ISSN: 2349-6002

AI Solar Dust Accumulation Predictor Using Operational Data

MS.J. Jesimaagnasrajamalar[1], AP Tamizh Kumaran M[2], Ragu S[3], Santhosh Kumar S[4]
[1,2,3,4] *Department of Computer Science and Engineering Surya Group of Institution, Vikravandi, TamilNadu, India*
doi.org/10.64643/IJIRTV12I11-201048-459

***Abstract*—Dust accumulation on photovoltaic (PV) solar panels significantly degrades energy output, causing losses of 20–40% in high-irradiance regions. This paper presents SolarSense, a machine learning-powered web application that detects and quantifies dust-induced performance degradation using standard operational sensor data. An XGBoost regression model is trained on 17,401 real-world 15-minute interval records to predict the Performance Ratio (PR) of solar panels under clean conditions. The predicted PR is compared against actual energy generation to compute energy loss, which is mapped to a three-level dust severity classification (Low, Medium, High) with actionable maintenance recommendations. The model achieves an R^2 score of 0.9412 and MAE of**
0.031. A Flask-based web interface supports manual data entry and CSV batch prediction modes with optional email alert dispatch. The framework requires no specialised soiling sensors, relying entirely on standard plant monitoring infrastructure.

***Index Terms*—solar panel soiling; dust detection; XGBoost; performance ratio; predictive maintenance; Flask; machine learning; photovoltaic systems**

I. INTRODUCTION

Solar photovoltaic (PV) technology has emerged as a cornerstone of the global clean energy transition. The International Energy Agency reports cumulative global PV installations exceeding 1.5 terawatts as of 2023, with annual additions reaching record highs [1]. As solar deployments proliferate across diverse geographies, maximising operational efficiency becomes critical for economic viability and environmental impact.

Among the foremost operational challenges facing PV installations is soiling — the accumulation of dust, dirt, pollen, and particulate matter on panel surfaces. In arid regions such as Rajasthan and the Middle East, soiling losses can reduce energy output by 20–40% within weeks [2]. Even in temperate climates, aerosols and urban pollution contribute to gradual efficiency degradation that compounds over time.

Traditional maintenance approaches rely on fixed-interval cleaning schedules or ad-hoc manual inspection, which are inherently reactive — leading to either over-cleaning (wasting water and labour) or under-cleaning (prolonging periods of reduced yield). For utility-scale farms spanning hundreds of acres, manual inspection is logistically impractical and cost-prohibitive [3].

The proliferation of Industrial IoT sensors in modern solar installations generates continuous streams of operational data: irradiance, temperature, wind speed, and power output. These rich time-series present an opportunity to apply machine learning for intelligent, demand-driven soiling detection without additional hardware.

This paper presents SolarSense, a full-stack ML system that processes operational data to predict dust-induced performance degradation. The contributions are: (i) a robust preprocessing and feature engineering pipeline; (ii) an XGBoost PR prediction model achieving $R^2 = 0.9412$;
(iii) a three-level dust classification scheme; and (iv) a deployable web application with email alert integration.

II. LITERATURE SURVEY

A. Traditional Soiling Mitigation

Conventional soiling mitigation involves manual water washing, robotic cleaning, and electrostatic dust removal. Ghazi et al. [2] found that manual

ISSN: 2349-6002

cleaning costs account for up to 15% of annual O&M budgets for utility-scale PV plants. Robotic systems reduce water consumption but operate on fixed schedules, failing to account for variable soiling rates across seasons and weather conditions.

B. Rule-Based Approaches

Zahrawi et al. [3] proposed a rule-based system flagging panels when output deviates beyond a predefined threshold. Such systems exhibit high false-positive rates and cannot distinguish soiling from shading, ageing, or inverter faults. Fathi et al. [6] applied Support Vector Machines for soiling severity classification, achieving 82% accuracy but lacking quantitative energy loss estimation.

C. Machine Learning Approaches

Abubakar et al. [5] explored artificial neural networks for PV soiling detection, outperforming rule-based systems. Harrou et al. [7] demonstrated Random Forest ensemble methods exceed single-model approaches for PV anomaly detection. However, most prior work frames soiling as classification rather than regression, limiting utility for scheduling. SolarSense addresses this gap by providing continuous PR prediction integrated into a real-time web dashboard.

III. PROPOSED SYSTEM

A. System Overview

SolarSense formulates dust detection as a regression problem. The system predicts the expected Performance Ratio under clean-panel conditions from eight operational features, then quantifies soiling loss as the percentage deviation between PR-predicted expected energy and actual measured energy generation.

B. Dust Classification Logic

The energy loss percentage drives a three-tier classification scheme:

- Low (<5% loss): No Cleaning Required — performance is optimal.
- Medium (5–15%): Schedule Cleaning Soon — yield is moderately reduced.
- High (>15%): Immediate Cleaning Required — severe loss detected.

C. Advantages Over Existing Systems

SolarSense offers four key advantages:
(i) continuous quantitative loss estimation rather than binary detection; (ii) zero additional hardware requirement; (iii) real-time web interface with CSV batch support; and (iv) interpretable feature importance for operator transparency.

IV. DATASET DESCRIPTION

The training dataset (solar_data.csv) comprises 17,401 rows of 15-minute interval sensor readings captured from a real PV installation beginning February 2022. Thirteen columns are recorded per row. Table I summarises the eight features used in model training after preprocessing.

TABLE I. MODEL INPUT FEATURES

Feature	Unit	Type
ambient_temp	°C	Raw
module_temp	°C	Raw
tilt_radiation	Wh/m²	Raw
peak_tilt_irradiation	Wh/m²	Raw
wind_speed	km/h	Raw
plant_peak_power	kW	Raw
temp_diff	°C	Engineered
radiation_ratio	ratio	Engineered

Rows with null values in essential columns are discarded. All columns are coerced to numeric type; non-parseable entries are dropped. A 1×10^{-6} epsilon is added to zero-valued denominators to prevent division errors.

V. METHODOLOGY

A. Performance Ratio Derivation

The target variable (pr) is re-derived from first principles for training-inference consistency, rather than using the plant-reported PLANT PR (%) column:

$$expected = plant\_peak\_power \times (tilt\_radiation / 1000)$$

$$PR = energy\_generation / (expected + 1\times10^{-6})$$

The derived PR is clipped to [0, 1.2] to remove physically impossible readings while retaining slightly super-nominal values arising from cool, high-

ISSN: 2349-6002

irradiance conditions.

B. Feature Engineering

Two features are engineered to enrich representational capacity:

$$temp\_diff = module\_temp - ambient\_temp$$

$$radiation\_ratio = tilt\_radiation / (peak\_tilt\_irradiation + \varepsilon)$$

temp_diff proxies direct solar heating and panel thermal stress. radiation_ratio captures the fraction of peak irradiance received, encoding combined effects of cloud cover and surface soiling.

C. Model Training

XGBoost (eXtreme Gradient Boosting) was selected for its superior handling of heterogeneous tabular data, built-in regularisation, and native missing-value support [4]. The dataset is partitioned 80/20 into training and test sets (seed=42).

Hyperparameters:

n_estimators=500, learning_rate=0.05, max_depth=5, n_jobs=−1.

D. Evaluation Metrics

Model performance is assessed via R^2 (proportion of PR variance explained; values → 1.0 are better) and Mean Absolute Error (MAE, average absolute PR prediction error in original units). Together these metrics characterise both overall fit quality and practical per-prediction accuracy.

VI. SYSTEM ARCHITECTURE

A. Training Pipeline (offline)

CSV ingestion → column renaming → null dropping → PR derivation and clipping → feature engineering → 80/20 split → XGBRegressor.fit() → test evaluation → joblib serialisation to model/pr_model.pkl.

B. Inference Pipeline (per request)

User input → build_features() → predict_pr() → calculate_loss() → classify_dust() → result.html rendered. Full chain e
compute_expected_energy()

C. Web Application Layers

Four functional layers: (i) Presentation — dark-themed Jinja2 templates (login, index, result); (ii) Routing — five Flask routes for auth, manual, and CSV prediction; (iii) ML — XGBoost model loaded into memory at startup; (iv) Notification — optional SMTP email alerts via Gmail TLS port 587.

D. Route Map

TABLE II. FLASK ROUTE MAP

Route	Method	Function
/login	GET/POST	Auth; session create
/logout	GET	Clear session
/	GET	Render dashboard
/predict_manual	POST	Single-row inference
/predict_csv	POST	Batch CSV inference

VII. IMPLEMENTATION

A. Technology Stack

TABLE III. TECHNOLOGY STACK

Component	Technology
ML Model	XGBoost 1.x (XGBRegressor)
Data	Pandas 1.4, NumPy 1.22
Serialisation	joblib
Web Backend	Flask 2.x + Jinja2
Frontend	Vanilla CSS + Bootstrap Icons
Fonts	Syne, JetBrains Mono (CDN)
Email	smtplib / MIMEText (stdlib)

B. Frontend Design

All three HTML templates share a unified dark design language: deep navy background (#05080F), amber accent (#F59E0B), and ambient radial glow orbs. Syne typeface is used for headings; JetBrains Mono for labels. The login card features a CSS pulse animation and a shake animation on authentication failure. The index.html dashboard implements a two-tab interface: Manual Input (seven numeric fields + optional email) and CSV Upload (drag-and-drop file picker with column listing). The result.html renders a verdict banner, a metrics grid (Predicted Energy, Actual Energy, Loss %, Confidence %), and an optional per-row CSV breakdown table.

ISSN: 2349-6002

C. CSV Batch Processing

The /predict_csv route reads the uploaded file into a Pandas DataFrame, normalises column names (strip, lowercase, replace spaces with underscores), validates required columns, drops invalid rows, and processes each row through the full inference chain. The row exhibiting maximum energy loss drives the overall system verdict; all per-row results populate the breakdown table in result.html.

D. Email Notification

When the user supplies an email address, send_email() constructs a plain-text summary (predicted energy, actual energy, loss%, dust level, action) and dispatches it via Gmail SMTP with STARTTLS on port 587 using credentials embedded in the source code.

VIII. RESULTS AND DISCUSSION

A. Model Performance

Table IV presents comparative performance of the XGBoost model against three baseline regressors on the 20% held-out test partition. All models were trained under identical data splits and preprocessing conditions.

TABLE IV. MODEL PERFORMANCE COMPARISON

Model	R^2 Score	MAE
XGBoost (Proposed)	0.9412	0.031
Random Forest	0.9104	0.047
Gradient Boosting	0.8987	0.052
Linear Regression	0.7831	0.094

XGBoost achieves $R^2 = 0.9412$ and MAE = 0.031, outperforming all baselines. The advantage over Random Forest ($R^2 = 0.9104$) reflects XGBoost's sequential residual correction strategy. Linear Regression's R^2 of 0.7831 confirms the inherently non-linear relationship between environmental conditions and PR.

B. Feature Importance

XGBoost gain-based feature importance reveals radiation_ratio as the most influential predictor (32.1%), reflecting its direct encoding of irradiance attenuation from soiling. tilt_radiation contributes 24.7%, ambient_temp 18.3%, module_temp 12.1%, temp_diff 7.4%, plant_peak_power 3.2%, wind_speed 1.4%, and peak_tilt_irradiation 0.8%. Dominance of radiation-based features validates the engineering decisions.

C. Practical Validation

Field validation over two weeks at a 500 kW rooftop installation in Chennai demonstrated 89.3% precision and 91.7% recall for cleaning requirement detection. Alerts were generated an average of 2.4 days earlier than conventional fixed-schedule approaches, preventing an estimated 3.2% additional energy loss. Extrapolated to a 10 MW plant, this corresponds to approximately 28,000 kWh of annually recoverable energy.

IX. ADVANTAGES AND LIMITATIONS

A. Advantages

- No specialised soiling sensors required; uses standard plant monitoring data only.
- Continuous quantitative loss estimation rather than binary cleaning flags.
- Demand-driven alerts reduce unnecessary cleaning cycles by up to 30%.
- Deployable on commodity hardware or cloud infrastructure; Docker-compatible.

B. Limitations

- Hardcoded credentials and SMTP password pose security risks in shared environments.
- No database backend; prediction history is not persisted across sessions.
- Confidence score is a heuristic (100 − loss%), not a statistically calibrated interval.
- Temporal features (hour-of-day, seasonality) are absent from the current model.

X. FUTURE SCOPE

Several extensions are planned. Integration of drone-captured imagery with convolutional neural networks will enable panel-level spatial soiling mapping [8]. Weather forecast API integration will support predictive soiling rate modelling for proactive

scheduling ahead of dust storm events.

Cyclic temporal encoding (sine/cosine of hour-of-day and day-of-year) will capture diurnal and seasonal patterns currently absent. Transfer learning will address the cold-start problem for new plants by fine-tuning from climatically similar sites. SCADA integration via Modbus TCP/IP will automate sensor ingestion, eliminating manual entry. The confidence heuristic will be replaced with quantile XGBoost prediction intervals for statistically rigorous uncertainty bounds.

XI. CONCLUSION

This paper has presented SolarSense, a machine learning-powered web application for data-driven detection and quantification of dust accumulation on photovoltaic solar panels. The XGBoost regression model, trained on 17,401 real-world operational records, predicts the Performance Ratio under clean-panel conditions with $R^2 = 0.9412$ and MAE = 0.031, significantly outperforming linear and ensemble baselines.

By comparing predicted against actual energy generation, the system delivers a three-level dust severity classification with actionable maintenance recommendations through a responsive Flask dashboard. Field validation demonstrates alerts issued 2.4 days earlier than fixed-schedule approaches, recovering an estimated 3.2% additional energy yield. The framework requires no specialised soiling sensors, making it immediately deployable across existing solar plant monitoring infrastructure. SolarSense demonstrates that commodity operational data, thoughtful feature engineering, and gradient-boosted regression can effectively replace expensive sensor arrays and manual inspection protocols in solar maintenance management.

ACKNOWLEDGMENT

The authors express sincere gratitude to Ms. Janani, Assistant Professor, Department of Computer Science and Engineering, Sri Venkateswara College of Engineering and Technology, for her invaluable guidance throughout this project. The authors also acknowledge the open-source communities behind Flask, XGBoost, Pandas, and Scikit-learn.

REFERENCES

[1] International Energy Agency, “Solar PV Global Supply Chains,” IEA Report, Paris, 2023.

[2] S. Ghazi, A. Sayigh, and K. Ip, “Dust effect on flat surfaces — a review,” Renew. Sustain. Energy Rev., vol. 54, pp. 1527–1536, 2016.

[3] M. Zahrawi et al., “Rule-based soiling monitoring for utility-scale PV,” in Proc. IEEE ICRERA, Istanbul, 2018, pp. 211–216.

[4] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. ACM KDD, San Francisco, 2016, pp. 785–794.

[5] M. Abubakar et al., “ANN-based soiling detection for PV systems,” Sol. Energy, vol. 210, pp. 60–71, 2020.

[6] M. Fathi et al., “Soiling severity classification using SVM,” in Proc. IEEE ISPE, Bari, 2019, pp. 1–6.

[7] F. Harrou et al., “Unsupervised PV monitoring using ensemble learning,” Energy Convers. Manag., vol. 258, p. 115511, 2022.

[8] A. Mehta et al., “Drone-based solar panel inspection using deep CNNs,” IEEE Trans. Ind. Informat., vol. 19, no. 3, pp. 2104–2113, 2023.