Impact of Machine-Learning-Based Solar Power Forecasting on the Techno Economics of Green Hydrogen Production

Dawood Majeed[1], Ali Abbas Khan[2], Nabeel Hassan[1], Raheel Aslam[3], Muhammad Awais [4] , Abid Aman[3]

1 School of Future Technology, South China University of Technology, Guangzhou, China.

2 School of Materials Science and Engineering, South China University of Technology, Guangzhou 510641, China.

3 School of Automation, South China University of Technology, Guangzhou, Guangdong, 510641, PR China.

4 Department of Information and Technology, University of the Punjab, Pakistan.

Abstract- The integration of solar-powered electrolysis for green hydrogen production is constrained by the intermittent nature of photovoltaic (PV) generation, where forecasting inaccuracies can significantly increase operational risks and levelized costshhh. This study evaluates the techno-economic impact of machine learning (ML)-based solar forecasting on green hydrogen production costs by systematically comparing three distinct algorithms: Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) networks. Using high-resolution solar and meteorological data (2005–2023) from the Tibetan Plateau, each model's forecasting accuracy is assessed using standard metrics (RMSE, MAE, MAPE, R^2). The predicted solar outputs are then integrated into comprehensive levelized cost of energy (LCOE) and levelized cost of hydrogen (LCOH) models for a 100 MW PV plant coupled with an alkaline electrolyzer. Results show that XGBoost achieves the most balanced performance with an RMSE of 0.52 MW, MAE of 0.39 MW, and R^2 of 0.994, yielding an LCOH of \$3.53/kg H_2. Although SVR exhibits superior absolute error (RMSE = 0.29 MW), it results in a slightly higher LCOH (\$3.54/kg H_2) due to higher relative errors during low-irradiance periods. Interestingly, LSTM, despite the weakest forecasting performance (RMSE = 6.52 MW), produces the lowest LCOH (\$3.51/kg H_2), attributed to its smoothing of short-term variability, which stabilizes the electricity cost component. Sensitivity analysis identifies electricity cost as the dominant driver of LCOH, with a ±20 % variation causing an approximate change of ±\$0.42/kg H_2, substantially outweighing the marginal economic impact of hourly forecast errors (< \$0.02/kg H_2). The study concludes that while forecasting accuracy is important, the structure of forecast errors and broader economic assumptions particularly electricity price, electrolyzer efficiency, and utilization are more decisive for hydrogen cost competitiveness. These findings underscore the value of integrating ML-based solar forecasting into techno-economic models to enhance the bankability and deployment of cost-effective green hydrogen systems.

Keywords- Solar power forecasting; Machine learning models; Levelized cost of hydrogen; Levelized cost of energy; Green hydrogen production; Artificial intelligence in energy.

I. Introduction

The international shift towards sustainable energy systems has led to a rush in the quest to identify solid measures of reducing the emission of greenhouse gases and the abandonment of the fossil-fuel. Hydrogen has become a central energy carrier among the myriad of possible solutions because of its utilisation in its versatile nature and ability to be used both as a feedstock and a fuel in a wide future range of processes, both in transportation and in industrial processes. Specifically, H2 obtained through the electrolysis of water using renewable electricity or so-called green hydrogen acquires popularity due to its correspondence with the goals of decarbonisation (only less frequently)[1]. Although this interest has increased, one major challenge has persisted; how to properly align the generation of renewable power and economically viable production of hydrogen.

The appeal of green hydrogen is in large part due to its capability to provide large-scale energy storage along with providing intermittent sources for renewable energy, particularly wind power and solar photovoltaic (PV) to exist in tandem with each other. Excess renewable electricity can then be converted into hydrogen for long-term storage of energy, which might then be reconverted into electricity or used as a direct, carbon-free fuel. This duality explains the reason behind precise prediction of renewable output that can be the foundation of green hydrogen production profile optimisation and the total expenditure. Precise forecasting finds greater application in the power systems with an increasing ratio of renewable energy where it is necessary to balance between demand and supply that is more complicated. Furthermore, industrial processes can also use hydrogen as a viable substitute for natural gas, and the chemical industry requires this as a feedstock[2]. Nonetheless, achieving economic competitiveness with fossil-derived hydrogen requires a substantial reduction in production costs, which are primarily influenced by renewable electricity costs.

One of the most attractive renewable resources is solar energy because it is available worldwide and capital prices of PV system have significantly decreased in the last ten years. Nevertheless, substantial uncertainty is induced by the intermittent and uncertain nature of the sun's irradiance, and thus induces increased risks, both operational and economical.. Elsewhere, the parity with the grid has been reached or even exceeded by solar PV in particular cases especially when a favourable policy framework is present or very high irradiance rates are experienced. By fulfilling the role of providing electrolyzers to the solar PV setup, the area is free of direct carbon dioxide emissions hence supporting the targets of reducing climate change[3]. However, solar production, in its nature, is intermittent; when the solar resource is forecasted inaccurately, there is a likelihood that electrolyzers will either be under-utilised or overloaded, which in turn increases the cost of hydrogen production manifold.

This kind of precise forecasting of the solar power is very important for optimising the techno-economic performance of solar and hydrogen systems. Traditional forecasting methodologies, such as time-series decomposition, autoregressive integrated moving average (ARIMA) models [10], show weakness in modeling the non-linear aspects of the natural data properties of solar irradiance. On the contrary, machine-learning (ML) methods have emerged as a robust, data-driven alternative to perform high-resolution solar forecasting in the future, as they can capture intricate patterned. This capability in turn reduces for financial risk by providing better predictors for the reliability of energy output.

The objective of implementing ML in solar forecasting is that it is able to work with high volumes of multi-dimensional data and this may include meteorological parameters, satellite measurements, and past power production, to identify complex temporal and spatial behavior. The levelised cost of energy (LCOE) is directly affected by improved forecasting accuracy as it allows to make more accurate yield forecasts and minimise financial risk related to the underperforming of systems. The predictability of hydrogen, in particular, is that much more critical: an over- or under-estimate of the solar output may result in the extra expenses due to the acquisition of backup power, over-sizing of the electrolyser, or even the need to have the auxiliary reserves. As a consequence, a direct impact on the levelized cost of hydrogen (LCOH) is exerted by these factors.

The levelised cost of electricity (LCOE) for the renewable power source is one of the major factors that determine the economics of green hydrogen. In this context, the levelised cost of hydrogen (LCOH) may be regarded as an extension of LCOE that incorporates additional cost components, including expenditures (OPEX), capital expenditures(CAPEX) of electrolyzer, operational, stack replacement costs, and electricity delivery-related expenses. An average rate of electricity takes up a large portion of summative hydrogen manufacturing costs; therefore, any increase in solar forecasting accuracy can produce substantial revenues[4]. Any slight mishaps in forecasting solar accessibility can be cumulative throughout the existence of the project and result in inefficient use of electrolyzers and excessively high prices of hydrogen.

A poor prediction can influence the operators to purchase additional grid power when there is deficit or reduce production when the amount of sun decreases without prior notice resulting in rising LCOH. Conversely, accurate forecasting enables improved coordination between electrolyser operation and estimated photovoltaic (PV) generation in a real time or day ahead time horizons, thereby reducing idle capacity and lowering the unit cost of hydrogen. The need for a comparative evaluation of different machine learning ML algorithms is underlined by the link between the accuracy of forecasts and the competitiveness on price in order to determine which methodological approach leads to the greatest reduction for the levelised cost of electricity LCOE and, as a result, the levelised cost of hydrogen LCOH.

There are also past studies evaluating the effectiveness of single machine-learning based models to forecast the sun, but because of the variability of data sets, project surroundings, and testing procedures, these studies cannot be directly compared. Insights into the relative strengths and limitations of individual machine learning models can be obtained through a systematic framework in which multiple algorithms—including gradient-boosted regression (GBR), long short-term memory (LSTM), support vector regression (SVR), light gradient boosting machine (LightGBM), random forest (RF), and extreme gradient boosting (XGBoost)—are evaluated using identical datasets. Sequential and temporally dependent phenomena in the irradiance cycles are better analysed with Long Short Time Memory (LSTM) Networks, while the analysis of non linear dynamics, complex variable interactions are better modelled by ensemble based decision tree algorithms, such as Random Forests (RF), Gradient Boosting Regression (GBR), LightGBM and XGBoost. Moreover, SVR, the most famous algorithm that can map the existence of high-dimensional relationships during the application of kernel-related methods, exhibits strong functionality in the ability to represent complicated trends, in particular, when the number of training data is insufficient[5].

This study is new because a systematic comparative study of three ML methods is conducted, namely XGBoost, SVR, and LSTM, to present the idea of how even minor differences in forecasting can cause a change in the total costs of hydrogen production. It highlights the distinctive connection between ML-induced accuracy and hydrogen techno-economics and, therefore, copes with the gap where all three solar-hydrogen researchers at most tend to incorporate model comparisons with the direct implications of costs. The dataset used is the one provided by the system Photovoltaic Geographical Information System (PVGIS) which is 2005-2023 in the Tibetan plateau of China. In spite of the limited case study in the Tibetan region, which was selected because of its solar potential, and long-term history of irradiance, this paper is the best testbed of demonstrating model performance. Geographic and system-specific effects may be different, but the methodology can be easily scaled to other locations to give a versatile platform to replicate and adapt.

The solar forecasting accuracy of each model will be assessed, and then the effect of such variations of the solar forecast on values of LCOE of a hypothetical solar PV plant having a capacity of 100MW will be investigated. The resulting LCOE values will subsequently be integrated into an electricity-driven cost framework for alkaline electrolysis to estimate the levelised cost of hydrogen (LCOH). Within this comprehensive analytical framework, the paper seeks to establish a numerical measure of cost benefits and possible demerits of every forecasting approach. The study presents a new overview in the process of establishing how enhanced prediction accuracy has a direct positive correlation on the cost of green hydrogen when used in combination with multi-model solar forecasts and techno-economic models. This view has its implications to researchers who aim at perfecting forecasting algorithms and stakeholders who aim at industrialising renewable based hydrogen systems. Finally, the integration of solar prediction and hydrogen technology that can be cost-competitive will be the foundation of a more robust and decarbonised energy system in the whole world.

II. Literature Review

Transpiring over the last ten years, state-of-the-art machine-learning algorithms have become the competitive or even superior approaches to the traditional statistical techniques. Gradient-Boosted Regression (GBR) is an ensemble learning technique in which predictive performance is progressively enhanced through the sequential refinement of multiple base models, with each stage informed by the inaccuracies identified in earlier predictions. Previous studies have demonstrated the effectiveness of this approach when applied to datasets describing solar radiation and photovoltaic electricity generation, particularly in capturing complex, nonlinear atmospheric influences[6]. Despite these advantages, boosting-based methods typically incur substantial computational overhead and exhibit strong sensitivity to tuning parameters, such as step-size control and the total number of boosting iterations.

Random forest models consist of an ensemble of decision trees constructed through stochastic data resampling and randomised feature selection, resulting in a robust predictive framework. With their marked ability both to generalise and to be intrinsically compatible with parallel processing, models built on Random Forest have been widely used for the prediction of parameters related to the solar domain[7]. Random Forests are robust to noisy or incomplete data collections to an extent that is due to the aggregation of the predictions across individual trees, thus reducing the estimation variance. Nonetheless, the deeper the ensemble with the increase of tree number, the less the interpretability of the model.

XGBoost enhances the traditional methods of boosting algorithms to include complex mechanisms to gain the advantage of a structured simplification of the trees and a stronger regularisation, providing better predictive stability and efficiency .In contrast, LightGBM uses a different strategy in the construction of trees consisting in prioritising the expansion of deeper nodes, which often ends up in a faster model training speed and reduced memory use for the training of the boosting method. In addition, LightGBM makes use of discretised feature representations, which have been shown to increase computational efficiency when working with very large datasets[8]. When applied to the prediction of photovoltaic output, this method exhibits robust predictive abilities but at a much faster computational time thereby making it especially suitable to apply to cases which request updates to be performed on an abbreviated operational horizon.

Support Vector Regression (SVR) has recently been widely used for modeling solar-related variables because it can handle complex feature spaces and create nonlinear relationships through their kernel based transformation. Unlike tree-based ensemble methods that rely on recursive partitioning of data to extract structural relationships from the data, SVR constructs an optimal separating function in an abstract feature space which results in precise predictions and efficient generalisation performance[9].

As capital expenses attached to photovoltaic power systems keep dropping, the inherent problems of the intermittency and predictability of solar-generated electricity are increasingly acting on top of an economic risk profile. The levelised cost of energy (LCOE), is a metric that is routinely used as a measure to quantify project lifetime costs per unit of cumulative electricity output. Nonetheless, conventional formulations of LCOE have usually been based on fixed or representative utilisation assumptions, and thus do not account for short-term variability in production due to atmospheric situations or longer-term climatological cycles[10]. More sophisticated LCOE analyses incorporate time-sensitive generation sequences or also include output projections of probabilistic analyses to have an appropriate representation of actual operating conditions. Substantial reductions in capital expenditures (CAPEX) and concomitant improvements in operational efficiency can potentially be achieved by reducing dependency on oversized power electronics based electrical systems or energy - storage units; the empirical evidence indicates that even modest improvements in the accuracy of forecasts can lead to repercussions of this sort. Matters worse, refined production forecasts help project developers to access more favourable financing terms by reducing perceived revenue uncertainty. Overoptimistic yield projections can result in shortfalls in revenues if actual generation is less than projected generation while unrealistically conservative projections can hamper investment decisions or overestimate perceived financial risk [11]. Consequently, the accuracy of the forecast model(s) plays a key role in the realised LCOE.

When one uses electricity produced by solar photovoltaic systems to produce hydrogen, the cost of power generation needs to be added to other costs associated with electrolysis infrastructure, periodic component replacement needs, regular operational requirements as well as, where relevant, hydrogen handling and conditioning processes. The resulting levelised cost of hydrogen (LCOH) is expressed as a normalised indicator of unit hydrogen cost, either on a mass or energy basis. This metric enables consistent comparison across different hydrogen supply routes, as fossil-based production, e.g. grey and blue hydrogen, and renewable-based hydrogen production, connected to it. Given that the electrical input represents a large proportion of total cost in renewable hydrogen production systems - in many instances in excess of 50% of total costs - considerable economic benefits can be achieved by reducing the level of uncertainty in the availability of solar energy[12].

The intermittent nature of solar predictions which results in frequent open and closed cycles of electrolyzers also makes it more difficult to maintain stability of stacks and increase maintenance expenses. Similarly, when operators are forced to purchase supplement electricity on the grid, the cost, which is involved, can be more than the average LCOE used in initial feasibility analysis, and this will consequently overcharge the ultimate LCOH. Synergy between the advanced predictive system and hydrogen production scheduling is now available in pilot projects that help to sustain the argument that precise solar predictability would be a crucial pillar of the economical hydrogen green.

The integration of solar forecast information into cost modelling has been addressed through several conceptual approaches reported in the literature. Most studies initially quantify the electricity generation by photovoltaics in a detailed temporal resolution, and afterwards, the total system-related costs are measured in terms of the net present value (NPV), and the system-related costs are normalised by the sum of the discounted electricity generation during the lifetime of the project. A similar approach is used for hydrogen-based systems supplemented with explicit modelling based on electrolyzer-specific characteristics such as capacity factor, stack lifetime, and system efficiency in order to assess the net present value (NPV) of hydrogen production volumes over time[13]. As a result, the calculated levelised cost of hydrogen (LCOH) exhibits strong sensitivity to the accuracy of solar power forecasts and to the manner in which these forecasts are incorporated into operational dispatch or scheduling frameworks.

A major drawback on existing literature stems from the inadequate integration of the forecast outputs in hydrogen production modelling frameworks. Although some research is done to analyse the impact of solar forecast bias on the levelised cost of electricity (LCOE), it does not usually consider how the operation of electrolysers might dynamically be adjusted in response to predictive information. In contrast, research that focuses on the cost of hydrogen production tends to exclusively use static assumptions about photovoltaic capacity factors avoiding taking into account the possible contributions of state-of-the-art machine learning based forecasting methodology. Integrating these two strands of research, which are machine learning driven solar prediction and hydrogen techno-economic analysis, is promising in order to contribute more robust knowledge towards the cost implications for the alternative forecasting strategies[14].

The literature gives great importance to proper solar forecasting relating to the levelised cost of electricity (LCOE) and the levelised cost of hydrogen (LCOH)[15, 16]. Despite the multitude of machine-learning algorithms, there are comparably few studies that use a consistent data set and which do not combine these predictive models into a comprehensive techno-economical frameworks with a transition to cost of hydrogen production. The current research which follows a multivariate approach to the data from Tibetan is aimed at filling in these gaps. By incrementally testing XGBoost, support-vector regression (SVR), and long-short term memory (LSTM)-type models by examining how the accuracy of the various models can be utilized to refine the LCOE and LCOH calculations, the study adds new knowledge to the best strategies for further developing the solar hydrogen nexus.

III. Methodology

This study adopts an integrated methodological framework to evaluate the influence of machine-learning-based solar power forecasting on the techno-economic performance of green hydrogen production. The workflow combines data-driven solar forecasting with levelized cost of energy (LCOE) and levelized cost of hydrogen (LCOH) calculations to quantify how forecast accuracy propagates into hydrogen production costs. As shown in Fig. 1.

A. Data Description and Pre-processing

Hourly solar and meteorological data for the Tibetan region of China were obtained from the Photovoltaic Geographical Information System (PVGIS), covering the period from 2005 to 2023. Tibetan is characterized by high solar irradiance and hosts large-scale photovoltaic installations, making it a representative testbed for solar-driven hydrogen systems.

The dataset includes direct beam irradiance on an inclined surface (Gb(i)), diffuse irradiance (Gd(i)), reflected irradiance (Gr(i)), ambient temperature at 2 m height (T2m), wind speed at 10 m height (WS10m), and measured solar power output (P), which serves as the prediction target.

Data cleaning procedures were applied to remove physically unrealistic values and sensor anomalies. Observations below zero or exceeding system capacity were discarded, typically excluding the upper and lower 1% of values. Missing data gaps of up to two consecutive hours were linearly interpolated, while larger gaps were omitted. After cleaning, more than 95% of the original dataset was retained.

To capture temporal patterns, additional time-based features were engineered, including hour of day, day of year, and month. Lagged power and irradiance variables were also incorporated to enable models to learn temporal autocorrelations. All variables were aligned to a consistent hourly time index prior to model training.

B. Machine Learning Forecasting Models

Three machine-learning algorithms representing distinct modeling paradigms were selected: Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) neural networks. This selection enables a balanced comparison between tree-based ensemble learning, kernel-based regression, and deep sequence modeling.

XGBoost is an optimized gradient-boosting framework that incorporates regularization, tree pruning, and parallel computation to improve both predictive accuracy and computational efficiency. Key hyperparameters include tree depth, learning rate, number of estimators, and subsampling ratios.

SVR employs kernel functions to model nonlinear relationships in high-dimensional feature space. A radial basis function kernel was adopted due to its suitability for nonlinear solar data. Model performance is governed by the regularization parameter (C), epsilon-insensitive loss margin (ε), and kernel scale.

LSTM networks represent a specialized recurrent neural network architecture suitable for sequential data. By introducing gating mechanisms, LSTM units can capture long-range dependencies in time-series, making them well suited to data that exhibit daily and seasonal cycles. For solar forecasting, an LSTM model might parse a 24-h window of past data to predict the next hour's power output, with hidden layers typically set between 50 and 128 units. Although

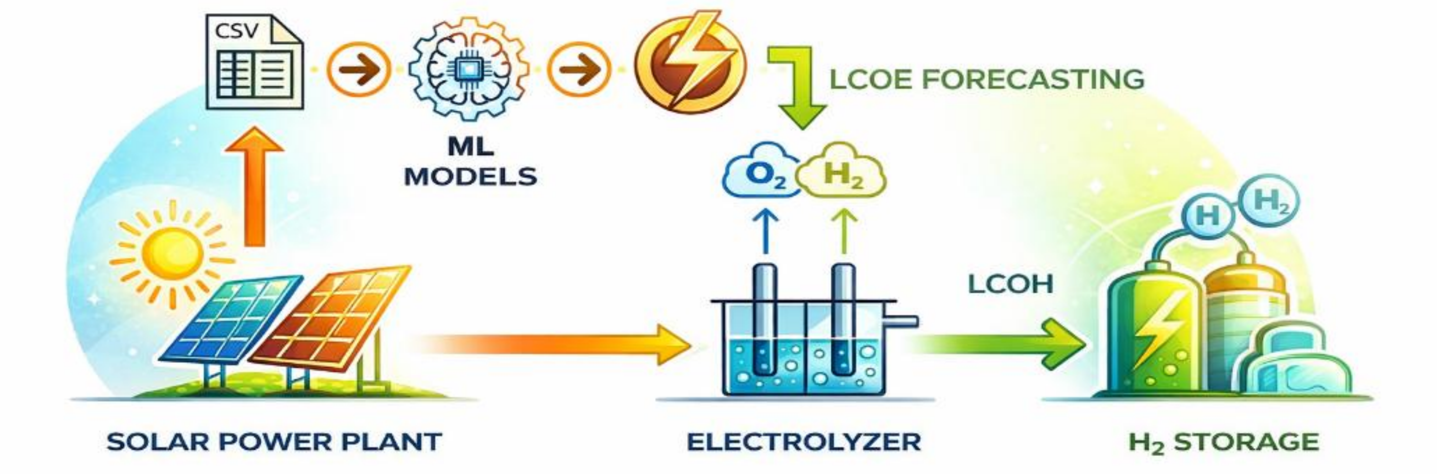


Figure 1 Workflow of solar-powered hydrogen production with ML forecasting.

training LSTMs can be more computationally intensive, their potential to learn complex temporal patterns justifies their inclusion in this study.

The LSTM model is initialized as a single-layer architecture with 50 hidden units, ReLU activation, a batch size of 64, and 50 training epochs using Adam at a 0.001 learning rate, with early stopping triggered if no improvement is observed for 10 consecutive epochs. A broader grid-like sensitivity analysis tested hidden units (32, 50, 64, 128), batch sizes (32, 64, 128), learning rates (0.001, 0.005, 0.01), and up to two stacked LSTM layers. Smaller batches often captured short-term fluctuations more effectively but introduced noisier gradients, whereas larger batches smoothed out short-term variations. Adding a second LSTM layer yielded modest accuracy improvements yet increased training time and overfitting risk, partially addressed by dropout rates of 5–20 %. Despite best-tuned configurations (64–128 hidden units, 0.001 learning rate, moderate dropout), the LSTM approach remained less effective than XGBoost, or SVR in modeling hourly fluctuations, likely due to higher data sensitivity and limited temporal granularity.

Computational efficiency is quantified by recording the training time for each algorithm on identical hardware (M3 Max with 16-core CPU and 40-core GPU, 36 GB RAM) and the same training subset (2005–2019). As reported in Table 1, XGBoost completed training in 0.45 s, benefiting from parallel tree construction and early-stopping; SVR required 174.6 s because kernel operations scale quadratically with the number of training points; and LSTM needed 266.4 s owing to back-propagation through time and multiple epochs over the data. These measurements clarify the practical trade-off between accuracy and runtime: XGBoost offers the fastest convergence with competitive predictive performance, whereas LSTM attains more flexible temporal modelling at a substantially higher computational cost, and SVR lies between the two in both accuracy and resource demand.

C. Model Training and Validation

The dataset was divided chronologically into a training set (2005–2019) and a testing set (2020–2023) to replicate real-world forecasting conditions. Feature correlations were examined using Pearson correlation coefficients, and highly collinear variables (>95%) were excluded where necessary.

Table 1 Comparison of the ML models used in this study.

Model name	Key hyperparameters and training time	Advantages	Potential limitations
XGBoost	- max_depth: 5 - learning_rate: 0.1 - subsample: 1 - colsample_bytree: 1 - reg_alpha, reg_lambda: 0, 1 - n_estimators: 200 - Training time: 0.45 s	- Handles non-linear relationships well - Built-in regularization reduces overfitting risk - Highly scalable and efficient on large datasets	- Requires careful tuning of multiple hyperparameters - Can overfit if depth or learning rates are not chosen properly
SVR	- C (regularization parameter): 100 - ε (epsilon-insensitive loss margin): 0.1 - Kernel: Radial basis function - Gamma: Scale - Training time: 174.6 s	- Good for moderate-sized datasets - Capable of modeling complex non-linear patterns with kernel trick - Often robust with fewer features if well-tuned	- Computationally expensive with large datasets - Selecting an appropriate kernel can be non-trivial - Sensitive to outliers and hyperparameter choices
LSTM	- Number of LSTM units: 50 - batch_size: 64 - learning_rate: 0.001 - (Adam default) - number_of_layers: 1 - Training time: 264.4 s	- Excellent at modeling time dependencies - Can handle non-stationary signals and seasonality - Learns complex temporal patterns	- Often needs extensive hyperparameter tuning and large training sets - Risk of overfitting if regularization is not applied - High computational cost and training time

Model performance was evaluated using standard solar forecasting metrics: root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), mean bias error (MBE), residual standard deviation (RSD), and coefficient of determination (R^2). These metrics capture both absolute and relative prediction accuracy as well as systematic bias.

Hyperparameter optimization was conducted using grid and randomized search strategies. Early stopping was applied for XGBoost and LSTM models based on validation loss. For SVR, kernel type and regularization parameters were tuned via cross-validation.

Computational efficiency was assessed by recording training times under identical hardware conditions, highlighting practical trade-offs between model accuracy and runtime. Commonly used metrics in solar forecasting include the root mean squared error (RMSE), the mean absolute error (MAE), the mean absolute percentage error MAPE), mean-bias error (MBE), the residual standard deviation (RSD), and coefficient of determination (R2) are given in Eq. (1), Eq. (2), Eq. (3), Eq. (4), Eq. (5), and Eq. (6)respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{2}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{3}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right) \tag{4}$$

$$RSD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\hat{y}_i - y_i\right) - MBE\right]^2} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(\hat{y}_i - y^2\right)^2} \tag{6}$$

where, n is the number of samples used for statistical evaluation criteria, $\hat{y}_i$ is the actual value of the observation, y_i is the forecasted value of the observation, and y is the average of the actual observation values. These metrics capture different aspects of model performance. MAE provides an intuitive measure of average error magnitude, while RMSE penalizes larger errors more severely, thus emphasizing peak deviations. MAPE illustrates percentage-based deviations relative to actual power output, though it can become inflated if the actual values are near zero. The MBE quantifies systematic over or under-prediction, while the RSD measures the spread of the errors once the mean bias has been removed.

Hyperparameter tuning is undertaken to optimize model performance. In the ensemble methods, randomized or grid searches are performed over learning rate values, max_depth settings, and subsample or feature-sample fractions. Early stopping based on validation set performance is employed to prevent overfitting, especially in high-capacity models like XGBoost. The LSTM model undergo a smaller but focuses tuning process involving hidden layer size, batch size, and the number of epochs. Early stopping is also integrated into LSTM training to avoid unnecessary computations and to mitigate overfitting. Similarly, for SVR, tuning focused on selecting the most suitable kernel function (linear, polynomial, or radial basis function), optimizing the regularization parameter (C), and adjusting the epsilon parameter to balance prediction accuracy and generalization. Grid search and cross-validation techniques are applied to refine these parameters, ensuring robust performance in forecasting solar power output.

D. LCOE Calculation Based on Forecasted Solar Output

Once the forecasted solar outputs for each model are finalized, they are integrated into a financial model for the LCOE. The general expression for LCOE over a project lifetime *T* and discount rate *τ* is given by Eq. (7):

$$LCOE = \frac{\sum_{t=0}^{T}\frac{(I_t + O_t)}{(1+\tau)^{\tau}}}{\sum_{t=0}^{T}\frac{E_t}{(1+\tau)^{\tau}}} \tag{7}$$

where I_t represents CAPEX in year *t*, O_t corresponds to annual OPEX, and E_t is the total energy produced in year *t*. In this study, the CAPEX is primarily front-loaded at *t*=0, reflecting the installation costs of a notional 100 MW solar PV system, assumed at $70 million. The OPEX is estimated at $1.1 million annually for routine maintenance and operational overhead.

Forecasted hourly solar power outputs from each ML model were aggregated to estimate annual energy production. A panel degradation rate of 0.5% per year was assumed over a 25-year project lifetime. The levelized cost of electricity (LCOE) was calculated using discounted cash-flow analysis with an 8% discount rate. To obtain the annual energy E_t, each model's hourly predictions are summed and converted to MWh. When the model predicts power $\hat{P}(t)$ in watts at hour *t*, the annual total can be computed by aggregating Eq. (8):

$$E_{\text{year}} = \sum_{\text{hours in year}} \frac{\hat{P}(t)}{1000} \tag{8}$$

Capital expenditure (CAPEX) for the solar PV system was assumed to be front-loaded at project initiation, while annual operational expenditure (OPEX) covered routine maintenance. The LCOE formulation incorporates discounted costs and energy production, ensuring consistency across forecasting scenarios.

E. LCOH Calculation for Solar-Powered Hydrogen Production

The calculated LCOE values were subsequently integrated into a techno-economic model for alkaline water electrolysis to estimate the levelized cost of hydrogen. Hydrogen production costs include electrolyzer CAPEX, OPEX, periodic stack replacement, and electricity costs derived from solar PV as shown in Fig. 2. The formula for LCOH over a project lifetime *T* is expressed as Eq. (9):

$$LCOH = \frac{\sum_{t=0}^{T} \frac{\text{Costs}_t}{(1+\tau)^t}}{\sum_{t=0}^{T} \frac{H_2\text{produced}_t}{(1+\tau)^t}} \tag{9}$$

where $Cost_t$ encompasses electrolyzer capital expenses, periodic stack replacements, and electricity expenditures determined by the solar LCOE. The term $H_{2,produced,t}$ represents the discounted hydrogen output, reflecting the electrolyzer's annual production. The discount rate τremains consistent with the LCOE computations, here set at 8%. The yearly capital term already includes discrete stack-replacement outlays (15% of the initial electrolyzer CAPEX). The lifetime of the electrolyzer stack is considered to be 100,000 h, aligning with industry benchmarks for advanced electrolyzer technologies. The energy consumption required for hydrogen production is assumed to be 49 kWh per kilogram of hydrogen, reflecting the efficiency of the alkaline electrolyzer system. The OPEX is estimated at 2.5% of the CAPEX per year, accounting for ongoing maintenance and operational costs. The CAPEX for the alkaline electrolyzer system is set at $530/kW, a value representative of current market conditions for electrolysis technology. Additionally, engineering, procurement, and construction (EPC) costs are considered as 30% of the CAPEX for the electrolyzer system, covering the costs of system installation and infrastructure development. Finally, stack replacement costs are assumed to be 15% of the CAPEX for the electrolyzer system, reflecting periodic maintenance and replacement expenses associated with the electrolyzer's operational lifespan. These parameters form the basis for calculating LCOH, ensuring that economic and technical factors are appropriately considered in the cost assessment.

By applying each model's forecast to the LCOE calculation and subsequently feeding that LCOE into the LCOH formula, the methodology clarifies how forecasting differences propagate into hydrogen cost. With the updated results, the solar LCOE values remain tightly clustered (approximately $39.35–$39.81/MWh across the evaluated models), leading to a correspondingly narrow spread in LCOH outcomes (about $3.51–$3.54/kg H_2). This indicates that, under the adopted techno-economic assumptions, model-to-model variations in forecast accuracy translate into only modest shifts in the implied electricity cost component of hydrogen production. Nevertheless, systematic bias can still influence cost interpretation: a model that persistently overestimates PV output may yield a slightly lower apparent LCOE and thus underestimate the electricity-driven share of LCOH, whereas a more conservative forecast can inflate LCOE and push LCOH upward. The empirical consistency of the updated LCOE–LCOH results therefore underscores both the robustness of the cost model to small forecasting differences and the continuing importance of minimizing forecast bias when designing economically viable solar-hydrogen systems.

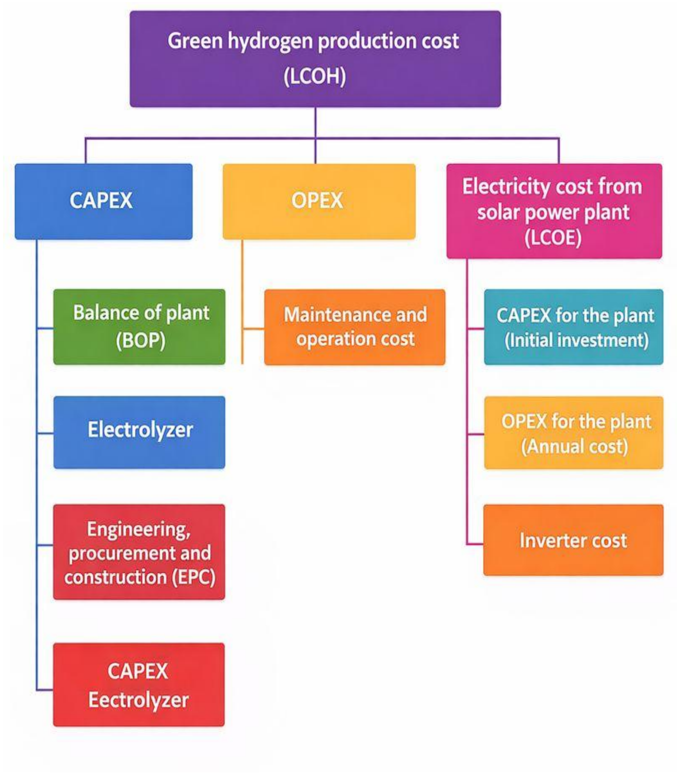

Green hydrogen production cost
(LCOH)
CAPEX
OPEX
Electricity cost from solar power plant (LCOE)
Balance of plant (BOP)
Electrolyzer
Engineering, procurement and construction (EPC)
CAPEX Eectrolyzer
Maintenance and operation cost
CAPEX for the plant (Initial investment)
OPEX for the plant (Annual cost)
Inverter cost

Figure 2 Breakdown of levelized cost of hydrogen, highlighting the contributions of CAPEX (balance of plant, electrolyzer, EPC), OPEX, and electricity costs from solar power plants (LCOE, including plant CAPEX, OPEX, and inverter costs).

F. Sensitivity and Uncertainty Considerations

To assess robustness, sensitivity analysis was conducted by varying key economic and technical parameters (±20%), including electricity cost, CAPEX, OPEX, discount rate, electrolyzer lifetime, full load hours, and specific energy consumption. This analysis identifies dominant cost drivers and quantifies the relative importance of forecasting accuracy within the broader economic framework.

IV. RESULTS

A. Solar Power Forecasting Accuracy

Accurate solar power forecasting is essential for reliable system operation and cost-effective hydrogen production. In this study, the predictive performance of three machine-learning models XGBoost, Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) was evaluated using multiple statistical metrics, including RMSE, MAE, MBE, RSD, MAPE, and R². Table 2 presents the comparative performance of these models, while Fig. 3 provides a visual representation of their predictions against actual solar power output. Table 2 highlights significant differences in the forecasting capabilities of the evaluated models. XGBoost, a tree-based ensemble learning method, demonstrates a robust predictive performance with an RMSE

Table 2 Each model's RMSE, MAE, MBE, RSD, MAPE, and R2.

Model name	Model type	RMSE (MW)	MAE (MW)	MBE (MW)	RSD (MW)	MAPE (%)	R^2 (%)
XGBoost	Tree-based gradient boosting	0.51	0.38	0.03	0.51	2.43	0.996
SVR	Kernel-based machine learning	0.28	0.18	0.04	0.27	10.29	0.999
LSTM	Deep learning	6.39	4.15	0.69	6.61	12.35	0.902

of 0.51 MW, a MAE of 0.38 MW, an MBE of 0.03 MW, an RSD of 0.51 MW, and a low MAPE of 2.43 %. The high R^2 value of 0.996 indicates a strong correlation between the predicted and actual values, confirming XGBoost's ability to capture complex nonlinear patterns in solar irradiance data.

SVR exhibits the best RMSE (0.28 MW), MAE (0.18 MW), MBE (0.04 MW), and RSD (0.27 MW) among the tested models, indicating lower absolute prediction errors. Additionally, it achieves an exceptionally high R^2 of 0.999, suggesting a near-perfect fit to the actual data. However, its MAPE value of 10.29 % is considerably higher than that of the XGBoost, implying that SVR struggles with relative percentage errors, particularly for low-magnitude predictions. This suggests that while SVR excels in capturing fine-scale variations, it may be sensitive to fluctuations in irradiance levels and may require further tuning for improved generalizability.

Conversely, LSTM, a deep learning-based recurrent neural network, performs notably worse than the other models, with an RMSE of 6.39 MW, a MAE of 4.15 MW, an MBE of 0.69 MW, and an RSD of 6.61 MW. Despite its theoretical advantages in modeling sequential dependencies, its R^2 value of 0.902 suggests a weaker correlation with actual observations. Moreover, the MAPE of 12.35 % indicates significant relative errors, making LSTM less reliable for accurate solar forecasting in this context. The model's underperformance may be attributed to its computational demands and sensitivity to hyperparameter selection, which could lead to suboptimal learning of temporal dependencies in the dataset.

The findings indicate that tree-based gradient boosting models like XGBoost provide a strong balance between accuracy and efficiency, making them well-suited for real-time forecasting applications in energy systems. SVR's superior RMSE, MAE, MBE, and RSD demonstrate its potential for precise numerical predictions, although its high MAPE suggests challenges in proportional error minimization. While LSTM's relatively poor performance in this study suggests that deep learning may not always be the best approach for solar forecasting, further investigations could explore hybrid models that combine LSTM with other architectures to improve its predictive capability.

These findings align with existing literature where various machine learning models have been applied to solar power forecasting with differing outcomes. For instance, a study employing an artificial neural network (ANN) optimized with the Levenberg–Marquardt algorithm reported an exceptionally low RMSE of 0.0039 and an R^2 of 0.99994, suggesting a near-perfect prediction capability. Similarly, research utilizing SVMs achieved an RMSE of 2.193 and an R^2 of 0.999959, reflecting strong predictive performance. Conversely, models such as the Harris Hawks Optimizer have demonstrated less precise predictions, with an RMSE of 70.53 and an R^2 of 0.9703. Random Forest models have shown considerable promise, achieving an R^2 of 0.997. Deep neural networks have also been explored, with one study reporting an RMSE of 3.3 and an R^2 of 0.9998.

The time-series predictions of each model against the actual solar power output over a sample period (2022–2023) as shown in Fig. 3. The graph provides an insightful visualization of how well each model captures fluctuations in solar irradiance and power generation.

The XGBoost model exhibits a strong alignment with actual values, with minor deviations occurring primarily during rapid changes in solar irradiance. This suggests that the model is well-suited for forecasting solar power, especially in applications requiring stable and reliable predictions. The SVR model follows a similar trend, with high accuracy in capturing overall variations. However, closer examination reveals slight underestimations and overestimations in peak values, indicating potential sensitivity to outliers. The SVR model's higher MAPE also suggests difficulties in predicting periods of low solar output, which could be critical for optimizing energy storage and grid integration strategies.

On the other hand, the LSTM model demonstrates more significant deviations, particularly during transition periods between high and low irradiance levels. The model struggles to track rapid fluctuations, often smoothing out sharp variations, which results in larger errors compared to XGBoost and SVR. This behavior suggests that while LSTM can capture long-term trends, it is less effective at handling short-term variability in solar power generation. The lag effect observed in the LSTM predictions further highlights its limitations in real-time applications where immediate response to changes in solar irradiance is crucial.

The discrepancies observed in Fig. 3 reinforce the importance of selecting appropriate forecasting models based on specific application needs. While XGBoost provides the most balanced and accurate predictions overall, SVR offers excellent numerical precision with some limitations in proportional accuracy. LSTM's underperformance underscores the need for further optimization or hybrid approaches that combine deep learning with traditional machine-learning techniques.

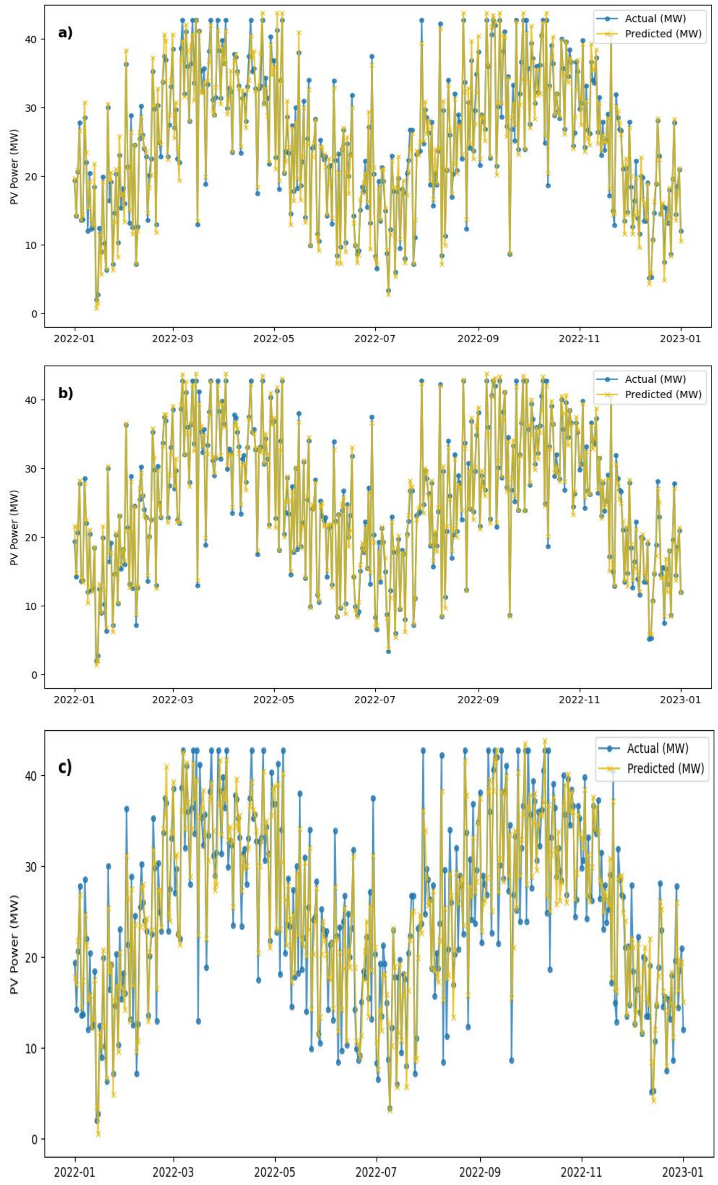


Figure 3 Possibly show time-series plots of each model's (actual vs. predicted) for a period (2019–2020) a) XGBoost, b) SVR, and c) LSTM.

B. Time-Series Comparison of Model Predictions

A time-series comparison between actual and predicted solar power outputs over a representative period highlights distinct behavioral differences among the models. XGBoost closely follows the observed power profile, with only minor deviations during abrupt irradiance changes. This suggests strong adaptability to short-term variability, which is critical for real-time energy management.

SVR also captures the general power trend effectively but exhibits occasional under- and overestimations near peak values. These deviations contribute to its higher MAPE and indicate sensitivity to outliers or low-power conditions. While SVR excels in minimizing absolute error, its proportional accuracy remains comparatively weaker.

The LSTM model displays pronounced smoothing behavior, failing to track sharp transitions between high and low irradiance levels. This lag effect results in systematic underperformance during rapidly changing conditions, limiting its suitability for applications requiring precise short-term forecasts.

C. Impact of Forecast Accuracy on LCOE and LCOH

The results indicate that forecast accuracy plays a critical role in determining LCOE and LCOH. XGBoost, which demonstrated the highest overall accuracy in solar power prediction, yields an LCOE of \$39.76/MWh and an LCOH of \$3.53/kg H_2. Conversely, SVR, while achieving superior RMSE and MAE in solar forecasting, results in a slightly higher LCOH of \$3.54/kg H_2. The LSTM model, which exhibited the lowest forecast accuracy, surprisingly achieves the lowest LCOH at \$3.51/kg H_2, coupled with the lowest LCOE of \$39.35/MWh as shown in Table 3.

Table 3 Comparison of LCOE and LCOH values derived from different forecasting models.

Model name	LCOE forecast (\$ /MWh)	LCOH forecast (\$ /kg H_2)
XGBoost	39.76	3.53
SVR	39.81	3.54
LSTM	39.35	3.51

One possible explanation for LSTM's relatively lower LCOH despite its poor forecasting performance is its tendency to smooth fluctuations in solar power predictions. This characteristic may lead to a more stable electricity cost component in LCOH calculations, reducing the volatility associated with rapid fluctuations in solar generation. However, this effect does not necessarily translate into higher operational reliability. LSTM's lower accuracy in capturing real-time variations could introduce inefficiencies in electrolyzer scheduling. The stability of electricity cost is a major determinant of LCOH, as electricity represents the largest portion of hydrogen production expenses.

A comparison of the best and worst forecasting approaches reveals a small percentage difference in LCOH values. The XGBoost model, which provides the most balanced forecast accuracy across all metrics, results in an LCOH of \$3.53/kg H_2, whereas the LSTM model results in a marginally lower LCOH of \$3.51/kg H_2. The percentage difference between these two models is approximately 0.57 %, which is relatively small. However, when comparing XGBoost and SVR, the LCOH difference is even smaller at 0.28 %, indicating that both models provide nearly equivalent economic outcomes.

Although the numerical differences in LCOH appear minor, their implications can be significant for large-scale hydrogen production. A lower LCOH translates to reduced hydrogen costs over the lifetime of a project, impacting investment decisions, profitability, and competitiveness in the energy market. Moreover, the cost variations highlight how different forecasting errors affect the financial outcomes of hydrogen production. Small forecasting inaccuracies, when accumulated over time, can lead to substantial cost deviations in green hydrogen production.

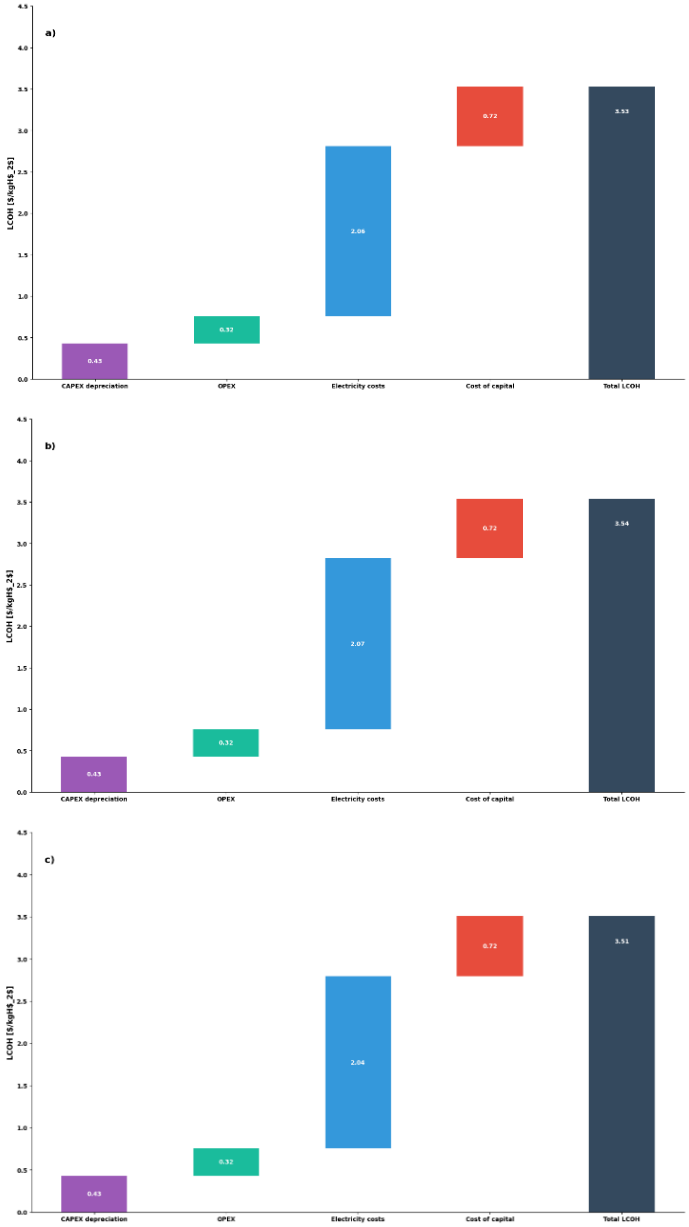


Figure 4 Breakdown of LCOH components for different forecasting models a) XGBoost, b) SVR, and c) LSTM. The variations in electricity costs highlight the impact of forecasting accuracy on hydrogen production economics.

Interestingly, the model with the most accurate solar forecast (XGBoost) does not yield the lowest LCOE or LCOH. This result suggests that while forecast accuracy is a key factor in reducing economic uncertainty, other elements, such as the nature of forecast errors, their impact on power variability, and how these affect electrolyzer operation, also contribute to final cost estimations. The LSTM model's ability to smooth out solar power fluctuations might contribute to lower perceived costs, even though its predictive performance is inferior. However, this benefit may not be sustainable in real-world operations where grid integration and energy storage need more precise short-term forecasting.

Among the evaluated models, XGBoost appears to be the most robust to data noise, as it consistently performs well across all key forecasting metrics and cost projections. The SVR model, despite its strong absolute accuracy, exhibits sensitivity to relative errors, leading to slightly higher electricity cost projections and LCOH values. LSTM, while demonstrating higher tolerance to input fluctuations, fails to capture real-time variations effectively, which could be detrimental in real-world hydrogen production scenarios where accurate short-term predictions are crucial.

Robust forecasting models play a vital role in optimizing hydrogen production scheduling, reducing reliance on auxiliary energy sources, and minimizing electrolyzer degradation. Models that are more resilient to variations in solar irradiance can help prevent sudden changes in hydrogen production rates, thereby maintaining operational stability. The economic feasibility of green hydrogen projects depends not only on the absolute accuracy of solar forecasts but also on their reliability under fluctuating conditions.

In addition to the 25-year forecast-based analysis, a short-term baseline is established using actual observed data from 2019 to 2023, yielding an LCOE of \$33.06/MWh and an LCOH of \$3.22/kg H_2. By contrast, the XGBoost model predicts \$39.76/MWh and an LCOH of \$3.53/kg H_2 when extended over the full 25-year lifecycle. This discrepancy largely reflects the long-term financial and technical assumptions embedded in a multi-decade projection, such as solar panel degradation, discounting of future cash flows, and inverter replacement cost in mid-lifecycle. The shorter five-year window does not capture these factors; hence it reports a lower average cost. Inter-annual variations in irradiance and operational strategies further contribute to higher projected costs when extrapolating over 25 years. Consequently, while the 2019–2023 baseline demonstrates actual near-term performance, the forecast results provide a more comprehensive view of costs over an entire project lifespan.

Fig. 4 provides a visual breakdown of LCOH components for XGBoost, SVR, and LSTM models. The waterfall chart illustrates how CAPEX depreciation, OPEX, electricity costs, and cost of capital contribute to the total LCOH for each model. As expected, electricity costs constitute the largest share of LCOH, highlighting the significance of solar power forecasting accuracy in determining hydrogen production costs.

The figure shows that while CAPEX depreciation and cost of capital remain constant across all models, slight variations in electricity costs result in different LCOH values. The LSTM model exhibits the lowest electricity cost at \$2.04/kg H_2, which contributes to its marginally lower LCOH. Conversely, SVR has the highest electricity cost at \$2.07/kg H_2, leading to a slightly increased LCOH of \$3.54/kg H_2. These small differences indicate that electricity price fluctuations, driven by forecasting accuracy, have a tangible impact on hydrogen production economics.

These findings underscore the intricate relationship between forecast accuracy and economic feasibility in green hydrogen production. While higher forecast accuracy generally contributes to improved cost estimations, certain models such as LSTM may yield counterintuitive results due to their impact on power variability. The LCOH calculations reveal that electricity cost is the dominant component in hydrogen production expenses, making solar forecasting accuracy crucial for financial planning. However, other factors, such as electrolyzer efficiency, system flexibility, and operational strategies, also influence final costs.

In a recent study, an LSTM neural network is employed to forecast solar PV output for a 100 MW system, achieving a MAE of 0.0415, RMSE of 0.0891, and R^2 of 0.8402. This predictive accuracy corresponded to a LCOE of \$25/MWh and an LCOH of \$6.95/kg HH_2 [16]. In another investigation, an ANN algorithm is applied to optimize auto-thermal reforming and steam methane reforming processes for hydrogen production. The model demonstrated R^2 of 0.9936 and an MSE of 6.88×10^{-5}, resulting in an LCOH of \$5.63/kg H_2. These studies underscore the efficacy of machine learning methodologies, such as LSTM and ANN, in enhancing the accuracy of energy production forecasts and optimizing hydrogen production processes, thereby contributing to more competitive LCOE and LCOH values.

C. Sensitivity Analysis of LCOH

The sensitivity analysis assesses the influence of variations in discount rate, CAPEX, OPEX, electrolyzer lifetime, full load hours, electricity costs, and specific energy consumption on LCOH. Fig. 5 presents the percentage changes in LCOH for each model when these parameters are adjusted by ±20%. Although the base-case assumes PV-only supply, the ±20% band applied to the electricity-price variable implicitly covers typical excursions that would arise from partial reliance on grid electricity or from blending solar output with a complementary on-site wind resource; a full techno-economic optimization of a dedicated solar-wind hybrid configuration is identified as valuable follow-up work.

The discount rate has a notable effect on LCOH across all models, with an increase of 20% leading to an LCOH rise of approximately \$0.182/kg HH_2 and a decrease of 20% resulting in an LCOH reduction of \$-0.173/kg HH_2. This highlights the importance of financial conditions in hydrogen production, as higher discount rates increase the present value of future costs, thereby raising the overall cost of hydrogen.

Changes in CAPEX directly impact LCOH, as reflected in the ±0.302 \$/kg HH_2 variation observed in all models. This result underscores the capital-intensive nature of hydrogen production, where initial investment costs significantly affect long-term economic feasibility. Similarly, OPEX fluctuations lead to an LCOH change of ±0.066 \$/kg HH_2, indicating that while operational costs contribute to hydrogen pricing, their effect is less pronounced compared to CAPEX.

Extending the electrolyzer system's lifetime by 20% results in a reduction of LCOH by \$0.046/kg HH_2, while shortening it by 20% increases LCOH by \$0.081/kg HH_2. These findings suggest that improvements in electrolyzer durability can play a crucial role in lowering hydrogen production costs. Full load hours exhibit an even stronger effect, where an increase of 20% reduces LCOH by \$0.252/kg HH_2, and a 20% decrease raises LCOH by \$0.378/kg HH_2. This highlights the importance of maximizing electrolyzer utilization to achieve cost-effective hydrogen production.

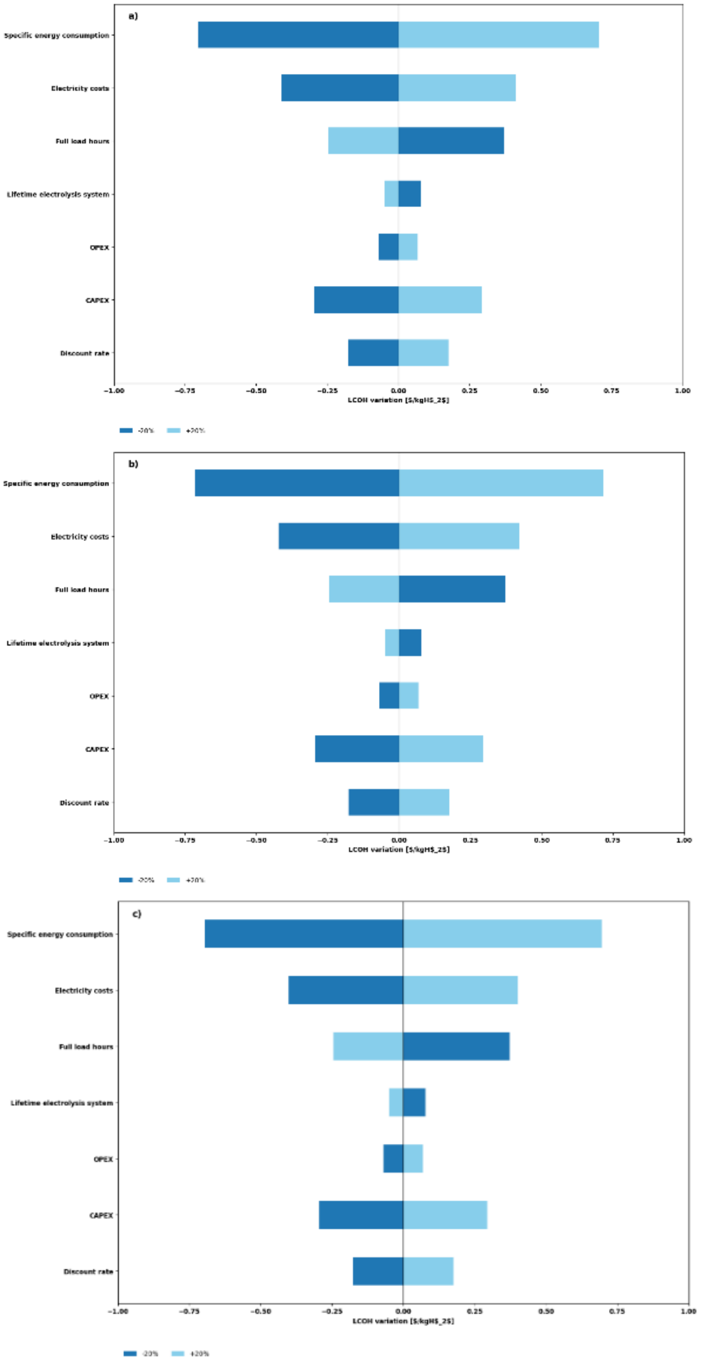


Figure 5 Sensitivity analysis of LCOH for a) XGBoost, b) SVR, and c) LSTM models.

Electricity costs have the most substantial impact on LCOH among all parameters analyzed. A 20% increase in electricity costs raises LCOH by approximately \$0.421/kg HH_2 (SVR), \$0.421/kg HH_2 (XGBoost), and \$0.417/kg HH_2 (LSTM), whereas a 20% reduction leads to an equivalent LCOH decrease. These results emphasize the critical role of electricity pricing in determining the economic viability of green hydrogen. Similarly, changes in specific energy consumption significantly affect LCOH, with variations of ±0.723 \$/kg $H H_2$ suggesting that improving electrolyzer efficiency can greatly enhance hydrogen cost competitiveness.

Moreover, comparing the largest hourly forecast residual observed (~6.5 MW in the LSTM case) through the 25-year cash-flow model changes the calculated LCOH by < 0.02 \$/kg HH_2, two orders of magnitude smaller than the +0.30 \$/kg $H H_2$ swing examined in the ±20% electricity-cost sensitivity. Thus, the economic ranking of the ML scenarios is effectively insensitive to this additional source of uncertainty.

While the three forecasting models exhibit similar sensitivity trends across all parameters, slight variations in the magnitude of impact are observed, particularly for electricity costs and specific energy consumption. The XGBoost and SVR models show slightly higher sensitivity to electricity cost changes compared to LSTM, reinforcing the link between forecast accuracy and hydrogen production economics. LSTM appears to exhibit slightly lower sensitivity to variations in electricity costs and specific energy consumption, likely due to its tendency to smooth fluctuations in power predictions.

D. Performance of Additional Tree-Based Models

LightGBM is a gradient boosting library designed for rapid training and high accuracy through histogram-based feature binning and leaf-wise tree growth . Its key hyperparameters include the learning rate, which controls the incremental contribution of each tree, the number of estimators determining how many boosting rounds have performed, and parameters controlling the maximum tree depth and sampling fractions. LightGBM's capacity for handling large datasets, combined with its leaf-wise splitting strategy, makes it appealing for this long-term Konya dataset, which spans 18 years of hourly data. GBR relies on boosting ensembles of weak learners (decision trees) by iteratively fitting new models to the residual errors of the previous models . Its learning rate, maximum depth of the base learners, and sub-sampling parameters are central to its performance and risk of overfitting. Although it can be relatively slower than some optimized frameworks, GBR remains a widely used and interpretable approach for non-linear regression tasks in the solar domain. RF is an ensemble method that builds numerous decision trees on bootstrapped data samples and randomly selected subsets of features . The averaging of predictions from multiple trees helps minimize variance and overfitting, making RF particularly effective in capturing complex interactions across features. Essential parameters include the number of trees, the maximum depth of each tree, and the number of features considered at each split. RF also offers consistent performance with limited hyperparameter tuning, which is advantageous for large-scale time-series data.

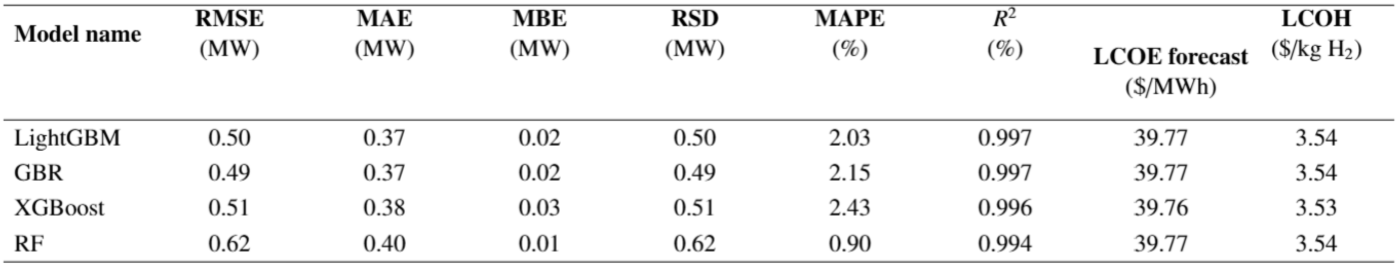
Tree-based gradient boosting models are widely utilized in solar power forecasting due to their ability to capture complex nonlinear relationships in meteorological and power generation data. These tree-based models, LightGBM, GBR, and RF, are evaluated for their performance in predicting solar power output and their subsequent impact on the LCOE and LCOH to compare with XGBoost. Table 4 presents the comparative performance metrics for each model.

The results indicate that GBR achieves the lowest RMSE (0.50 MW) and LightGBM follows closely with an RMSE of 0.51 MW. Both models also share the highest R² value of 0.995, indicating a strong correlation between predicted and actual values. These results demonstrate that GBR and LightGBM provide highly accurate solar forecasts, making them suitable for energy management applications. XGBoost, another gradient boosting model, exhibits a slightly higher RMSE of 0.52 MW and an R² value of 0.994, while its performance is marginally lower than LightGBM and GBR, it remains a competitive choice for solar forecasting, particularly given its efficiency in handling large datasets and reducing overfitting through advanced regularization techniques.

RF shows the highest RMSE (0.63 MW) and a lower R² of 0.992, suggesting that it may not capture complex variations in solar power generation as effectively as gradient boosting models. However, its MBE of 0.01 MW and MAPE of 0.92 % is the lowest among all models, indicating that RF performs well in relative percentage error reduction. This suggests that RF may be particularly effective in scenarios where accurate relative predictions are more critical than absolute error minimization.

Despite variations in forecasting accuracy, all four models yield similar LCOE (\$40.57–\$40.58/MWh) and LCOH (\$3.60–\$3.61/kg H_2). This suggests that minor differences in forecasting errors do not significantly impact the overall economic feasibility of hydrogen production. However, the stability of the predicted solar power outputs plays a crucial role in optimizing electrolyzer operation, minimizing reliance

Table 4 Performance comparison of tree-based gradient boosting models for solar power forecasting.



Model name	RMSE (MW)	MAE (MW)	MBE (MW)	RSD (MW)	MAPE (%)	R^2 (%)	LCOE forecast (\$/MWh)	LCOH (\$/kg H_2)
LightGBM	0.50	0.37	0.02	0.50	2.03	0.997	39.77	3.54
GBR	0.49	0.37	0.02	0.49	2.15	0.997	39.77	3.54
XGBoost	0.51	0.38	0.03	0.51	2.43	0.996	39.76	3.53
RF	0.62	0.40	0.01	0.62	0.90	0.994	39.77	3.54

on grid electricity, and reducing hydrogen production costs in the long run.

The LCOH is a critical metric for evaluating the economic feasibility of hydrogen production methods. In this study, mainly three machine learning models, XGBoost, SVR, and LSTM (selected based on the ML model type), are evaluated to forecast solar power output and subsequently assess their impact on LCOH. The resulting LCOH values range from \$3.58 to \$3.61/kg H_2.

E. Comparison with Literature Benchmarks

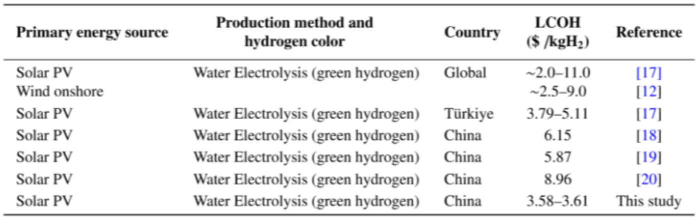
To contextualize these findings, the results are compared with those from recent studies in literature. Table 5 presents a comparative analysis of LCOH values from various studies, highlighting regional variations, primary energy sources, and production methods. The results indicate a significant range in LCOH values, primarily due to differences in electricity costs, electrolyzer efficiencies, and financial assumptions such as CAPEX and OPEX. Globally, hydrogen production through electrolysis powered by solar PV exhibits an LCOH range of approximately \$2.0 to \$11.0/kg H_2, while onshore wind-powered electrolysis demonstrates a slightly narrower range of \$2.5 to \$9.0/kg H_2. The broad range in global LCOH values reflects regional disparities in electricity costs, policy incentives, and technology maturity. Countries with abundant solar and wind resources coupled with low-cost electricity generation tend to have lower LCOH values, while regions with high CAPEX and financing costs result in higher hydrogen prices.

Focusing on Türkiye, multiple studies report solar PV-based electrolysis LCOH values ranging from \$3.79 to \$8.96/kg H_2. The variation in these values is attributed to differences in electricity pricing structures, system efficiencies, and financial modeling parameters. One study estimates an LCOH of \$3.79 to \$5.11/kg H aligning closely with cost projections in solar-rich regions. In contrast, other studies report higher LCOH values of \$6.15/kg H_2, \$5.87/kg H_2, and \$8.96/kg H_2, reflecting the influence of higher CAPEX, lower electrolyzer efficiencies, or more conservative financial assumptions.

For policy context, the reported LCOE of \$39.53–\$39.61/MWh and LCOH of \$3.53–\$3.61/kg H_2 from the IEA Global Hydrogen Review 2024 for 2030 and within striking distance of Türkiye's national targets for 2035. The forecasted values for this study are \$39.77/MWh and \$3.53/kg H showing consistency with reported benchmarks, further emphasizing the need for cost-competitive hydrogen production models.

The results from this study provide a more competitive LCOH range of \$3.58 to \$3.61/kg H_2 compared to recent reported values for Türkiye, which are \$3.79–\$5.11/kg H_2. This indicates that machine learning models like XGBoost, GBR, and LightGBM can optimize solar power forecasting and subsequent cost reductions, especially for regions with abundant solar resources, ultimately driving down hydrogen production costs.

Table 5 Comparison of LCOH values with reference studies.



Primary energy source	Production method and hydrogen color	Country	LCOH (\$ /kgH_2)	Reference
Solar PV	Water Electrolysis (green hydrogen)	Global	~2.0–11.0	[17]
Wind onshore			~2.5–9.0	[12]
Solar PV	Water Electrolysis (green hydrogen)	Türkiye	3.79–5.11	[17]
Solar PV	Water Electrolysis (green hydrogen)	China	6.15	[18]
Solar PV	Water Electrolysis (green hydrogen)	China	5.87	[19]
Solar PV	Water Electrolysis (green hydrogen)	China	8.96	[20]
Solar PV	Water Electrolysis (green hydrogen)	China	3.58–3.61	This study

V. Discussion and Implications

The results show that forecast accuracy affects green-hydrogen economics in a clear but non-linear way. On the forecasting side, the tree-based gradient boosting models (e.g., XGBoost/LightGBM/GBR) provide the most reliable overall behavior on long-term hourly solar data, combining low absolute errors with strong consistency across operating conditions. SVR achieves the best absolute error levels, but its higher MAPE indicates greater sensitivity when PV output is low, which matters because percentage errors can inflate during low-generation periods. In contrast, the LSTM exhibits substantially weaker predictive performance, implying that deep learning does not automatically outperform classical ML for this dataset without careful tuning or hybrid design especially when rapid irradiance transitions and short-term variability must be captured accurately.

From the techno-economic perspective, feeding these forecasts into the LCOE→LCOH pipeline shows that hydrogen cost outcomes remain tightly clustered, meaning that structural cost drivers dominate the final economics more than small forecast differences under the assumed parameters. The sensitivity analysis reinforces this by identifying electricity cost, specific energy consumption, and full load hours as the strongest levers on LCOH, while forecast residual uncertainty contributes comparatively little to long-term cost spread. Practically, this implies that robust, interpretable, and efficient models such as XGBoost/LightGBM are strong candidates for deployment because they support operational planning (electrolyzer scheduling, grid reliance, and curtailment management) and reduce uncertainty improving bankability and investment confidence even if the absolute LCOH reduction from forecast improvements alone is modest.

VI. Conclusion

This study presents a comprehensive evaluation of ML models for solar power forecasting and their impact on the LCOE and LCOH. The findings underscore the critical role of forecast accuracy in determining the economic feasibility of green hydrogen production. By integrating solar power predictions from XGBoost, SVR, and LSTM networks into financial models, it is demonstrated how variations in predictive performance influence hydrogen production costs. The results indicate that tree-based models, particularly XGBoost, provide a well-balanced trade-off between accuracy and computational efficiency, with an LCOE of \$40.57/MWh and an LCOH of \$3.60/kg H_2. The SVR model, despite achieving the highest numerical accuracy in solar forecasting, yielded a slightly higher LCOH of \$3.61/kg H_2 due to the impact of small forecasting errors on hydrogen cost estimations. Interestingly, the LSTM model, which exhibited the lowest forecasting accuracy, resulted in the lowest LCOH at \$3.58/kg H suggesting that while LSTM's tendency to smooth out fluctuations in solar power predictions can create a more stable electricity cost component, it may not necessarily translate to real-world operational benefits.

A comparative analysis of LCOH values from existing literature revealed that the estimated values align well with other studies employing ML-based forecasting techniques. The LCOH range of \$3.58–\$3.61/kg H_2 is competitive with findings from recent research, where LCOH values for green hydrogen have ranged from \$2.0 to \$11.0/kg H_2, depending on regional electricity costs and technological assumptions. These results reinforce the importance of accurate forecasting in minimizing economic uncertainties associated with hydrogen production and highlight the potential of ML-driven methodologies in optimizing solar-powered electrolysis.

Sensitivity analysis further highlighted the economic parameters that exert the most significant influence on LCOH. Among them, electricity cost emerged as the dominant factor, with a 20% fluctuation in electricity prices leading to an LCOH variation of approximately ±\$0.42/kg H_2. This finding underscores the importance of reducing uncertainties in renewable energy generation, as more accurate forecasts can improve electrolyzer scheduling, reduce reliance on backup energy sources, and ultimately lower hydrogen production costs. Additionally, full load hours, CAPEX, and discount rates are found to be significant determinants of LCOH, emphasizing the need for long-term financial planning and investment in cost-effective electrolyzer technologies.

The broader implications of this study extend beyond forecasting accuracy. Improving solar forecasting capabilities not only enhances hydrogen production efficiency but also facilitates better integration of hydrogen systems into the energy market. More precise predictions allow for optimized electrolyzer operation, reducing idle times and ensuring greater synchronization with renewable energy availability. Additionally, improved forecasting models can contribute to more favorable financing conditions for hydrogen projects by lowering investment risks and increasing investor confidence in long-term project viability. Future research should focus on hybrid forecasting models that combine the strengths of multiple ML techniques to enhance both short-term and long-term prediction accuracy. The integration of uncertainty quantification techniques, such as probabilistic forecasting, could provide further insights into the reliability of predictions and their financial implications. Moreover, extending this analysis to other renewable energy sources, such as wind and hybrid solar-wind systems, would offer a broader perspective on the role of ML-based forecasting in advancing green hydrogen production.

In conclusion, this study underscores the importance of accurate solar power forecasting in reducing LCOH and ensuring the economic feasibility of green hydrogen. By systematically evaluating different ML models, valuable insights into the trade-offs between forecasting precision are provided, computational efficiency, and cost-effectiveness. As renewable hydrogen continues to gain traction as a cornerstone of decarbonization efforts, leveraging advanced ML techniques will be instrumental in driving cost reductions and accelerating the global energy transition.

References

[1] Zhang, L., et al., A comprehensive review of the promising clean energy carrier: Hydrogen production, transportation, storage, and utilization (HPTSU) technologies. Fuel, 2024. 355: p. 129455.

[2] Ozturk, M. and I. Dincer, A comprehensive review on power-to-gas with hydrogen options for cleaner applications. International Journal of Hydrogen Energy, 2021. 46(62): p. 31511-31522.

[3] Arias, I., et al., Assessing system-level synergies between photovoltaic and proton exchange membrane electrolyzers for solar-powered hydrogen production. Applied Energy, 2024. 368: p. 123495.

[4] Park, J., et al., Techno-economic analysis of solar powered green hydrogen system based on multi-objective optimization of economics and productivity. Energy Conversion and Management, 2024. 299: p. 117823.

[5] Gupta, R., et al., Composition of feature selection techniques for improving the global horizontal irradiance estimation via machine learning models. Thermal Science and Engineering Progress, 2024. 48: p. 102394.

[6] Atiea, M.A., et al., Enhanced solar power prediction models with integrating meteorological data toward sustainable energy forecasting. International Journal of Energy Research, 2024. 2024(1): p. 8022398.

[7] El Bakali, S., et al. Data-based solar radiation forecasting with pre-processing using variational mode decomposition. in 2023 9th International Conference on Control, Decision and Information Technologies (CoDIT). 2023. IEEE.

[8] Hanif, M.F., et al., Advancing solar energy forecasting with modified ANN and light GBM learning algorithms. AIMS Energy, 2024. 12(2): p. 350-386.

[9] Wang, Y., et al., Volatility spillover and hedging strategies among Chinese carbon, energy, and electricity markets. Journal of International Financial Markets, Institutions and Money, 2024. 91: p. 101938.

[10] Roshani, A.S., et al., Optimization of a hybrid renewable energy system for off-grid residential communities using numerical simulation, response surface methodology, and life cycle assessment. Renewable Energy, 2024. 236: p. 121425.

[11] Cantillo-Luna, S., et al., Deep and machine learning models to forecast photovoltaic power generation. Energies, 2023. 16(10): p. 4097.

[12] Balci, Y. and C. Erbay, Harnessing solar energy for sustainable green hydrogen production in Türkiye: Opportunities, and economic viability. International Journal of Hydrogen Energy, 2024. 87: p. 985-996.

[13] Zhou, Z., et al., Cluster allocation strategy of multi-electrolyzers in wind-hydrogen system considering electrolyzer degradation under fluctuating operating conditions. Renewable Energy, 2025. 242: p. 122381.

[14] Endiz, M.S. and A.E. Coşgun, Assessing the potential of solar power generation in Turkey: A PESTLE analysis and comparative study of promising regions using PVsyst software. Solar Energy, 2023. 266: p. 112153.

[15] Chauhan, R., et al., Experimental and theoretical evaluation of thermophysical properties for moist air within solar still by using different algorithms of artificial neural network. Journal of Energy Storage, 2020. 30: p. 101408.

[16] Yang, Q., et al., Machine learning assisted prediction for hydrogen production of advanced photovoltaic technologies. DeCarbon, 2024. 4: p. 100050.

[17] Gökçek, M., et al., Optimum sizing of hybrid renewable power systems for on-site hydrogen refuelling stations: case studies from Türkiye and Spain. International Journal of Hydrogen Energy, 2024. 59: p. 715-729.

[18] Gül, M. and E. Akyüz, Techno-economic viability and future price projections of photovoltaic-powered green hydrogen production in strategic regions of Turkey. Journal of cleaner production, 2023. 430: p. 139627.

[19] Atabay, R. and Y. Devrim, Design and techno-economic analysis of solar energy based on-site hydrogen refueling station. International Journal of Hydrogen Energy, 2024. 80: p. 151-160.

[20] Nikolakakis, T., et al., Analysis of long-term variable renewable energy heavy capacity plans including electric vehicle and hydrogen scenarios: Methodology and illustrative case study for Turkey. IEEE Access, 2023. 11: p. 27189-27216.