Aalekh **Roy**

Data Scientist · Researcher: Data Science
New Delhi, India
(+91) 8800961923 roy.aalekh@gmail.com https://royaalekh.github.io RoyAalekh Aalekh Roy

Experience

Data Scientist — Noida, India

SmartHelio(YC-22) Jan 2024 - Oct 2025

- Built and scaled the Data Science function, growing the team from 3 to 10 members and taking end-to-end ownership of hiring, onboarding, technical mentorship, sprint planning, and execution discipline.
- Drove SmartHelio's transition from an early-stage analytics offering to a scalable solar-intelligence platform; supporting company growth from single-digit to triple-digit portfolio.
- Architected and delivered major components of the solar intelligence engine, owning core algorithm designs across fault detection, performance modeling, downtime diagnostics, and weather simulation models.
- Brought engineering rigor into the Data Science landscape by introducing software engineering principles, testing processes, CI/CD pipelines, and modular analytics libraries, improving reliability at scale.
- Acted as cross-functional lead, collaborating with Product, Engineering, and Customer Success for feature scoping, technical roadmaps, and deployment strategies.
- Worked closely with CXOs on strategic alignment, prioritization, and organizational processes, contributing to improved operational clarity, efficiency, and cross-team coordination.
- **Impact Summary**: Led the maturation of SmartHelio's Data Science function and analytics engine, 3x delivery velocity, scaling diagnostics to 3000+ PV assets, reducing pipeline regressions by 40%, and strengthening organizational product–data alignment through close CXO collaboration.

Associate Data Scientist — Noida, India

SmartHelio(YC-22) July 2023 - Dec 2023

- Designed SmartHelio's first predictive-maintenance solution for soiling-rate forecasting, enabling data-driven and cost-efficient cleaning-cycle optimization for utility-scale PV plants.
- Created a unified internal Python analytics library used across teams, standardizing statistical and physics-based modeling workflows and improving consistency across codebases.
- Contributed to a high-impact R&D initiative on PV-module connector health classification, applying signal processing, frequency-domain analysis, and machine learning.
- Worked closely with Customer Success to translate analytical outputs into actionable operational insights for field and operations teams.
- **Impact Summary**: Delivered SmartHelio's first soiling-forecasting and cleaning-optimization workflow, standardized internal analytics practices, and strengthened diagnostic accuracy, directly improving O&M decision-making and reducing energy losses across early client portfolios.

Data Analyst — Noida, India

SmartHelio(YC-22) Sep 2022 - June 2023

- Joined SmartHelio as the 7th employee and 3rd member of the Data Science team, contributing to foundational analytics work that supported the company's early rapid scaling.
- Developed statistical and physics-informed algorithms for fault detection, downtime diagnostics, and performance modeling across utility-scale solar PV systems.
- Supported the transition from locally executed analytical scripts to cloud-deployed pipelines and micro-services, enabling more reliable and scalable diagnostics infrastructure.
- **Impact Summary**: Built foundational analytical components that enabled early automation, improved diagnostic stability, and supported successful enterprise-client onboarding during SmartHelio's initial growth and product expansion phase.

Projects

Fast Diagnostic Service (Cloud-Based PV Health Assessment Platform)

SmartHelio (YC-22)

Designer/Developer — Jan 2025

- Designed and implemented a cloud-deployed Python microservice that allows users to upload AC-side PV generation data (CSV/Excel) and receive instant automated diagnostics, including fault detection, performance KPIs, energy-loss quantification, weather-based simulations, and benchmarking against comparable assets.
- Integrated robust input validation, missing-timestamp repair, and performance optimizations to handle large plant datasets efficiently.
- Enabled Marketing and Sales to close 300+ new client engagements within three months by replacing multi-hour manual analysis with a reliable, self-serve diagnostic workflow.
- **Tools & Methods**: Python, Dash, Pandas, Numpy, SciPy, Polars, DuckDB, PVlib, AWS S3, microservice design.

formatify-py (Python Library for Messy Datetime Inference)

PieceWiseProjects/formatify

Designer/Developer — May 2025 – Present

- Designed and developed a pure-Python library (published on **PyPI**) for inferring datetime formats from messy, heterogeneous, and incomplete timestamp data.
- Implements a structural parsing engine capable of detecting delimiters, textual months, numeric roles, and mixed-format sequences.
- Provides coverage far beyond standard parsing utilities, with full validation suite and CI pipeline ensuring correctness across thousands of timestamp variants.
- **Tools & Methods**: Pure Python, custom tokenization and parsing, regex, heuristical modelling, CI/CD, python package development.

Cleaning Optimization (Cost–Loss Optimization Component)

SmartHelio (YC-22)

Designer/Developer — 2024

- Designed and implemented SmartHelio's cleaning-schedule optimization engine as a live commercial offering, now deployed across 300+ utility-scale PV assets and used by O&M teams to plan data-driven, cost-efficient cleaning campaigns.
- Combined four predictive modules into a single decision pipeline: (1) historical soiling estimation using RDTools SRR, (2) rainfall forecasting using 10 years of daily rain data per calendar day, (3) soiling accumulation projection from the regression-based accumulation model, and (4) short-term PV generation forecasting to estimate future energy.
- For each candidate cleaning date in the planning horizon, modeled counterfactual future trajectories by simulating "no cleaning", "clean now", and "clean later" scenarios—propagating forward soiling accumulation, applying predicted rain resets, and combining these with expected PV generation to compute future energy output under each scenario.
- Defined a cost–loss objective function that integrates expected soiling-induced energy loss, cleaning cost (per event), operational constraints (minimum gap between cleanings, maximum cleanings per year), and site-level rules (rainy-day exclusions, blackout periods).
- Solved the resulting discrete optimization problem using constrained search with heuristic pruning to identify a small set of high-value cleaning dates that maximize net savings while respecting operational constraints.
- Delivered **8–12%** annual reduction in soiling-related energy losses and **15–20%** reduction in O&M cost per site, turning the methodology into a core revenue-generating product within SmartHelio's portfolio.
- **Tools & Methods**: Python, NumPy, SciPy, RDTools SRR, 10-year rainfall regression modeling, forward scenario simulation, cost–loss function design, constrained search with heuristic pruning, integration of multiple forecasting modules into a single optimization engine.

Soiling-Rate Forecasting (Time-Series Modeling Component)

SmartHelio (YC-22)

Designer/Developer — 2023

- Addressed the core challenge that RDTools SRR returns highly intermittent, saw-tooth soiling curves (with rain/cleaning resets and small decimal values) that are unsuitable for naive time-series forecasting, by designing a dedicated accumulation-modeling pipeline.
- Estimated historical soiling using RDTools SRR and then stabilized the signal via missing-data repair, temporal alignment, frequency aggregation, smoothing, and segmentation of the curve into monotonic accumulation phases versus reset events.
- Framed the problem as forecasting **future accumulation speed** (daily soiling-rate slope) rather than raw SRR values, and modeled this accumulation rate using regression on environmental, seasonal, and operational features (irradiance gradients, clear-sky index, month-of-year, time-since-last-clean, recent rainfall).

- Integrated a rainfall-reset model that uses 10 years of historical daily rain; for each calendar day, regression over the 10 historical values yields an expected rainfall, and thresholds on this expected rainfall are used to trigger stochastic "rain reset" events in the future trajectory.
- Generated scenario-based forward soiling trajectories by numerically integrating the predicted accumulation rates day-by-day, applying rain-driven resets and (later) cleaning-driven resets, and evaluated these trajectories horizon-wise using error metrics on accumulated loss and detection of high-soiling windows.
- Achieved stable forward projections with **2–3%** error on weekly soiling-loss trends and correctly flagged **92%** of high-soiling intervals, providing a robust basis for downstream cleaning optimization rather than brittle direct SRR forecasting.
- **Tools & Methods**: Python, Pandas, NumPy, SciPy, RDTools SRR, custom accumulation-rate regression, feature engineering (seasonality, irradiance gradients, time-since-clean), rainfall-reset modeling, scenario simulation via numerical integration.

Short-Term Solar Irradiance Forecasting (Clear-Sky Modeling + ML)

SmartHelio (YC-22)

Designer/Developer

2023

- Developed a short-term (15-min to 24-hour horizon) solar irradiance forecasting pipeline to estimate expected GHI and clear-sky–normalized irradiance, used as a critical input for predicting future energy availability and quantifying soiling-induced losses in SmartHelio's cleaning optimization engine.
- Preprocessed irradiance and plant power data by repairing missing timestamps, aligning sampling frequencies, and generating irradiance-specific features including lagged GHI values, ramp rates, irradiance gradients, rolling variability, and clear-sky index (CSI).
- Integrated clear-sky modeling using pvlib to compute theoretical clear-sky irradiance and normalize observed GHI, enabling separation of environmental variability from plant- or sensor-specific effects.
- Benchmarked statistical baselines (Persistence, ETS, ARIMA) against machine-learning regressors (XGBoost, RandomForest), treating irradiance forecasting as a supervised regression problem over engineered temporal and environmental features.
- Applied rolling-origin, horizon-wise time-series cross-validation and selected the best-performing models per horizon based on RMSE/MAE, achieving **>15%** RMSE improvement over the persistence benchmark for day-ahead irradiance forecasts.
- **Tools & Methods**: Python, Pandas, NumPy, pvlib (clear-sky modeling), feature engineering (lagged GHI, gradients, CSI, variability metrics), scikit-learn, XGBoost, RandomForest, ETS/ARIMA baselines, rolling-origin time-series cross-validation.

Education

Institute Of Mathematics And Applications

Bhubaneswar, Orissa, India

M.Sc. in Mathematics with Data Science

August 2020 - August 2022

- Awarded NBHM scholarship for academic excellence (Top 3 rankers).
- Founder and Head Coordinator of the Data Science Society, promoting interdisciplinary collaboration and technical workshops.
- **Selected Coursework:** Graduate Algebra, Linear Algebra, Vector Calculus, Functional Analysis, Topology, Stochastic Calculus, Machine Learning, Statistical Learning, Numerical Optimization, Deep Learning, Python for Data Science.

Dr. B.R. Ambedkar University Delhi (AUD)

Delhi, India

BA (Honors) in Mathematics

July 2015 - July 2020

- Received semester-wise merit scholarships for academic distinction (Top 5 rankers).
- Interdisciplinary liberal arts education combining mathematical rigor with social sciences and humanities.
- **Selected Coursework:** Real Analysis, Calculus, Abstract Algebra, Optimization Theory, Game Theory, Social Theory, Psychology, Statistical Methods in Economics.

Skills

Programming Languages	**Python**, SQL, R, JS
Libraries	**Pandas**, **Numpy**, **Plotly**, **scikit-learn**, **scipy**, sktime, XGBoost, PVLib, matplotlib, pytorch, keras, Tenserflow, polars, fastAPI
Tools	**Git**, **AWS**, **CI/CD**, Docker, DVC, Jira, Notion
Databases	**MySQL**, **MongoDB**, **S3**, AWSTimestreamDB, PostgreSQL, SQLite, DuckDB
Technical Domains	**TabularML**, **Statistical Learning and Modeling**, **Time Series Analysis**, **Network Science**
Spoken Languages	**English**, **Hindi**