Predicting Solar Energy Potential Using Machine Learning: A Comparative Analysis of Classification Models on Aswan Weather Data

ABSTRACT:

Solar power is now at the forefront of sustainable development strategies due to the world's shift to renewable energy. However, the stochastic and intermittent nature of solar radiation poses a significant obstacle to the integration of solar Photovoltaic (PV) systems into national power grids. Solar generation, in contrast to traditional fossil-fuel power plants, is highly reliant on changing meteorological conditions, particularly ambient temperature, humidity, cloud cover, and wind speed. Because grid operators must balance supply and demand in real-time, this volatility poses serious challenges. An abrupt, unforeseen decline in solar output can result in frequency instability or an expensive reliance on spinning reserves.

In order to solve this issue, this research project creates a strong, data-driven Machine Learning (ML) framework that can categorize daily solar energy potential into three operational categories: "Low," "Medium," and "High." We give grid operators a more transparent decision-making tool for energy storage and dispatch planning by breaking down complex, continuous output data into discrete classes.

We developed a thorough data science lifecycle using historical weather data from Aswan, Egypt, an area with high solar irradiance. The approach includes sophisticated preprocessing methods, such as quantile binning to address target variable imbalance and mean imputation for missing values. Six different supervised learning algorithms—Linear Discriminant Analysis (LDA), Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Decision Trees, Logistic Regression, and Multi-Layer Perceptron (MLP)—were thoroughly compared.

The MLP Classifier (Neural Network) achieved a classification accuracy of 70.0%, outperforming all other models, according to our experimental results. Additionally, the most statistically significant predictors of solar generation in arid climates are average temperature and dew point, according to statistical analysis (ANOVA and correlation heatmaps). The study comes to the conclusion that non-linear machine learning models provide a feasible route for automated, highly accurate solar forecasting systems because they are better at simulating weather-energy dynamics than linear baselines.

Ziad Elsayed Fouad
ziad.320230025@ejust.edu.eg

Mohamed Hussien Kamal
mohamed.320230008@ejust.edu.eg

1. INTRODUCTION

1.1. Problem Definition: The core problem addressed in this study is the **uncertainty of solar power generation**. Inefficiencies in energy management systems (EMS) result from the inability to precisely forecast whether tomorrow's solar output will be adequate (High) or insufficient (Low). Operators waste renewable energy (curtailment) if they underestimate generation. They will experience power shortages if they overestimate. Current atmospheric physics-based physical models are frequently too computationally costly for real-time use. This project proposes a **Machine Learning Classification approach** as a computationally efficient alternative to solve this uncertainty.

1.2. Techniques Used: We employed a full stack of Data Science techniques:

- **Statistical Analysis:** Skewness, Kurtosis, and Z-tests to understand data distribution.
- **Preprocessing:** StandardScaler for normalization and LabelEncoder for categorical handling.
- **Supervised Learning:** We utilized a diverse set of algorithms—**Geometric** (KNN, LDA), **Probabilistic** (Naive Bayes), **Logical** (Decision Tree), and **Connectionist** (MLP)—to evaluate which mathematical approach best models the weather-solar relationship.

1.3. Main Contribution: The primary contributions of this research are:

1. **Imbalance Handling:** Implementation of Quantile Binning (qcut) to transform skewed solar data into balanced classes, preventing the "accuracy paradox" where models only predict the majority class.
2. **Feature Reduction Analysis:** A critique of Linear Discriminant Analysis (LDA) versus Principal Component Analysis (PCA) for meteorological data.
3. **Regional Specificity:** providing a tailored model for the Aswan region, validating that temperature—not just sunlight hours—is a critical driver of PV efficiency in hot climates.

1.4. Organization: The rest of this document is structured as follows: A tabular review of 20 related studies is presented in Section 2. The algorithms and methodology are described in Section 3. The suggested model architecture is explained in Section 4. Comprehensive results, including confusion matrices and statistical tests, are presented in Section 5. Future research directions are discussed at the end of Section 6.

2. RELATED WORK

Machine learning has been widely applied to renewable energy forecasting. This section reviews twenty significant studies in the field, highlighting the diverse methodologies and key findings that inform our current research.

Foundational Studies (2008–2015)

Early research focused on establishing the viability of neural networks and support vector machines for solar prediction. Mellit and Kalogirou (2008) provided a comprehensive review of AI techniques, confirming that Artificial Neural Networks (ANN) were the dominant approach for modeling solar radiation [16]. Building on this, Benghanem et al. (2009) demonstrated that Radial Basis Function (RBF) networks outperformed standard Multi-Layer Perceptrons (MLP) in arid climates by better capturing non-linear temperature effects [6]. Similarly, Mellit and Pavan (2010) achieved a 98% correlation coefficient for 24-hour forecasting using MLP, setting a high benchmark for daily prediction [3].

Comparing different architectures, **Chen et al. (2011)** found that Support Vector Machines (SVM) offered better generalization on small datasets compared to RBF networks [4]. **Linares-Rodríguez et al. (2013)** showed that while complex models are powerful, simple Multiple Linear Regression was often sufficient for very short-term horizons [5]. However, **Amrouche and Le Pivert (2014)** argued for data enrichment, proving that hybrid models combining satellite imagery with ground data improved accuracy by 12% [7]. **Olatomiwa et al. (2015)** further validated SVMs, showing they outperformed ANNs in Nigerian solar studies [14]. **Wan et al. (2015)** emphasized the importance of probabilistic forecasting for smart grid management [19].

Deep Learning & Recent Advances (2016–2025)

The last decade has seen a shift toward Deep Learning. Gensler et al. (2016) compared Long Short-Term Memory (LSTM) networks against standard ANNs, finding that LSTMs reduced Root Mean Squared Error (RMSE) by 15% by capturing temporal dependencies [2]. Raza et al. (2016) reviewed these emerging PV output forecast methods, noting a trend toward hybrid models [17]. Alzahrani et al. (2017) successfully applied deep neural networks to irradiance forecasting [12], while Voyant et al. (2017) noted that while time-series methods like ARIMA are effective for 1-hour predictions, machine learning is superior for daily classification [8].

Das et al. (2018) provided a critical review of model optimization, suggesting that preprocessing is as vital as model selection [13]. **Wang et al. (2018)** utilized Deep Belief Networks (DBN) to automate feature extraction, outperforming Support Vector Regression [9]. **Sobri et al. (2018)** categorized these forecasting methods, highlighting the trade-off between computational speed and accuracy [18]. **Al-Dahidi et al. (2019)** introduced Extreme Learning Machines (ELM), achieving a remarkable 92% classification accuracy [1].

In recent years, **Alomari et al. (2020)** validated the use of ANNs specifically for the Jordanian climate, which shares similarities with Aswan [11]. **Li et al. (2020)** developed a hybrid deep learning model that combined varying time-scales for short-term forecasting [15], and **Zang et al. (2020)** used Convolutional Neural Networks (CNN) to process weather maps directly [10]. Most recently, **Khadeeja et al. (2025)** conducted a comparative study of multiple models, reaffirming ANN as a highly effective tool with a Mean Absolute Percentage Error (MAPE) of just 5.26% [20].

3. METHODOLOGY

The project utilized the following six algorithms:

1. **Linear Discriminant Analysis (LDA):** A dimensionality reduction technique that projects data to maximize the distance between means of different classes while minimizing variance within each class.
2. **Gaussian Naive Bayes:** A probabilistic classifier based on Bayes' Theorem, assuming independence between predictors (e.g., assuming Wind is independent of Temperature).
3. **K-Nearest Neighbors (KNN):** A non-parametric method that classifies a data point based on the majority class of its 'k' closest neighbors in the feature space.
4. **Decision Tree:** A tree-like model of decisions that splits data based on feature thresholds (e.g., "If Temp > 30°C") to maximize information gain.
5. **Multi-Layer Perceptron (MLP):** A feedforward artificial neural network that uses backpropagation to learn non-linear relationships through hidden layers of neurons.
6. **Logistic Regression:** A statistical model that uses a logistic function to model the probability of a certain class or event.

4. PROPOSED MODEL

Phase 1: Preprocessing

- **Missing Values:** We identified missing entries in the Humidity and Wind columns. These were treated using **Mean Imputation** to preserve the central tendency of the data.
- **Binning:** The continuous Solar (PV) target was discretized into three classes (Low, Medium, High) using pd.qcut(q=3), ensuring a balanced distribution (approx. 33% of data in each class).

Phase 2: Feature Selection & Reduction

- **Correlation Analysis:** A Pearson Correlation Heatmap was generated. AvgTemperture showed the highest positive correlation (0.64) with the target.
- **LDA:** We applied Linear Discriminant Analysis to project the 5-dimensional feature space into 2 dimensions to visualize class separability.

Phase 3: Classification

- **Training:** The dataset was split into 80% Training and 20% Testing sets.
- **Scaling:** We used StandardScaler to normalize features to mean=0 and variance=1, which is critical for the convergence of the MLP and KNN algorithms.

Phase 4: Evaluation

- Models were evaluated using **Accuracy** (overall correctness), **F1-Score** (balance of precision/recall), and the **Confusion Matrix** (error analysis).

5. RESULTS AND DISCUSSION

5.1. Data Set Description The dataset consists of **398 daily observations** with five independent variables (Average Temperature, Dew Point, Humidity, Wind Speed, Pressure) and one dependent variable (Solar PV generation), which was categorized into High, Low, and Medium classes.

5.2. Statistical Analysis Results We performed a descriptive statistical analysis on the raw data:

- **Central Tendency:** The Average Temperature had a mean of 82.5°F, while Solar PV output had a mean of 21.4.

- **Skewness & Kurtosis:** The Solar PV data showed a skewness of **-0.21**, indicating a slight left skew (tendency toward higher generation). The Kurtosis was **-0.85**, indicating a platykurtic distribution with fewer outliers than a normal distribution.
- **Correlation:** The Heatmap analysis revealed a strong positive correlation (**+0.64**) between Temperature and Solar output, and a strong negative correlation (**-0.58**) between Humidity and Solar output. This confirms that high heat and low humidity are the ideal conditions for generation in Aswan.

5.3. Feature Reduction Results (LDA) Applying LDA reduced our features to 2 components. The first component (LD1) accounted for 78% of the variance. Plotting the data on LD1 vs LD2 revealed that "High" and "Low" classes were linearly separable, but the "Medium" class showed significant overlap, explaining why models struggled with "Medium" predictions.

5.4. Classification Results The models were ranked based on accuracy and F1-Score:

1. **MLP Classifier (Neural Network):** Achieved the highest **Accuracy of 70.0%** and an F1-Score of 0.6949.
2. **Decision Tree:** Achieved an Accuracy of 67.5% and an F1-Score of 0.6724.
3. **Naive Bayes:** Achieved an Accuracy of 65.0%.
4. **K-Nearest Neighbors (KNN):** Achieved an Accuracy of 65.0%.
5. **Logistic Regression:** Achieved an Accuracy of 63.8%.
6. **LDA:** Achieved an Accuracy of 63.8%.

Analysis: The **MLP Classifier** achieved the highest accuracy. The non-linear models (MLP, Decision Tree) consistently outperformed the linear models (Logistic Regression, LDA), confirming that the relationship between weather and solar output is non-linear.

5.5. Confusion Matrix Analysis (Best Model: MLP) The confusion matrix for the MLP classifier provided deep insights into the model's performance:

- **High Class Prediction:** The model correctly identified 22 "High" days and made zero errors in confusing them with "Low" days.
- **Low Class Prediction:** The model correctly identified 19 "Low" days.
- **Medium Class Challenge:** The majority of errors occurred in the "Medium" class, where the model correctly identified 15 days but misclassified 14 others as either High or Low.
- **Conclusion:** The model has **0% critical error** (confusing High for Low), making it highly safe for grid operations.

6. CONCLUSION AND FUTURE WORK

6.1. Conclusion This research project set out to address one of the most critical challenges in the modern renewable energy landscape: the unpredictability of solar power generation. By applying a data-driven Machine Learning approach to historical weather data from Aswan, Egypt, we successfully demonstrated that standard meteorological forecasts can be translated into actionable operational insights for power grid management.

The core finding of this study is that **non-linear machine learning models**, specifically the Multi-Layer Perceptron (MLP) Neural Network, significantly outperform traditional linear statistical methods in predicting solar energy potential. Our MLP model achieved a classification accuracy of **70.0%**, effectively categorizing daily output into "High," "Medium," and "Low" tiers. This performance superiority—approximately 6-7% higher than linear baselines like Logistic Regression—validates the hypothesis that the relationship between weather variables (such as temperature, humidity, and atmospheric pressure) and photovoltaic efficiency is complex and non-linear.

Statistically, our analysis highlighted the dominant role of **Average Temperature** and **Dew Point/Humidity** as the primary predictors of solar output in arid climates. We observed a strong positive correlation between temperature and generation, but crucially, this relationship is modulated by humidity levels; clear, hot days produce the most power, while humid days, often indicative of haze or cloud cover, reduce efficiency.

6.2. Future Work To further enhance the predictive accuracy and operational utility of this system, several avenues for future research are proposed:

1. **Temporal & Sequential Modeling:** The current model treats each day as an independent event. Future work should implement **Long Short-Term Memory (LSTM)** networks, a type of Recurrent Neural Network (RNN) designed for time-series data. This would allow the model to learn from weather trends over the preceding days (e.g., a cooling trend over 3 days) to better predict the next day's output.

2. **Granularity Improvement:** Moving from daily averages to **hourly data** would significantly increase the resolution of predictions. This would allow for "intra-day" forecasting, helping operators manage the famous "duck curve" of energy demand more effectively.

3. **Hybrid Ensemble Methods:** Combining the interpretability of Decision Trees with the predictive power of Neural Networks (e.g., using a **Stacking Classifier**) could help resolve the misclassifications seen in the "Medium" category.

7. REFERENCES

[1] Linares-Rodríguez, A., Ruiz-Arias, J. A., Pozo-Vázquez, D., & Tovar-Pescador, J. (2013). Generation of synthetic daily global solar radiation data based on ERA-Interim reanalysis and artificial neural networks. *Energy*, *52*, 317-325.

[2] Gensler, A., Henze, J., Sick, B., & Raabe, N. (2016). Deep learning for solar power forecasting—A comparison of a LSTM and a feedforward NN. *2016 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 1-8.

[3] Mellit, A., & Pavan, A. M. (2010). A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy. *Solar Energy*, *84*(5), 807-821.

[4] Chen, C., Duan, S., Cai, T., & Liu, B. (2011). Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Solar Energy*, *85*(11), 2856-2870.

[5] Al-Dahidi, S., Ayadi, O., & Adeeb, J. (2019). Solar photovoltaic power classification using extreme learning machine. *Energies*, *12*(15), 2990.

[6] Benghanem, M., Mellit, A., & Alamri, S. N. (2009). ANN-based modelling and estimation of daily global solar radiation data: A case study. *Energy Conversion and Management*, *50*(7), 1644-1655.

[7] Amrouche, B., & Le Pivert, X. (2014). Artificial neural network based daily local forecasting for global solar radiation. *Applied Energy*, *130*, 333-341.

[8] Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569-582.

[9] Wang, K., Qi, X., & Liu, H. (2018). Photovoltaic power forecasting based on deep belief network and seasonal exponential smoothing method. *IEEE Access*, *6*, 38354-38363.

[10] Zang, H., Cheng, L., Ding, T., Cheung, K. W., Wei, Z., & Sun, G. (2020). Day-ahead photovoltaic power forecasting approach based on deep convolutional neural networks and meta-learning. *International Journal of Electrical Power & Energy Systems*, *118*, 105790.

[11] Alomari, M. H., Adeeb, J., & Al-Dahidi, S. (2020). Solar photovoltaic power forecasting in Jordan using artificial neural networks. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(1), 587-596.

[12] Alzahrani, A., Shamsi, P., Dagli, C., & Ferdowsi, M. (2017). Solar irradiance forecasting using deep neural networks. *Procedia Computer Science*, *114*, 304-313.

[13] Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., ... & Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, *81*, 912-928.

[14] Olatomiwa, L., Mekhilef, S., & Shamshirband, S. (2015). Hybrid support vector regression with bat algorithm for solar radiation prediction in Nigeria. *International Journal of Electrical Power & Energy Systems*, *73*, 269-277.

[15] Li, P., Zhou, K., Lu, X., & Yang, S. (2020). A hybrid deep learning model for short-term PV power forecasting. *Applied Soft Computing*, *86*, 105824.

[16] Mellit, A., & Kalogirou, S. A. (2008). Artificial intelligence techniques for photovoltaic applications: A review. *Progress in Energy and Combustion Science*, *34*(5), 574-632.

[17] Raza, M. Q., Nadarajah, M., & Ekanayake, C. (2016). On recent advances in PV output power forecast. *Solar Energy*, *136*, 125-144.

[18] Sobri, S., Koohi-Kamali, S., & Rahim, N. A. (2018). Solar photovoltaic generation forecasting methods: A review. *Energy Conversion and Management*, *156*, 459-497.

[19] Wan, C., Zhao, J., Song, Y., Xu, Z., Lin, J., & Hu, Z. (2015). Photovoltaic and solar power forecasting for smart grid energy management. *CSEE Journal of Power and Energy Systems*, *1*(4), 38-46.

[20] Khadeeja, F. P., Radha, M., Vanitha, G., Nirmala, D. M., Mahendiran, R., & Vishnu, S. S. (2025). Daily solar power prediction using machine learning: A model wise comparative study. *Plant Science Today*, *12*(sp1). https://horizonepublishing.com/index.php/PST/issue/view/63