Forecasting Solar Irradiance at An-Najah National University

Computer Science Apprenticeship by the Faculty of Engineering

Supervised by:

Dr. Adnan Salman

Eng. Asem Saleh

Prepared by:

Nemeh Abu Issa *12114863*

Islam Sad Al Deen *12220096*

Abstract

*This project explores the feasibility of forecasting hourly solar irradiance at **An-Najah National University** to support informed decision-making around photovoltaic system deployment on campus. One year of hourly **NASA POWER** data was collected for the university's geographic location in Nablus and processed through timestamp reconstruction, missing-value handling, and engineered temporal features that capture daily and seasonal patterns.*

*Five forecasting models were developed and evaluated, including **SARIMA, SARIMAX, Prophet, Random Forest, and XGBoost,** using strictly time-ordered train–test splits. To assess true predictive capability, walk-forward validation was additionally applied to the machine learning models. Results show that **the Random Forest model under walk-forward validation clearly outperformed all other approaches**, demonstrating superior accuracy and consistency. These findings confirm that feature-rich, non-linear models provide the strongest foundation for reliable irradiance forecasting and can directly support the university's ongoing renewable energy planning initiatives.*

Introduction

***An-Najah National University** is steadily moving toward a more sustainable energy future, driven by the need to reduce electricity costs and lessen dependence on conventional power sources with rising financial and environmental burdens. With a campus that includes multiple academic buildings, research laboratories, and service facilities operating around the clock, the university faces a growing demand for reliable, affordable energy.*

*To support this transition, the ability to forecast **solar irradiance** with high precision becomes a key requirement. Accurate predictions allow the institution to evaluate the performance potential of photovoltaic installations and to determine the feasibility of expanding solar infrastructure across campus.*

This project develops a forecasting framework trained on meteorological data captured specifically for the university's location in Nablus, allowing engineers and decision-makers to estimate expected solar energy production at both daily and monthly scales. Such insights help optimize budget allocation, estimate electricity savings, and plan installations based on evidence rather than assumptions.

*Beyond its immediate operational impact, the project positions **An-Najah National University** as a leader in sustainable energy research within Palestine. It enables faculty and students to engage in applied, data-driven work and lays the foundation for future innovations in renewable systems that support both the university and the surrounding community.*

Dataset

*This study utilizes one full year of hourly meteorological data collected from the N**ASA POWER** database for the geographic coordinates of **An-Najah National University in Nablus (32.2326°N, 35.2465°E).** The dataset serves as the core input for all forecasting models, providing both the target variable—solar irradiance—as well as the atmospheric factors believed to influence its variation throughout the day and across seasons.*

*The dataset contains approximately **8,736 hourly records**, corresponding to 365 days × 24 hours, and includes the following key features:*

- ***YEAR, MONTH, DAY, HOUR** – time components used to reconstruct a continuous timestamp*
- ***ALLSKY_SFC_SW_DWN** – the target variable representing incoming solar irradiance (W/m²)*
- ***T2M** – air temperature at 2 meters*
- ***RH2M** – relative humidity*
- ***PS** – surface pressure*
- ***WS10M** – wind speed at 10 meters*
- ***T2MDEW** – dew-point temperature*
- ***SZA** – solar zenith angle*

Preprocessing

Before applying the forecasting models, several preparation steps were carried out:

Timestamp Construction

The YEAR–MO–DY–HR columns were combined into a single datetime index to preserve chronology.

Cleaning and Filtering

Missing irradiance values were removed, and all numerical features were converted to appropriate formats.

Feature Engineering

Since irradiance naturally follows daily and seasonal cycles, sinusoidal encodings were generated for:

hour of day (sin & cos), and

day of year (sin & cos).

Temporal Memory Features

To allow models to learn from recent history, lagged irradiance features were created at:

1h, 2h, 3h, 6h, and 24h delays,
in addition to a rolling 24-hour average.

This processed dataset provides a rich structure suitable for both classical and machine learning forecasting models. The inclusion of weather variables, cyclical time patterns, and lagged history enables the models to learn both short-term fluctuations and long-term seasonal behavior relevant to solar generation on campus.

Methodology

This project aims to assess the feasibility of forecasting solar irradiance to support energy planning at ***An-Najah National University,*** *using a multi-model approach applied to one full year of hourly meteorological data collected for the university's geographic location in Nablus. The processing pipeline begins with accurate timestamp reconstruction, removal of missing values, and chronological ordering of all observations to preserve the temporal structure required for forecasting tasks.*

Feature engineering plays a crucial role in improving model performance. Daily and seasonal patterns are represented using sinusoidal transformations (sin/cos) for hour-of-day and day-of-year, while temporal memory is introduced through lagged irradiance values at 1, 2, 3, 6, and 24-hour intervals, along with a rolling 24-hour average. For models that leverage external atmospheric information, additional predictors—including temperature, relative humidity, surface pressure, wind speed, dew point temperature, and solar zenith angle—are incorporated.

Five models were developed to reflect different forecasting philosophies in time-series analysis and machine learning:
• ***SARIMA*** *– captures autoregressive seasonal structure using only irradiance history*
• ***SARIMAX*** *– integrates exogenous meteorological variables alongside the irradiance series*
• ***Prophet*** *– decomposes the signal into trend, seasonal, and residual components*
• ***Random Forest*** *– a non-linear ensemble capable of modeling complex interactions across all engineered features*
• ***XGBoost*** *– a gradient-boosted tree framework optimized for structured tabular data*

To ensure a fair and realistic evaluation, all models were trained on earlier time intervals and tested on later unseen periods, simulating true forecasting conditions. In addition, because machine learning models are prone to overfitting, we applied ***Walk-Forward Validation*** *to both the* ***Random Forest*** *and* ***XGBoost*** *models. In this procedure, the model is repeatedly retrained on expanding segments of past data and used to predict the next horizon. This mirrors real deployment and measures how well the models generalize when forecasting ahead.*

Through this validation strategy, the reported performance reflects genuine predictive capability rather than reproduction of previously seen values.

Results

The evaluation of the five forecasting models revealed clear differences in predictive performance across statistical and machine learning approaches. As shown in ***Figure 1****, the classical time-series models—****SARIMA, SARIMAX, and Prophet****—captured the broad daily and seasonal patterns of solar irradiance, yet their accuracy dropped noticeably during periods of irregular or rapidly changing weather. This resulted in higher MAE and RMSE values and lower* R^2 *scores, making them less reliable for day-ahead planning scenarios.*

In contrast, the machine learning models delivered substantially stronger and more stable results. ***Random Forest and XGBoost****, which were trained on a richer feature set—including meteorological variables, lagged irradiance history, and cyclical time encodings—achieved far lower error scores and higher predictive strength compared to the classical models (Figure 1). These models were better able to account for short-term variability and complex non-linear relationships that traditional time-series structures could not model effectively.*

A closer look at the generated predictions, shown in ***Figure 2****, highlights the accuracy of the Random Forest model in reproducing real irradiance behavior over the final seven days of the test period. The model almost perfectly captures both the daytime peak shape and the nighttime zero levels, demonstrating strong alignment with the actual recorded values.*

Finally, ***Figure 3*** *presents the results of the walk-forward validation procedure, where the Random Forest and XGBoost models were repeatedly retrained on expanding time windows. Across all folds, Random Forest consistently achieved lower RMSE than XGBoost, confirming its superior generalization capability when forecasting truly unseen data.*

Together, these findings establish that feature-rich, non-linear models—particularly ***Random Forest****—are the most suitable for operational short-term irradiance forecasting at An-Najah National University, and provide a reliable analytical basis for future photovoltaic planning and decision-making on campus.*

Main Test Results Table

Model	MAE	RMSE	R^2
Random Forest	**16.61**	**30.88**	**0.992**
XGBoost	25.31	41.01	0.986
SARIMAX	69.83	83.93	0.941
SARIMA	56.37	87.79	0.936
Prophet	98.55	117.66	0.884

Walk-Forward Validation Summary

Model	Mean MAE	Mean RMSE	Mean R^2	Std RMSE
Random Forest	37.85	74.90	0.934	24.36
XGBoost	45.17	77.93	0.926	23.35

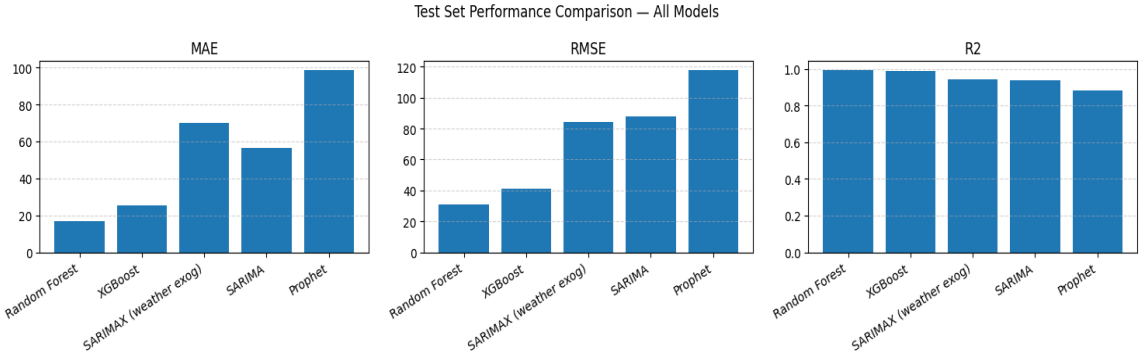

Figure 1 – Model Comparison: MAE, RMSE, and R^2

Figure 1 shows a comparison of test-set performance across the five forecasting models. Random Forest and XGBoost achieved the lowest MAE and RMSE values and the highest R^2 scores, indicating superior predictive accuracy relative to SARIMA, SARIMAX, and Prophet.

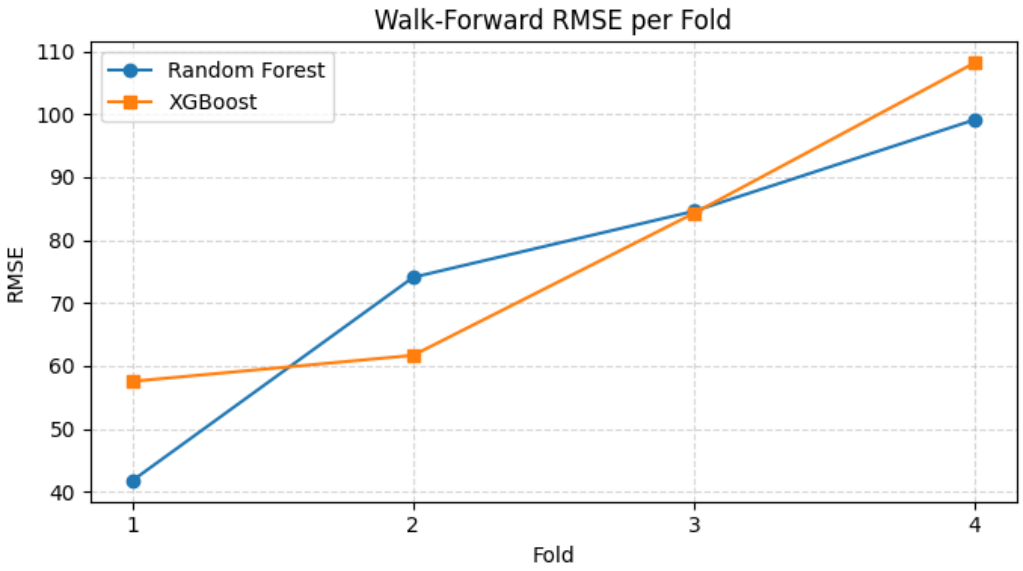

Figure 2 – Walk-Forward Validation (Random Forest vs XGBoost)

Figure 2 displays RMSE values over four sequential walk-forward folds. Although both models experience increasing error as the forecast horizon progresses, Random Forest consistently records marginally lower RMSE values, confirming its stronger generalization ability under real deployment conditions.

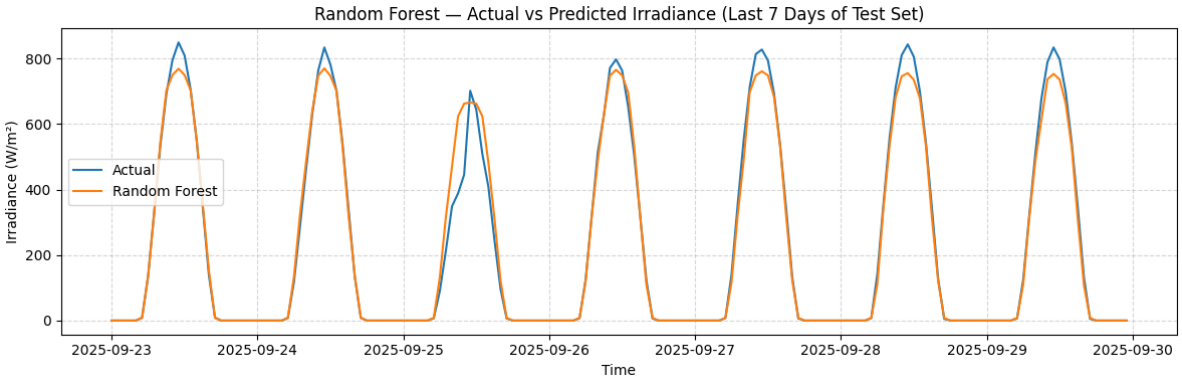

Figure 3 – Actual vs Predicted Irradiance (Random Forest)

Figure 3 presents the Actual vs Predicted irradiance values for the last seven days of the test set using the Random Forest model.
The close alignment between the two curves demonstrates that the Random Forest model effectively tracks both the daily peak structure and nighttime zero irradiance pattern.

Conclusion

*The results of this project demonstrate that machine learning models—especially **Random Forest**—offer the most accurate and reliable approach for forecasting hourly solar irradiance at **An-Najah National University**. Compared to classical time-series methods, Random Forest maintained superior performance, particularly under walk-forward validation, which confirmed its ability to generalize effectively to unseen future data.*

These findings highlight the value of non-linear, feature-rich models that leverage both weather variables and historical irradiance patterns. The outcome also indicates strong potential for applying such models to real energy planning, where accurate forecasts can guide solar system sizing and long-term decision-making. Furthermore, this work provides a solid starting point for future developments, including integrating new data sources, improving predictive accuracy, or extending forecasting to full photovoltaic energy production.

References

1. NASA POWER Data Access Viewer. NASA Prediction of Worldwide Energy Resources (POWER). NASA POWER | Data Access Viewer (DAV)

2. Tandon et al. (2025) – Rajasthan solar forecasting paper Machine learning-driven solar irradiance prediction: advancing renewable energy in Rajasthan | Discover Applied Sciences

3. Scikit-learn documentation scikit-learn: machine learning in Python — scikit-learn 1.8.0 documentation