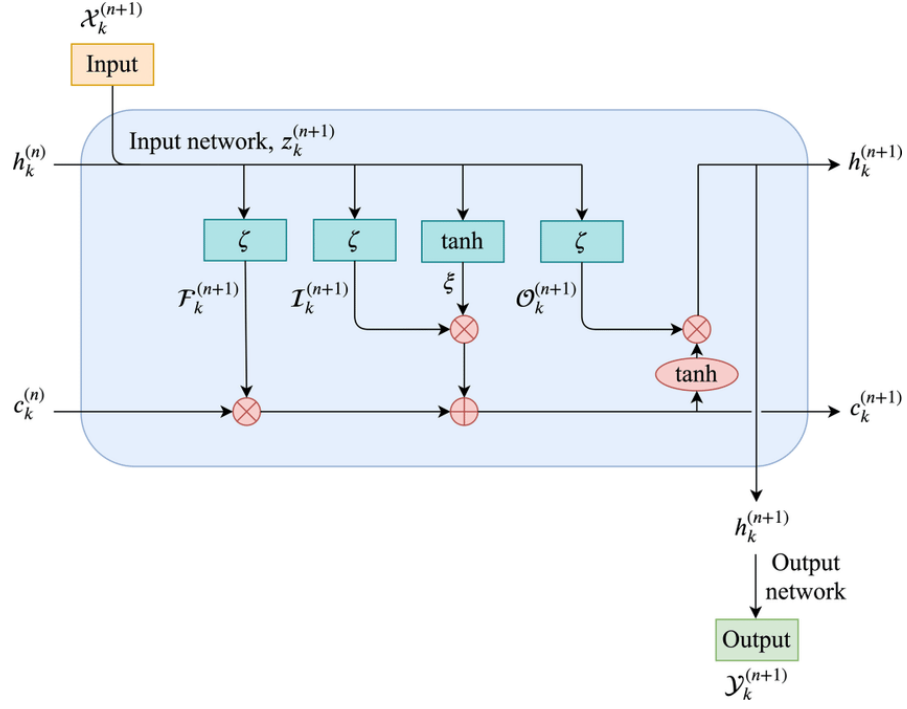


Fig. 1: The SLTM network structure with its general information flowchart (Rahman et al., 2019).

The LSTM gates in the network can then be commuted between the input and output sequence with the following equations:

$$fg = \text{sigmoid}(W_{fg}X_t + W_{hfg}h_{t-1} + b_{fg}) \quad (2.1)$$

$$ig = \text{sigmoid}(W_{ig}X_t + W_{hig}h_{t-1} + b_{ig}) \quad (2.2)$$

$$og = \text{sigmoid}(W_{og}X_t + W_{hog}h_{t-1} + b_{og}) \quad (2.3)$$

$$(C)_t = (C)_{t-1} * (fg)_t + (ig)_t * (\tanh(W_cX_t + W_{hc}h_{t-1} + b_c)) \quad (2.4)$$

$$h_t = og * \tanh((C)_{t-1}) \quad (2.5)$$

Where in the above equations, fg , ig and og stand for forget gate, input gate and output gate respectively. The W and b represent the weights and bias variables of the three gates mentioned respectively and a memory cell. The h_{t-1} symbolizes the prior hidden layers units that element-wise adding with weights of three gates. After the processing of Eq. 2.4, $(C)_t$ turns into current memory cell unit. Eq. 2.5 shows the element wise multiplication of prior hidden unit outputs and previous memory cell unit. Add the non-linearity on top of the three gates in the form of tanh and sigmoid activation functions, which is shown in Eqs. (2.1–2.5). Here, $t - 1$ and t are previous and current time steps. However, there is a limitation to this model, and that is it only preserves information of the past and cannot use the future one.

2.3.3.2 Bi-LSTM

To overcome the limitation addressed with the LSTM models, the bidirectional LSTM will run the model inputs in two ways, one from the past to future, and the second from future to past the thing that differs from LSTM. With this network, the previous and future information is exploited in output layer. Fig. 2 is an example of a Bi-LSTM network.

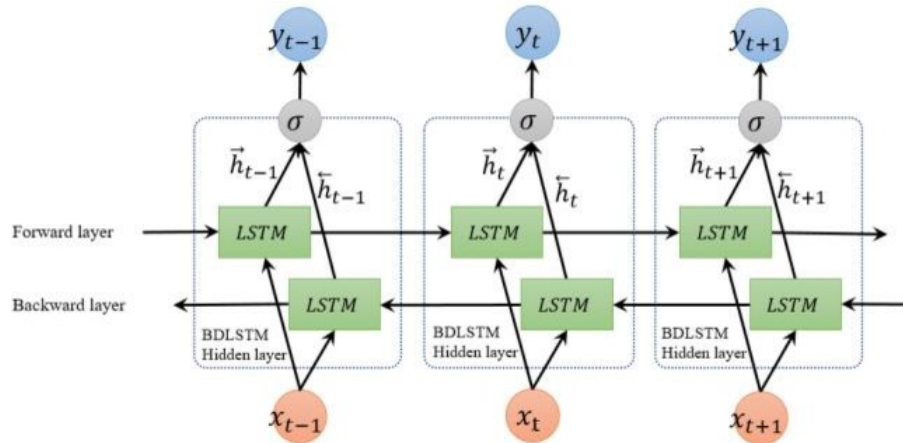


Fig. 2: Bi-LSTM network architecture shown unfolded in time for three time-steps (Cui et al., 2018).

An input sequence $X = (X_1, X_2, \dots, X_n)$ in Bi-LSTM is calculated in forward direction as $hft = (hf_1, hf_2, \dots, hf_n)$ and backward directions

as $hbt = (hb_1, hb_2, \dots, hb_n)$. Where hf represent the feedforward and hb feedback ward direction. The final out of this cell y_t is formed by both ht and hbt , the final sequence of out looks like $y = (y_1, y_2, \dots, y_t, \dots, y_n)$.

2.3.3.3 GRU

Gate Recurrent Unit is a special type of RNN based network (Mateus et al., 2021). Similar to the internal unit of LSTM, GRU also uses gates to control the flow of information, however it differs in combining the incoming port and the forgetting port in LSTM into a single update port. This makes the GRU much less complicated in comparison with the LSTM since it has less number of gates making it easier to train as less calculation is required to update the internal states (Busari & Lim, 2021). Therefore, when it comes to small dataset the GRU is preferred, otherwise the LSTM works best for a larger dataset. As shown in Fig. 3, GRU is much simpler than LSTM since it includes only two gates, the reset gate and update gate, that control the information flow inside the units.

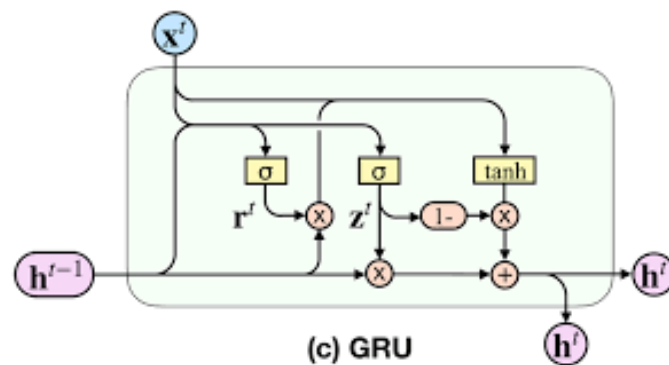


Fig. 3: Structure of a GRU network (Huang et al., 2019).

The mathematical functions used to control the mechanism in the GRU cell is then given as follows:

$$z_t = \sigma(x_t W_z + h_{t-1} U_z + b_z) \quad (2.6)$$

$$r_t = \sigma(x_t W_r + h_{t-1} U_r + b_r) \quad (2.7)$$

$$\bar{h}t = \tan(rt \times ht - 1U + xtW + b) \quad (2.8)$$

$$ht = (1 - zt) \times \bar{h}t + zt \times ht - 1 \quad (2.9)$$

Where Wz , Wr and W denote the weights for the corresponding connected input vector. While the Uz , Ur and U represent the weights of the previous time step. The b denotes the bias and the σ is the logistic sigmoid function.

2.3.3.4 Bi-GRU

The Bi-GRU is an enhanced version of a GRU network (Li et al., 2020). Similar to the Bi-LSTM, the two-layer structure allows the use of forward and backward information in the network. The only difference between the two is in the cyclic unit. Because the GRU network is simpler than the LSTM, then hence the Bi-GRU model is simpler than Bi-LSTM. Fig. 4 illustrates the structure of a Bi-GRU network.

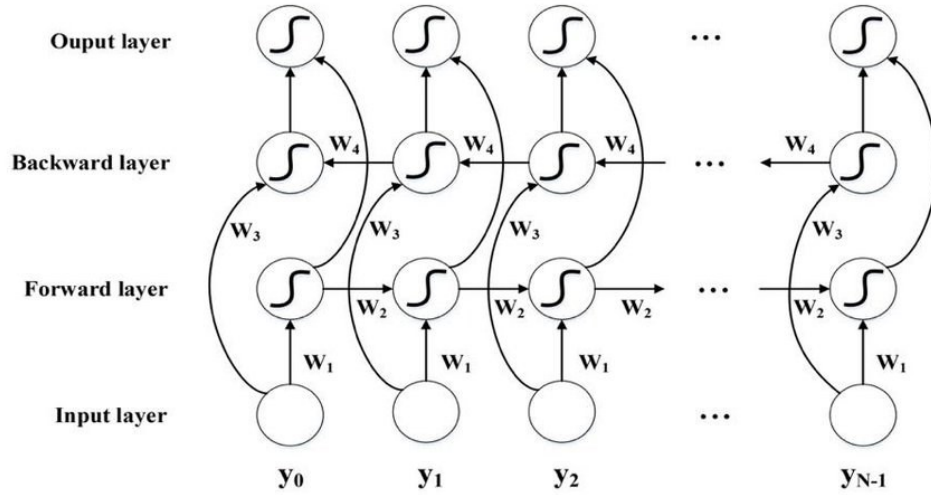


Fig. 4: Architecture of a Bi-GRU network structure (Meng et al., 2020).

The mathematical calculation for forward and backward layer can then be calculated as follows:

$$\vec{h}_t^1 = f \left(W_{xh1} \rightarrow x_t + W_{h1h1} \rightarrow \vec{h}_{t-1}^1 + b_{h1} \rightarrow \right) \quad (2.10)$$

$$\overleftarrow{h}_t^1 = f \left(W_{xh1} \leftarrow x_t + W_{h1h1} \leftarrow \overleftarrow{h}_{t-1}^1 + b_{h1} \leftarrow \right) \quad (2.11)$$

$$\vec{h}_t^2 = f \left(W_{\rightarrow xh2} \vec{h}_t^1 + W_{h2h2} \rightarrow \vec{h}_{t-2}^2 + b_{h2} \rightarrow \right) \quad (2.12)$$

$$\overleftarrow{h}_t^2 = f \left(W_{\leftarrow xh2} \overleftarrow{h}_t^1 + W_{h2h2} \leftarrow \overleftarrow{h}_{t-2}^2 + b_{h2} \leftarrow \right) \quad (2.13)$$

$$y_t = g \left(W_{h2y} \rightarrow \vec{h}_t^2 + W_{h2y} \leftarrow \overleftarrow{h}_t^2 + b_y \right) \quad (2.14)$$

Where $\vec{h}_t^1$ and $\overleftarrow{h}_t^2$ are the output vectors of the hidden layer of the forward layer in the first and second layers of the Bi-GRU at time t , and h is the number of units in the GRU cell. Consequently, the $\overleftarrow{h}_t^1$ and $\overleftarrow{h}_t^2$ are the output vectors in the first and second backward layer.

Although LSTM, Bi-LSTM, and GRU have been firmly established as state-of-the-art models in sequence modelling, LSTM, in particular, has received broader coverage in research (Sherstinsky, 2020). However, the other RNN variants have their strengths in time-series forecasting, which, to the best of our knowledge, has not been investigated yet in the forecasting of Solar power.

Chapter Three

Methods and Data

3.1. Raw Data

3.2. Pre-Processing

3.3. Performance Metrics

3.4. Forecasting Models

3 Chapter Three

Methods and Procedures

In this chapter, the data collection, description, pre-processing step and the method implemented in this thesis will be discussed in this chapter. Further, the evaluation metrics of the models will also be discussed.

In this work, all the steps taken, from processing data to training the model and making prediction have been implemented in the following experimental setup:

- Programming language: Python version 3.7.5,
- Libraries: Scikit-learn version 0.23.2, TensorFlow version 2.7 and Keras library version 2.2.4.
- Configuration of the computer: Intel Core i7-8565U CPU, 16 GB RAM, 64bit-Windows 10.

3.1 Raw Data

3.1.1 Data Collection & Description

The data to be used in this thesis was provided by the Renewable Energy Lab at the campus of UoB that is located in Sakhir, Kingdom of Bahrain at location of 26.0510°, 50.5134°.



Fig. 5: Location of the campus from which the data is extracted, and the solar-grid modules deployed at UoB- part of 5mw BAPCO pilot project.

A total of 3 datasets have been collected, all of which were of very high resolution, sampled at one-minute intervals, that covering different time frames. The first dataset captures the Solar radiation, the second records the current and voltage produced from the solar radiation, and the last dataset contains meteorological data.

3.1.2 Solar Data

The solar irradiance was taken using the pyranometer, measuring the global radiation. The solar data consisted only of the global radiation, recorded for approximately two years. Negative values were recorded in the data, which were observed during nighttime. This often happens due to the Thermal Dome effect caused by the thermal offset generated from the temperature difference between the outer and inner dome, which is opposite to the daytime situation, hence casing negative (Zo et al., 2017). Table 3 shows the first and last two rows of the data as a sample, along with the graph, which shows the pattern over two years.

Table (4): Solar dataset showing the start and end date for logged data and the number of instances.

	TIME_STAMP	Global Radiation
0	2018-11-04 14:48:00	343.00000
1	2018-11-04 14:49:00	398.60000
.	.	.
.	.	.
616022	2022-01-03 23:58:00	-9.090000
616023	2020-11-04 00:00:00	-9.050000

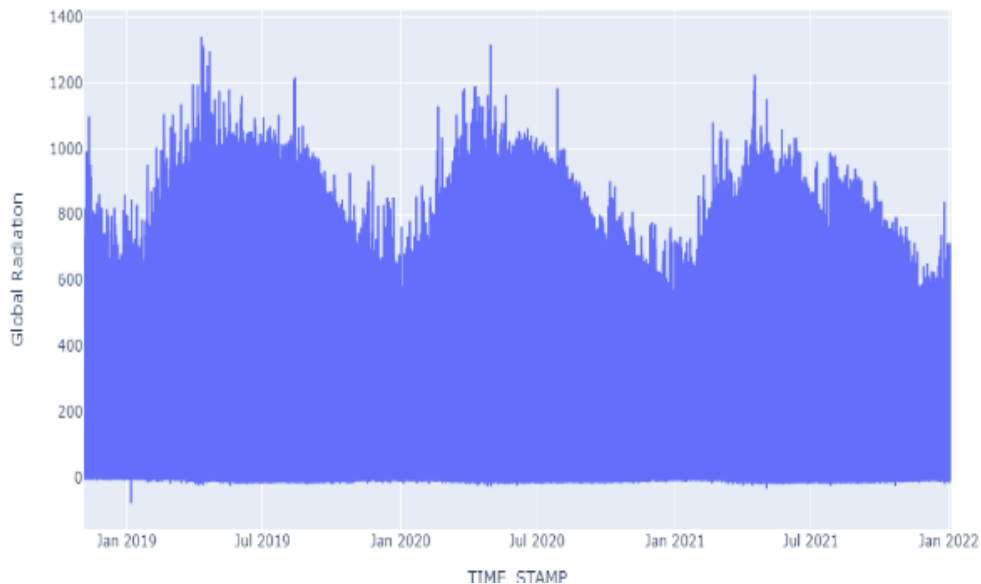


Fig. 6: Hourly reading of Global Solar Irradiance from Jan 2019 to Jan 2020

One can see a clear pattern where global solar irradiance recorded the highest reading between April and August, with some distinct spikes at the beginning of April. The highest solar irradiance recorded was on the 4th of April 2019, with 1340 Watt. The negative values can also be seen in the graph.

3.1.3 Poly Data

The poly data is named after the type of solar cells in the panel, the Polycrystalline Solar Cells. The poly data consists of two columns, the voltage at maximum power (V_{mpp}) and the current at maximum power (I_{mpp}). The maximum amount of power generated is configured by getting the product of the V_{mpp} and I_{mpp} as described in section 2.0 (Ela et al., 2018).

Unlike the other datasets, the poly does not capture the total 24 hours of the day. Instead, the recording of data starts from 3 am till 7 pm. The data is captured for nearly three years, where Fig.7 shows the trend found in the V_{mpp} and I_{mpp} and a table that includes the data's first and last two rows along with the dates.

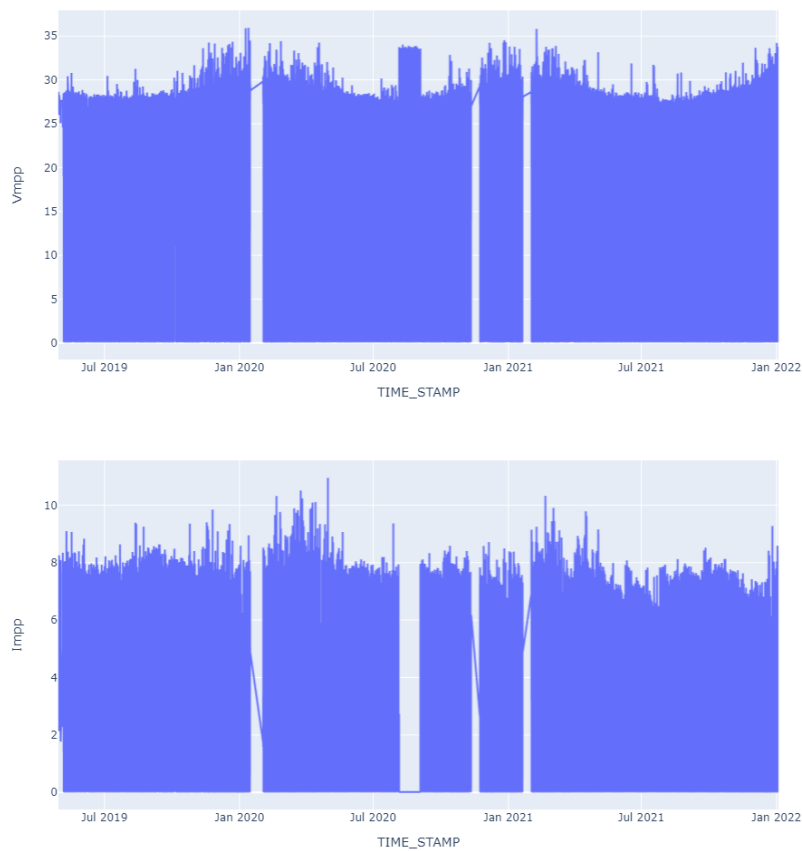


Fig. 7: Daily reading of current and voltage captured from the solar panels.

Table (5): Poly dataset showing the start and end date for logged data and the number of instances

	TIME_STAMP	Imp	Vmp
0	2019-04-29 08:49:00	5.733000	27.573000
1	2019-04-29 08:50:00	5.654000	28.004000
.	.	.	.
.	.	.	.
938795	2022-01-03 19:58:00	0.015000	0.107000
938796	2022-01-03 19:59:00	0.015000	0.107000

Similar to the solar data, the high peaks were recorded in April and August, with an unexpected rise in January. However, there appears to be missing data for several months, such as August 2020. Nonetheless, after further investigation, they are found to be small values close to zero, reflecting an equipment failure, as the Imp reading was constant at 0.015000 for the whole month. This was also the case for the Vmp values recorded.

3.1.4 Meteorological Data

The weather dataset comprises a wide set of meteorological features but recorded less historical data than the rest of the datasets, covering only five months, starting from July to November. This was collected from the solar energy lab. For a more profound explanation of the weather dataset, table (6) provides the variables and their respective description.

Table (6): Meteorological variables and their respective description.

	Variable Name	Description
1	Wind direction [°]	direction from which Wind originates
2	Wind direction 10min [°]	Wind speed direction at 10m
3	Wind speed 10 min [kt]	Wind speed at 10m
4	Wind speed knots [kt]	Wind speed
5	Wind gust speed [kt]	brief increase in speed of the wind
6	Dry temperature [°C]	Ambient temperature
7	Dew point [°C]	temperature the air needs to be cooled to in order to achieve a relative humidity
8	Wet temperature [°C]	Temperature of adiabatic saturation
9	Relative humidity [%]	measure of the water vapor content of air
10	Vapor pressure [hPa]	partial pressure of water vapour in the atmosphere
11	Apparent temperature [°C]	Perceived temperature
12	Global Radiation [W/m ²]	Instantaneous flux of global radiation
13	UV Radiation [W/m ²]	strength of the sunburn-producing ultraviolet
14	Precipitation [mm]	Accumulated precipitation
15	Pressure QNH [hPa]	mean sea level pressure

Initially, the model was trained on the whole five months of the data provided. But after conducting the visual inspection, it was worth investigating different historical times to train the model and test if it improves the model's ability to predict, similar to what was done (Prema & Uma Rao, 2015). As seen in Fig. 6, there is a clear trend of solar irradiance recorded in summer and winter. However, due to the shortage of historical data in the meteorological data, it was difficult to get a specific pattern from that dataset since the whole year cycle was not captured. Based on this, and the fact that operators can't always obtain all the necessary historical data to train the model successfully (AlShafeey & Csáki, 2021), the five months have been further broken down to two months of historical data and three months.

To link this with the pattern observed in the solar irradiance in Fig. (6), the choice of the time frame was specific and based on the solar irradiance data. The global solar irradiance data shows the highest peaks

were from February till September, and the lowest reading was captured from September till January. Accordingly, the two months data covered a summer season, from July to August inclusive, and the three months data covered the winter season from September to November. The two datasets are then referred to as summer data and winter data.

3.2 Data Pre-processing

All datasets were cleaned of missing and/or NaN values, where the main part of the data pre-processing went into sorting and studying the characteristics of the data. General steps were carried out to merge the datasets as follows:

- 1- The date column consisted of date and time and therefore had to be separated into two separate columns, a column for date only and another for the time.
- 2- The format of the new date columns has been changed to the DateTime format, where this column along with time has been set as index, putting the date column first index.
- 3- The PV column was created by taking the product of the current and voltage, and the current and voltage columns were dropped. Making the PV power generated the target.
- 4- The values were reduced to their mean value per hour.
- 5- The datasets were merged, and NaN values for PV power were set to zero. This is where the time before 3am and after 7pm did not capture values for current and voltage.
- 6- Another necessary step was taken and that is replacing the negative values in the Global Radiation to zero in both datasets.

The final dataset consisted of 8 attributes, seven meteorological variables, and the 8th is the target to predict the generated PV power values. The dataset consisted of a total of 3,673 records.

For the faulty recordings captured in the poly data, a proposed solution is to replace the 2019 data with 2020. August 2020 would be replaced with the 2019 poly data only for the current and voltage reading. Although the issue was only in the current reading for that month, both measures were replaced since power is the product of the two. Because there is a shortage in historical data for the weather variables, one way of verifying the accuracy and validity of the proposed solution was from the solar irradiance reading. This variable is most correlated to the PV power and has the same characteristics.

Further, solar irradiance is nothing but an energy conversion into electricity to get the power, meaning they are nearly the same. Therefore, one can observe that 2019 and 2020 had a similar pattern. To further validate this, the solar irradiance values for 2019 and 2020 were similar, as shown in Table (7). Based on this approach, the values were replaced only by the current and voltage, not the weather data, since they were unavailable.

Table (7): Average solar irradiance readings for the 5 months in 2019 and 2022

	2019	2020	% Difference
July	314.71	297.32	5.68%
August	291.33	276.32	5.29%
September	262.58	234.74	11.2%
October	209.86	207.11	1.32%
November	161.05	159.59	0.91%

For some days with similar issues, the value of the day was replaced by considering the value of the consecutive days. For example, for September, the first three days had only the value of Impp constant at 0.015000; therefore, the first two days were replaced with the last day of August values, current, and voltage, and the third day of September was replaced with the fourth day values of the same month. This is done to ensure more accuracy of the data and avoid biases. The figure below illustrates the PV power output for September.

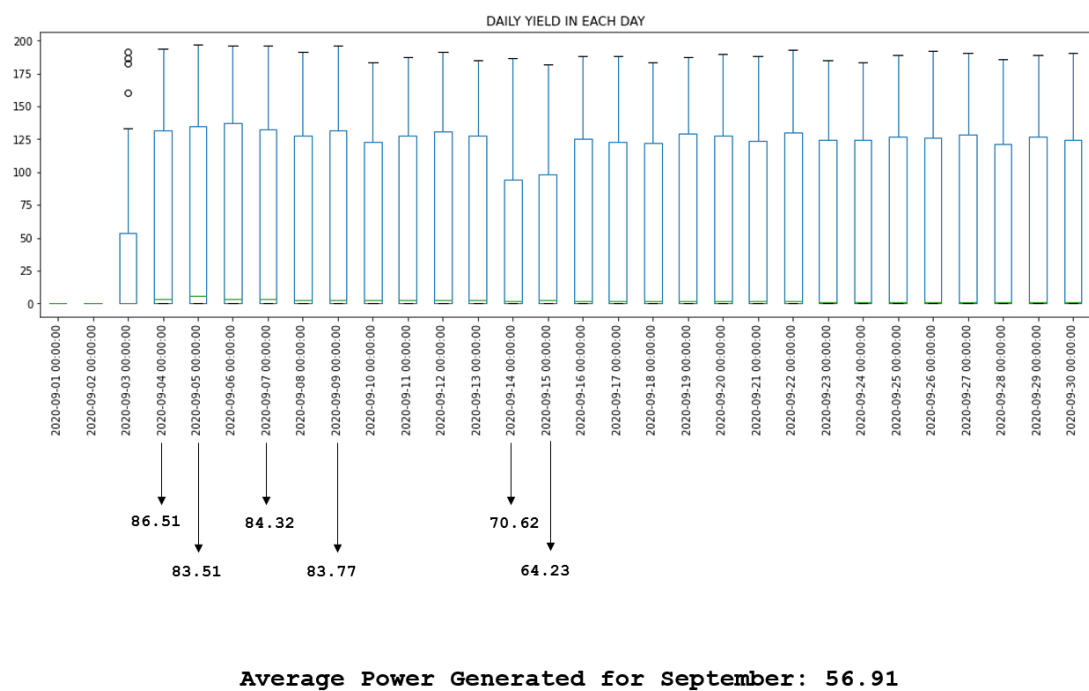


Fig. 8: Boxplot of the PV power output for the month of September, with mean values calculated for some of the days.

After this step was done, the average PV power output for the 5 months was calculated, along with the same 5 months but for year 2019 for comparison.

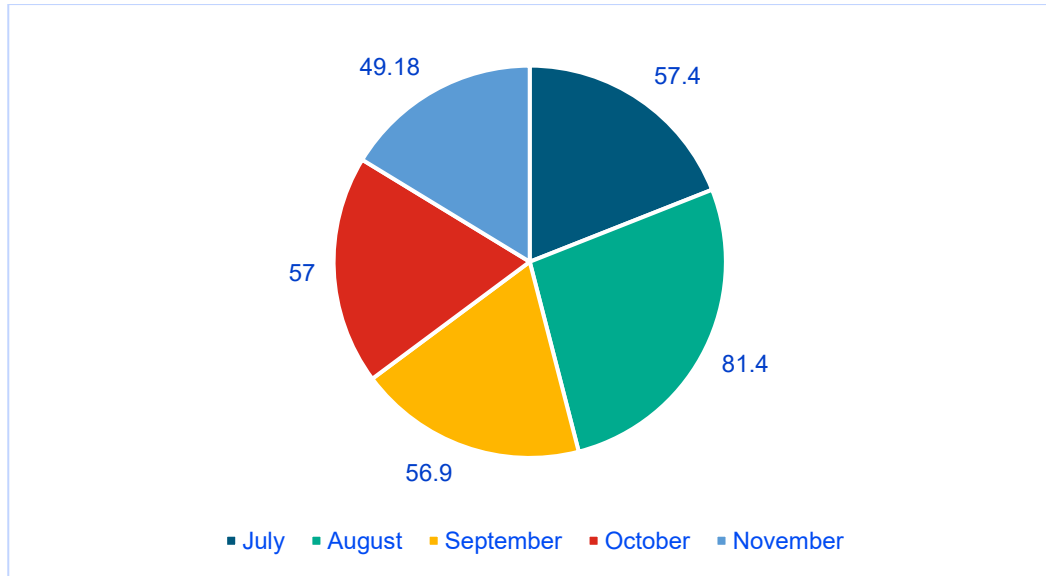


Fig. 9: Average PV power Output for each month per Watts

August has the highest average in solar power produced than the rest of the months. Although the number seem to be off quietly highly, it may be due to some weather conditions back in 2019 that caused such high values in August.

3.2.1 Feature Analysis

Feature Selection is one of the most prominent preprocessing steps in many machine learning tasks (Blessie & Karthikeyan, 2012). This technique reduces the features by identifying the most relevant features according to an evaluation (Soares et al., 2018). This facilitates model interpretation and reduce the chances of data being overfitted and hence produce better results due to the elimination of features that may confound the uncovering of patterns and trends, not to mention reducing the measurement and storage requirements, reducing training and utilization times.

There are three main types of correlation coefficients: Pearson's product moment correlation coefficient (PCC), Spearman's rank correlation coefficient (SCC) and Kendall Rank correlation coefficient (KCC). PCC is

one of the most commonly used technique (Mukaka, 2012). This technique relies on four key assumptions, which are listed as follows:

1. The variables are linearly related.
2. The data is normally distributed.

This coefficient is heavily affected by extreme values or outliers affect heavily the correlation test, and therefore may exaggerate or dampen the strength of relationship. Hence this coefficient becomes inappropriate when either or both variables are not normally distributed and the relationship is not linear. For a correlation between variables x and y , the formula for calculating the sample Pearson's correlation coefficient is given by:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^n (x_i - \bar{x})^2][\sum_{i=1}^n (y_i - \bar{y})^2]}} \quad (3.1)$$

Where

r = correlation coefficient

x_i = values of the x -variable in a sample

$\bar{x}$ = mean of the values of the x -variable

y_i = values of the y -variable in a sample

$\bar{y}$ = mean of the values of the y -variable

In contrast, the Spearman on the other does not rely on normality, nor that the data is linearly related CC (Akoglu, 2018). Thus, it's a non-parametric test. Moreover, it is a robust coefficient when extreme values are present. For a correlation between variables x and y , the formula for calculating the sample Spearman's correlation coefficient is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \quad (3.2)$$

Where

ρ = Spearman's rank correlation coefficient

d_i = is the difference between the two ranks of each observation

n = number of observations

Lastly, KCC is similar to the SCC, that is being non-parametric (Puth et al., 2015). It can be used for ordinal or continuous data. This coefficient evaluates the degree of similarity between two sets of ranks given to a same set of objects (Abdi, 2008). The key difference between these two non-parametric measures, is that Spearman's measure returns values closer to the desired coverage rates (Puth et al., 2015), whereas Kendall's results differ more and more from the desired level as the number of ties increases, especially for large correlation values. (Fredricks & Nelsen, 2007)

As concluded in past literature, the relationship between PV power and the meteorological variables is monotonic, meaning they have a non-linear relationship, violating the first assumption of PCC. To further examine the data for this thesis, the weather data were plotted against each other; there were no linearly related variables. Additionally, the distribution and Q-Q graph were plotted, and the Shapiro-Wilk test was carried out. This test is used for normality and the goodness of fit in the dataset (Jurečková & Pícek, 2007). Fig. 10 illustrates the visual inspection and the statistical test for normality. Moreover, the pair plot of all-weather variables can be found in Appendix 1.

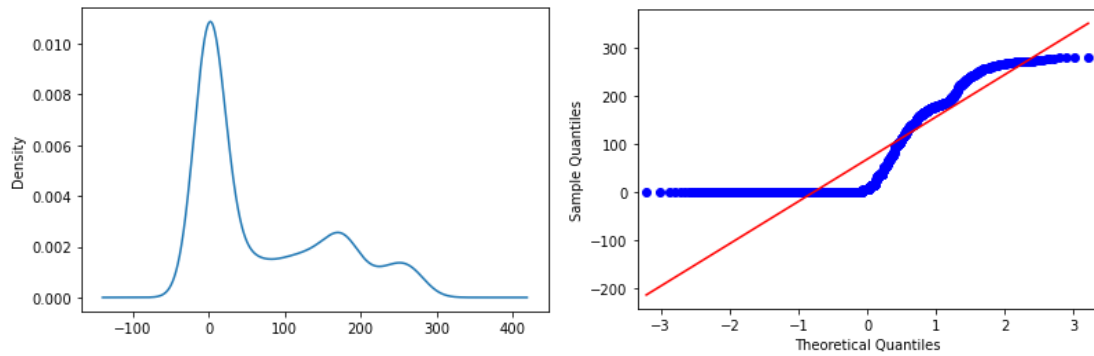


Fig. 10: Distribution and Q-Q plot of the data, where the Shapiro test gave 0.767, this means rejecting hypothesis that data is normally distributed, the data is not gaussian

Table (8) shows the results obtained using the Spearman correlation coefficient on the features in the weather data.

Table (8): Spearman correlation coefficient results for the meteorological al variables.

Attributes	Power
Wind direction [°]	0.076914
Wind direction 10min [°]	0.090402
Wind speed 10 min [kt]	0.600485
Wind speed knots [kt]	0.600375
Wind gust speed [kt]	0.601105
Dry temperature [°C]0.	0.646615
Dew point [°C]	-0.300967
Wet temperature [°C]	0.020928
Relative humidity [%]	-0.710431
Vapour pressure [hPa]	-0.29811
Apparent temperature [°C]	0.401795
Global Radiation [W/m2]	0.638872
UV Radiation [W/m2]	0.636921
Precipitation [mm]	-0.003647
Pressure QNH [hPa]	-0.190171
Global Radiation	0.812630
Power	1.00000

After conducting the correlation test, the next step would be determining the strengths of these correlated features (Akoglu, 2018). Many benchmarks are used to name strengths and choose based on a scale of the most critical variables. Table (9) outlines Spearman and Pearson's most commonly used interpretations of the correlation coefficients.

Table (9): Interpretation of Spearman's correlation coefficients.

Correlation Coefficient		Dancey & Reidy (Phycology)	Quinnipiac University (Politics)	Chan YH (Medicine)
+1	-1	Perfect	Perfect	Perfect
+0.9	-0.9	Strong	Very Strong	Very Strong
+0.8	-0.8	Strong	Very Strong	Very Strong
+0.7	-0.7	Strong	Very Strong	Moderate
+0.6	-0.6	Moderate	Strong	Moderate
+0.5	-0.5	Moderate	Strong	Fair
+0.4	-0.4	Moderate	Strong	Fair
+0.3	-0.3	Weak	Moderate	Fair
+0.2	-0.2	Weak	Weak	Poor
+0.1	-0.1	Weak	Negligible	Poor
0	0	Zero	None	None

In this context, the Chan YH will be the reference of identifying the most correlated features to the PV power output. The only attributes that would be selected are considered either: Perfect, Very Strong, and Moderate correlation. Accordingly, the selected attributes are the following:

- 1- Global Radiation W/m^2
- 2- Relative humidity [%]
- 3- Wind speed 10 min [kt]
- 4- Wind speed knots [kt]
- 5- Wind gust speed [kt]
- 6- UV Radiation [W/m²]
- 7- Dry temperature [°C]

3.2.2 Handling Outliers

After the data was sorted and the important variables selected, visual analysis was done to have an initial inspection of the data for any outliers for the weather variables. Outliers were found with varying intensity for each attribute, most notably in the Relative Humidity, as shown in Fig. 11. All box plots of meteorological variables can be found in Appendix 2.

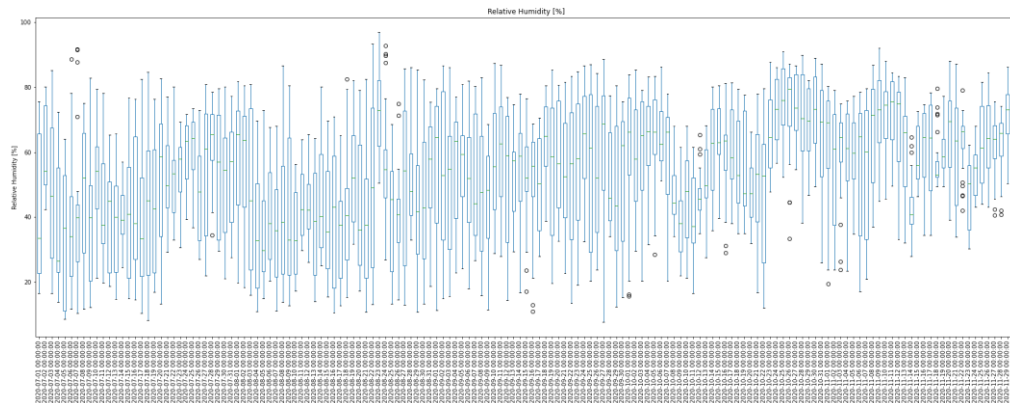


Fig. 11: Box-plot of hourly reading of Relative Humidity for the whole 5 months.

These outliers were not treated since this approach was considered not accessible since few outliers would be treatable such as altering the mean performance; however, what was found in the dataset was far more than just a few outliers. Further, the treatment of the outliers is rather complicated, and there is no unanimously accepted theoretical framework for the treatment of these outliers; hence would require formulating a specific approach based on the problem at hand, which is out of the scope of this thesis. Moreover, The PV power depends considerably on the variation and the irreducible errors of the weather data. Hence, the reading captured for the I_{mpp} and V_{mpp} , the power generated, was the product of the fact that these natural noises of data were present and existed in the first place to give the presented values of power generated. That and the fact the ML techniques are very robust when it comes to outliers, a finding concluded by many researchers in past literature, they were eventually left untreated.

3.3 Performance Metrics

Some standard metrics can evaluate the performance of the model. These metrics are based on the observed and predicted values of the model. Metrics give numerical values and make it easy to compare with other models. There are numerous performance matrices such as R², Root Means Squared Error, F1 score, etc.

As explained earlier, using the proper performance metric depends heavily on the nature and characteristics of the data. That is to say, R^2 is an invalid goodness-of-fit for non-linear data; as such, it will be excluded. The metrics used in this thesis are as follows.

3.3.1 Root Mean Squared Error

One of the common performance metrics of the forecasting model is the root mean squared error which can be given by equation

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.3)$$

where $\hat{y}_i$ means the prediction rating and y_i means the true rating in testing data set; N is the number of rating prediction pairs between the testing data and prediction result.

3.3.2 Mean Absolute Error

Another performance metrics of the forecasting model is the mean absolute error (MAE) which is given by equation

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3.4)$$

where, y_i is the true value and $\hat{y}_i$ is the forecasted value. For MAE, absolute values of the difference between observed and forecasted value are averaged.

3.4 Forecasting Models

For the proposed model, the RNN variants' structure will consist of an input layer with an Adam optimizer, one hidden layer, and a dense layer in the output with ReLu being the activation function, as illustrated in Fig.12. The input layer receives the meteorological variables, which would be processed in the hidden layer to be finally sent as computed information from the output layer, to give the predictor which is the PV power generated.

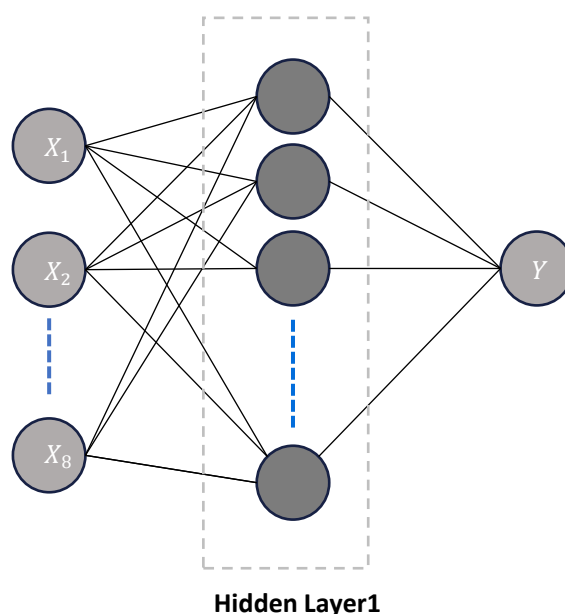


Fig. 12: Structure of the proposed vanilla network for all models, X_n are the variables as input and Y is the output which is the PV power generated

A similar structure is proposed for the hybrid model, which incorporates the CNN layer first before the RNN variants, as illustrated in Fig.13. However, for the CNN layer, the hyperparameters will be left the same throughout the experiment, as the aim is to test the effect of the CNN layer on the other RNN variants, GRU, Bidirectional GRU, and LSTM since it has been proven to enhance the accuracy of the LSTM model in

past literature. The CNN layer will have a kernel size of 3, followed by a Maxpool layer, with a size of 2.

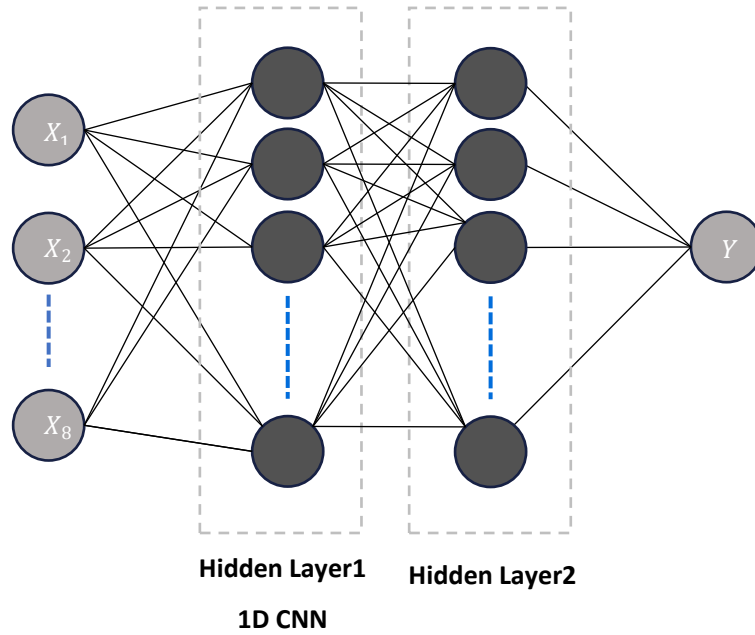


Fig. 13: Structure of the proposed hybrid network for all models, X_n are the variables as input and Y is the output which is the PV power generated

Both vanilla and hybrid models will be tested using 32, 64, and then 128 neurons per layer in the structure, with the number of hidden layers remaining constant. No more combinations of parameters were tested due to the extensive computational time needed to run the ML models. Table (10) summarizes the hyperparameter setting for the Vanilla and Hybrid models. Finally, the data was split into 60% for training data, 20% for validation, and 20% for the test set.

Table (10): Hyperparameter values for the vanilla and hybrid network

Method	Hyperparameters	Value
Vanilla Model	# Hidden layer/s	1
	#Neurons	32, 64, 128
	Input time steps	24
	Input feature dimensions	8
	Learning Rate	0.001
	Optimizer	Adam Optimizer
	#Epochs	500
	Batch Size	32
	# Nodes in ReLu layer	1
Hybrid – CNN Layer	# Hidden layer/s	1
	#Neurons	32, 64, 128
	Kernel Size	3
	Pool Layer	Maxpool, size 2

Chapter Four

Results and Discussion

4.1. Results

4.2. Discussion

4 Chapter Four

Results and Discussion

The results would be divided in separate sections based on the experiment conducted in this thesis.

4.1 Results

4.1.1 Case 1- Selection of data time

Before conducting the actual experiment, a minor investigation was carried out initially to test the model's accuracy based on the data time cycle, which is whether to use the entire day data, including the nighttime data, or the half-day data, excluding the nighttime data. The poly data had the data collected from 3 am till 7 pm. The fact that there is no solar power at night gave an initial assumption that using the half-day data would provide adequate accuracy and efficient computational time. Moreover, Isaksson & Conde (2018) also addressed this idea. Nonetheless, this was not justified by an actual experiment in their research. Therefore, this little experiment was carried out. The results contradict the initial assumption, where using the entire day data gave lower RMSE than using the half-day data.

Table (11): Result for the full day cycle and half day cycle.

		Half-day Data			Full Day Data		
	#Neurons	32	62	128	32	64	128
RMSE	LSTM	107.0	124.63	127.77	96.56	96.70	98.98
	Bi-LSTM	32.37	39.41	29.92	27.82	31.63	33.89
	GRU	29.21	30.42	30.25	22.71	22.96	22.64
	Bi-GRU	33.69	29.85	28.78	28.40	23.69	23.18
MAE	LSTM	88.88	93.81	96.18	66.98	67.14	68.42
	Bi-LSTM	19.5	24.75	18.64	13.49	15.66	18.28
	GRU	15.99	18.94	18.45	10.78	11.38	10.62
	Bi-GRU	20.84	19.31	17.72	13.91	11.58	10.43

As can be seen, when the night data was included in the model, the RMSE and MAE reduced in all models significantly in the LSTM models. Accordingly, the full-day cycle was included in the training of the rest of the models, and no further models were tested for the half-day data.

4.1.2 Case 2 – PV Power Forecasting on Full Dataset

Table 12 shows the results obtained for both the vanilla models and the CNN form varying the number of neurons on one single hidden layer.

Table (12): RMSE and MAE values for models trained on 5 months data.

		Vanilla Models			Hybrid Models		
	# Neurons	32	62	128	32	62	128
RMSE	LSTM	96.56	96.703	98.98	42.86	23.22	27.33
	Bi-LSTM	27.82	31.63	33.89	27.99	25.81	25.46
	GRU	22.71	22.96	22.64	20.68	22.44	21.51
	Bi-GRU	28.40	23.69	23.18	<u>20.74</u>	27.22	24.61
MAE	LSTM	66.98	67.13	68.42	24.55	12.57	14.73
	Bi-LSTM	13.49	15.657	18.28	14.52	13.22	13.10

GRU	10.78	11.380	10.692	9.23	10.06	9.73
Bi-GRU	13.91	11.58	10.427	9.64	13.66	12.25

It was observed that changing the number of neurons in the vanilla models did not significantly affect the RMSE and MAE. The only case with a notable change was in the BI-LSTM model, where increasing neurons caused an increase in the RMSE value.

Surprisingly, when it comes to the vanilla models, LSTM gave the lowest accuracy, highest RMSE, and MAE compared to all other models. Other models achieved similar results between the GRU and Bi-GRU, followed by Bi-LSTM. Adding the CNN layer significantly improved the performance of the LSTM, with a decrease of 80% in RMSE using the 62 neurons. In contrast to the other models, the effect was not as substantial, particularly in the GRU and Bi-GRU. However, it contributed to the model's performance nonetheless. The best model was using the CNN-Bi-GRU model with 32 neurons.

4.1.3 Case 3 – PV Power Forecasting Based on Weather Seasons

Tables 13 and 14 show the results obtained for both the summer and winter data.

Table (13): RMSE and MAE values for models trained on summer data.

Vanilla Models					Hybrid Models		
RMSE	# Neurons	32	62	128	32	62	128
	LSTM	139.22	118.99	122.77	84.66	85.22	53.72
	Bi-LSTM	35.90	40.40	29.42	11.8	8.17	9.06
	GRU	11.94	16.31	16.79	19.85	12.55	8.88
	Bi-GRU	17.15	15.40	12.81	78.59	45.46	35.32

MAE	LSTM	102.17	91.98	90.69	48.43	48.85	29.03
	Bi-LSTM	22.32	23.99	17.42	6.47	<u>4.23</u>	4.75
	GRU	6.56	9.75	9.38	10.41	7.28	4.89
	Bi-GRU	9.50	8.38	7.00	44.4	23.83	18.19

Table (14): RMSE and MAE values for models trained on winter data.

Vanilla Models					Hybrid Models		
RMSE	#Neurons	32	62	128	32	62	128
	LSTM	79.85	96.39	87.83	56.88	35.56	29.04
	Bi-LSTM	39.54	41.25	33.51	22.15	<u>21.72</u>	22.85
	GRU	27.22	29.02	29.35	23.53	29.88	23.71
	Bi-GRU	32.67	33.12	33.51	34.19	27.54	27.22
MAE	LSTM	54.79	66.78	62.78	30.42	20.34	14.13
	Bi-LSTM	20.70	20.90	19.07	10.39	<u>10.32</u>	10.70
	GRU	14.50	13.95	14.21	13.97	14.21	11.94
	Bi-GRU	18.92	17.01	19.07	18.27	13.53	12.97

One can observe from the tables above for both summer and winter data that there is an acute increase in the model's accuracy when increasing the number of neurons, which was more notably when using the five months data. Additionally, some models' accuracy increased with an increasing number of neurons, and others decreased depending on the dataset. For instance, one can observe the Bi-GRU model's performance using the summer data. The accuracy is enhanced, which was not the case when using the winter data, where the RMSE did not change significantly. On the other hand, for both datasets, the GRU networks performance tends to decay with an increasing number of neurons in the Bi-LSTM model, whereas in the winter dataset, the performance was enhanced.

As for the vanilla models, LSTM overall gave poor performance compared to other models, similar to the results noted when using the five months data. Moreover, the GRU network with 32 neurons best performed for summer and winter data with RMSE 11.94 and 27.22, respectively. Unlike in case 2, adding the CNN layer did not enhance all models' performances. For instance, the CNN - GRU increased the RMSE significantly when using the summer data, while the winter was improved only a little. On the contrary, the effect on Bi-LSTM enhanced the most when adding the CNN layer. The best model for both datasets, nonetheless, was of the same network, and that is the CNN-Bi-LSTM using the 62 neurons, with the summer data giving least RMSE than the winter data.

4.1.4. Results Summary

The Bi-LSTM, GRU, and Bi-GRU produced much lower RMSE and MAE values than the LSTM model. Further, adding the CNN layer enhanced all models' performances for all datasets, with few exceptions. While changing the number of neurons did not affect all models, some models become more accurate when increasing the number of neurons using the summer data and less accurate on winter data. Nonetheless, most models for the three datasets gave better performance when using 32 neurons for the vanilla models and 62 with the CNN layer. Splitting the data based on weather conditions showed good performance, with the summer data giving the least RMSE value in all models except LSTM.

Although the RMSE and MAE cannot be compared as they measure the error magnitude differently, as explained in section 3, one can observe a clear trend in results where these matrices are coherent with each other. Meaning when the RMSE decreases, MAE decreases as well. For easier visualization, the graph below illustrates the trend for the three datasets and

their respective RMSE and MAE results with the number of neurons.

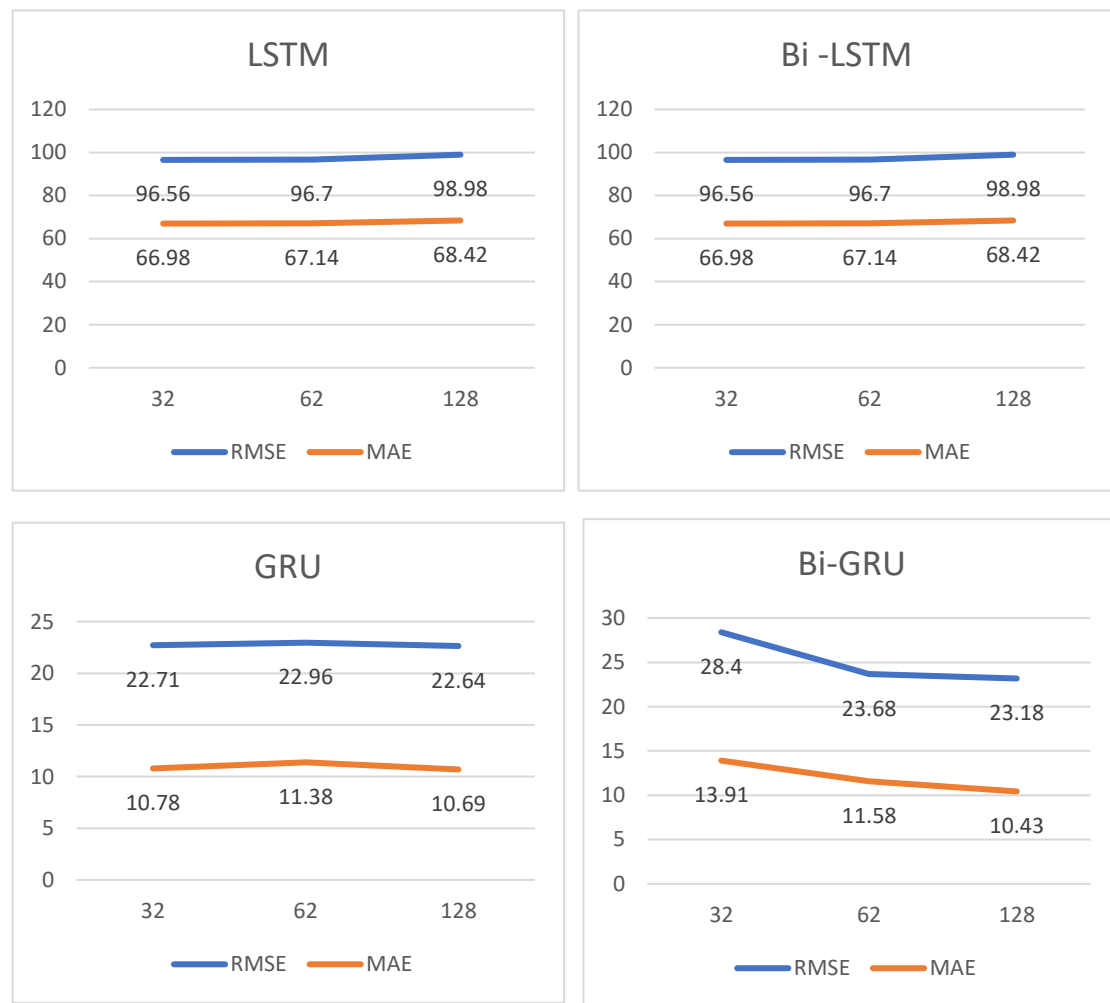


Fig. 14: RMSE and MAE values for 5 months data.

Although both matrices were in line with each other, the best models chosen for each case were ultimately based on the RMSE since it penalizes the deviations, which is the aim of our best models not to have predictions that diverge significantly from the actual value.

Finally, to answer the Researchers question which was addressed in chapter 1, the answers are listed below in the order of the questions:

- For the meteorological attributes most correlated with Solar Power are the following: Global Radiation, Relative humidity [%], Wind speed 10 min [kt], Wind speed knots [kt], Wind gust speed [kt], UV Radiation [W/m²], and Dry temperature [°C].

- CNN-Bi-GRU gave best results using the 32 neurons, which gave RMSE value of 20.74.
- Splitting the data to two months of summer season enhanced the performance significantly, where the best model was CNN-Bi-LSTM using 64 neurons which gave RMSE of 8.17.

4.2. Discussion

In this section, the results obtained will be analyzed and discussed. The first section will handle the results of the PV power forecasting by the ML techniques, followed by the different datasets used in Case 2 and Case 3. In the later chapters, the discussion is directed toward the general issues of the study as a whole.

4.2.1 Machine Learning Techniques

The achievements on the LSTM model, both the vanilla and hybrid network, were disappointing as the validation measures were extremely high compared to the rest of the models. Contrary to all mentioned Researchers in section 2, the LSTM gave the worse performance, and the other RNN variants achieved higher performance. Nonetheless, adding the CNN layer enhanced all models considerably in LSTM models, which agrees with all studies in the literature that incorporated the CNN layer into their model.

Several reasons might explain this result. First, the LSTM is known to work better on long historical data, and the data given consisted of only five months. Looking at past Researchers, all had one year or more worth of data, which would then be suitable to train for the LSTM. Further, the GRU is known to work best on small data, this might explain the lowest RMSE was noted when trained on summer data, but the significant change between LSTM and other models is still very substantial. The finding

aligns with (Louzazni et al., 2020), where the study used one month of historical data and compared multiple ML models, and scholars noted that the NARX performed best, not LSTM. Another research that confirmed using less historical data would give satisfactory results (Prema & Uma Rao, 2015). Two months showed the best performance than four months and one month; however, this was used on simple statistical tools.

Another issue may be the outliers present in the meteorological variables captured by the Bi-LSTM and Bi-GRU more adequately than LSTM. Since both networks have in common that additional layer of information in their network, which helped capture the information more adequately.

The RNN variants differ in their network. Accordingly, not all will perform equally using the same parameter. It was observed that some models overfitted at early epochs while others did not. This means it would be more adequate to custom the parameters accordingly to each model. Not to mention that the same model would perform differently on different datasets.

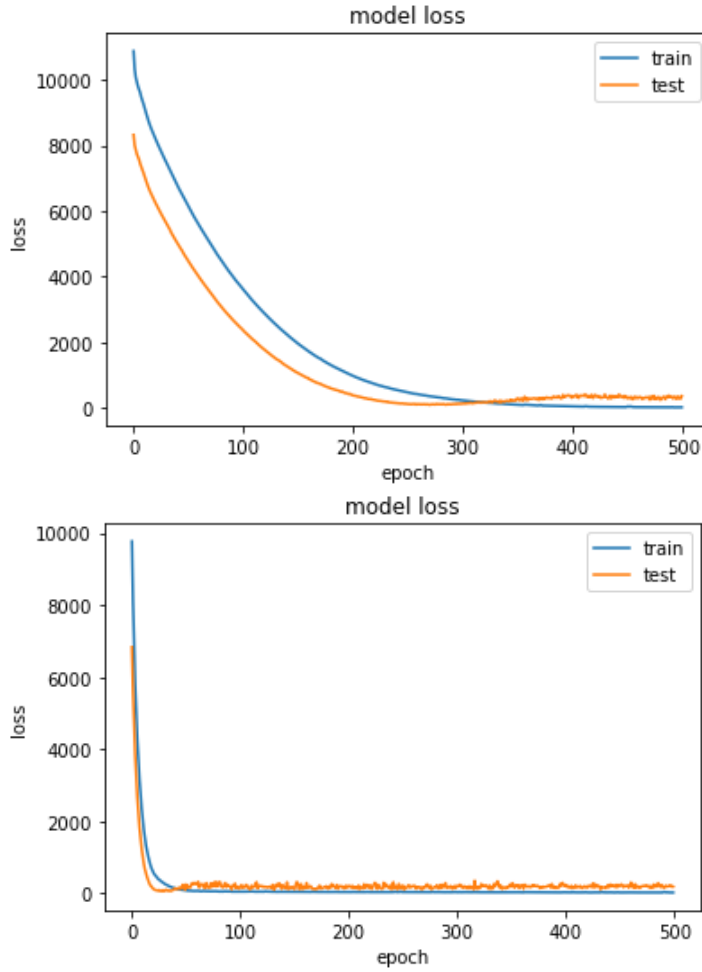


Fig. 15: Top to down: CNN-LSTM and CNN-Bi-LSTM using 128 neurons for the 5 months data.

The CNN-LSTM seems not to be generalizing well and is learning well on the training but not as much as for the test, which is diverging at epoch 300, in contrast to CNN-Bi-LSTM at epoch 50. This means that the CNN-Bi-LSTM model learned faster than CNN-LSTM and didn't need that much of epochs. Therefore, it would have made a perfect model for CNN-LSTM when using epoch 300 and epoch 100 for CNN-Bi-LSTM. Fig. 16 shows the same model, CNN-LSTM, for the three datasets.

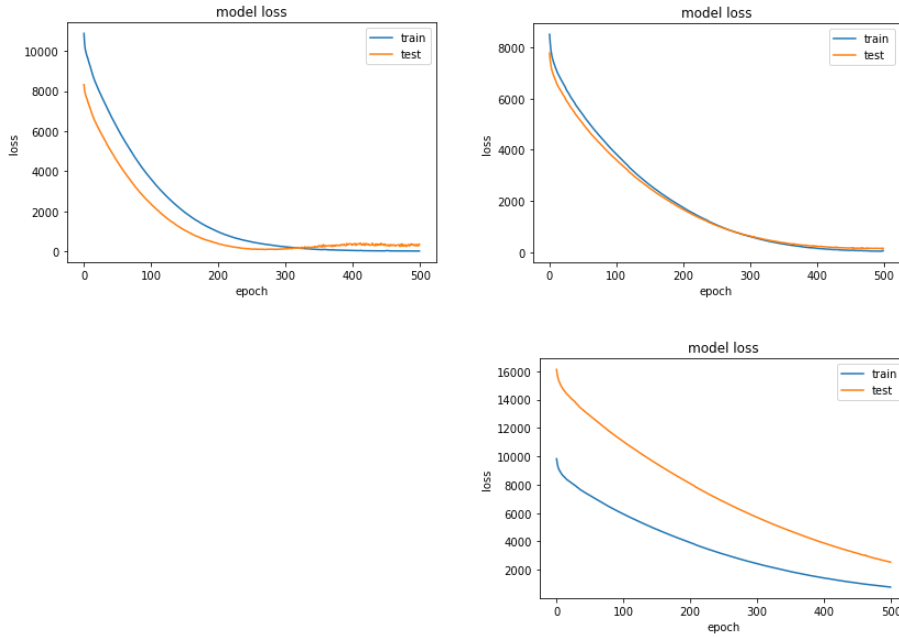


Fig. 16: Clockwise: 5 months, winter and summer data loss graph using the CNN-LSTM 128 neurons model.

Similar to what was discussed above, the performance of the same model varied based on the data. Further, for five months and winter data, the model gave similar RMSE, 27.33 and 29.04, respectively, but the epoch at which the model started to overfit is different. For the five months, the model would have performed better at epoch 300 while the winter data at 200. Although the epoch value is quite similar, the level of divergence is notable, and the five-month data seem to be diverging more significantly than winter data. This means there is room for better improvement for both models using specific hyperparameters for each dataset. Furthermore, the summer data already has overfitting, suggesting that this model is not suitable from the start for the summer data.

The models exhibit different performances for the same hyperparameters between each vanilla model and hybrid form and differ based on the same dataset. This means that the hyperparameter selection needs to be specifically tailored for each model based on their characteristics in their network structure.

4.2.2 Data

Splitting the five months' data helped improve the results for all models trained with few exceptions based on the weather season in contrast to (Najibi et al., 2021) and (Chen et al., 2011). Specifically, the GRU and Bi-GRU gave the best performance when using the summer data. To connect this to the fact that GRU, Bi-GRU, and Bi-LSTM work better on small data and the fact LSTM model gave a poor performance for all datasets means that splitting the data made these models further effective in capturing the pattern present. In addition, using the five months of data, the model was forced to generalize over varying weather conditions. The five months are captured from July till November, consisting of Summer and beginning of winter weather. This might explain why these models performed well once the data was split based on the season, giving these models the ability to learn better and focus on these specific weather conditions and not have to make a trade-off between learning the different pattern in summer and winter data. Although the season-based datasets gave higher performance than using the whole five months, this, on the other hand, models which are build based on specific climate conditions would mean changing models from season to season accordingly, which may be inefficient in an operational setting.

Not enough historical data affected identifying the actual performance of the model. As mentioned earlier, LSTM work best with long data, which was trained for five months. Moreover, the summer and winter data did not cover the corresponding weather season, and they were at least three months short on the actual cycle of summer and winter. Evidently, having the whole year cycle, summer and winter seasons would be more adequate in evaluating the model's performances.

Alternatively, typical days are observed in both the summer and winter data. This means there seem to be cloudy days in July and in October. The day in July may be due to a dust storm that caused a drop in the efficiency of PV panels, similar to the scenario of having cloudy days, which reflect the light and hence reduce the production of power generated. Fig. 17 shows the PV power generated for one of the days in July and October and the PV power generated, which are very similar for each day.

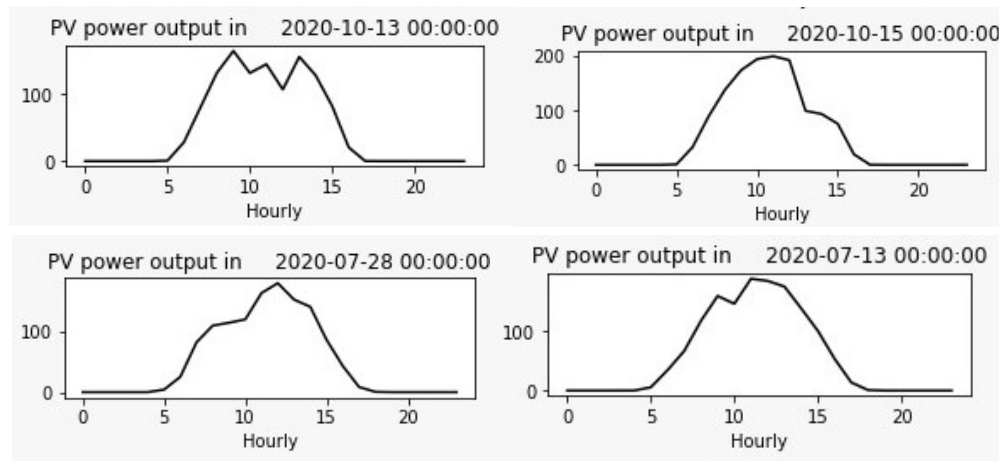


Fig. 17: First row shows two days in the month of October, second row two days from July.

Similar to Chen et al. (2011) approach, perhaps grouping the similar days would have enhanced the model's accuracy. Since the idea behind splitting data based on weather conditions, summer and winter, in Bahrain, as there aren't many seasons and prominent weather variations as opposed to other countries, so will be the case in splitting based on only sunny and cloudy days. As for rainy days, Bahrain does not receive much rainfall, where the average annual rain is only nine days (Directorate of Enviromental Assessment and Planning, 2009). This can be dealt with in a particular case, perhaps an outlier, and it is negligible considering 365 days per year. To compare the approach between what was proposed in this and in and (Chen et al., 2011), more historical data is needed to justify and

reinforce which would be more adequate and suitable in the case of Bahrain.

It was observed in table 10 that summer data was performing significantly well than winter data, especially in model GRU and Bi-GRU. A possible reason is that the outliers were present more heavily in the months covering the winter season. Fig. 18 illustrates the two dataset's humidity and dry temperature attributes.

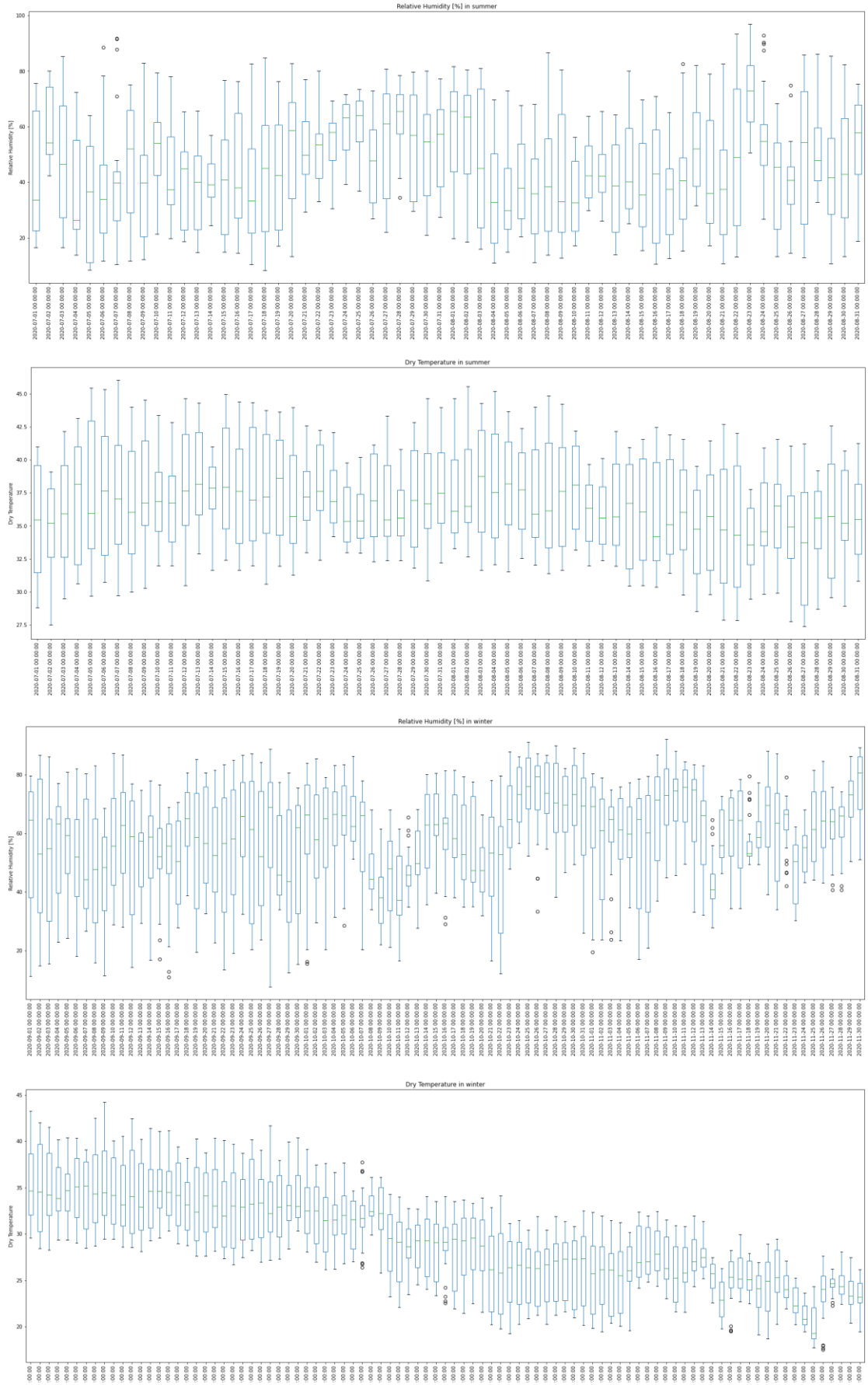


Fig. 18: First two figures are the Relative Humidity and Dry Temperature during July August, summer data, and the last two during September to November, winter data.

This might be why the winter data gave better results than summer data when trained on the LSTM, similar to what was found in (Nguyen et al., 2021). In addition, scholars noted that the LSTM model could identify and learn the noises present in the weather data, which might have been the case here with the winter data since it consisted of more outliers than summer data.

Another aspect that might affect the model's performance is replacing the 2019 PV power values with 2020. In the pre-processing step, the faulty equipment issue discussed earlier had to be replaced within months of 2019. This included biasness to the model as the data is not authentic anymore, which means that the August PV power was not produced for the exact weather conditions 2020, rather 2019. Thus, this is not ideal, as accordingly, the weather variables should have also been.

Chapter Five

Conclusion and Recommendations

5.0. Conclusions and Recommendations

5 Chapter Five

Conclusion and Recommendations

5.1 Conclusions and Recommendations

In this thesis, the relationship of the meteorological variables with the PV power generated in Bahrain has been investigated using the Spearman correlation coefficient for accounting for the non-linear relationship. Out of 15 attributes, seven were most correlated with PV power output. Three models were tested, and their performance was compared against the LSTM model: GRU Bi-LSTM and Bi-GRU. The comparison used a simple network structure, and the second used the hybrid form, including a CNN layer. To further test the model's ability to learn and accurately predict smaller historical data, a problem often associated with meteorological data, the five months of data have been divided into a smaller dataset that covers different seasonal times summer, which consists of two months and winter of three months.

After analyzing the results, it was confirmed that the other RNN variants outperformed LSTM in both the simple and hybrid structures across all data sets. It's interesting to note that although the LSTM gave the worst performance, adding the CNN layer significantly enhanced the LSTM model. Further, splitting the dataset enhanced the accuracy of CNN-Bi-LSTM and CNN-GRU when trained on summer data. In contrast, CNN-Bi-GRU's performance deteriorated when trained on fewer data. However, this finding was inconsistent when trained on the winter dataset due to the heavy presence of outliers.

In light of this, having good data quality that is clean of outliers or faulty equipment is crucial to conclude the performance of these models. Another limitation is the availability of historical data used in this research. More historical data would give more room to test the models' strengths and shortcomings on the requirements of historical data for an accurate prediction based on different scenarios of historical data. In addition, the time needed to train the model was another limitation to this research, as it restricted the chance to further investigate the model's performance before and after conducting the correlation analysis as another case study experiment. Table 13 Shows the time training for each model using the dataset after feature selection.

Table (15): Time recorded in minutes for the models to train, both vanilla and hybrid network

Model		Run-time (minutes)		
	#Neurons	32	62	128
Vanilla	LSTM	~40	~40	~240
	Bi-LSTM	~20	~30	~50
	GRU	~15	~15	~30
	Bi-GRU	~300	~1080	~1080
Hybrid	LSTM	~30	~50	~200
	Bi-LSTM	~40	~40	~100
	GRU	~60	~120	~120
	Bi-GRU	~800	~1200	~1200

Attribute selection is most suitable for creating models that are aimed to be easily interpreted; such algorithms are Least Square Regression and RF (Shmueli, 2010). On the other hand, algorithms such as k-nearest-neighbors and NNs are more concerned with predicting new observations than shedding light on underlying relationships. With this in mind, and the fact that NNs can capture complicated associations, perhaps eliminating the attributes might have removed valuable information that

the models could have captured and learned the patterns, thereby leading to better accuracy. As such, comparing the model's accuracy before and after feature selection to test the robustness of these models would be essential and can be tested in future work. It's worth mentioning that few models were trained on the dataset with full features; however, due to the very long time to train the model, this experiment was not run.

As Reindl et al., (2017) concludes that correctly converting energy into power is a challenge by itself and hence an important aspect to consider. This produces uncertainties that need to be investigated to not only predict accurately the PV power generation, but to better understand their characteristics. Perhaps accounting for the energy conversion efficiency as a variable or a factor would improve the accuracy of the model.

Another experiment could be done to further expand the work which predicts one day ahead with smaller resolutions, for example, 15 minutes. Since all datasets were of very high resolution, averaging 1 min recording of data into an hour might have affected the model's accuracy. In the future, one can replicate the study with higher resolution to predict day ahead PV power generation.

Lastly, continuing to explore the parameters of the model as what has been tested in this thesis was relatively simple and general. Adding more layers might also enhance the model's ability to learn and capture the complicated relations of the meteorological variables and thus give better results. Further, combining the Autoencoder or the Genetic Algorithm Optimizer will show promising results.

References

- Abdi, H. (2008). Kendall Rank Correlation Coefficient. *The Concise Encyclopedia of Statistics*, 278–281. https://doi.org/10.1007/978-0-387-32833-1_211
- Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. <https://doi.org/10.1109/ICMIRA.2013.45>
- Ahmad, T., Zhang, H., & Yan, B. (2020). A review on renewable energy and electricity requirement forecasting models for smart grid and buildings. *Sustainable Cities and Society*, 55(October 2019), 102052. <https://doi.org/10.1016/j.scs.2020.102052>
- Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124(March), 109792. <https://doi.org/10.1016/j.rser.2020.109792>
- Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Al-Badi, A., & AlMubarak, I. (2019). Growing energy demand in the GCC countries. *Arab Journal of Basic and Applied Sciences*, 26(1), 488–496. <https://doi.org/10.1080/25765299.2019.1687396>
- Al-Maamary, H. M. S., Kazem, H. A., & Chaichan, M. T. (2017). Climate change: The game changer in the Gulf Cooperation Council Region. *Renewable and Sustainable Energy Reviews*, 76(January 2016), 555–576. <https://doi.org/10.1016/j.rser.2017.03.048>
- Al-Sadah, F. H., & Ragab, F. M. (1991). Study of global daily solar

- radiation and its relation to sunshine duration in Bahrain. *Solar Energy (Journal of Solar Energy Science and Engineering); (United States)*, 47:2(2), 115–119. [https://doi.org/10.1016/0038-092X\(91\)90042-U](https://doi.org/10.1016/0038-092X(91)90042-U)
- Al-Sadah, F. H., Ragab, F. M., & Arshad, M. K. (1990). Hourly solar radiation over Bahrain. *Energy*, 15(5), 395–402. [https://doi.org/10.1016/0360-5442\(90\)90036-2](https://doi.org/10.1016/0360-5442(90)90036-2)
- Alhakeem, D., Mandal, P., Haque, A. U., Yona, A., Senjyu, T., & Tseng, T. L. (2015). A new strategy to quantify uncertainties of wavelet-GRNN-PSO based solar PV power forecasts using bootstrap confidence intervals. *IEEE Power and Energy Society General Meeting, 2015-Septe*, 0–4. <https://doi.org/10.1109/PESGM.2015.7286233>
- Alharbi, F. R., & Csala, D. (2020). GCC Countries' Renewable Energy Penetration and the Progress of Their Energy Sector Projects. *IEEE Access*, 8, 211986–212002. <https://doi.org/10.1109/ACCESS.2020.3039936>
- Alnaser, N.W., & Alnaser, W. E. (2019). The impact of the rise of using solar energy in GCC countries. *Renewable Energy and Environmental Sustainability*, 4, 7. <https://doi.org/10.1051/rees/2019004>
- Alnaser, Naser Waheeb, & Alnaser, W. E. (2019). Analyzing the Impact of Bapco 5 MW Solar PV Grid-Connected Project on Bahrain's Outlook for Energy-Mix Production. *International Journal of Modern Nonlinear Theory and Application*, 08(03), 72–91. <https://doi.org/10.4236/ijmnta.2019.83006>
- Alnaser, W. E. (1989). Empirical correlation for total and diffuse radiation in Bahrain. *Energy*, 14(7), 409–414. [https://doi.org/10.1016/0360-5442\(89\)90136-9](https://doi.org/10.1016/0360-5442(89)90136-9)
- Alnaser, W. E. (1993). New model to estimate the solar global irradiation using astronomical and meteorological parameters. *Renewable*

Energy, 3(2–3), 175–177. [https://doi.org/10.1016/0960-1481\(93\)90016-A](https://doi.org/10.1016/0960-1481(93)90016-A)

Alnaser, Waheeb Essa, & Alnaser, N. W. (2019). *Bahrain ' s BAPCO 5MWp PV Grid – Connected Solar Project. July*, 71–84.

Alomari, M. H., Younisy, O., & Hayajnehz, S. M. A. (2018). A predictive model for solar photovoltaic power using the Levenberg-Marquardt and Bayesian regularization algorithms and real-time weather data. *International Journal of Advanced Computer Science and Applications*, 9(1), 347–353. <https://doi.org/10.14569/IJACSA.2018.090148>

Alsammarraee, A., & Alshareeda, N. (2019). *A Proposed Program to audit the Sustainable Development Goals in the Kingdom of Bahrain A Proposed Program to audit the Sustainable Development Goals in the Kingdom of Bahrain. February*.

AlShafeey, M., & Csáki, C. (2021). Evaluating neural network and linear regression photovoltaic power forecasting models based on different input methods. *Energy Reports*, 7, 7601–7614. <https://doi.org/10.1016/j.egyr.2021.10.125>

AlSkaif, T., Dev, S., Visser, L., Hossari, M., & van Sark, W. (2020). A systematic analysis of meteorological variables for PV output power estimation. *Renewable Energy*, 153, 12–22. <https://doi.org/10.1016/j.renene.2020.01.150>

Arabia, S., & Town, I. (1993). *Global Solar Radiation With Meteorological Data for Bahrain*. 34(1), 1–6.

Aslam, M., Lee, S. J., Khang, S. H., & Hong, S. (2021). Two-Stage Attention over LSTM with Bayesian Optimization for Day-Ahead Solar Power Forecasting. *IEEE Access*, 9, 107387–107398. <https://doi.org/10.1109/ACCESS.2021.3100105>

Bexell, M., & Jönsson, K. (2017). Responsibility and the United Nations'

- Sustainable Development Goals. *Forum for Development Studies*, 44(1), 13–29. <https://doi.org/10.1080/08039410.2016.1252424>
- Bin Shams, M., Haji, S., Salman, A., Abdali, H., & Alsaffar, A. (2016). Time series analysis of Bahrain's first hybrid renewable energy system. *Energy*, 103, 1–15. <https://doi.org/10.1016/j.energy.2016.02.136>
- Blessie, E. C., & Karthikeyan, E. (2012). Sigmis: A feature selection algorithm using correlation based method. *Journal of Algorithms and Computational Technology*, 6(3), 385–394. <https://doi.org/10.1260/1748-3018.6.3.385>
- Busari, G. A., & Lim, D. H. (2021). Crude oil price prediction: A comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. *Computers and Chemical Engineering*, 155, 107513. <https://doi.org/10.1016/j.compchemeng.2021.107513>
- Cang, S., & Seetaram, N. (2012). Time series analysis. *Handbook of Research Methods in Tourism: Quantitative and Qualitative Approaches*, June, 47–70. <https://doi.org/10.4337/9781781001295.00009>
- Carrera, B., & Kim, K. (2020). Comparison analysis of machine learning techniques for photovoltaic prediction using weather sensor data. *Sensors (Switzerland)*, 20(11). <https://doi.org/10.3390/s20113129>
- Casaca de Rocha Vaz, A. G. (2014). Photovoltaic Forecasting with Artificial Neural Network. *FACULDADE DE CIÊNCIAS\DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA*, 86. http://repositorio.ul.pt/bitstream/10451/11405/1/ulfc107351_tm_Andre_Vaz.pdf
- Chen, C., Duan, S., Cai, T., & Liu, B. (2011). Online 24-h solar power

- forecasting based on weather type classification using artificial neural network. *Solar Energy*, 85(11), 2856–2870. <https://doi.org/10.1016/j.solener.2011.08.027>
- Cui, Z., Ke, R., Pu, Z., & Wang, Y. (2018). *Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction*. 1–11. <http://arxiv.org/abs/1801.02143>
- Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., Horan, B., & Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, 81(June 2017), 912–928. <https://doi.org/10.1016/j.rser.2017.08.017>
- Directorate of Environmental Assessment and Planning. (2009). State of the Environment in the Kingdom of Bahrain. *Public Commission for the Protection of Marine Resources, Environment and Wildlife*, 1–131. <https://www.sce.gov.bh/en/Media/Downloads/reports/pdf/11> Eng Final Version of the report.pdf
- Ela, S., Grotendorst, T., Ehrler, B., & Futscher, M. (2018). *The Correlation between Temperature and Properties of Solar Panels*.
- Elgabiri, M., Palmer, D., Al Buflasa, H., & Thomson, M. (2021). Offshore wind energy potential for Bahrain via multi-criteria evaluation. *Wind Engineering*, 45(4), 838–856. <https://doi.org/10.1177/0309524X20925399>
- Fredricks, G. A., & Nelsen, R. B. (2007). On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. *Journal of Statistical Planning and Inference*, 137(7), 2143–2150. <https://doi.org/10.1016/j.jspi.2006.06.045>
- Frroukhi, R., Khalid, A., Diala, H., Divyam, N., El-Katiri, L., Fthenakis, V., & Al-Fara, A. (2016). Renewable Energy Market Analysis The GCC Region. *Irena*, 1–96.

https://www.irena.org/DocumentDownloads/Publications/IRENA_Market_GCC_2016.pdf

- Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164–181. <https://doi.org/10.1016/j.engappai.2010.09.007>
- Gensler, A., Henze, J., Sick, B., & Raabe, N. (2017). Deep Learning for solar power forecasting - An approach using AutoEncoder and LSTM Neural Networks. *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings*, 2858–2865. <https://doi.org/10.1109/SMC.2016.7844673>
- Herzog, A. V, Lipman, T. E., & Kammen, D. M. (2001). Renewable energy sources: A variable choice. *Environment: Science and Policy for Sustainable Development*, 8–20.
- Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1), 388–427. <https://doi.org/10.1016/j.ijforecast.2020.06.008>
- Huang, Z., Yang, F., Xu, F., Song, X., & Tsui, K. L. (2019). Convolutional Gated Recurrent Unit-Recurrent Neural Network for State-of-Charge Estimation of Lithium-Ion Batteries. *IEEE Access*, 7, 93139–93149. <https://doi.org/10.1109/ACCESS.2019.2928037>
- Isaksson, E., & Conde, M. K. (2018). *Solar Power Forecasting with Machine Learning Techniques*. 46.
- Jurečková, J., & Picek, J. (2007). Shapiro-Wilk-type test of normality under nuisance regression and scale. *Computational Statistics and Data Analysis*, 51(10), 5184–5191. <https://doi.org/10.1016/j.csda.2006.08.026>
- Karaca, M., Alvarado, M. M., Gahrooei, M. R., Bihorac, A., & Pardalos, P. M. (2022). Frequent pattern mining from multivariate time series

- data. *Expert Systems with Applications*, 194(January), 116435.
<https://doi.org/10.1016/j.eswa.2021.116435>
- KAVAKÇI, G. (2021). *TIME SERIES PREDICTION OF SOLAR POWER GENERATION USING*. February.
- Lai, J. P., Chang, Y. M., Chen, C. H., & Pai, P. F. (2020). A survey of machine learning models in renewable energy predictions. *Applied Sciences (Switzerland)*, 10(17). <https://doi.org/10.3390/app10175975>
- Li, P., Luo, A., Liu, J., Wang, Y., Zhu, J., Deng, Y., & Zhang, J. (2020). Bidirectional gated recurrent unit neural network for Chinese address element segmentation. *ISPRS International Journal of Geo-Information*, 9(11). <https://doi.org/10.3390/ijgi9110635>
- Louzazni, M., Mosalam, H., & Khouya, A. (2020). A non-linear autoregressive exogenous method to forecast the photovoltaic power output. *Sustainable Energy Technologies and Assessments*, 38(December 2019), 100670.
<https://doi.org/10.1016/j.seta.2020.100670>
- Malik, M. M. (2020). *A Hierarchy of Limitations in Machine Learning*. 1–68. <http://arxiv.org/abs/2002.05193>
- Mateus, B. C., Mendes, M., Farinha, J. T., Assis, R., & Cardoso, A. M. (2021). Comparing LSTM and GRU models to predict the condition of a pulp paper press. *Energies*, 14(21), 1–21.
<https://doi.org/10.3390/en14216958>
- Meng, S., Jiang, X. Q., Gao, Y., Hai, H., & Hou, J. (2020). Performance Evaluation of Channel Decoder based on Recurrent Neural Network. *Journal of Physics: Conference Series*, 1438(1).
<https://doi.org/10.1088/1742-6596/1438/1/012001>
- Mohammadi, K., Shamshirband, S., Tong, C. W., Arif, M., Petković, D., & Sudheer, C. (2015). A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation.

Energy Conversion and Management, 92, 162–171.
<https://doi.org/10.1016/j.enconman.2014.12.050>

Monteiro, C., Fernandez-Jimenez, L. A., Ramirez-Rosado, I. J., Muñoz-Jimenez, A., & Lara-Santillan, P. M. (2013). Short-term forecasting models for photovoltaic plants: Analytical versus soft-computing techniques. *Mathematical Problems in Engineering*, 2013. <https://doi.org/10.1155/2013/767284>

Mukaka, M. M. (2012). Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. In *Malawi Medical Journal* (Vol. 24, Issue 3). www.mmj.medcol.mw

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 1–21. <https://doi.org/10.1186/s40537-014-0007-7>

Najibi, F., Apostolopoulou, D., & Alonso, E. (2021). Enhanced performance Gaussian process regression for probabilistic short-term solar output forecast. *International Journal of Electrical Power and Energy Systems*, 130(February), 106916. <https://doi.org/10.1016/j.ijepes.2021.106916>

Nguyen, N. Q., Bui, L. D., Doan, B. Van, Sanseverino, E. R., Cara, D. Di, & Nguyen, Q. D. (2021). A new method for forecasting energy output of a large-scale solar power plant based on long short-term memory networks a case study in Vietnam. *Electric Power Systems Research*, 199(January), 107427. <https://doi.org/10.1016/j.epsr.2021.107427>

Nwaigwe, K. N., Mutabilwa, P., & Dintwa, E. (2019). An overview of solar power (PV systems) integration into electricity grids. *Materials Science for Energy Technologies*, 2(3), 629–633. <https://doi.org/10.1016/j.mset.2019.07.002>

Poudel, P., & Jang, B. (2017). *Solar Power Prediction Using Deep*

Learning Technique. November, 148–151.
<https://doi.org/10.14257/astl.2017.146.26>

Prema, V., & Uma Rao, K. (2015). Development of statistical time series models for solar power prediction. *Renewable Energy*, 83, 100–109.
<https://doi.org/10.1016/j.renene.2015.03.038>

Puth, M. T., Neuhäuser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, 102, 77–84.
<https://doi.org/10.1016/j.anbehav.2015.01.010>

Radhi, H. (2009). A comparison of the accuracy of building energy analysis in Bahrain using data from different weather periods. *Renewable Energy*, 34(3), 869–875.
<https://doi.org/10.1016/j.renene.2008.06.008>

Rahman, S. M., Pawar, S., San, O., Rasheed, A., & Iliescu, T. (2019). Nonintrusive reduced order modeling framework for quasigeostrophic turbulence. *Physical Review E*, 100(5).
<https://doi.org/10.1103/PhysRevE.100.053306>

Ravaee, H., Farahat, S., & Sarhaddi, F. (2012). Artificial Neural Network Based Model Of Photovoltaic Thermal (pvt) Collector. *Journal of Mathematics and Computer Science*, 04(03), 411–417.
<https://doi.org/10.22436/jmcs.04.03.15>

Raza, M. Q., Nadarajah, M., & Ekanayake, C. (2016). On recent advances in PV output power forecast. *Solar Energy*, 136, 125–144.
<https://doi.org/10.1016/j.solener.2016.06.073>

Reikard, G. (2009). Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, 83(3), 342–349.
<https://doi.org/10.1016/j.solener.2008.08.007>

Reindl, T., Walsh, W., Yanqin, Z., & Bieri, M. (2017). Energy meteorology for accurate forecasting of PV power output on different time horizons.

- Energy Procedia*, 130, 130–138.
<https://doi.org/10.1016/j.egypro.2017.09.415>
- Rhodes, C. J. (2010). Solar energy: Principles and possibilities. *Science Progress*, 93(1), 37–112.
<https://doi.org/10.3184/003685010X12626410325807>
- Rogier, J. K., & Mohamudally, N. (2019). Forecasting photovoltaic power generation via an IoT network using nonlinear autoregressive neural network. *Procedia Computer Science*, 151(2018), 643–650.
<https://doi.org/10.1016/j.procs.2019.04.086>
- Santra, A. S., & Lin, J. L. (2019). Integrating long short-term memory and genetic algorithm for short-term load forecasting. *Energies*, 12(11), 1–11. <https://doi.org/10.3390/en12112040>
- Schmidt, J., Marques, M. R. G., Botti, S., & Marques, M. A. L. (2019). Recent advances and applications of machine learning in solid-state materials science. *Npj Computational Materials*, 5(1).
<https://doi.org/10.1038/s41524-019-0221-0>
- SEU TEAM. (2017). *The Kingdom of Bahrain National Renewable Energy Action Plan (NREAP)*. 76. http://www.seu.gov.bh/wp-content/uploads/2018/04/02_NREAP-Full-Report.pdf
- Sharma, A., Dharwal, M., & Kumari, T. (2021). Materials Today : Proceedings Renewable energy for sustainable development : A comparative study of india and china. *Materials Today: Proceedings*, xxxx. <https://doi.org/10.1016/j.matpr.2021.09.242>
- Shboul, B., AL-Arfi, I., Michailos, S., Ingham, D., Ma, L., Hughes, K. J., & Pourkashanian, M. (2021). A new ANN model for hourly solar radiation and wind speed prediction: A case study over the north & south of the Arabian Peninsula. *Sustainable Energy Technologies and Assessments*, 46(February), 101248.
<https://doi.org/10.1016/j.seta.2021.101248>

- Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, 132306. <https://doi.org/10.1016/j.physd.2019.132306>
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>
- Singh, G. K. (2013). Solar power generation by PV (photovoltaic) technology: A review. *Energy*, 53, 1–13. <https://doi.org/10.1016/j.energy.2013.02.057>
- Soares, E., Costa, P., Costa, B., & Leite, D. (2018). Ensemble of evolving data clouds and fuzzy models for weather time series prediction. *Applied Soft Computing Journal*, 64, 445–453. <https://doi.org/10.1016/j.asoc.2017.12.032>
- Son, N., & Jung, M. (2021). Analysis of meteorological factor multivariate models for medium- and long-term photovoltaic solar power forecasting using long short-term memory. *Applied Sciences (Switzerland)*, 11(1), 1–20. <https://doi.org/10.3390/app11010316>
- Stanojevic, A. B. (2021, July). *Application of photovoltaic technology in the use of solar energy*. <https://www.peertechzpublications.com/articles/AEST-5-144.php>
- Sun, Ying. (2020). Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems. In *Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems*. <https://doi.org/10.5772/intechopen.85999>
- Sun, Yujing, Wang, F., Wang, B., Chen, Q., Engerer, N. A., & Mi, Z. (2017). Correlation feature selection and mutual information theory based quantitative research on meteorological impact factors of

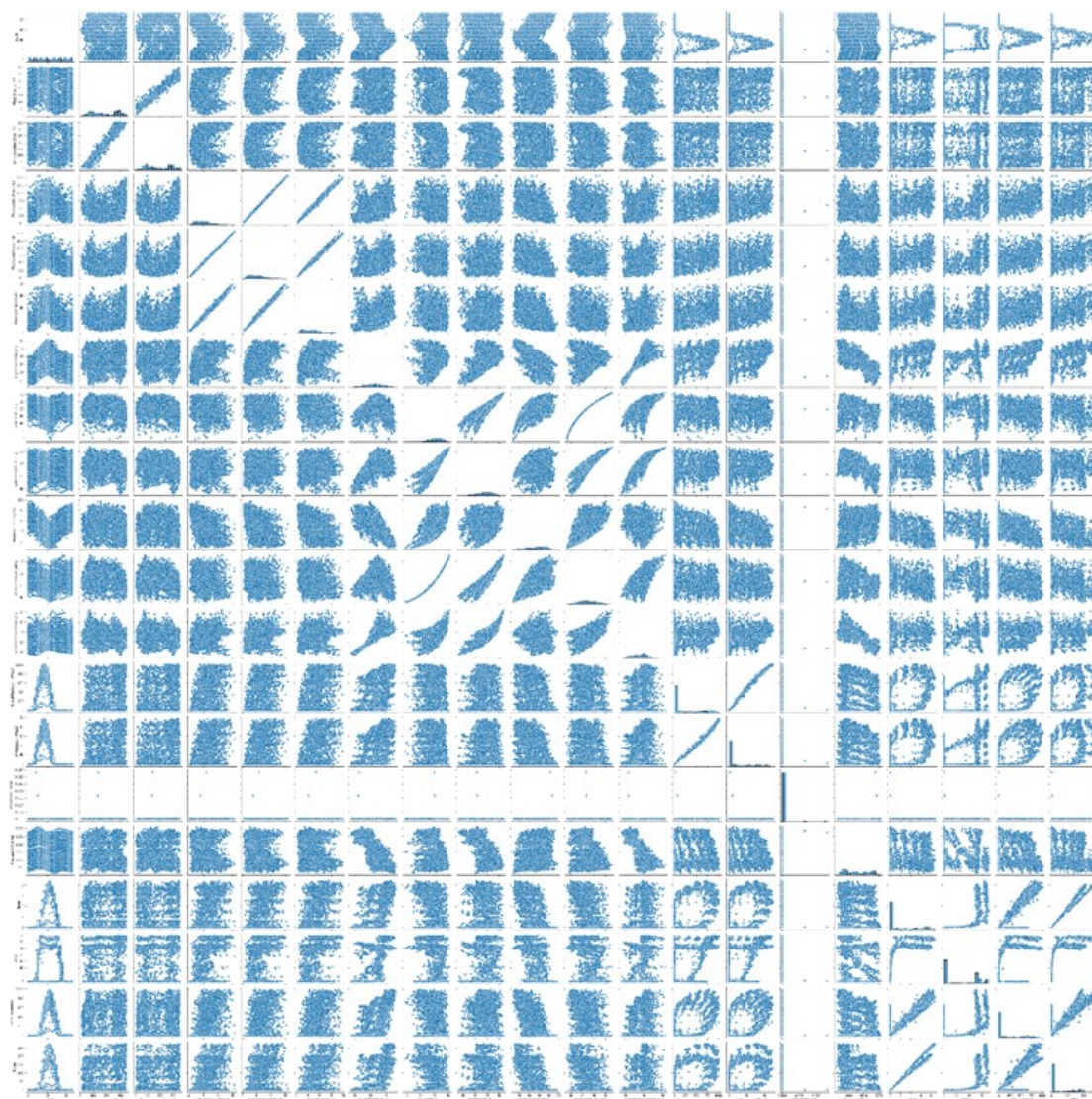
- module temperature for solar photovoltaic systems. *Energies*, 10(1).
<https://doi.org/10.3390/en10010007>
- Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language processing. *Interspeech 2012*, 194–197.
https://www.isca-speech.org/archive/interspeech_2012/i12_0194.html
- Torino, P. D. I. (2020). *Department of Control and Computer Engineering Master Thesis Robotic arm pick-and-place tasks. October.*
- Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., & Troncoso, A. (2021). Deep Learning for Time Series Forecasting: A Survey. *Big Data*, 9(1), 3–21. <https://doi.org/10.1089/big.2020.0159>
- Wang, F., Zhang, Z., Liu, C., Yu, Y., Pang, S., Duić, N., Shafie-khah, M., & Catalão, J. P. S. (2019). Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting. *Energy Conversion and Management*, 181(August 2018), 443–462.
<https://doi.org/10.1016/j.enconman.2018.11.074>
- Wang, F., Zhen, Z., Mi, Z., Sun, H., Su, S., & Yang, G. (2015). Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting. *Energy and Buildings*, 86, 427–438.
<https://doi.org/10.1016/j.enbuild.2014.10.002>
- Wang, F., Zhen, Z., Wang, B., & Mi, Z. (2017). Comparative study on KNN and SVM based weather classification models for day ahead short term solar PV power forecasting. *Applied Sciences (Switzerland)*, 8(1). <https://doi.org/10.3390/app8010028>
- Wang, Z. (2019). *SOLAR POWER FORECASTING* [The University of Sydney].
https://ses.library.usyd.edu.au/bitstream/handle/2123/21248/zheng_z

w_thesis.pdf?sequence=2&isAllowed=y

- Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2021). *Time Series Data Augmentation for Deep Learning: A Survey*. 4653–4660. <https://doi.org/10.24963/ijcai.2021/631>
- Wolff, B., Sauer, J., Kramer, O., & Parisi, J. (2017). *Support Vector Regression for Solar Power Prediction*. April.
- Zabihian, F. (2021). Solar Energy and Photovoltaic Solar Units. *Power Plant Engineering*, June, 759–904. <https://doi.org/10.1201/9780429069451-11>
- Zanero, S., & Savaresi, S. M. (2004). Unsupervised learning techniques for an intrusion detection system. *Proceedings of the ACM Symposium on Applied Computing*, 1, 412–419. <https://doi.org/10.1145/967900.967988>
- Zeman, M. (2014). *Solar Cells Theory*.
- Zheng, J., Zhang, H., Dai, Y., Wang, B., Zheng, T., Liao, Q., Liang, Y., Zhang, F., & Song, X. (2020). Time series prediction for output of multi-region solar power plants. *Applied Energy*, 257(June 2019), 114001. <https://doi.org/10.1016/j.apenergy.2019.114001>
- Ziane, A., Necaibia, A., Sahouane, N., Dabou, R., Mostefaoui, M., Bouraiou, A., Khelifi, S., Rouabhia, A., & Blal, M. (2021). Photovoltaic output power performance assessment and forecasting: Impact of meteorological variables. *Solar Energy*, 220(March), 745–757. <https://doi.org/10.1016/j.solener.2021.04.004>
- Zo, I., Jee, J., Kim, B., & Lee, K. (2017). Analysis of the thermal dome effect from global solar radiation observed with a modified pyranometer. *Current Optics and Photonics*, 1(4), 263–270. <https://doi.org/10.3807/COPP.2017.1.4.263>

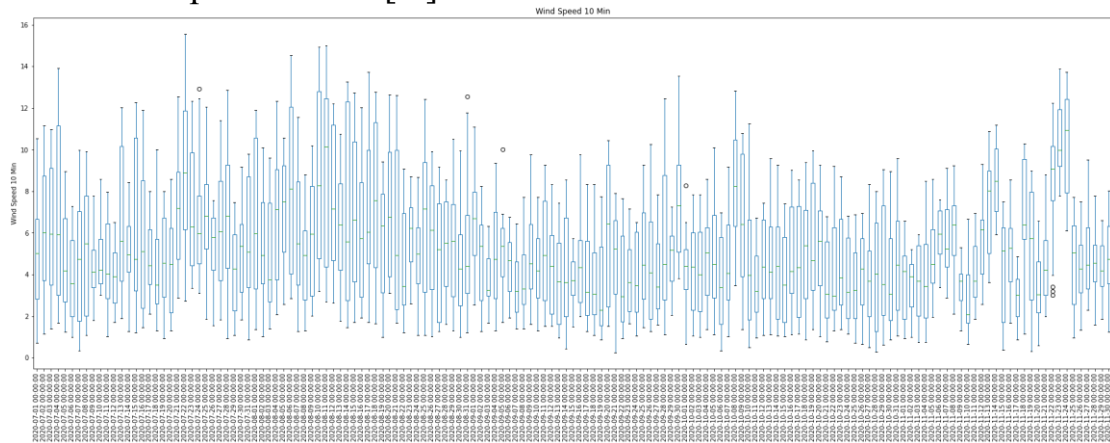
Appendices

Pair of plot of all meteorological attributes

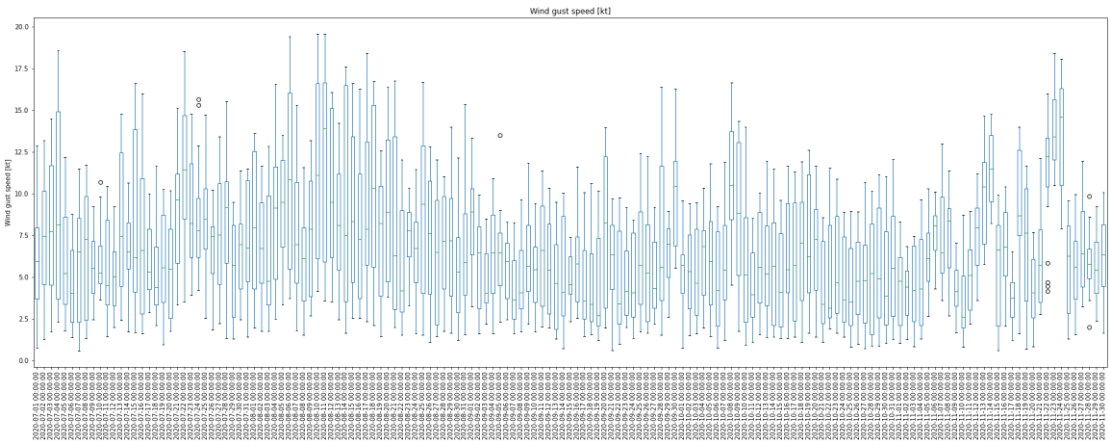


Box plot for all 8 meteorological attributes

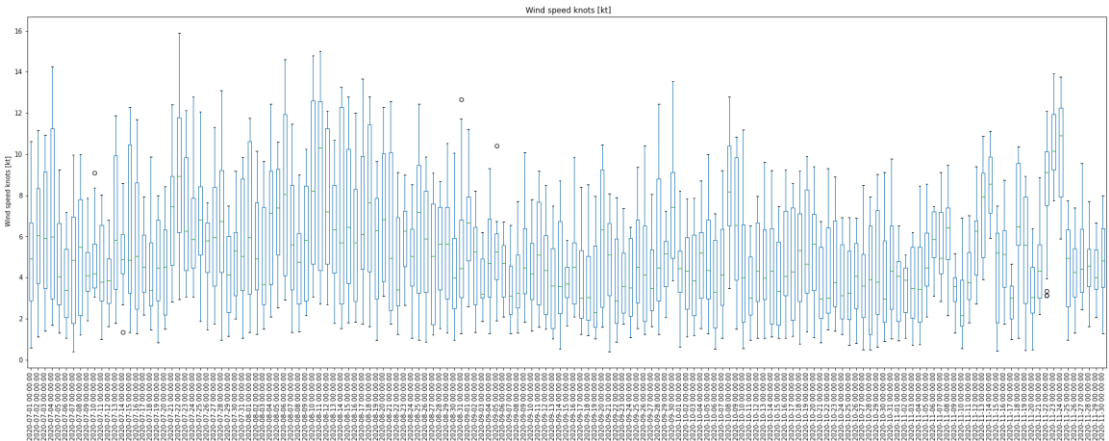
Wind speed 10 min [kt]



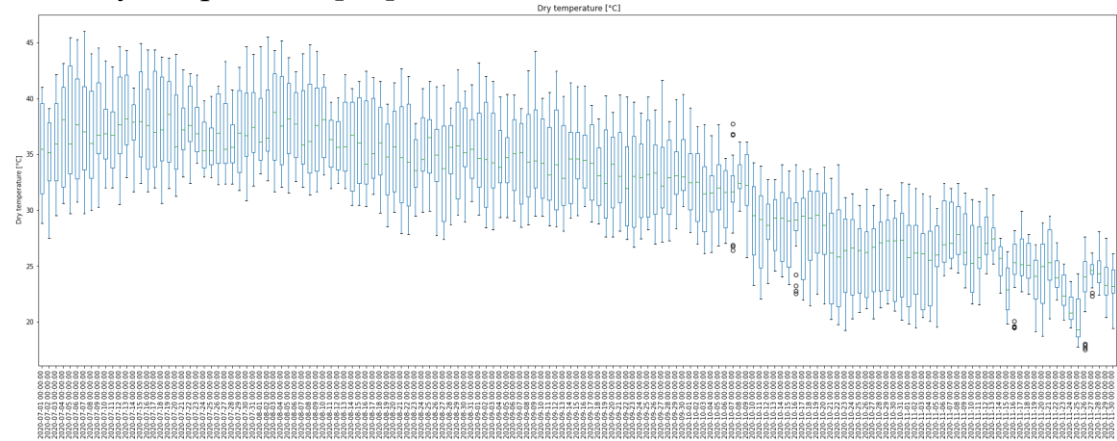
Wind speed knots [kt]



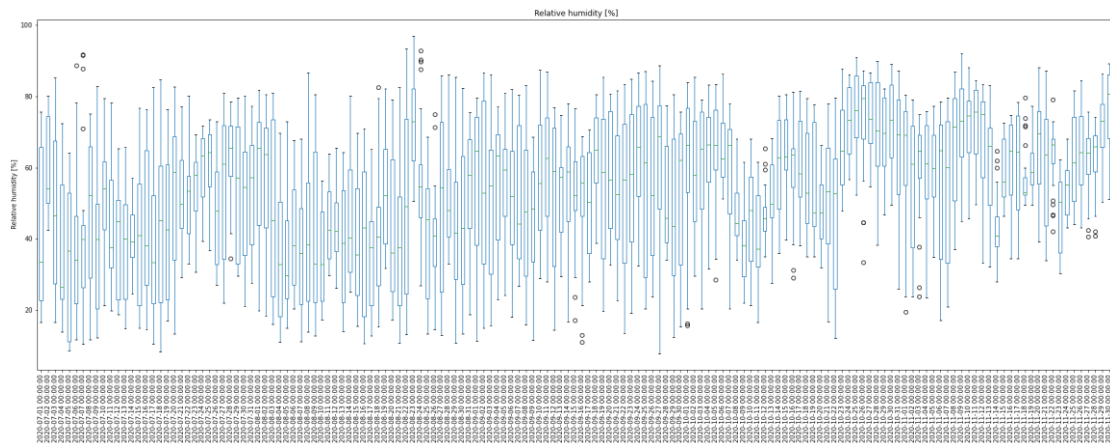
Wind gust speed [kt]



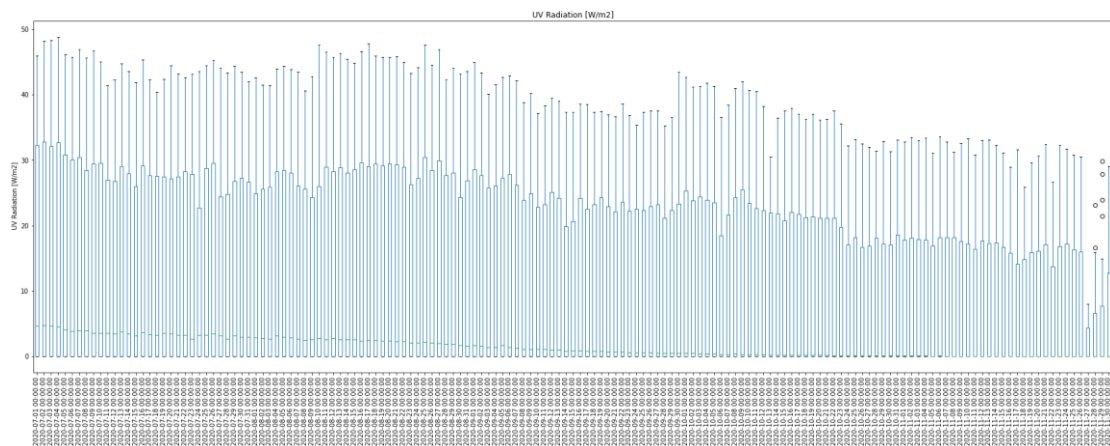
Dry temperature [°C]



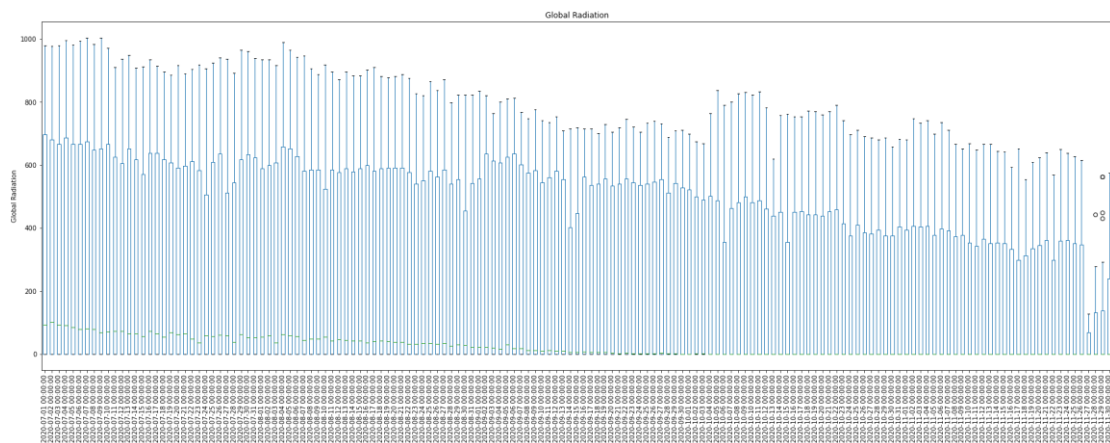
Relative humidity [%]



UV Radiation [W/m2]



Global Radiation



الملخص

نظرًا لارتفاع الكبير في الطلب على الطاقة في جميع البلدان، فقد زاد الإقبال على توليد الطاقة الكهروضوئية بشكل كبير وتنامي الطلب نظرًا لأنه يوفر طاقة نظيفة ويمكن الوصول إليها. ومع ذلك، فإن التحدي المتمثل في الطاقة الشمسية يكمن في تقلبها في التوليد لأنها تعتمد بشكل كبير على متغيرات الطقس. وبالتالي، فإن التنبؤ بتوليد الطاقة الكهروضوئية قبل يوم أمر ضروري لجدولة أنظمة الشبكة الذكية وتشغيلها لدى موفري الطاقة الشمسية. تهدف هذه الرسالة إلى تحديد العلاقة بين متغيرات الأرصاد الجوية التي تؤثر في الغالب على توليد الطاقة الشمسية في مملكة البحرين. ثانيًا، التحقق من تدريب خوارزميات التعلم الآلي التي لم يتم اختبارها في الأدبيات السابقة للتنبؤ بإنتاج الطاقة الكهروضوئية، وهي GRU و Bi-LSTM و Bi-GRU. في هذه الدراسة، تم استخدام معامل سبيرمان للارتباط لقياس واكتشاف الأنماط المحتملة للبيانات بشكل أفضل بين الطاقة الكهروضوئية ومتغيرات الطقس. لخوارزميات التعلم الآلي، تم استخدام بيئة بسيطة، والآخر معقد أي يحتوي على طبقة واحدة من CNN، وتم اختبار النماذج للخوارزميات ومقارنتها مع LSTM. لاختبار أداء النماذج، تم إجراء تجربة أخرى لتدريب النماذج على بيانات موسم محدد من الطقس، التي تحتوي على بيانات صيفية والآخر شتوي، والتي تم تقسيمها من بيانات الأشهر الخمسة التي تم جمعها من حرم جامعة البحرين. عبر جميع مجموعات البيانات، أعطى LSTM أدنى أداء على عكس النماذج الأخرى. عزز الشكل المعقد الأداء الأكثر في LSTM عن باقي النماذج الأخرى ومع ذلك لا يزال ينتج أعلى RMSE و MAE. على الرغم من أن النتائج كانت قابلة للمقارنة بين GRU و Bi-LSTM و Bi-GRU، كلاهما هيكل المعقد وفانيليا، فإن أفضل نموذج أعطى أفضل النتائج لمجموعة البيانات الكاملة كان CNN-Bi-GRU مع RMSE 20.74 و MAE 9.64. بالنسبة لبيانات الصيف والشتاء، أعطت CNN-Bi-LSTM أفضل أداء لكليهما مع RMSE بقيمة 8.17 و 21.72 و MAE من 4.23 و 10.32 على التوالي.

جامعة البحرين

قسم برامج الدراسات العليا

كلية العلوم



التنبؤ القصير المدى لمخرجات الطاقة الكهروضوئية باستخدام

خوارزميات التعلم الآلي

إطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير
في علوم وتحليل البيانات الضخمة

إعداد

هاجر البحري

20134920

إشراف

د. هشام العمّال

استاذ مساعد

جامعة البحرين

مملكة البحرين

مايو 2022