Machine Learning Driven Cloud Detection Using Solar Power Production

David Mueller
University of the Pacific, School of Engineering and Computer Science
Stockton, CA, USA
dmueller@pacific.edu
Umith Chandra, Vivek Pallipuram, Houman Kamran, Tapadhir Das
University of the Pacific, School of Engineering and Computer Science

Abstract: This study presents a machine learning (ML) model designed to classify cloud cover using solar power production data. The model was trained using the ratio of actual power output to simulated clear-sky output, generated using pvlib-python, and is trained using cloud class labels derived from local Meteorological Aerodrome Reports (METARs). The model achieved 82.4% accuracy, demonstrating robust performance in detecting cloud conditions categorized as clear, few, scattered, broken, and overcast. This approach enhances solar forecasting, supports grid stability, and reduces the cost of solar resource monitoring.

Keywords: cloud cover detection, solar power production, machine learning, neural network

1. Introduction

Accurate solar forecasting is critical for grid stability and renewable energy integration. Cloud cover introduces significant variability in solar energy production, making real-time detection essential. Traditional cloud monitoring tools such as all-sky cameras, ceilometers, and sky imagers are expensive, see Table 1, and difficult to deploy at scale. Traditional cloud detection systems are installed regionally, at a nearby airport for example, whereas our solution allows for localized insights into cloud coverage. This study contributes a scalable, low-cost solution by classifying cloud cover using machine learning applied to photovoltaic (PV) system production data and open-source modeling tools. The benefits include reduced cost, portability, and scalability across various solar installation sizes, while the tradeoffs include no cloud base height, thickness, and motion information – the latter we plan to implement in future work.

Table 1: Comparison of typical cost of traditional cloud detection hardware and our ML-based cloud detection method.

Device	Cost (USD)	Function	Notes
Traditional Tools			
All-Sky Camera	$3,000 – $15,000	Captures images of the sky to identify cloud types	Requires image processing software; sensitive to weather and lens issues
Ceilometer	$6,000 – $30,000	Uses LIDAR to measure cloud base height	Highly accurate but expensive and power-hungry
Sky Imager	$4,000 – $12,000	High-resolution hemispheric sky photos	Advanced models include thermal or IR filters
ML Cloud Classifier			
PV System	Already installed	Power output data source	No added hardware cost
pvlib-python / simulation	Free (open source)	Simulate clear-sky output	Needs metadata (tilt, azimuth, etc.)
ML Cloud Classifier	compute cost	Predict cloud type from power ratio features	Lightweight model, easily deployed in embedded systems

https://doi.org/10.52202/083252-0008

2. **Methodology**

To classify cloud cover using solar power production data, we constructed a novel dataset, as shown in Figure 1, that integrates actual PV system output, simulated clear-sky power data, and cloud classification labels derived from Meteorological Aerodrome Reports (METARs) (Lui, 2014). The actual PV output data was sourced from a 5.35 MW photovoltaic system located at the University of the Pacific (Mueller, 2024). Clear-sky power output was simulated using the open-source pvlib-python library (Anderson, 2023; Holmgren, 2018).

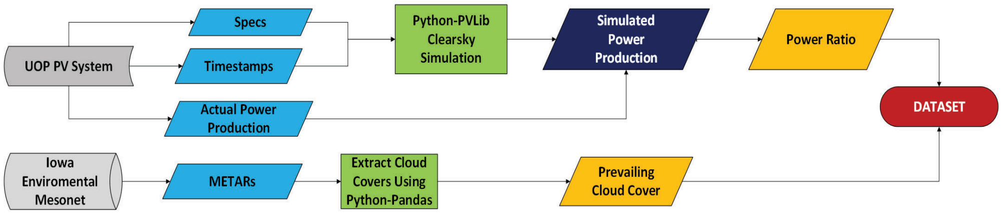
UOP PV System
Specs
Timestamps
Actual Power Production
Python-PVLib Clearsky Simulation
Simulated Power Production
Power Ratio
DATASET
Iowa Enviromental Mesonet
METARs
Extract Cloud Covers Using Python-Pandas
Prevailing Cloud Cover

Figure 1: Flowchart for Machine Learning (ML) Dataset Generation

Hourly cloud cover labels were obtained from the METARs over an 18-month period (July 2022 to December 2023) from the Stockton, CA airport, located near the PV installation. Each hour of PV data was aggregated and labeled with one of five METAR cloud classes: Clear (CLR), Few (FEW), Scattered (SCT), Broken (BKN), and Overcast (OVC).

Feature engineering, as shown in Figure 2 (left), involved computing the ratio of actual PV production power output to simulated clear-sky production power output at 15-minute intervals, resulting in four power ratio features per hour. To capture temporal dynamics, we added four sinusoidal features representing the hour of day and day of year. These features help the model account for seasonal and diurnal variations in solar irradiance. The final input vector for each hour consisted of eight features.

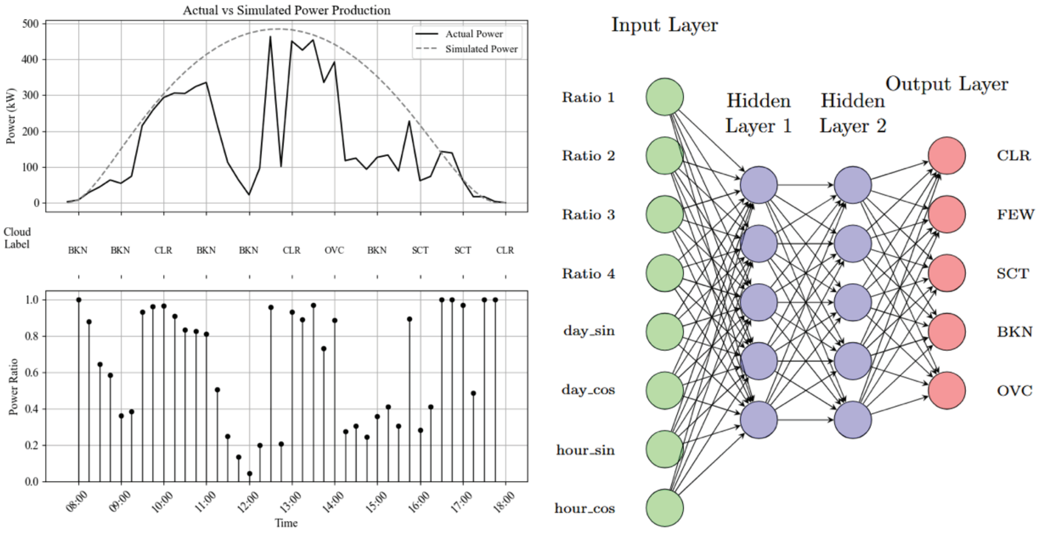

Figure 2: (left) Example of one day of labeled power production and power ratios for November 9th, 2022.

(right) Schematic of the neural network architecture with input features and output labels.

We implemented a feedforward neural network (NN), as shown in Figure 2 (right), with two hidden layers, each containing five nodes, and an output layer with five nodes corresponding to the cloud classes. The model was trained using labeled hourly blocks of PV data, with regularization (λ = 2) applied to prevent overfitting and improve generalization.

3. Results

The machine learning model achieved an average classification accuracy of 82.4% across the five cloud classes, with an F1-score exceeding 80%. The model demonstrated strong performance in distinguishing between clear (CLR) and overcast (OVC) conditions, which exhibit distinct power output profiles. However, intermediate classes such as FEW, SCT, and BKN showed greater overlap in power ratios, making them more challenging to classify.

Incorporating temporal features resulted in a nearly 7% improvement in classification accuracy. Temporal features were calculated using:

Day of Year Temporal Features

$$Day_{sin} = \sin sin\left(2\pi\frac{day}{365}\right) \quad Day_{cos} = \cos cos\left(2\pi\frac{day}{365}\right)$$

Hour of Day Temporal Features

$$Hour_{sin} = \sin sin\left(2\pi\frac{hour}{24}\right) \qquad Hour_{cos} = \cos cos\left(2\pi\frac{hour}{24}\right)$$

This suggests that time-based patterns play a significant role in cloud cover detection. Additionally, the variance of power ratios within each hour proved to be a useful feature for differentiating between dense and scattered cloud formations. Box-and-whisker plots of average and variance power ratios by cloud class revealed clear separability for CLR and OVC, as shown in Figure 3, while intermediate classes exhibited wider distributions.

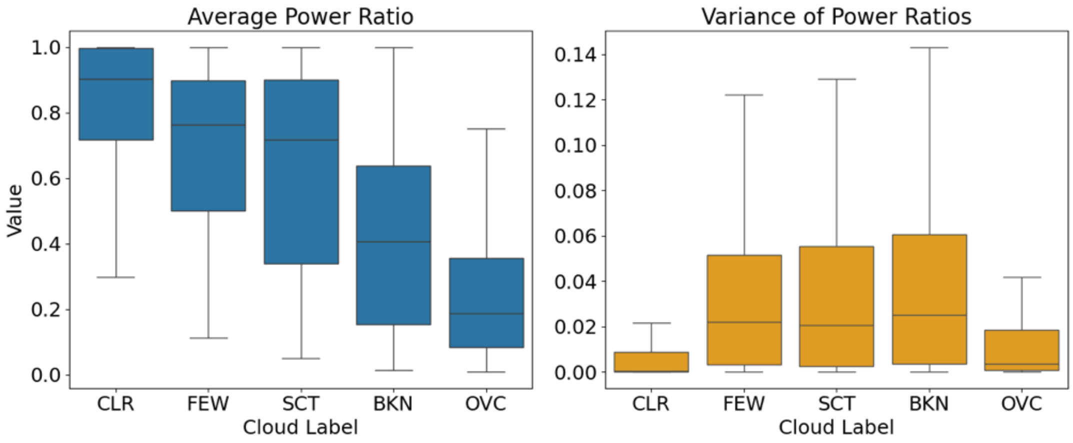

Figure 3: Box and Whisker Plot for Average and Variance Power Ratios by Cloud Labels

4. Conclusions and Future Work

This study demonstrates that machine learning models can effectively classify cloud cover using only solar power production data and simulated clear-sky output. By leveraging open-source tools and existing PV infrastructure, this approach potentially offers a scalable and cost-effective solution for cloud detection in solar forecasting applications. The model's >82% cloud class prediction accuracy and robustness make it suitable for deployment in both small-scale rooftop systems and large grid-connected solar farms. Moreover, the use of temporal and statistical features enhances the model's ability to capture complex cloud dynamics, supporting improved forecasting and grid stability.

Future research will focus on extending the model to forecast power loss based on predicted cloud class, enabling more proactive energy management. By using power production from multiple inverters along with precise mapping and temporal information, we are working to enhance our model with cloud motion information. Additionally, cloud classification could be used to diagnose performance anomalies or identify maintenance needs in PV systems. Integration with real-time energy management platforms will be explored to support dynamic grid operations and demand response strategies.

Further improvements may include the use of ensemble learning methods, such as gradient boosting or random forests, to enhance classification performance. Finally, expanding the dataset to include multiple geographic locations and PV system configurations would improve generalizability and support broader deployment.

5. Acknowledgments

This work was supported by a strategic investment grant from the School of Engineering and Computer Science at the University of the Pacific. System level information and data was kindly provided by the Office of Sustainability at the University of the Pacific. METAR data sourced from National Oceanic and Atmospheric Administration (NOAA) and Iowa State University.

6. References

Anderson, K., Hansen, C., Holmgren, W., Jensen, A., Mikofski, M., & Driesse, A. (2023). pvlib-python: 2023 Project Update. Journal of Open Source Software, 8(92), 5994. https://doi.org/10.21105/joss.05994

Holmgren, W. F., Hansen, C. W., & Mikofski, M. A. (2018). pvlib-python: A Python Package for Modeling Solar Energy Systems. Journal of Open Source Software, 3(29), 884. https://doi.org/10.21105/joss.00884

Liu, M. C. M. (2014). Complete Decoding and Reporting of Aviation Routine Weather Reports (METARs). NASA/TM-2014-218385.

Mueller, D., & Notra, H. (2024). Modeling and Production Performance Analysis of a Campus 5 MW Solar Installation in the California San Joaquin Valley. American Solar Energy Society. https://doi.org/10.52202/077496-0008