- Leveraging AI and Data-Driven Techniques for Solar Power Prediction -

MINOR PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the

degree Of

-BACHELOR OF TECHNOLOGY-

In

-ECE-AI-

By

Anjali - 0130118022
Himanshi Kataria - 04101182022
Livya Jajoriya - 06901182022

Guided by

-Prof. Nidhi Goel-
- Professor-
-Dept. Of ECE-

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

NEW DELHI – 110006

- December 2025 –

PROJECT COMPLETION CERTIFICATE

This is to hereby certify that **Anjali (01301182022), Himanshi Kataria (0411182022), Livya Jajoria (06901182022)**, students of **Indira Gandhi Delhi Technical University for Women**, has successfully completed the project titled "**Leveraging AI and Data-Driven Techniques for Solar Power Prediction**" under my supervision and guidance. The candidate has shown excellent dedication, sincere effort, and professional competence throughout the project duration.

I extend my best wishes for their future academic and professional endeavors.

Prof. Nidhi Goel
Project Mentor

ACCEPTANCE CERTIFICATE

This is to certify that the research paper entitled **Leveraging AI and Data-Driven Techniques for Solar Power** Predictionauthored by **Ms. Anjali (Roll No.: 01301182022)** B.Tech (Electronics and Communication Engineering – Specialization in Artificial Intelligence) Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi has been accepted at the **7th International Conference on Artificial Intelligence and Speech Technology (AIST-2025)**, organized by the Department of Information Technology, Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi.

Supervisor's Name: Nidhi Goel

Designation: Professor, Department of ECE (AI)

Indira Gandhi Delhi Technical University for Women, Delhi

Student's Signature:

Anjali (01301182022)

Himanshi Kataria (04101182022)

Livya Jajoria (06901182022)

Date: ____________________________

Place: Delhi

UNDERTAKING REGARDING ANTI-PLAGIARISM

We, hereby, declare that the material/ content presented in the report is free from plagiarism and is properly cited and written in my own words. In case plagiarism is detected at any stage, I shall be solely responsible for it.

Anjali (01301182022)

Himanshi Kataria (04101182022)

Livya Jajoria (06901182022)

ACKNOWLEDGEMENT

I would like to express my profound gratitude to **Prof. Nidhi Goel** , of ECE (Professor) department for their contributions to the completion of my project titled **Leveraging AI and Data-Driven Techniques for Solar Power Prediction**.

I would like to express my special thanks to our mentor his time and efforts She provided throughout the project. Your useful advice and suggestions were really helpful to me during the project's completion. In this aspect, I am eternally grateful to you.

I would like to acknowledge that this project was completed entirely by me and not by someone else.

Sincerely,

Anjali (01301182022)

Himanshi Kataria (04101182022)

Livya Jajoria (06901182022)

DECLARATION

I, solemnly declare that the project report, **Leveraging AI and Data-Driven Techniques for Solar Power Prediction ,** is based on my own work carried out during the course of our study under the supervision of **Prof. Nidhi Goel , Professor, AI & DS Department, IGDTUW**. I assert the statements made and conclusions drawn are an outcome of my research work.

I further certify that:

I. The work contained in the report is original and has been done by me under the supervision of my supervisor.

II. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.

III. We have followed the guidelines provided by the university in writing the report.

IV. Whenever we have used materials (text, data, theoretical analysis/equations, codes/program, figures, tables, pictures, text etc.) from other sources, we have given due credit to them in the report and have also given their details in the references.

Anjali (01301182022)

Himanshi Kataria (04101182022)

Livya Jajoria (06901182022)

LIST OF ABBREVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
PV	Photovoltaic
DC	Direct Current
ANN	Artificial Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
CNN	Convolutional Neural Network
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
R^2	Coefficient of Determination

TABLE OF FIGURES

ABSTRACT/SUMMARY

The growing global demand for clean and sustainable energy has accelerated research in solar power generation. However, the intermittent and weather-dependent nature of solar energy poses significant challenges in accurately predicting power output. This project, "Leveraging AI and Data-Driven Techniques for Solar Power Prediction," aims to develop and evaluate advanced machine learning and deep learning models for forecasting solar power generation using real-world data. The dataset, obtained from Kaggle, includes key meteorological and temporal parameters such as irradiance, temperature, humidity, and wind speed.

A comprehensive data preprocessing and feature engineering pipeline was implemented, followed by experimentation with multiple predictive models Linear Regression, Random Forest, XGBoost, Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), 1D Convolutional Neural Network (1D-CNN), and Transformer. The models were assessed using standard evaluation metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

Results indicate that among machine learning approaches, Random Forest achieved robust performance with high interpretability, while among deep learning models, GRU outperformed others, achieving the lowest MAE and RMSE values. The findings underscore the potential of data-driven and AI-based techniques for enhancing the accuracy and reliability of solar power prediction. The project also proposes future directions such as model optimization, cloud deployment, and the integration of explainable AI (XAI) for improved transparency and scalability in real-world energy management systems.

INDEX

Chapter 1: INTRODUCTION

This project explores the use of Artificial Intelligence (AI) and data-driven methods to accurately forecast solar power generation. With the growing global demand for clean energy, precise prediction of solar output is vital for efficient grid management and sustainable energy planning.

1.1 Background and Motivation

Solar energy has emerged as one of the most promising renewable sources due to its abundance and environmental benefits. However, its generation is highly dependent on varying weather conditions such as temperature, irradiation, and humidity, making accurate forecasting difficult. Reliable prediction of solar power output helps in grid stability, energy planning, and resource optimization. Recent advances in AI and Machine Learning (ML) have made it possible to build accurate data-driven models that capture these complex patterns and improve prediction reliability.

1.2 Problem Definition

Conventional mathematical models fail to handle the nonlinear and uncertain nature of solar energy data. There is a need for a data-driven predictive system that can accurately forecast solar power generation based on meteorological and temporal inputs. This project develops and compares several ML and Deep Learning (DL) models to predict solar power efficiently and identify the best performing model for real-time applications.

1.3 Objectives of the Study

1. To preprocess and analyze real-world solar and weather datasets.
2. To implement ML and DL algorithms for solar power forecasting.
3. To evaluate models using MAE, RMSE, and R^2 metrics.
4. To identify the most accurate and efficient model for deployment.
5. To integrate the trained model with an API for real-time prediction.

1.4 Scope of the Project

The project uses the Kaggle "Solar Power Generation Data" dataset. It covers multiple algorithms Linear Regression, Random Forest, XGBoost, ANN, LSTM, GRU, 1D-CNN, and Transformer and compares their performance for forecasting. The work focuses on software-based analysis and prediction, not on physical implementation with solar panels.

1.5 Organization of the Report

The report begins with an introduction outlining the background, objectives, and scope of the project. It then presents a review of existing research and techniques related to solar power prediction, followed by a detailed description of the dataset and methodology adopted for model development. The subsequent section discusses the experimental results and performance analysis of different models, and the report concludes with key findings and suggestions for future work.

Chapter 2 : LITERATURE REVIEW

2.1 Overview of Solar Power Forecasting Techniques

The rapid growth of renewable energy systems has intensified the need for reliable solar power forecasting. Traditional methods such as statistical and physical models often rely on historical averages, irradiance equations, and meteorological parameters to estimate solar output. However, these methods struggle to handle the nonlinearity and uncertainty of atmospheric variations. To overcome these limitations, data-driven approaches using Machine Learning (ML) and Deep Learning (DL) have gained prominence due to their ability to learn complex patterns and correlations from large datasets. These models enable more accurate short-term and long-term forecasts, which are crucial for energy scheduling, grid balancing, and solar farm optimization.

2.2 Machine Learning Approaches for Energy Prediction

Machine Learning algorithms have been widely adopted for solar power forecasting because of their flexibility and interpretability. Linear Regression (LR) is one of the earliest techniques used to model the relationship between solar irradiance and power output, but its linear nature restricts its predictive strength for nonlinear weather data. Random Forest (RF), an ensemble learning technique, enhances accuracy by combining multiple decision trees, effectively handling multivariate dependencies and noise. XGBoost, a gradient boosting method, has demonstrated superior performance in many forecasting tasks due to its optimization and regularization capabilities.

These ML techniques are computationally efficient and often serve as strong baselines before applying more complex deep learning architectures.

2.3 Deep Learning Models for Time-Series Forecasting

Deep Learning models have recently shown exceptional performance in solar power prediction tasks, especially for time-series data. Artificial Neural Networks (ANNs) capture nonlinear relationships by learning hidden feature representations. Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are particularly effective in modeling temporal dependencies in sequential data. LSTM networks manage long-term dependencies using memory cells and gating mechanisms, while GRU models offer a simplified yet efficient alternative. Convolutional Neural Networks (1D-CNN) have also been utilized to capture local temporal patterns in power generation sequences. More recently, Transformer-based architectures have emerged, leveraging attention mechanisms to learn global dependencies across time steps, offering state-of-the-art accuracy in time-series forecasting applications.

2.4 Research Gaps and Challenges Identified

While many studies have demonstrated the effectiveness of AI models for solar forecasting, certain gaps and challenges remain. Most existing research focuses on optimizing individual models rather than developing hybrid or ensemble approaches that combine ML and DL strengths. Additionally, model performance can vary significantly across different geographic regions and weather conditions, indicating a need for adaptive models. Data quality and feature selection also

Chapter 3 : METHODOLOGY

This chapter explains the step-by-step methodology adopted for developing the solar power prediction system. The approach includes data acquisition, preprocessing, model development, performance evaluation, and integration for real-time prediction. Both Machine Learning (ML) and Deep Learning (DL) models were implemented and compared to determine the most suitable technique for accurate and efficient forecasting.

3.1 System Architecture

The proposed system follows a modular architecture that integrates data collection, preprocessing, model training, evaluation, and deployment stages. The workflow begins with acquiring solar power generation and meteorological data, followed by cleaning and transformation into a usable format. The processed dataset is then used to train multiple Machine Learning (ML) and Deep Learning (DL) models to predict solar power output based on environmental features. The best-performing model is selected based on performance metrics and later integrated with an external weather API for real-time predictions. This modular design ensures scalability, adaptability, and the potential for future cloud deployment.The overall data processing and modeling workflow is illustrated in Figure 3.1.

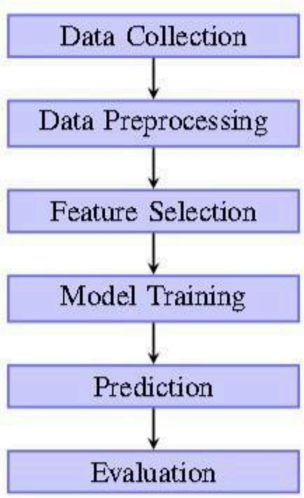
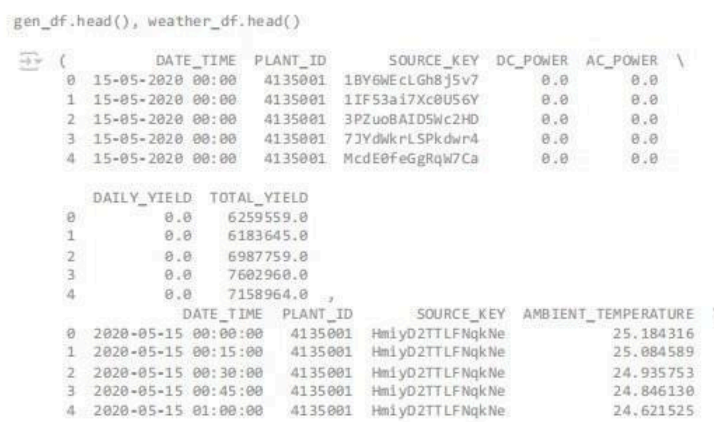


Fig. 3.1: Overall System Architecture for Solar Power Prediction



```
gen_df.head(), weather_df.head()

(          DATE_TIME  PLANT_ID       SOURCE_KEY  DC_POWER  AC_POWER  \
 0  15-05-2020 00:00   4135001  1BY6WEcLGh8j5v7       0.0       0.0
 1  15-05-2020 00:00   4135001  1IF53ai7Xc0U56Y       0.0       0.0
 2  15-05-2020 00:00   4135001  3PZuoBAID5Wc2HD       0.0       0.0
 3  15-05-2020 00:00   4135001  7JYdWkrLSPkdwr4       0.0       0.0
 4  15-05-2020 00:00   4135001  McdE0feGgRqW7Ca       0.0       0.0

    DAILY_YIELD  TOTAL_YIELD
 0          0.0    6259559.0
 1          0.0    6183645.0
 2          0.0    6987759.0
 3          0.0    7602960.0
 4          0.0    7158964.0  ,
              DATE_TIME  PLANT_ID       SOURCE_KEY  AMBIENT_TEMPERATURE  \
 0  2020-05-15 00:00:00   4135001  HmiyD2TTLFNqkNe            25.184316
 1  2020-05-15 00:15:00   4135001  HmiyD2TTLFNqkNe            25.084589
 2  2020-05-15 00:30:00   4135001  HmiyD2TTLFNqkNe            24.935753
 3  2020-05-15 00:45:00   4135001  HmiyD2TTLFNqkNe            24.846130
 4  2020-05-15 01:00:00   4135001  HmiyD2TTLFNqkNe            24.621525
```

Fig. 3.2: Sample Dataset Showing Weather and Power Generation Attributes

3.2 Dataset Description

The dataset used in this study was obtained from the Kaggle repository titled "Solar Power Generation Data." It contains historical records from multiple photovoltaic (PV) plants, including both weather data and power generation parameters. Key features include ambient temperature, module temperature, solar irradiance, wind speed, and relative humidity, along with corresponding DC and AC power values. Time stamps are recorded at regular intervals, allowing temporal analysis of generation patterns. The dataset was divided into training (80%) and testing (20%) subsets, with an additional 5-fold cross-validation applied to ensure the robustness of the models.A sample of the dataset containing key weather and power generation attributes is shown in Figure 3.2.

3.3 Data Preprocessing and Feature Engineering
Data preprocessing is a crucial step to ensure the accuracy and efficiency of model training.Missing values were handled using interpolation, and irrelevant or redundant features were removed.The date and time columns were converted into numerical formats such as hour, day, and month to capture temporal patterns. Feature scaling was applied using Min-Max normalization to standardize input variables across all models. Correlation analysis was conducted to identify key factors influencing solar power output, revealing that solar irradiance and module temperature were the most significant predictors. Feature engineering also included combining environmental and temporal attributes to enhance predictive performance.

3.4 Model Selection and Implementation
A combination of ML and DL algorithms was implemented to evaluate performance differences and identify the optimal approach for solar power forecasting. The selected models include Linear Regression (LR), Random Forest (RF), XGBoost, Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), 1D Convolutional Neural Network (1D-CNN), and Transformer.
Machine learning models such as RF and XGBoost were trained on tabular data using scikit-learn libraries. These models are known for their interpretability and robustness against overfitting.
Deep learning models were implemented using TensorFlow and Keras frameworks. The ANN consisted of multiple dense layers with ReLU activation, while the LSTM and GRU networks were designed to capture sequential temporal dependencies in power generation data. The 1D-CNN model extracted localized temporal features, and the Transformer model utilized self-attention mechanisms to capture long-range dependencies. Each model was trained for multiple epochs, with the Adam optimizer and early stopping to prevent overfitting.

3.5 Evaluation Metrics
To evaluate model performance, three standard statistical metrics were used:
- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values, providing a direct interpretation of model accuracy.
- Root Mean Square Error (RMSE):Emphasizes larger errors and penalizes models that produce outlier predictions.
- Coefficient of Determination (R^2): Represents the proportion of variance in the target variable that can be explained by the model.

These metrics were computed for both the training and testing datasets to assess model generalization. Lower MAE and RMSE values and higher R^2 scores indicate better predictive performance.

3.6 Experimental Setup
All experiments were conducted in a Google Colab environment using Python 3. Libraries such as NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow, and Keras were utilized for data handling, visualization, and model training. Each model was trained on the preprocessed dataset for 50–100 epochs depending on convergence behavior, using batch normalization and dropout regularization to improve generalization. Model hyperparameters such as learning rate, number of layers, and

hidden units were fine-tuned through iterative testing.

After training, each model's predictions were compared visually using line graphs and scatter plots of actual versus predicted solar power. Random Forest emerged as the top-performing ML model, while GRU achieved the best performance among DL models. The GRU model was further integrated with the Open-Meteo API to enable real-time solar power prediction based on live weather inputs.

Chapter 4 : RESULTS AND DISCUSSION

This chapter presents the experimental results obtained from various Machine Learning (ML) and Deep Learning (DL) models developed for solar power prediction. The performance of each model was analyzed using the evaluation metrics described earlier—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). The comparison of actual versus predicted power values and training curves provides a comprehensive understanding of each model's accuracy, learning behavior, and generalization capability.

4.1 Model Performance Comparison

Multiple ML and DL models were trained on the preprocessed dataset to forecast DC power output. For machine learning models, Random Forest (RF) and XGBoost demonstrated superior performance compared to Linear Regression, owing to their ability to capture nonlinear dependencies between meteorological features and power generation. In deep learning models, the GRU achieved the best results, followed closely by LSTM and ANN.

The comparative performance of all implemented models based on MAE, RMSE, and R^2 scores is presented in Table 4.1. The Random Forest model achieved a high R^2 value of 0.9864, while the GRU model achieved the lowest MAE among deep learning approaches, indicating precise forecasting ability.

Model	MAE (kW)	RMSE (kW)	R^2 Score
Random Forest	168.38	469.05	0.9864
XGBoost	187.36	472.26	0.9862
Linear Regression	268.50	567.03	0.9801
GRU	195.27	557.08	0.9779
LSTM	195.79	556.98	0.9779
ANN	211.31	527.86	0.9828
1D-CNN	236.20	607.46	0.9737
Transformer	506.31	795.55	0.9548

TABLE I: Performance Summary of Machine Learning Models

4.2 Analysis of Machine Learning Models

Machine learning algorithms provided robust baseline predictions. The Random Forest model, in particular, captured complex nonlinear relationships effectively and showed minimal overfitting. XGBoost also produced competitive results due to its gradient boosting mechanism, though it required careful hyperparameter tuning. Linear Regression, on the other hand, underperformed as it could not adapt to nonlinearity in solar data.

The scatter plots of actual versus predicted power values validate these findings. In the Random Forest model (Figure 4.1) and XGBoost model (Figure 4.2), the data points closely align with the diagonal line, indicating a strong correlation between actual and predicted outputs. Additionally, the time-series comparison of actual and predicted DC power (Figure 4.3) illustrates the Random Forest model's ability to follow the real generation trend with minimal deviation.

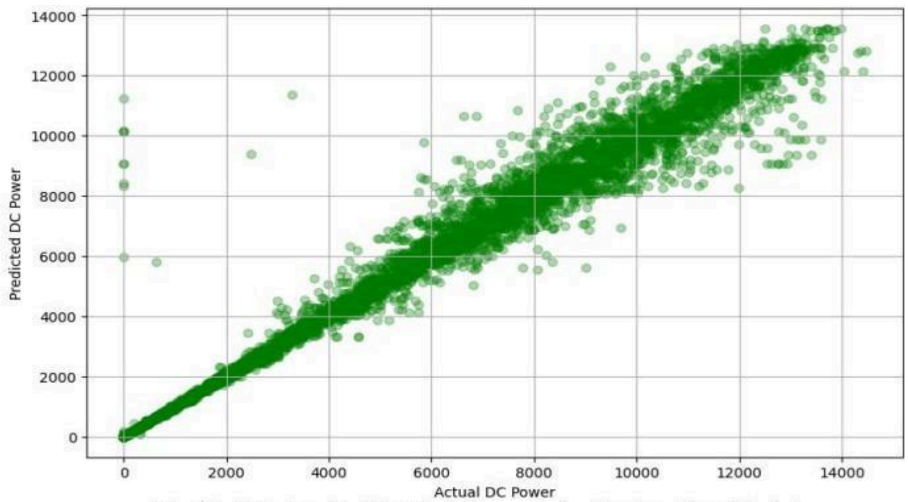


Fig. 4.1: Actual vs Predicted Solar Power using Random Forest Model

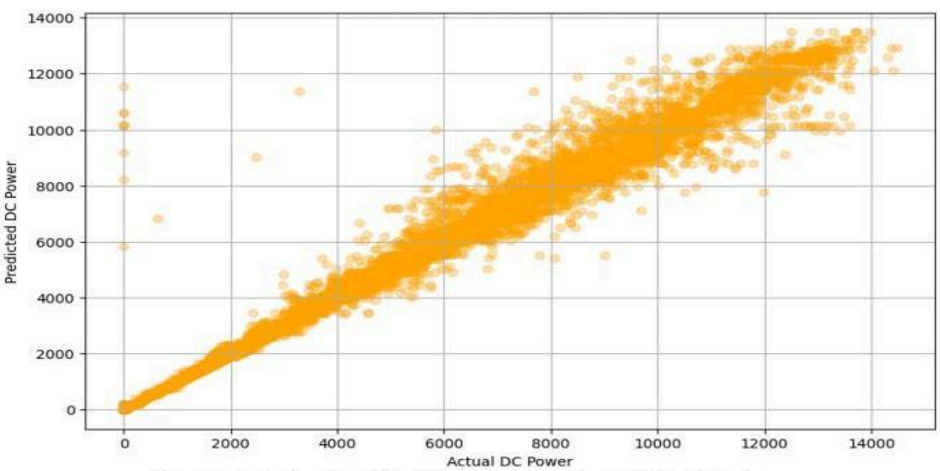


Fig. 4.2: Actual vs Predicted Solar Power using XGBoost Mode

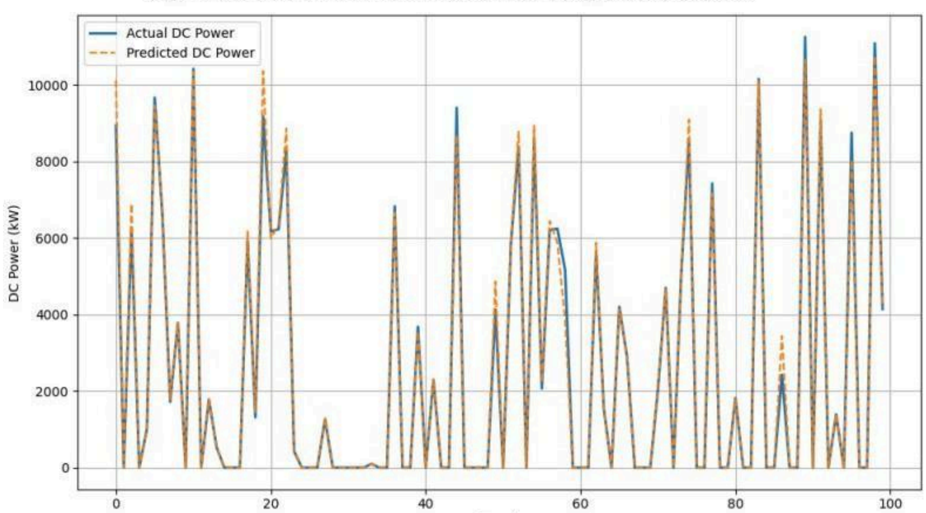


Fig. 4.3: Comparison of Actual and Predicted DC Power on Test Data

4.3 Analysis of Deep Learning Models

Deep learning models were employed to capture temporal dependencies within the time-series data. The Artificial Neural Network (ANN) provided a reasonable baseline, while Recurrent Neural Network (RNN)-based architectures such as LSTM and GRU offered significantly improved accuracy due to their sequential memory units. The GRU model outperformed others in terms of faster convergence and lower error rates.

Figures 4.4 to 4.8 display the training and validation MAE curves for all DL models. The gradual reduction in loss and minimal gap between training and validation errors confirm that the models learned effectively without overfitting. Among these, the GRU model exhibited the smoothest convergence curve and achieved the best balance between accuracy and efficiency. The Transf

4.4 Performance Metrics

Table 4.1 presents a comparative summary of model performance. Among the ML models, Random Forest achieved an MAE of 168.38 kW and RMSE of 469.05 kW, while XGBoost achieved MAE 187.36 kW and RMSE 472.26 kW. In DL models, GRU recorded an MAE of approximately 195.27 kW with strong generalization, outperforming LSTM (MAE ≈ 195.79 kW) and ANN (MAE ≈ 211.31 kW). These results demonstrate that both ensemble learning and sequential neural networks are well-suited for solar power forecasting tasks.

The evaluation results confirm that data-driven approaches can effectively model the nonlinear and time-dependent behavior of solar generation. The correlation between weather conditions and power output is accurately captured by the trained models, enabling reliable short-term predictions.

4.5 Key Observations and Insights

From the comparative study, several observations were made:

- Machine learning models such as Random Forest are fast, interpretable, and deliver highly accurate results with minimal parameter tuning.
- Deep learning models, particularly GRU, achieve superior prediction accuracy for time-dependent data but require longer training times and computational power.
- The combination of environmental and temporal features significantly improves the accuracy of all models.
- The integration of the GRU model with the Open-Meteo API successfully enabled real-time solar power forecasting based on live weather inputs.

These findings highlight the effectiveness of AI-driven approaches in renewable energy forecasting. The GRU model, in particular, demonstrates strong potential for deployment in real-world energy management systems due to its accuracy and adaptability.

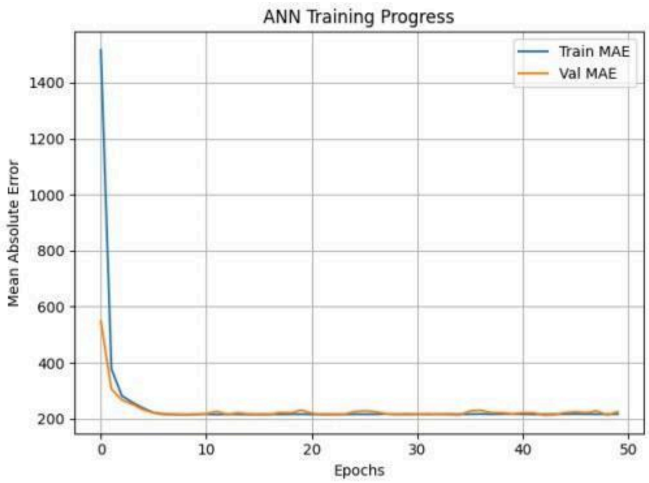


Fig. 4.4: Training and Validation MAE Curve for ANN Model

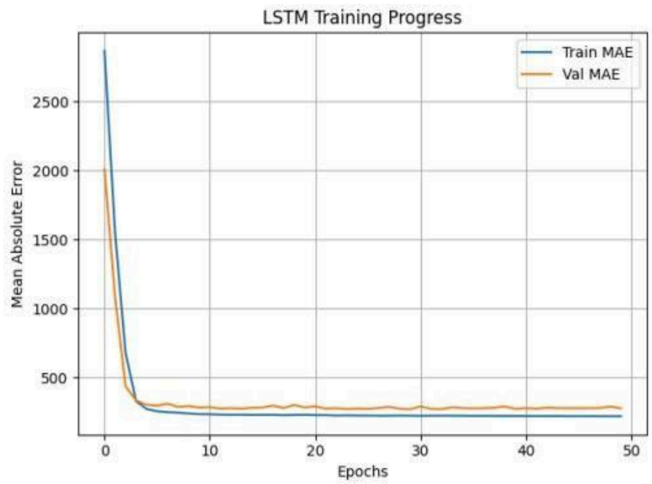


Fig. 4.5: Training and Validation MAE Curve for LSTM Model

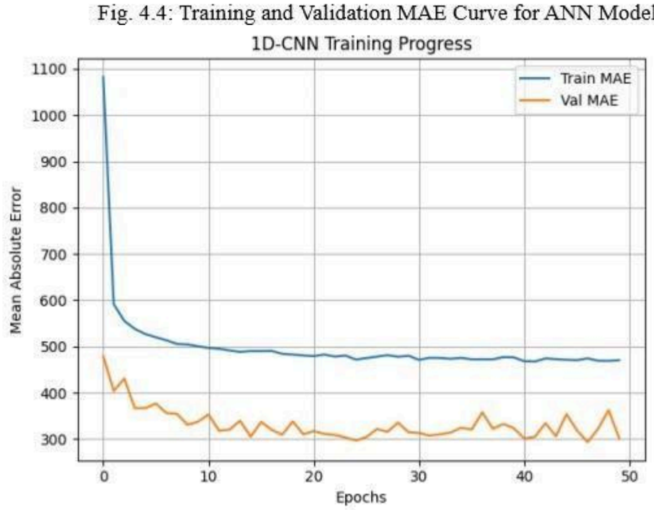


Fig. 4.6: Training and Validation MAE Curve for 1D-CNN Model

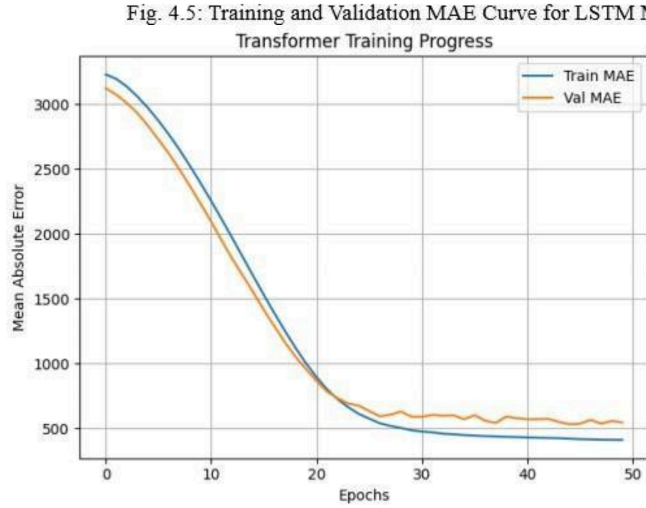


Fig. 4.7: Training and Validation MAE Curve for Transformer Model

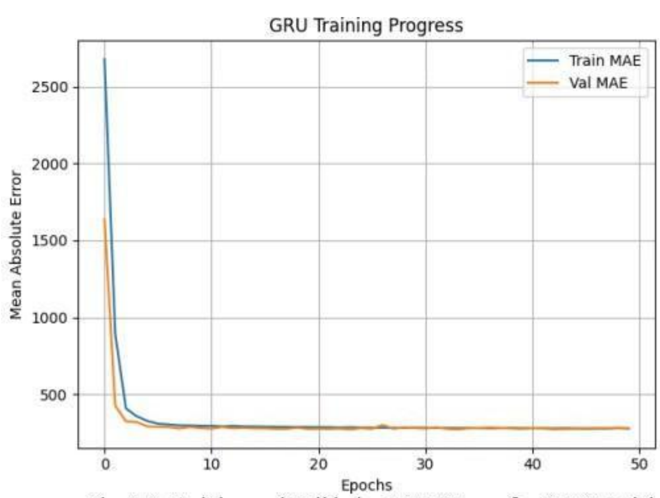


Fig. 4.8: Training and Validation MAE Curve for GRU Model

Chapter 5 : CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The project successfully demonstrated the effectiveness of Artificial Intelligence (AI) and data-driven approaches in predicting solar power generation. Various Machine Learning (ML) and Deep Learning (DL) models were implemented, analyzed, and compared using real-world data obtained from the Kaggle Solar Power Generation dataset.

From the experimental results, it was observed that ensemble-based ML models such as Random Forest performed exceptionally well in capturing nonlinear dependencies between meteorological parameters and power output. Among the DL models, the Gated Recurrent Unit (GRU) achieved the best overall performance with the lowest MAE and RMSE values, reflecting its strong ability to learn temporal patterns in time-series data.

The integration of the GRU model with the Open-Meteo API enabled real-time solar power forecasting using live weather data, demonstrating the project's applicability to practical renewable energy systems. Overall, the study highlights that AI-based predictive modeling can play a significant role in enhancing the reliability and efficiency of solar energy production planning.

5.2 Future Scope

While the results are promising, there remain several opportunities for further improvement and expansion:

- Incorporation of additional meteorological variables and satellite imagery data to enhance accuracy.
- Development of hybrid or ensemble models combining ML and DL architectures for better robustness.
- Integration of Explainable AI (XAI) techniques to improve model transparency and interpretability.
- Deployment of the trained model as a cloud-based API or Progressive Web App (PWA) for broader accessibility.
- Implementation of lightweight edge-compatible versions for on-site solar systems.

With these future advancements, the system can evolve into a scalable and intelligent solar forecasting solution capable of supporting large-scale renewable energy management and smart grid applications.

REFERENCES

[1] Kaggle, "Solar Power Generation Data,"
Available: https://www.kaggle.com/datasets/anikannal/solar-power-generation-data.
[Accessed: Oct. 2024].

[2] T. Pedro, R. L. Tavares, and J. B. Soares,
"Short-Term Solar Power Forecasting Using Machine Learning Techniques,"
Energy Reports, vol. 7, pp. 679–689, 2021.

[3] S. S. Soman, H. Zareipour, O. Malik, and P. Mandal,
"A Review of Wind Power and Solar Power Forecasting Methods,"
Renewable and Sustainable Energy Reviews, vol. 16, no. 7, pp. 4606–4619, 2012.

[4] A. Nespoli, G. Manzolini, and D. Moser,
"Machine Learning for Solar Energy Forecasting: A Review,"
Renewable Energy, vol. 201, pp. 1205–1223, 2023.

[5] J. Brownlee, "Deep Learning for Time Series Forecasting,"
Machine Learning Mastery, 2020.

[6] T. Lim, A. R. Jang, and Y. J. Kim,
"Deep Learning Models for Short-Term Solar Power Forecasting,"
IEEE Access, vol. 9, pp. 95653–95666, 2021.

[7] Open-Meteo, "Free Weather API for
Forecasting," Available: https://open-meteo.com/.
[Accessed: Oct. 2024].

[8] F. Chollet et al., "Keras: Deep Learning
Framework," Available: https://keras.io/.
[Accessed: Sept. 2024].

[9] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python,"
Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[10] TensorFlow, "An Open Source Machine Learning
Platform," Available: https://www.tensorflow.org/.
[Accessed: Sept. 2024].