POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Executive Summary of the Thesis

Solar PV Power Forecasting Using Machine Learning

Laurea Magistrale in Electrical Engineering - Ingegneria Elettrica

Author: Saloni Dhingra

Advisor: Prof. Giancarlo Storti Gajani

Academic year: 2022-2023

1. Introduction

Amid concerns about greenhouse gas emissions and environmental pollution caused by excessive fossil fuel consumption, there is a growing focus on utilizing renewable energy sources for power generation. Solar photovoltaics (PV) has emerged as a prominent option due to its rapid growth and potential. However, the power output of PV panels is heavily influenced by meteorological factors like solar irradiance, air temperature, and relative humidity. Consequently, the power generated by PV panels exhibits variability based on these parameters. Therefore, accurate solar PV power prediction has become critical for ensuring reliable and cost-effective grid operation, facilitating the deployment of large-scale PV plants, and optimizing the performance of solar power systems [8]. This study demonstrates the potential of using machine learning techniques for accurate and reliable solar PV power forecasting.

2. Data Exploration and Pre-Processing

The data are sourced from the National Renewable Energy Laboratory Photovoltaic Data Acquisition (NREL PVDAQ), which is a large-scale time-series database containing system metadata and performance data from a variety of experimental PV sites and commercial public PV sites [5]. Photovoltaic field array data are made up of time-series, raw performance data collected by a number of sensors linked to a PV system. Two datasets are utilized - One ranging from 2013 to 2018 with one-minute sampling interval and the other from 2011 to 2019 with 15-minute sampling interval.

The quality of input data is crucial for accurate and reliable forecasting. The data may contain intermittent static or spike elements caused by weather or seasonal variations, electricity demand fluctuations and power system failures. Moreover, data may also sometimes be corrupted or missing due to sensor defects or erroneous recordings. Therefore, it is imperative to pre-process distorted input data by reconstruction using decomposition, interpolation or seasonal adjustments(i.e. data cleansing and structure change).

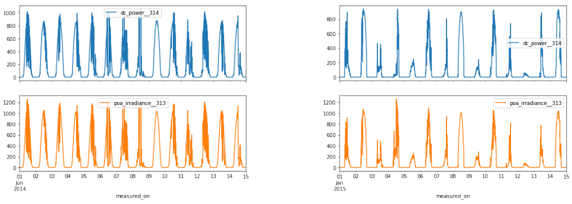
Data from each year is considered to transform the dataset into a usable format. To address discrepancies in the data, the datasets corresponding to each year are combined, and the unnecessary columns are removed. To further sanitize the data, negative solar irradiance and missing associated power values due to solar irradiance sensors offset and inverter failures, respectively are set to zero.

Table 1 shows the statistical measures used to better comprehend the data.

	DC Power	Irradiance
Count	349275.0000	349275.0000
Mean	193.6086	239.3601
Std	295.8135	355.2808
Min	0.0000	0.0000
25%	0.0168	0.0000
50%	7.3763	13.4476
75%	310.2527	393.8758
Max	1184.8890	1442.0380

Table 1: Statistical Measures, Year 2018

Figure 1 depicts the Temporal progression of generated power and solar irradiance. Power and solar irradiance have comparable trends, with higher values in the summer and lower values in the winter, as expected.



(a) Power and Irradiance (June 2014)

(b) Power and Irradiance (January 2015)

Figure 1: Temporal Progression of generated power and solar irradiance

3. Evaluation

The Evaluation metrics employed in this study are presented in this section.

3.1. Prediction Horizons

The prediction horizon must be specified in order to select an appropriate approach. The prediction horizon is the period of time in the future for which PV output power is forecasted. Based on specified prediction horizons, statistical methods in particular Artificial Neural Networks are selected for PV Power prediction tasks [1]. Applications of different prediction horizons are depicted in Figure 2.

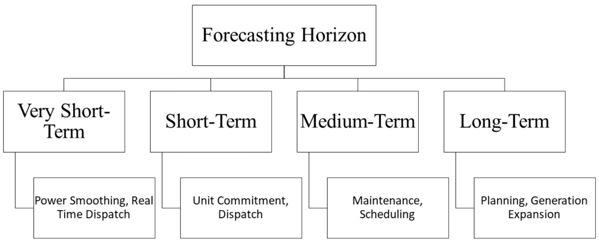

Forecasting Horizon
Very Short-Term
Short-Term
Medium-Term
Long-Term
Power Smoothing, Real Time Dispatch
Unit Commitment, Dispatch
Maintenance, Scheduling
Planning, Generation Expansion

Figure 2: Different prediction horizons and their applications

3.2. Performance Metrics

Performance metrics are statistical measures used to evaluate the quality and accuracy of a model. They provide a way to determine the effectiveness of a model by comparing its predictions to actual results. In this case, following performance metrics are used:

- **Mean Squared Error (MSE)**

Mean Squared Error (MSE) is a commonly used loss function in supervised machine learning, which is defined as the average of the squared differences between the predicted values and the true values as calculated in Equation 1.
Supposing power time series is $Y_i = Y_1, Y_2, ..., Y_n$ and $\hat{Y_i}$ is the predicted time series, and i indicates time.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2 \tag{1}$$

where n is the number of instances in the dataset and the $\sum$ symbol represents the sum over all instances. The MSE is commonly used as it is easy to compute and differentiable, making it suitable for optimization with gradient-based methods.

- **Mean Absolute Error (MAE)**

Mean Absolute Error (MAE) is a loss function used in regression problems, where the goal is to predict a continuous value output. It is calculated as the average of the absolute differences between the true values and the predicted values as shown in Equation 2.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \tag{2}$$

where y_i is the prediction and x_i is true value.

- **Root Mean Squared Error (RMSE)**

Root Mean Squared Error (RMSE) is calculated as the square root of the average of the squared differences between the actual and predicted values as represented in Equation 3.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}} \tag{3}$$

where $y_i = y_1, y_2, ..., y_n$ are observed values, $\hat{y_i}$ is a predicted time series and i indicates time.

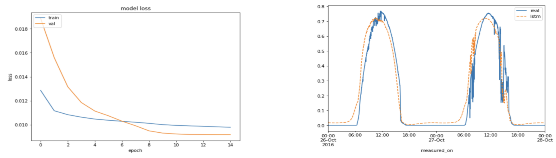
3.3. Hyperparameters Tuning

Hyperparameter tuning is the process of finding the optimal set of hyperparameters for a machine learning model to achieve the best performance on a given dataset. Learning rate, regularization strength, and the number of hidden layers in a neural network are all examples of hyperparameters [3].

4. Forecasting Results

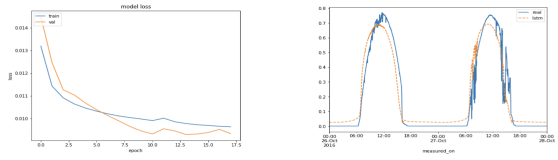
4.1. Short-Term Prediction with Hyperparameters Tuning

Figures 3 - 7 show the outcomes of short-term prediction with one-hour prediction horizon and using the dataset with one-minute sampling interval for the LSTM architecture.



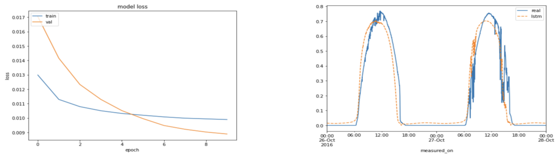
(a) Learning Curve (b) Simulation Results

Figure 3: Learning and simulation results, Learning rate = 0.0001, Epochs = 15



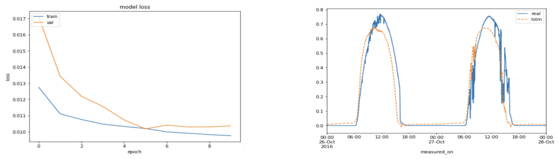
(a) Learning Curve (b) Simulation Results

Figure 4: Learning and simulation results, Learning rate = 0.001, Epochs = 18



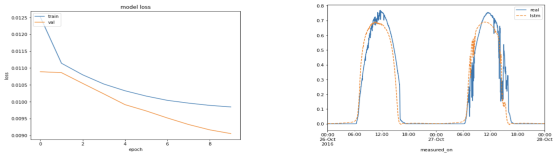
(a) Learning Curve (b) Simulation Results

Figure 5: Learning and simulation results, Learning rate = 0.01, Epochs = 10



(a) Learning Curve (b) Simulation Results

Figure 6: Learning and simulation results, Learning rate = 0.001 for epoch ≤ 6; Learning rate = 0.0001 for epoch > 6, Total Epochs = 10



(a) Learning Curve (b) Simulation Results

Figure 7: Learning and simulation results, Exponentially Decaying Learning rate; Initial value = 0.01; Decay steps = 1000 and Decay rate = 0.9, Epochs = 10

Table 2 summarizes short-term prediction results with hyperparameter adjustment for the LSTM architecture.

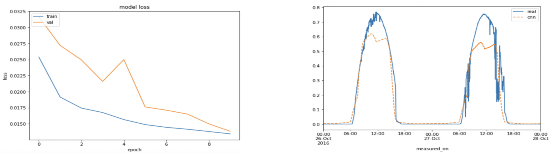
L.Rate	Epochs	MSE	MAE	RMSE
0.0001	15	0.0099	0.0512	0.0997
0.001	18	0.0100	0.0528	0.1002
0.01	10	0.0100	0.0521	0.1000
Varying	10	0.0099	0.0500	0.0995
Decaying	10	0.0100	0.0510	0.1002

Table 2: Short-Term Prediction Results

Based on the findings presented in this table, it is evident that the performance of a machine learning model can be significantly influenced by various combinations of learning rates. Notably, employing lower and varying learning rates tends to yield superior performance on the test dataset, as demonstrated by decreased Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values.

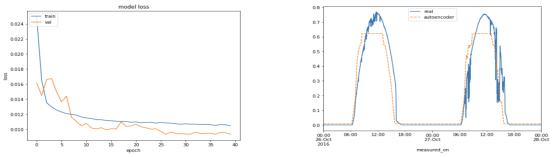
4.2. Comparison Between Different Neural Networks

A brief comparison between four popular neural network architectures [4]: convolutional neural networks, autoencoders, long-short term memory and gated-recurrent unit are shown in Figures 8 - 11.



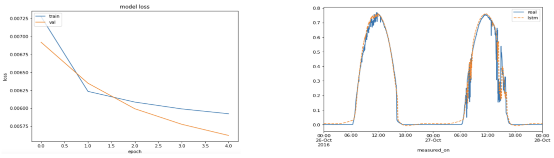
(a) Learning Curve (b) Simulation Results

Figure 8: Convolutional Neural Network (CNN)



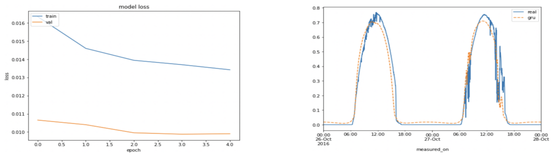
(a) Learning Curve (b) Simulation Results

Figure 9: Autoencoders



(a) Learning Curve (b) Simulation Results

Figure 10: Long-Short Term Memory (LSTM)



(a) Learning Curve (b) Simulation Results

Figure 11: Gated-Recurrent Unit (GRU)

When comparing different neural networks for prediction tasks, several factors need to be considered, including the architecture, training process, and performance metrics. The choice of neural network depends on the specific task at hand, the nature of the data, and the available resources. Experimentation and empirical evaluation are necessary to determine the most suitable neural network architecture for a given prediction problem. Table 3 summarizes results for short-time prediction to compare different neural networks considering window length of 12 hours and prediction horizon of 1 hour in each case.

Network	MSE	MAE	RMSE
CNN	0.0134	0.0649	0.1174
Autoencoders	0.0109	0.0534	0.1162
LSTM	0.0101	0.0463	0.1025
GRU	0.0104	0.0498	0.1159

Table 3: Comparison of Different Networks

Overall, **LSTM** and **GRU** models excel in capturing long- term dependencies in time series data, making them effective for prediction tasks but they typically require more processing power compared to Autoencoder and CNN models [6]. The complex architecture and memory-intensive operations of LSTMs and GRUs can increase the computational demands during training and inference.

Autoencoder models are useful for feature extraction and anomaly detection, but their performance may vary depending on the dataset but they have lower processing power requirements since they involve simpler neural network structures.

While **CNN** models are powerful for spatial data analysis, they may not be as effective for time series prediction due to their limited ability to capture long-term dependencies and also, they can be computationally demanding, especially when dealing with high-dimensional time series data. Therefore, it is important to consider the available processing power when selecting the appropriate model for time series prediction tasks.

5. Ageing

Solar panels are known to degrade over time due to exposure to the environment, temperature variations, and other factors. This degradation is commonly referred to as "ageing" and can have a significant impact on the power output of the solar panels. As solar panels age, their efficiency in converting sunlight into electricity decreases gradually. In this study, two methods are used to investigate the effect of ageing using the dataset with 15-minutes sampling interval.

5.1. Predictive Analysis

LSTM is used to train the model to observe the effect of ageing on solar PV power output as LSTM can model complex non-linear relationships between input and output signals, allowing it to capture the complex dynamics of the signal in panel ageing [2]. Data is retrieved from freshly installed PV panels (i.e. "old data" with respect to the present day) and the model is compared with more recent data. It was expected that predictions at times immediately after the data used to create the model should be predicted quite accurately, while predictions made at times that are more in the future should overestimate real data. Thus, the difference between the actual and predicted values over several years in future can be used to observe the effect of ageing.

The simulation results are represented in Figure 12. The graph spans from January 2015 to December 2019, and for each month within that time frame, the table provides the average difference between the actual and predicted values over the time span.

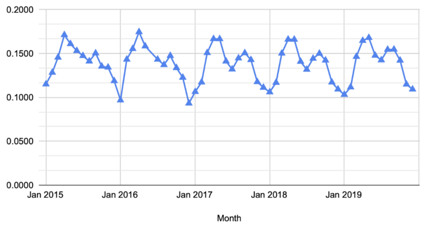


Figure 12: Ageing by Predictive Analysis

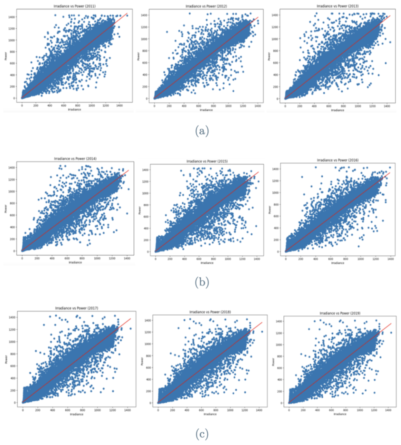
Based on the graph, we can infer that the accuracy of the predictions varies over time. The average difference between the actual and predicted values is quite small for some months (e.g., January 2015, January 2016), while for other months, the difference is relatively large. There are many factors that can contribute to the fluctuating difference between actual and predicted values in different months such as changes in external factors such as weather or maintenance patterns. Improving prediction accuracy by clustering the dataset for different weather conditions is discussed in detail in section 6.

5.2. Data Analysis

Scatter plots are used to study ageing using data analysis. Analyzing solar panel ageing using the slope of a scatter plot between irradiance and power during different years can provide insights into how the panels are performing over time. In general, the slope of a scatter plot between irradiance and power can indicate the efficiency of the solar panel, with a steeper slope indicating a more efficient panel.

5.2.1 Linear Interpolation

To analyze ageing, a scatter plot corresponding to each year along with the interpolating line is shown in Figure 13.



(a)

(b)

(c)

Figure 13: Ageing using Linear Interpolation

The slope for each year is graphically represented in Figure 14.

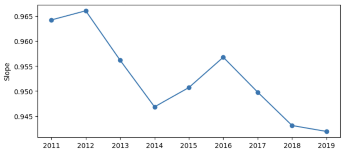
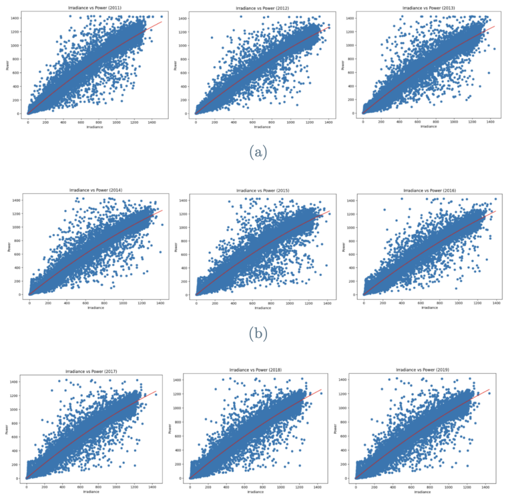


Figure 14: Slope of each year using linear interpolation

Based on Figure 14, there is a clear evidence of a consistent decrease in slope values over the 9-year period covered by the data. While there is some variability in slope values from year to year but there is a certain trend indicating a decrease in slope values over time.

5.2.2 Quadratic Interpolation

A scatter plot corresponding to each year along with Quadratic interpolation is shown in Figure 15.



(a)

(b)

(c)

Figure 15: Ageing using quadratic interpolation

The quadratic equation for each year is given in Table 4.

Year	Quadratic Fit
2011	$y = -0.00015x^2 + 1.12x - 1.70$
2012	$y = -0.00015x^2 + 1.13x - 1.90$
2013	$y = -0.00016x^2 + 1.14x - 1.91$
2014	$y = -0.00016x^2 + 1.11x - 1.30$
2015	$y = -0.00018x^2 + 1.13x - 1.00$
2016	$y = -0.00016x^2 + 1.10x - 0.41$
2017	$y = -0.00016x^2 + 1.10x - 0.73$
2018	$y = -0.00015x^2 + 1.08x - 1.61$
2019	$y = -0.00015x^2 + 1.08x - 1.63$

Table 4: Quadratic Equation for each year

Moreover, to understand the ageing, years vs PV power output graph considering quadratic fitting is plotted in Figure 16 for a fixed value of irradiance (1200 W$/m^2$).

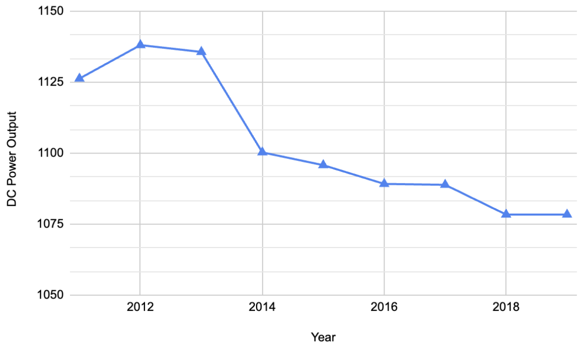


Figure 16: Power Output using quadratic interpolation with Irradiance = 1200 W$/m^2$

It is evident that the output power is gradually decreasing each year with the same value of irradiance, as observed through quadratic interpolation. This indicates that a quadratic fit may provide a more suitable description of the correlation between irradiance and power, and also aid in assessing the effects of ageing.

6. Model Performance Improvements - Dataset Clustering

The objective of dataset clustering is to systematically aggregate days exhibiting analogous patterns in photovoltaic (PV) power generation relative to solar irradiance. To address the challenge of uncertainty in power values on a day-to-day basis, the dataset is divided into two categories: sunny days and overcast days [7]. Each category is then trained separately using dedicated models. By grouping days with similar power production characteristics into the same category, the variability within each set is reduced. This clustering approach allows for a more targeted and accurate prediction by tailoring the models to specific weather conditions. To split the dataset into cloudy and sunny days, the daily mean irradiance values for each day in the dataset are obtained. Then, a threshold value is determined that will serve as the cut-off between cloudy and sunny days. In order to determine the most suitable threshold value, various thresholds were evaluated. This process resulted in the creation of a new dataset where days were grouped into either sunny or cloudy categories based on the threshold comparison. Flowchart of the above process is illustrated in Figure 17.

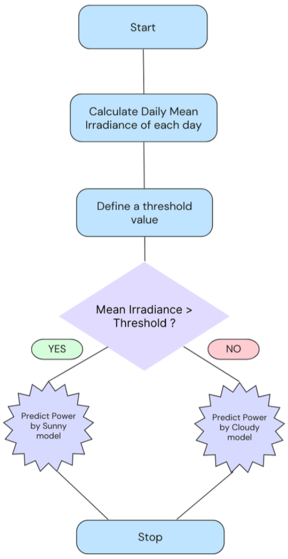

Start
Calculate Daily Mean Irradiance of each day
Define a threshold value
Mean Irradiance > Threshold ?
YES
NO
Predict Power by Sunny model
Predict Power by Cloudy model
Stop

Figure 17: Flowchart for Dataset clustering

The simulation results for cloudy and sunny days with a threshold value of mean daily irradiance equal to 250 W/m^2 are shown below from Figures 18 - 19.

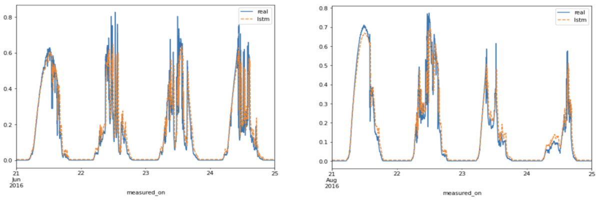


Figure 18: Sunny Days Results, MAE = 0.0324

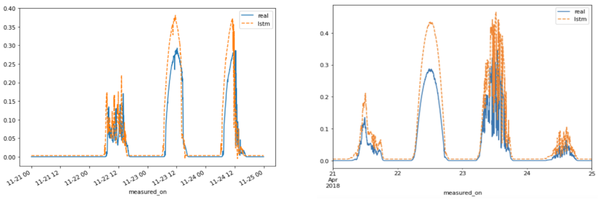


Figure 19: Cloudy Days Results, MAE = 0.0621

The clustering of the dataset has been shown to effectively reduce prediction errors. It is worth noting that the predictive accuracy is significantly lower for overcast days compared to sunny days. One possible explanation for this disparity is the substantial variability in power values within the overcast dataset, whereas the sunny dataset tends to exhibit more consistent patterns. Consequently, training models on the sunny dataset leads to more repetitive and reliable forecasts.

The performance of utilizing a model trained on one weather condition (sunny or overcast) for predicting power values on the opposite condition (overcast or sunny, respectively) can also be assessed. Figures 20 - 21 shows the simulation results of employing a model trained on one weather condition for predicting power values on the opposite condition.

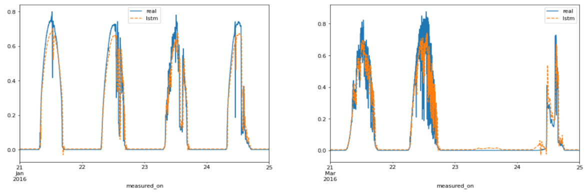


Figure 20: Cloudy Day model used for Sunny days, MAE = 0.0378

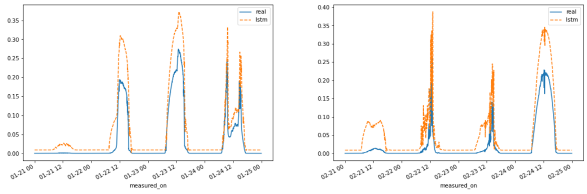


Figure 21: Sunny Day model used for Cloudy days, MAE = 0.0854

The results demonstrate that using a model trained on the sunny dataset and vice-versa, to predict power values on overcast days and sunny days, respectively led to higher MAE. These findings indicate that there are notable differences in power production characteristics between sunny and overcast conditions. The models trained on their respective datasets have learned specific patterns and relationships relevant to the corresponding weather conditions. When applied to opposite weather conditions, the models struggled to capture the nuanced dynamics, leading to decreased performance.

These findings highlight the importance of tailoring models to specific weather conditions for accurate power predictions. Therefore, it is recommended to employ separate models trained specifically for sunny and overcast conditions to address the challenge of uncertainty in power values on a day-to-day basis effectively.

7. Conclusions

The results suggests that LSTM architecture is the most effective model for predicting power generation from a solar PV system and it outperformed all other networks in terms of prediction accuracy. Additionally, the analysis of PV system aging provides valuable insights into the deterioration of prediction accuracy over time, suggesting the need for periodic recalibration or retraining of the prediction model to account for the changing characteristics of aging PV panels. Furthermore, by employing the clustering approach to classify the dataset into sunny and cloudy days and developing individual prediction models for each category lead to an improvement in prediction accuracy. The locations with more consistent and stable sunny weather conditions, with fewer occurrences of cloudy days throughout the year, achieved higher prediction accuracy.

8. Acknowledgements

I extend my utmost gratitude to my esteemed supervisor, Professor Giancarlo Storti Gajani for his guidance, invaluable support, and immense knowledge. His expertise and dedication have been instrumental in shaping the direction and quality of this research.
I am also indebted to the faculty members of Politecnico di Milano, whose wisdom, expertise, and passion for their respective fields have enriched my academic experience.

References

[1] R. Ahmed, V. Sreeram, Y. Mishra, and M.D. Arif. A review and evaluation of the state-of-the-art in pv solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124, 2020.

[2] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[3] C. Chen, S. Duan, T. Cai, and B. Liu. Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Solar energy*, 85(11):2856–2870, 2011.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] NREL. Solar power data for studies, 2022.

[6] M. Phi. Illustrated guide to lstm's and gru's: A step by step explanation, 2018.

[7] Talaye Talakoobi. Solar power forecast using artificial neural network techniques. Master's thesis, Politecnico di Torino, 10 2020. Department of Control and Computer Engineering.

[8] A. Tuohy, J. Zack, S. E. Haupt, and J. Sharp. Solar forecasting: methods, challenges, and performance. *IEEE Power and Energy Magazine*, 13(6):50–59, 2015.