CHAPTER I: Introduction

Solar energy, derived from the sun through photovoltaic (PV) panels, has become a pivotal component in the global shift towards renewable energy. This transition is primarily driven by the urgent need to mitigate climate change and reduce carbon emissions, as traditional energy sources like coal and natural gas are phased out in favour of cleaner alternatives. The growing investment in solar infrastructure worldwide highlights its critical role in achieving a low-carbon future, as nations recognize the significant potential of solar energy to provide sustainable and clean electricity.

Accurate prediction of solar power output is essential for optimizing energy management and ensuring the reliability of solar power systems. Variability in weather conditions and other meteorological factors such as solar irradiance, temperature, humidity, and wind speed introduce significant uncertainty in solar energy production. Recent studies have demonstrated that these uncertainties can be effectively managed through the application of machine learning (ML) models [1]. These models capture the complex interactions between various environmental variables, thereby improving the accuracy of solar power predictions.

Traditional methods of predicting solar energy output often rely on empirical formulas and manual calculations, which are not only time-consuming but also prone to human error [6]. In contrast, machine learning models offer a more efficient approach by automatically adjusting to changing conditions and providing more precise forecasts. For example, other research has shown that ML-based approaches can significantly enhance prediction accuracy compared to conventional methods, especially in regions with high weather variability.

This study aims to conduct a comprehensive analysis of solar panel energy output prediction based on storage capacity. We will utilize time series forecasting and regression methods, comparing their performance to identify the most effective approach for improving energy management in solar power systems. This work builds on the existing literature, including on the critical role of weather data in solar energy forecasting, to advance the application of machine learning in the renewable energy sector [5].

CHAPTER II: Theoretical Framework

A. Solar Power Production and Environmental Factors

Solar power production is primarily driven by the photovoltaic effect, where solar panels convert sunlight into electricity. Key factors influencing this process include solar irradiance, temperature, humidity, wind speed, and cloud cover. Solar irradiance is the most crucial, with higher levels generally boosting energy output. However, panel efficiency is also affected by temperature (with higher temperatures reducing efficiency), humidity, cloud cover, and wind speed, which can cool the panels but also introduce mechanical stress.

B. Time Series Analysis

Given the temporal nature of the data, time series analysis is essential for understanding patterns and trends in solar power production. Various methods are employed to forecast solar energy output accurately.

1. Prophet Forecasting Model

The Prophet model, developed by Facebook, is well-suited for handling daily, weekly, and yearly seasonality, as well as managing missing data and outliers, making it robust for solar power forecasting, particularly in the face of weather variability [9]. The general idea of the model is similar to a generalized additive model. The "Prophet Equation" fits, as mentioned above. This is given by,

$$\mathbf{y(t) = g(t) + s(t) + h(t) + e(t)} \quad (1)$$

Prophet, while user-friendly and effective for handling time series with strong seasonal patterns, has notable weaknesses. Prophet's reliance on predefined seasonalities may limit its accuracy when dealing with more complex or changing seasonal patterns, making it less flexible than other models like LSTM for certain types of data [8].

2. LSTM Networks

Long Short-Term Memory (LSTM), a type of recurrent neural network (RNN), is effective for sequential data, capturing complex patterns in solar power production by considering both short-term fluctuations and longer-term trends. The structure of the LSTM is shown in Figure 1.

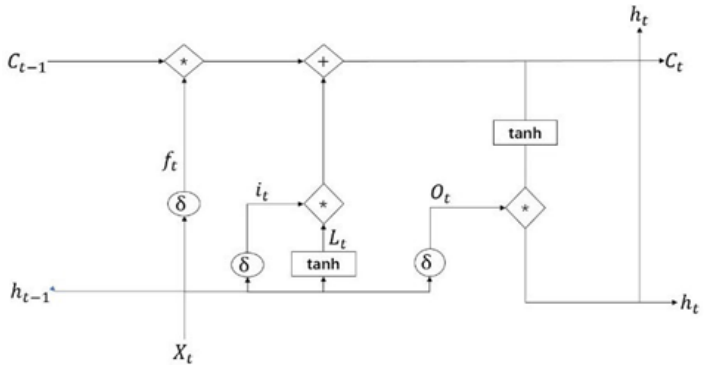

Figure 1. Long and short-term memory network LSTM model.

The LSTM architecture features a memory cell regulated by three gates: the input gate, the forget gate, and the output gate. The equation for the forget gate is:

$$\boldsymbol{f_t = \delta(W_f\,[h_{t-1}, X_t\,] + b_f\,)} \quad (2)$$

Here, δ represents the sigmoid activation function, Xt denotes the input vector, ht−1 refers to the hidden vector from the previous layer, and Wf and bf correspond to the weights and bias values of the inputs, respectively. The input gate is defined as follows:

$$\mathbf{i_t = \delta(W_i[h_{t-1}, X_t\,] + b_i)} \quad (3)$$

The output gate, which manages the output of the cell state values, is given by:

$$\mathbf{O_t = \delta(W_o[h_{t-1}, X_t\,] + b_o)} \quad (4)$$

Along with the three gates, a candidate memory cell is expressed as:

$$\boldsymbol{L_t = \tanh(W_c[h_{t-1}, X_t\,] + b_c)} \quad (5)$$

The current cell state is updated as follows:

$$C_t = f_t \times C_{t-1} + i_t \times L_t \quad (6)$$

Finally, the output of the current hidden state and its application as input to the next LSTM layer is represented by:

$$h_t = O_t \times \tanh(C_t) \quad (7)$$

Thus, it is evident from the LSTM equations and structure that LSTM is highly effective at mitigating gradient explosion or vanishing, giving it an edge in time series forecasting.

C. Regression Analysis

Regression analysis plays a vital role in solar power production by helping to measure how environmental factors like sunlight and temperature affect energy output[4].

1. XGBoost Baseline

XGBoost (XGB) is a powerful algorithm that handles complex data and large datasets efficiently [11], making it ideal for forecasting solar power. It's particularly good at understanding the complicated relationships between various environmental factors and how they influence energy production.

2. Ensemble Models

Ensemble learning methods can be employed to enhance the accuracy and reliability of solar power output predictions, as they combine multiple models to generate more robust predictions than any single model alone. Stacking, for instance, involves training various base models and using a meta-model to merge their predictions, leveraging the strengths of LGBM to improve overall accuracy. Boosting algorithms, including CatBoost (CAT), Gradient Boosting Machine (GBM), and eXtreme Gradient Boosting (XGB), iteratively train models to correct previous errors, making them particularly effective in managing complex datasets with non-linear relationships, which are common in solar power production data.

CHAPTER III: Analytical Steps

A. Data Preparation

The datasets (train, weather, solar irradiance) were aligned across identical time ranges (2014-2017) to ensure consistency. We conducted data quality checks, including type validation and missing data identification, to avoid leakage and maintain model integrity.

B. Exploratory Data Analysis (EDA)

EDA was critical in understanding data distributions and correlations, guiding feature engineering. The steps included targeted data visualizations and clustering analyses to extract meaningful insights.

C. Data Preprocessing

We merged and cleaned the datasets, addressing missing values with advanced techniques like LGBM-based iterative imputation and categorical clustering. Features were scaled using MinMaxScaler to improve model training efficiency.

D. Feature Engineering

New features were developed based on EDA and domain research. Feature selection was informed by baseline model performance, primarily leveraging XGBoost.

E. Model Development

Baseline models (XGBoost, CatBoost, GBM) were established with hyperparameter tuning. Time series models (LSTM, Prophet) were compared to regression models. Finally, AutoML was applied with a two-level stacking ensemble for optimal performance.

The workflow for preparing raw data for exploratory data analysis is illustrated in the following diagram. This diagram outlines the sequential steps taken to transform raw datasets into a structured format suitable for analysis and modelling.

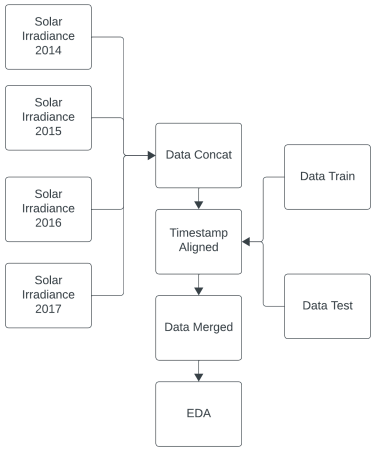

Figure 2. Model Development

CHAPTER IV: Analysis of Results

This study aims to identify the most effective machine learning models for predicting solar energy output based on storage capacity and to derive actionable insights. Analysis of % Baseline (percentage of energy generated in one hour) was resampled daily to uncover trends from 2014 to 2017.

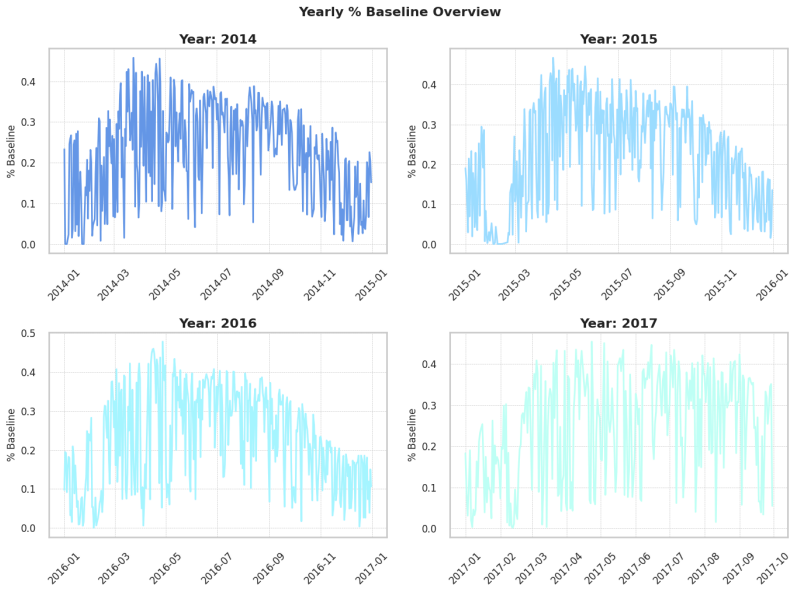

Figure 3. Target Distributions

The results show pronounced seasonal variability, with peak energy production during spring and summer and declines in winter. This variability underscores the impact of meteorological factors on solar energy output.

To better understand the impact of weather patterns on solar irradiance and energy production, we conducted an advanced cluster analysis using a comprehensive set of meteorological features. These features include maximum and minimum temperatures ('maxtempC', 'mintempC'), total sunshine hours ('sunHour'), UV index ('uvIndex'), dew point ('DewPointC'), and other atmospheric conditions such as cloud cover ('cloudcover'), humidity ('humidity'), precipitation ('precipMM'), pressure ('pressure'), air temperature ('tempC'), visibility ('visibility'), wind direction ('winddirDegree'), and wind speed ('windspeedKmph'). The goal was to examine how these weather variables contribute to variations in solar energy output.

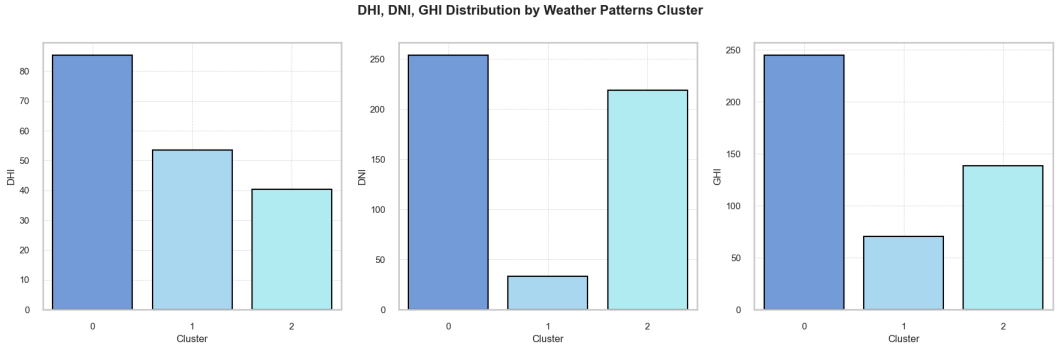

Figure 4.. Cluster Visualization

The K-means++ algorithm, optimized using the elbow method with the kelbowvisualizer library, determined that the optimal number of clusters is three. These clusters represent distinct weather patterns:

- Cluster 1: Cooler conditions with moderate cloud cover and relatively low temperatures.
- Cluster 2: Slightly warmer conditions with higher cloud cover and humidity.
- Cluster 0: The mildest conditions with moderate cloud cover and relatively higher temperatures.

Cluster profiling revealed that the mildest conditions (Cluster 0) are more favorable for maximizing solar energy production due to better direct sunlight exposure, suggesting that moderate cloud cover and higher temperatures lead to higher solar output. This insight highlights the importance of incorporating weather pattern analysis into predictive models to enhance the accuracy of solar energy forecasting, ultimately improving energy management strategies.

Other cluster analysis we combined both features from weather and solar irradiance to capture more complex the impact with energy solar output the features used weather patterns and additional features from solar irradiance, for instance DHI, GHI, DNI

A. Models Evaluation

1. Time Series Forecasting Models

The evaluation presents a robust approach to time series forecasting by integrating both classical and advanced modeling techniques.

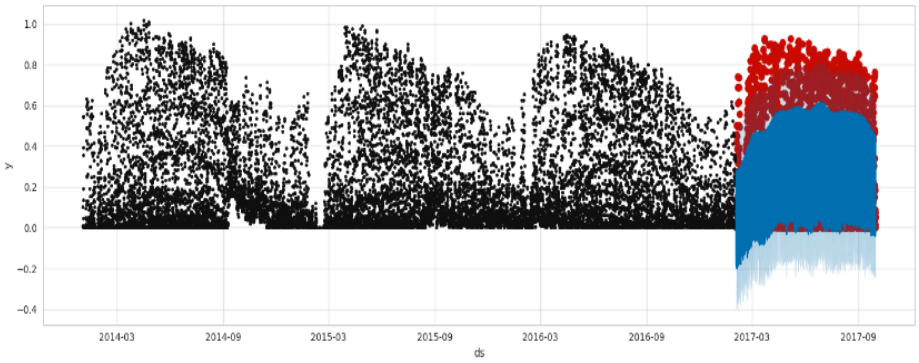

Figure 7. Prophet forecast accuracy and predictions vs actual on test validation data

Based on the visualization from the Prophet model, it is evident that the model does not fully capture the future values with high accuracy. These findings indicate that while Prophet offers valuable insights, there may be room for improvement in capturing the underlying patterns and trends necessary for more precise future predictions.

However, compared with the Prophet model, LSTM model leverages lagged features to capture more complex temporal dependencies within the data [10]. The lagged feature engineering, combined with LSTM's ability to handle sequential data, enhances the model's predictive power. The LSTM's performance, evaluated through RMSE, demonstrates its effectiveness in capturing underlying patterns, with results indicating a low error on both training and testing datasets [2][3].

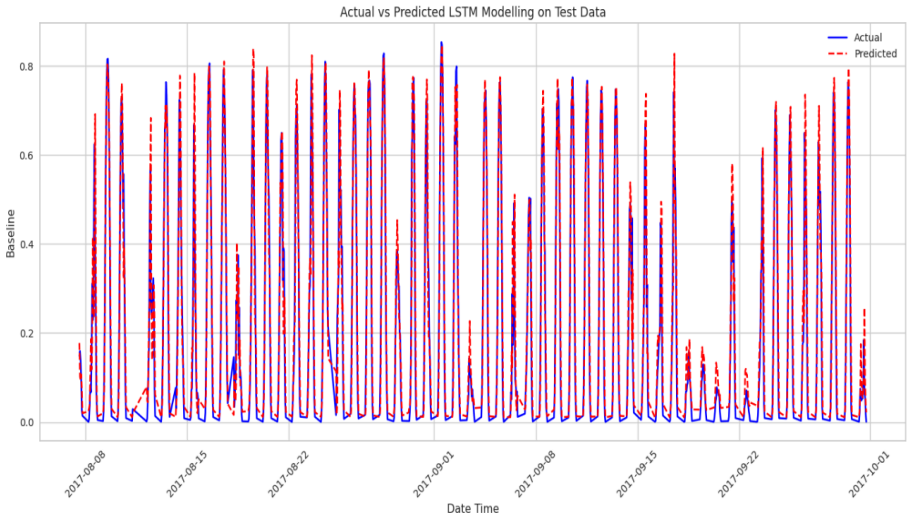

Figure 8. LSTM RMSE accuracy and predictions vs actual on test validation data

The visual comparison of actual versus predicted values further validates the model's accuracy and reveals its capability to model and predict time series data effectively. This dual-model approach not only reinforces the forecasting accuracy but also offers a comprehensive evaluation of different forecasting methodologies.

2. Baseline XGBoost and Ensemble Regression Models

The provided models showcase an effective use of XGBoost for regression tasks. Initially, the performance of a baseline model is assessed to establish a benchmark for comparison with the result of mean validation of RMSE is 0.08210. The AutoGluon model leverages ensemble techniques, combining the strengths of CatBoost, GBM, and XGBoost to improve predictive performance.

	model	score_val
0	XGBoost_BAG_L2/T2	-0.061105
1	XGBoost_BAG_L2/T1	-0.061271
2	WeightedEnsemble_L3	-0.060812
3	CatBoost_BAG_L2/T1	-0.061321
4	XGBoost_BAG_L2/T6	-0.061165
5	XGBoost_BAG_L2/T4	-0.061177
6	LightGBM_BAG_L2/T8	-0.061351
7	CatBoost_BAG_L2/T3	-0.061245
8	LightGBMLarge_BAG_L1	-0.063845
9	LightGBMXT_BAG_L1/T1	-0.063338
10	LightGBM_BAG_L1/T1	-0.064648
11	XGBoost_BAG_L1/T2	-0.068451
12	XGBoost_BAG_L1/T1	-0.069202
13	CatBoost_BAG_L1/T1	-0.068574

Figure 9. Score leaderboards for ensemble models

The leaderboard results illustrate the performance of various models, with ensemble methods like the WeightedEnsemble_L3 showing marginally better performance compared to individual models. This highlights the efficacy of combining predictions from multiple models. Overall, the leaderboard not only offers a comparative view of model performance but also demonstrates AutoGluon's capability in automating the model selection process, ensuring that the final model is both high-quality and well-suited for deployment.

The feature importance analysis highlights that certain features, like "Hour," significantly impact the model's predictions, indicating its strong correlation with solar energy production.

	importance	stddev	p_value	n	p99_high	p99_low
Hour	0.167768	0.012857	4.106488e-06	5	0.194240	0.141296
Solar Zenith Angle	0.065375	0.001799	6.873981e-08	5	0.069079	0.061671
humidity	0.043097	0.006018	4.444669e-05	5	0.055488	0.030707
cloudcover	0.027332	0.001804	2.262731e-06	5	0.031045	0.023618
sunset_minutes	0.019282	0.003328	1.023568e-04	5	0.026134	0.012430
sunHour	0.008943	0.001969	2.646208e-04	5	0.012998	0.004889
tempC	0.000048	0.000655	4.392587e-01	5	0.001395	-0.001300
TempRange	0.000022	0.000429	4.568556e-01	5	0.000905	-0.000861
totalSnow_cm	0.000005	0.000044	4.031572e-01	5	0.000096	-0.000086
WindChillC	-0.000093	0.000585	6.292325e-01	5	0.001112	-0.001297
HeatIndexC	-0.000121	0.000734	6.349097e-01	5	0.001389	-0.001632
Wind Speed	-0.000915	0.000778	9.708836e-01	5	0.000687	-0.002517

Figure 10. Top 6 vs bottom 6 of features importance

Other key features include "Solar Zenith Angle" and "humidity," which also play crucial roles in the model's accuracy. In contrast, features like "Wind Speed" and "Heat IndexC" show minimal importance, suggesting they are less relevant to the prediction task. Focusing on these influential features can enhance both the model's performance and interpretability.

CHAPTER V: Conclusion and Recommendation

The analysis highlights that refining the model by focusing on influential features enhances both its performance and interpretability. Understanding seasonal patterns and weather conditions proves essential for accurate solar energy forecasting [7]. The study demonstrates the value of advanced clustering techniques and sophisticated feature engineering in improving predictive accuracy. Integrating classical and modern modeling approaches, alongside AutoGluon automated model selection, ensemble methods and time series such as LSTM, offers significant potential for boosting forecasting accuracy and operational efficiency. To enhance the accuracy of solar energy forecasting, future predictive models should integrate detailed weather pattern analysis, emphasizing features such as temperature, cloud cover, and humidity to capture output variability more effectively.

Additionally, further exploration and refinement of feature engineering techniques, including time-related, historical, and cyclic features, are recommended to improve model performance. Prioritizing ensemble methods can lead to superior results by combining predictions from multiple models. Finally, maintaining forecasting accuracy requires the regular

updating and refining of models to adapt to new data and evolving solar energy production patterns.

REFERENCES

[1]
A. O. M. Maka and J. M. Alabid, “Solar energy technology and its roles in sustainable development,” *Clean Energy*, vol. 6, no. 3, 2022, doi: 10.1093/ce/zkac023.

[2]
C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Climate Research*, vol. 30, no. 1, 2005, doi: 10.3354/cr030079.

[3]
D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ Computer Science*, vol. 7, 2021, doi: 10.7717/PEERJ-CS.623.

[4]
S. Huang, “Linear regression analysis,” in *International Encyclopedia of Education: Fourth Edition*, 2022. doi: 10.1016/B978-0-12-818630-5.10067-3.

[5]
M. K. Abdelrazik, S. E. Abdelaziz, M. F. Hassan, and T. M. Hatem, “Climate action: Prospects of solar energy in Africa,” *Energy Reports*, vol. 8, 2022, doi: 10.1016/j.egyr.2022.08.252.

[6]
N. Novas, R. M. Garcia, J. M. Camacho, and A. Alcayde, “Advances in solar energy towards efficient and sustainable energy,” *Sustainability (Switzerland)*, vol. 13, no. 11, 2021, doi: 10.3390/su13116295.

[7]
H. H. Pourasl, R. V. Barenji, and V. M. Khojastehnezhad, “Solar energy status in the world: A comprehensive review,” *Energy Reports*, vol. 10. 2023. doi: 10.1016/j.egyr.2023.10.022.

[8]
G. Chiarot and C. Silvestri, “Time Series Compression Survey,” *ACM Computing Surveys*, vol. 55, no. 10, 2023, doi: 10.1145/3560814.

[9]
B. Lim and S. Zohren, “Time-series forecasting with deep learning: A survey,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2194. 2021. doi: 10.1098/rsta.2020.0209.

[10]
H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, “Deep learning for time series classification: a review,” *Data Mining and Knowledge Discovery*, vol. 33, no. 4, 2019, doi: 10.1007/s10618-019-00619-1.

[11]

H. T. Wen, H. Y. Wu, and K. C. Liao, “Using XGBoost Regression to Analyze the Importance of Input Features Applied to an Artificial Intelligence Model for the Biomass Gasification System,” *Inventions*, vol. 7, no. 4, 2022, doi: 10.3390/inventions7040126.