1 Introduction

Partial shading is a critical challenge for photovoltaic (PV) systems, causing severe power loss and creating damaging hotspots [1]. The economic impact is substantial, as even minor shading can drastically reduce performance. Therefore, accurate real-time shading quantification is essential for preventing system failures and maximizing energy yield [2], [3]. However, existing approaches either rely on expensive sensors or oversimplified models that fail to capture the complex, non-linear dynamics of PV systems. Current solutions for detecting partial shading are often inadequate. Hardwarebased approaches, such as thermal imaging or additional irradiance sensors [4], provide accuracy but at a high cost and maintenance effort, often failing to capture shading variations across large installations. Conversely, simple analytical software methods make too many assumptions, leading to failures in dynamic conditions and false positives from environmental fluctuations [5]. Accurately predicting shading percentage is a complex task due to the intricate, non-linear interactions between numerous factors. External conditions like solar irradiance and ambient temperature, along with internal system parameters such as panel configuration (series vs. parallel), combine to create unique electrical signatures. Extracting meaningful patterns from the resulting currentvoltage (I-V) measurements requires sophisticated analysis techniques capable of handling these high-dimensional, non-linear relationships [6]. Machine learning offers a promising data-driven solution, as its algorithms excel at automatically learning the complex, nonlinear patterns present in highdimensional datasets where traditional models fail. However, the effectiveness of machine learning relies on large and comprehensive data to learn these underlying relationships. The large-scale, curated dataset proposed in this work is valuable for developing and validating machine learning models for tasks such as shading prediction, power optimization, and system configuration. In addition, we present an end-to-end machine learning approach to predict shading percentage in PV systems using regression. The proposed methodology uses a MATLAB Simulink tool [7] to generate a comprehensive data set of more than 24 million data points by systematically varying irradiance (100-1000 W/m2), temperature (10-50°C), series-parallel configurations and shading levels (0-100%). This large-scale dataset, specifically designed for shading analysis, provides a robust foundation for training and validating multiple regression algorithms. 1.1 Research Objectives: The main objectives of this research are multifaceted and attempt to address both theoretical and practical aspects of shading percent prediction in solar PV cells: Modeling the Impact of Partial Shading on Photovoltaic Panels 3 1. Objective 1: Collection of High-quality Dataset: One of the most significant objectives is to curate a comprehensive dataset of more than 24 million data points with the help of MATLAB simulation software that contains the diverse combinations of irradiance levels, temperature conditions, system configurations, and electrical parameters to create a solid foundation not only for shading prediction but also for other tasks such as power optimization and predictions, proactive measure, etc. 2. Objective 2: Benchmark Introduction: Performance and testing with multiple regression models to accurately predict shading using environmental and electrical parameters as input features, with the main focus on models suitable for real-time deployment in resource-constrained environments. 3. Objective 3: Feature Engineering and Importance: To analyze the importance of the features in the dataset that would help to reduce complexity and also introduce new features if required, that might provide further insight about the dataset and the interactions of the features within the dataset.

2 Literature Review Solar energy is essential for the global transition to clean and renewable energy sources, as it offers an inexhaustible supply of clean energy. Advancements in solar technology have significantly improved efficiency and reduced costs, making it a viable and sustainable alternative to traditional energy sources. Analyzing the effectiveness of these PV cells and modules lets us optimize them for energy yield and long-term value. This reveals structural or manufacturing defects or early degradation. International programs such as IEA-PVPS Task 13 [8], [9] compile real data across differing conditions to improve future models and reliability. Among commonly studied configurations, 60-cell solar panels are widely used in residential applications due to their size, cost, and compatibility with typical household energy demands. Several studies since 2016 have explored individual performance factors—such as the linear impact of irradiance on PV output [10] and the detrimental effects of high operating temperatures [11, 12]. The effects of partial shading have also been characterized in isolation [3]. However, the combined influence of these factors—irradiance, temperature, and shading—especially across varying system configurations, remains underexplored. Most studies isolate these variables, limiting insights into their interdependent behavior under real-world conditions. Furthermore, there is limited research addressing adaptive or reconfigurable panel architectures that could improve performance under such dynamic environmental influences, particularly for standard 60-cell residential panels. Most studies isolate these variables, limiting insights into their interdependent behavior under real-world conditions. Furthermore, there is limited research addressing adaptive or reconfigurable panel architectures that could improve performance under such dynamic environmental influences, particularly for standard 60-cell residential panels. 4 J. Olivares et al.
XXXXX

3 Proposed Methodology This section outlines our approach for generating a large-scale dataset to analyze the effects of shading on PV cells. The dataset includes key features such as temperature, irradiance, shading percentage, series-parallel configuration, voltage, current, and power. With this comprehensive set of inputs, regression models can be trained to predict the shading class of PV cells. 3.1 Simulation Setup in MATLAB A MATLAB script (as illustrated in Figure 1), developed using key mathematical models based on reference [7], allows users to modify essential parameters and conditions, including series-parallel configurations, reference open-circuit voltage (Voc), short-circuit current (Isc), maximum power voltage (Vmp), maximum power current (Imp), temperature, irradiance, shadow temperature, and shadow irradiance. Fig. 1. Simulink Simulation to Illustrate the I-V and P-V Module Output Characteristics [Source: [7]] 3.2 Dataset Generation The number of PV modules and strings is specified to display a 60-cell solar panel array, typically used in California residential homes, in which there are 12 different possible configurations. Temperature, irradiance, shadow temperature, and shadow irradiance are specified for each configuration. The temperature Modeling the Impact of Partial Shading on Photovoltaic Panels 5 coefficient and the shadow temperature coefficient are the same, while the shadow irradiance is 80% of the irradiance. In this work, the experiments are conducted using twelve configurations (1s50p, 2s30p, 3s20p, 4s15p, 5s12p, 6s10p, 10s6p, 12s5p, 15s4p,

20s3p, 30s2p, and 60s1p) where the first number denotes the number of cells in series and the second number denotes the number of cells in parallel; nine temperature values (10°C, 15°C, 20°C, 25°C, 30°C, 35°C, 40°C, 45°C, 50°C), five irradiance values (1000 W/m2 , 800 W/m2 , 600 W/m2 , 400 W/m2 , 200 W/m2), and eleven shading percentages (0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%). Due to a software limitation, 1s60p is replaced by 1s50p. Each dataset consistently reflect a specific configuration at different temperatures, irradiances, and shading percentages. As a result, there are approximately 24 million datapoints, each consisting of voltage, current, and power values (4096 datapoints * 5940 datasets = ~24 million datapoints). 3.3 Regression-based Predictive Modeling All features and corresponding data are compiled into separate CSV files for each temperature value, resulting in nine CSV files used to train machine learning model. A supervised learning approach is employed, using regression analysis to predict the shading percentage associated with the maximum power output in each combined dataset. The regression model is trained using a comprehensive set of input variables (features), derived from our merged dataset. These features include key operational and environmental parameters such as series, parallel, temperature, irradiance, shading, voltage, and current. The model's core task is to learn the intricate mathematical relationship between these input features and the target output value. 4 Experimental Methodology The experimental methodology of this study is designed to systematically evaluate the efficacy of machine learning models in predicting shading percentage on PV systems. This section details the process of generating a comprehensive synthetic dataset, the feature engineering techniques employed, the selection and training of regression models, and the metrics used for performance evaluation. 4.1 Data Generation and Collection A core contribution of this research is the creation of a large-scale, high-fidelity dataset for shading analysis. To overcome the logistical and financial constraints of real-world data collection, a simulation-based approach is executed using a custom MATLAB script. This approach provided precise control over all relevant variables, enabling the generation of a comprehensive and balanced dataset essential for model training. 6 J. Olivares et al. 4.2 Simulation Model and Environment The dataset is generated using a custom MATLAB script implementing a physicsbased vertical shading model. The underlying methodology for this script is adapted from the modeling principles for studying partial shading effects presented by Fezzani et al. in [7]. The core of the simulation is the single-diode equivalent circuit model, which describes the current-voltage (I-V) relationship of a PV cell with the following equation [13]: I = Iph − I0 exp V + IRs Vt − 1 − V + IRs Rsh (1) where Iph is the photocurrent, I0 is the diode reverse saturation current, Rs and Rsh the series and shunt resistances, and Vt is the thermal voltage [6]. Unlike commercial black-box software, this script provides a transparent and adaptable environment. The model's methodology is twofold: 1. Parameter Derivation: First, the electrical characteristics of a single, unshaded module are calculated. This is done by using full-array performance data at a known reference irradiance and temperature, then scaling it according to the specified array configuration, with (Ns) modules in series and (Np) strings in parallel. 2. Performance Translation: Second, the script employs a validated physics model, incorporating temperature coefficients for voltage (β_Voc) and current (α_Isc), to accurately translate these reference module parameters to any new operational conditions of irradiance (Gop) and temperature (Top). This translation is governed by the following key relations: Isc_op = Gop Gref Isc_ref + αIsc(Top − Tref) (2) Voc_op = Voc_ref + βV oc(Top − Tref) + Vth ln Gop

Gref (3) This approach enables the simulation of distinct "sunny" and "shaded" modules within a single series string. Following the method outlined by Fezzani et al., the simulation engine calculates the voltage contribution of each module in the string. It explicitly models the activation of bypass diodes under reverse-bias conditions, which is essential for accurately reproducing the complex, multi-peaked P–V curves observed in partially shaded arrays [3]. 4.3 Parameter Space Definition To ensure the dataset captures a wide spectrum of real-world operating conditions, a multi-dimensional parameter space is defined for a 60-cell solar panel array. The data generation process involved systematically iterating through a nested loop of five key parameters: solar irradiance, ambient temperature, shading percentage, and series-parallel. The specific ranges and steps, consistent with the research objectives, are detailed in Table 1. Modeling the Impact of Partial Shading on Photovoltaic Panels 7 Table 1. Description of Features Parameter Description Range / Values Step Solar Irradiance (G) The incident solar irradiance on the PV panel in Watts per Square Meter (W/m2) 200 – 1000 W/m2 200 W/m2 Ambient Temperature (T) The operational temperature at which the PV panel operates in Celsius (°C) 10 – 50°C 5°C Shading Percentage The percentage of series-connected modules affected by shading 0% – 100% 10% Series The total number of PV modules connected in series in one string 1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60 N/A Parallel The total number of PV strings connected in parallel 50, 30, 20, 15, 12, 10, 6, 5, 4, 3, 2, 1 N/A 4.4 Data Consolidation and Processing The MATLAB script generate a large number of individual .csv files, each containing the discrete I-V curve from a single simulation run, organized into temperature-specific folders. To consolidate these outputs into unified datasets, a custom Python post-processing script is developed. The script recursively traversed the directory structure, loading each .csv file. It employs regular expressions to extract encoded input parameters (series-parallel configuration, temperature, irradiance, shading) from the filenames. These parameters are appended as additional columns to the corresponding I-V curve data. All processed dataframes are concatenated using the pandas library into a single master file per temperature. This workflow effectively merged thousands of discrete simulation results into consolidated datasets, preserving raw I-V data alongside associated simulation parameters for comprehensive analysis. 4.5 Feature Extraction and Dataset Finalization With the raw I-V data consolidated, the next step involves extracting key electrical features essential for the machine learning models. Unique simulation runs programmatically segment the dataset, and for each I-V curve, the following performance metrics are computed: • Voltage (Vmp), and Current (Imp): Determined by finding the maximum value in the power column and its corresponding voltage and current. 8 J. Olivares et al. • Open-Circuit Voltage (Voc): The voltage at which the current is zero. • Short-Circuit Current (Isc): The current at which the voltage is zero. The final, consolidated dataset for this study is created by aggregating these extracted features. Each row in the final dataset represents a unique operational state and contains the input parameters that defined the simulation (Irradiance, Temperature, Shading Percentage, Series-Parallel). This structured process resulted in a final dataset of over 24 million entries, providing a comprehensive foundation for training and validating machine learning regression models. 5 Data Preparation and Modeling Workflow This section outlines the data preprocessing pipeline employed in this study, including the steps taken to structure, engineer, and prepare the data for modeling. 5.1 Data Structure and Initial Analysis The dataset comprises 5,269 individual I-V curves generated from over 100 unique PV system configurations, varying in temperature, irradiance, shading percentage, and series-parallel panel arrangements. Each configuration systematically introduces variation in electrical topology while maintaining controlled environmental conditions. Specifically, 12 distinct series-parallel configurations are derived from 60 panels, including setups such as the low-voltage, high-current 1s50p configuration (one in series, 50 in parallel). Each configuration is tested under 11 discrete shading levels, ranging from 0% to 100% in 10% increments. This results in a comprehensive dataset that captures the full operational behavior of PV systems under diverse partial shading scenarios. Environmental parameters are systematically varied across all configurations to emulate realistic operating conditions. Temperature ranged from 10°C to 50°C in 5°C increments (i.e., 10, 15, . . . , 50°C), while irradiance levels spanned from 200 W/m2 to 1000 W/m2 in 200 W/m2 steps. This structured variation covers a wide range of operational scenarios, from low-light to standard test conditions. Each configuration is simulated to generate 4,096 I–V data points, capturing the full current-voltage profile from short-circuit to open-circuit conditions. This dense sampling ensures the dataset includes critical regions such as the maximum power point (MPP), knee, and saturation zones—key areas for accurately detecting and quantifying shading effects under varying environmental conditions. Before feature engineering, an extensive data quality assessment is performed to ensure consistency and integrity across all configurations. The dataset exhibits excellent quality, with less than 0.1% missing values. Validation protocols included checks for physical constraints (ensuring voltage ≥ 0 and current ≥ 0), numerical consistency (verifying Power_W = Voltage_V × Current_A), and parameter verification (confirming expected panel counts based on series-parallel Modeling the Impact of Partial Shading on Photovoltaic Panels 9 configuration). Additionally, we cross-validated metadata encoded in filenames against shading levels and verified data integrity during the merging process to ensure correctness throughout the pipeline. Fig. 2. Comparison of Minimum, Average, and Maximum Voltage (V) and Current (A) Statistics Across Different PV Array Configurations. 5.2 Feature Engineering and Normalization A key contribution of the preprocessing phase is the implementation of contextaware feature engineering that captures the inherent electrical behavior of each I–V curve, enabling effective learning across diverse PV configurations. This approach directly addresses the challenge posed by the wide variation in voltage and current ranges across different panel setups, which makes raw electrical measurements unsuitable for machine learning. For instance, a 60s1p configuration can reach up to 2160 V and 3.15 A, whereas a 1s50p configuration operates around 36 V and 188 A that highlights the need for features that are independent of absolute scale. The contextual feature engineering phase begins by using a groupby operation on the filename identifier to analyze each complete I-V curve and extract three fundamental factors: the maximum power point (P_max_curve), the open-circuit voltage (V_oc_curve), and the short-circuit current (I_sc_curve). These factors respectively define the curve's peak power output and its theoretical voltage and current limits. Based on these extracted parameters, a comprehensive normalization technique generates ratio-based features that represent the relative position of each measurement point. The voltage_ratio (Voltage_V / V_oc_curve) and current_ratio (Current_A / I_sc_curve) convert absolute measurements to a normalized 0-1 scale, indicating a point's position between short-circuit and open-circuit states. Similarly, the power_ratio (Power_W / 10 J. Olivares et al. P_max_curve) signifies the point's output relative to its curve's maximum generation capability. This normalization approach offers several advantages. First, it achieves scale invariance by eliminating magnitude differences across configurations, allowing the model to focus on underlying electrical behavior rather than absolute values. Second, it preserves the characteristic shape of I–V curves and shading-induced distortions, ensuring that key shading features remain detectable regardless of the configuration. Third, it enables cross-configuration abstraction by allowing the model to learn generalizable shading patterns independent of specific voltage and current ranges, thereby enhancing its ability to perform on previously unseen configurations. 5.3 Feature Selection and Partitioning The final feature selection phase emphasizes choosing variables that are both predictive and non-redundant. Key electrical features—Voltage_V and Current_A are kept because they directly represent the PV system's electrical behavior. System configuration parameters such as series and parallel counts, temperature, and irradiance are also included to inform the model of the experimental and environmental context, enabling it to learn configuration-specific patterns [14]. The three introduced features, i.e., voltage-ratio, current-ratio, & powerratio, form the core of our normalized feature set, providing a scale-invariant depiction of the electrical behavior that enables cross-configuration learning. These features capture the main electrical characteristics of shading effects and maintain compatibility through the diverse range of panel configurations. Several features are manually excluded from the training feature matrix based on redundancy and potential cases of data leakage. Raw power measurements (Power_W), curve-level parameters (P_max_curve, V_oc_curve, I_sc_curve), and filename identifiers are removed for various corresponding reasons. But essential configuration data is preserved through explicit series, parallel, and environmental parameter features. Modeling the Impact of Partial Shading on Photovoltaic Panels 11 Fig. 3. Feature Importance Scores for Various Parameters 5.4 Model Implementation The predictive model is implemented in Python within a supervised learning framework to estimate shading percentage from the engineered features. We use the XGBoost Regressor, a highly efficient gradient boosting algorithm known for its strong performance and robustness on tabular data [15]. The workflow leverages Pandas for data handling, Scikit-learn for data splitting and evaluation metrics, and the XGBoost library for model training and prediction. To ensure robust training and an unbiased evaluation, we partitioned the feature-engineered dataset using stratified sampling into training (80%), validation (10%), and test (10%) sets. The XGBoost model is trained on the training data, utilizing the validation set for an early stopping mechanism that halts the process after 50 rounds without improvement to prevent overfitting. The model's final performance is assessed on the unseen test set to report the R-squared and Mean Absolute Error metrics. 6 Experimental Results This section presents the results of our study along with a discussion of the key findings. The scatter plot of predicted versus actual shading percentages is shown in Figure 4, illustrating a high concentration of points closely aligned with the "Perfect Prediction" line. This demonstrates a strong positive correlation between 12 J. Olivares et al. the model's outputs and the ground truth data. The XGBoost model achieved an R2 value of 0.937, indicating that approximately 93.7% of the variance in shading percentage is explained by the model. Further validating its accuracy, the model yields a Mean Absolute Error (MAE) of 5.19% and a Root Mean Squared Error (RMSE) of 7.76%. These low error metrics confirm the model's robustness and precision in predicting the degree of shading. The vertical clustering of data points visible in the plot corresponds to the discrete 10% increments in which the actual shading data is recorded. The model, being a regressor, generates continuous predictions that are distributed around these true values, which is the expected and desired behavior for this type of analysis.

Fig. 4. Predicted vs. Actual Shading Percentage

7 Conclusion

This research successfully addresses the critical challenge of quantifying partial shading in photovoltaic (PV) systems through a comprehensive, data-driven methodology. We generated an extensive dataset of over 24 million data points via systematic simulation, covering a broad spectrum of operational conditions. Our novel context-aware feature engineering, which normalizes voltage, current, and power, enables a single machine learning model to generalize effectively across diverse PV array configurations. The resulting XGBoost regression model achieves strong predictive performance, with an R2 of 0.937 and a Mean Absolute Error of 5.19, demonstrating that properly processed electrical I–V curve data alone can accurately quantify shading without needing costly specialized hardware. Our approach offers a practical tool for optimizing residential PV systems.

8 Future Work

While the current model demonstrates strong predictive performance, there are several avenues for future research that could enhance its capabilities and broaden its applicability. Our future efforts will concentrate on enriching the feature set with non-linear characteristics. The current model relies on features that capture linear relationships within the data. However, the physical phenomena influencing solar panel output, such as the interplay between shading, temperature, and irradiance, are inherently non-linear. To better capture these complex interactions, our next step is to incorporate non-linear features into the dataset. By engineering new features, such as polynomial terms or interaction features derived from the existing inputs (e.g., voltage_ratio * current_ratio), we can provide the machine learning model with a more comprehensive representation of the underlying physics. This will allow the model to learn more intricate and generalized patterns that are not apparent with linear features alone. We hypothesize that this will reduce the remaining prediction errors and improve the model's robustness, particularly in accurately predicting shading under nuanced or rapidly changing environmental conditions [16].

Acknowledgements

The authors are grateful to the Undergraduate Research Opportunity Center (UROC) at California State University, Fullerton (CSUF) for the monetary and professional development support to the undergraduate students and faculty mentors involved in this work. Additionally, we are grateful to Southern California Edison for providing financial support through the Edison Sustainability Grant, which enabled the creation of the dataset utilized in this study. The authors also appreciate the Instructionally Related Activities (IRA) for funding the registration fee associated with submitting this work.

References

1. A. DJALAB, N. BESSOUS, M. M. REZAOUI, and I. MERZOUK, "Study of the effects of partial shading on pv array," in 2018 International Conference on Communications and Electrical Engineering (ICCEE), 2018, pp. 1–5.
2. H. Patel and V. Agarwal, "Matlab-based modeling to study the effects of partial shading on pv array characteristics," IEEE transactions on energy conversion, vol. 23, no. 1, pp. 302–310, 2008.
3. C. Deline, J. Meydbray, and M. Donovan, "Photovoltaic shading testbed for module-level power electronics: 2016 performance data update," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2016.
4. J. A. Tsanakas, L. Ha, and C. Buerhop, "Faults and infrared thermographic diagnosis in operating c-si photovoltaic modules: A review of research and future challenges," Renewable and sustainable energy reviews, vol. 62, pp. 695–709, 2016.

5. F. Belhachat and C. Larbes, "A review of global maximum power point tracking techniques of photovoltaic system under partial shading conditions," Renewable and Sustainable Energy Reviews, vol. 92, pp. 513–553,

2018. 6. M. G. Villalva, J. R. Gazoli, and E. Ruppert Filho, “Comprehensive approach to modeling and simulation of photovoltaic arrays,” IEEE Transactions on power electronics, vol. 24, no. 5, pp. 1198–1208, 2009. 7. A. Fezzani, I. H. Mahammed, and S. Said, “Matlab-based modeling of shading effects in photovoltaic arrays,” in 2014 15th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA). IEEE, 2014, pp. 781–787. 8. B. Meyers and M. Mikofski, “Accurate modeling of partially shaded pv arrays,” in 2017 IEEE 44th Photovoltaic Specialist Conference (PVSC). IEEE, 2017, pp. 3354–3359. 9. International Energy Agency Photovoltaic Power Systems Programme (IEA-PVPS), “Task 13: Performance, operation and reliability of photovoltaic systems,”
https://iea-pvps.org/research-tasks/
performance-operation-and-reliability-of-photovoltaic-systems/, 2025, last accessed 12 July 2025. 10. J. Omoriare, E. Ogherohwo, and J. Zhimwang, “Investigating the influence of solar irradiance variability on the output power of photovoltaic (pv) systems in akure, nigeria,” World Journal of Applied Science & Technology, vol. 16, no. 1, pp. 18–22, 2024. 11. P. Sharma, P. Goyal et al., “Analysing the effects of solar insolation and temperature on pv cell characteristics,” Materials Today: Proceedings, vol. 45, pp. 5539– 5543, 2021. 12. N. Zaini, M. Ab Kadir, M. Izadi, N. Ahmad, M. Radzi, and N. Azis, “The effect of temperature on a mono-crystalline solar pv panel,” in 2015 IEEE Conference on Energy Conversion (CENCON). IEEE, 2015, pp. 249–253. 13. R. Mahto and R. John, “Modeling of photovoltaic module,” in Solar Cells: Theory, Materials and Recent Advances. IntechOpen, 2021, pp. 19–28, published 22 September 2021. 14. F. Suliman, F. Anayi, and M. Packianather, “Electrical faults analysis and detection in photovoltaic arrays based on machine learning classifiers,” Sustainability, vol. 16, no. 3, p. 1102, 2024. 15. T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 785–794. 16. J. Devaraj, R. Madurai Elavarasan, G. Shafiullah, T. Jamal, and I. Khan, “A holistic review on energy forecasting using big data and deep learning models,” International journal of energy research, vol. 45, no. 9, pp. 13 489–13 530, 2021.