Journal of Renewable Energy and Sustainable Development (RESD)
Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

Received: 10 March 2026 Accepted: 14 April 2026 Published: 28 April 2026

# Machine Learning-Based Performance Prediction for Continuous Power Generation of Solar Water Pumping System

**Said M. A. Ibrahim [1], Alhasan M. Azouz [2*], and Hamdy H. El-Ghetany [3]**

[1,2*] Al-Azhar University, Faculty of Engineering, Mechanical Engineering Department, Cairo, Egypt

[2*] Omar Al-Mukhtar University, Faculty of Engineering, Sustainable and Renewable Energy Engineering Department, El-Beida, Libya

[3] National Research Center, Solar Energy Department, Dokki, Giza, Egypt

prof.dr.said@hotmail.com, alhassan.mohammed@omu.edu.ly, hamdy.elghetany@gmail.com

## Abstract

*Integrating photovoltaic (PV) systems with pumped-water storage (PWS) enables continuous power generation and efficient water management in off-grid applications. Despite the growing use of machine learning (ML) in renewable energy, its application in hybrid PV–PWS configurations remains limited, particularly for real-time prediction. This study develops an ML-based framework to predict the electro-hydraulic behavior of a PV–PWS system. Input features include solar irradiance, ambient temperature, wind speed, PV cell temperature, and pump flow rate, targeting hydro turbine power and water head. A 1000-sample dataset is generated via validated MATLAB numerical simulations. Five algorithms—Random Forest, SVR, AdaBoost, CatBoost, and XGBoost—are trained using Python and evaluated via $R^2$, MAE, RMSE, and MAPE. The XGBoost model achieves the highest accuracy, with $R^2$ values of 0.9768 and 0.949 for power output and water head, respectively. SHAP analysis identifies pump flow rate and solar irradiance as the most influential features. The proposed framework demonstrates ML's effectiveness for real-time prediction and operational improvement of hybrid PV–PWS systems under varying weather conditions.*

***Index-words:*** *Renewable Energy, Solar Water Pumping, Hydro Turbine, Hybrid System, Machine Learning.*

## I. Introduction

The rapid growth of the global population, coupled with accelerated technological advancement, has led to a substantial rise in energy demand, exerting considerable pressure on the existing energy infrastructure, which remains largely dependent on fossil fuels such as coal, oil, and natural gas [1]. Consequences encompass elevated temperatures, rising sea levels, modified weather patterns, and pollution, in addition to detrimental health impacts and ecosystem damage [2]. Therefore, an urgent transition to sustainable energy sources is essential to meet the escalating electricity demand and address environmental challenges [3, 4]. The transition to renewable energy technologies supports the global effort to reduce carbon emissions and lessen environmental impacts, leading to a more sustainable and cost-effective energy system [5, 6]. A photovoltaic water pumping system converts solar energy into electricity to Power a water pump, providing an effective solution for regions with limited access to the electrical grid [7, 8]. Photovoltaic panels can be designed to match daily water demand patterns, allowing the system to operate fully autonomously without manual intervention. In such systems, energy storage devices like batteries may be replaced by water storage tanks, or the pump can directly transport water as needed [9]. In a hybrid solar water pumping system, an energy storage device plays a vital role in enhancing overall system efficiency and reliability. Energy storage ensures a dependable supply of backup Power in cases where both solar and auxiliary energy sources are unavailable. This capability is particularly important in critical applications such as providing water for agriculture, livestock, or human consumption [10].

However, a significant challenge in solar photovoltaic systems with Pumped-Water Storage lies in the efficient collection, monitoring,

* Corresponding author



Copyright © 2026, authors

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

and analysis of operational data from various components to maximize system performance and reliability. Machine Learning (ML) techniques have proven highly effective in addressing data-processing challenges in renewable energy systems, enabling intelligent performance prediction and optimization [11, 12]. However, while ML is widely deployed for solar irradiance forecasting and photovoltaic output prediction, its application specifically in hybrid PV–PWS configurations for real-time prediction of electro-hydraulic behavior remains underexplored.

Moreover, researchers have begun utilizing ML methodologies to predict diverse models and parameters, with the objective of improving the accuracy and reliability of predictive models [13, 14].

## A. Evolution of forecast models

Predictive modeling for renewable energy systems has evolved substantially over recent decades, progressing from simple statistical regression approaches to sophisticated nonlinear and ensemble machine learning (ML) architectures. Early predictive models relied primarily on **linear regression** techniques, which offer interpretability and computational simplicity but are fundamentally limited by their inability to capture nonlinear system dynamics [15, 16]. To overcome this limitation, nonlinear models such as Support Vector Regression (SVR), Decision Trees (DTs), and Gaussian Processes (GPs) were subsequently developed, offering higher predictive accuracy and greater robustness to outliers without imposing linearity assumptions [17]. SVR applies structural risk minimization to construct an optimal regression hyperplane, while GP models provide probabilistic predictions with uncertainty estimates through a Bayesian framework, making them particularly suited to small datasets [18].

More recently, ensemble ML models have become the dominant paradigm for high-accuracy prediction tasks. Bagging approaches, such as Random Forest (RF), reduce variance by training multiple decision trees on bootstrapped data subsets and averaging their outputs [19, 20]. Boosting approaches—including Gradient Boosting (GB), AdaBoost, CatBoost, and XGBoost—employ sequential learning, where each successive model corrects the residual errors of its predecessor using gradient descent optimization [21, 22]. Among these, XGBoost has demonstrated particularly superior performance across diverse engineering prediction tasks due to its built-in regularization, parallel computation capability, and ability to handle sparse or missing data [23]. Despite the rising prominence of deep learning (DL) methods such as LSTM and CNNs—which offer minimal manual feature engineering and strong performance on time-series data—their practical deployment is constrained by large computational resource requirements and dataset size demands [24]. For moderate-sized, structured datasets such as those generated by numerical simulations of small-scale energy systems, ensemble ML methods consistently achieve competitive or superior accuracy at substantially lower computational cost, motivating their selection in the present study.

## B. Literature review

Haddad et al. [25] developed artificial neural network (ANN) models to predict the hourly flow rate of a photovoltaic water-pumping system (PVWPS) installed in Madinah, Saudi Arabia. Input variables included hourly solar irradiation and air temperature. The models were experimentally validated and demonstrated high prediction accuracy, making them applicable for system control, water-demand management in remote/off-grid areas, and fault detection.

Sapitang et al. [26] devised and evaluated Machine Learning models to forecast fluctuations in reservoir water levels in Malaysia based on two input scenarios. Scenario 1 used rainfall and water level, while scenario 2 used rainfall, water level, and discharged water. The prediction horizons for both scenarios were between one and seven days for the four evaluated models: Boosted Decision Tree Regression (BDTR), Bayesian Linear Regression (BLR), Decision Forest Regression (DFR), and Neural Network Regression (NNR). The BLR model exhibited superior performance in Scenario 1, with a coefficient of determination ($R^2$ = 0.998952), while the BDTR excelled in Scenario 2 ($R^2$ = 0.99992). The findings validated the superior precision and efficacy of machine learning models, specifically BLR and BDTR, in predicting reservoir water levels.

Dehghani et al. [27] developed a hybrid model integrating Grey Wolf Optimization (GWO) with the Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict hydropower generation using inputs such as dam inflow, precipitation, and historical data. In most cases in the Dez basin, the GWO-ANFIS

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

model was more accurate in making predictions than the standalone ANFIS model. Wang et al. [13] conducted a comparative analysis of ML models for hydropower generation prediction, confirming the advantage of ensemble approaches over single-algorithm models.

Although AI and ML are widely used to optimize solar energy systems, their application in hybrid photovoltaic water pumping systems with pumped-water storage (PWS) remains underexplored, particularly in predicting power output and turbine head during continuous power delivery. The present study addresses this gap.

### C. Summary of related work

Table 1 summarizes the key related works reviewed in the literature, highlighting the system description, forecast methodology, primary contributions, and limitations, thereby identifying the research gap addressed in the present research.

Table 1: Summary of related work on ML-based prediction for PV and hydropower systems, identifying the research gap addressed by the present study

| Reference | Year | System Description | Methodology | Contribution | Limitation |
|---|---|---|---|---|---|
| [25] | 2015 | PVWPS, Madinah | ANN | Hourly flow-rate prediction; experimental validation | Single output; no ensemble comparison |
| [26] | 2020 | Reservoir level, Malaysia | BDTR, BLR, DFR, NNR | Multi-horizon (1–7 day) forecasting | No SHAP/interpretability; no hybrid PV–storage system |
| [27] | 2019 | Dez basin hydropower | GWO-ANFIS hybrid | Improved accuracy over standalone ANFIS | No PV integration; no ensemble boosting |
| [13] | 2025 | Hydropower, comparative ML | RF, GB, XGBoost, DL | Comprehensive ML comparison for hydropower | No PV–PWS integration; no SHAP analysis |
| Present research | 2026 | Hybrid PV–PWS, Egypt | RF, SVR, AdaBoost, CatBoost, XGBoost, and SHAP | Dual-output (Pt & Ht) prediction; SHAP interpretability; 5-fold CV | Simulation-based dataset; limited to one climate region |

### D. Study objectives and novelty

To address the detected research gaps in the field, this study proposes a comprehensive ML-based modeling framework for predicting the electro-hydraulic performance of a hybrid PV–PWS system. The novelty of this study can be summarized as follows:

- Development of an integrated hybrid modeling framework that combines MATLAB-based mathematical simulation with Python-based machine learning (ML) techniques for performance prediction of a hybrid PV–PWS system.
- Simultaneous prediction of hydro turbine power $(P_t)$ and turbine head $(H_t)$ under dynamic climatic conditions, unlike previous works, is limited to steady-state or single-variable analyses.
- Implementation and comparative evaluation of five ML algorithms (RF, SVR, AdaBoost, CatBoost, and XGBoost) to identify the most effective model for electro-hydraulic prediction.
- Introduction of an interpretable ML framework using SHAP analysis, which offers physical information about how solar irradiance, flow rate, and temperature variations affect system performance.
- Establishment of a data-driven alternative to computationally expensive numerical simulations, enabling faster and more reliable performance forecasting for real-time system management.

## II. Modeling and implementation of the system under consideration

### A. System layout and components

In the initial phase of this study, the configuration, components, and design parameters reported by Pali and Vadhera [28], shown in Fig. 1 and Table 2, were adopted to ensure model consistency and model validation. The system consists of a solar

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

photovoltaic (PV) array, a solar water pump (SWP), a hydro turbine coupled with a permanent magnet generator, an upper reservoir, and an open well that acts as the lower reservoir. The PV array has a total installed capacity of 6.0 kWp, with a total collector area of 39 m². The generated DC power is supplied directly to the centrifugal-type SWP, which operates with a total dynamic head of 17 m, a flow rate of 0.0094 m³/s, and an overall efficiency of approximately 60%. The SWP lifts water from the open well, having a depth of 14 m, and stores it in the upper reservoir (UR) of 180 m² internal area and 1 m height. The stored water in the UR flows downward through a penstock (50 mm diameter) and drives the turbine, which operates under a hydraulic head ranging from 13 to 14 m, producing a discharge rate of 0.00313 m³/s. The turbine is coupled with a 0.3 kW single-phase permanent magnet generator rated at 230 V, 50 Hz, and a power factor of unity. The overall efficiency of the turbine-generator assembly, including hydraulic and mechanical losses, is approximately 70%.

The system is designed to operate in a closed water loop between the open well and the upper reservoir, ensuring continuous power generation and stable voltage output during both solar and non-solar periods. The Power generated is supplied to a resistive load of 176 Ω, representing the electrical demand. This configuration provides the basis for the subsequent numerical modeling and machine learning (ML) prediction of the system's electro-hydraulic performance, enabling the evaluation of generated Power and turbine head under varying meteorological and operating conditions.

Table 2: System parameters and ratings for the hybrid PV–PWS configuration adopted from Pali and Vadhera [28]

| Component | Parameter | Value / Rating |
|---|---|---|
| PV Array | Installed capacity | 6.0 kWp |
| | Collector area | 39 m² |
| Solar Water Pump (SWP) | Total dynamic Head | 17 m |
| | Flow rate | 0.0094 m³/s |
| | Overall efficiency | ~60% |
| Open Well (Lower Reservoir) | Depth | 14 m |
| Upper Reservoir | Internal area | 180 m² |
| | Height | 1 m |
| Penstock | Diameter | 50 mm |
| Hydro Turbine | Operating head range | 13–14 m |
| | Discharge rate | 0.00313 m³/s |
| PMG | Rated Power | 0.3 kW |
| | Voltage / Frequency | 230 V / 50 Hz |
| | Power factor | Unity |
| Turbine-Generator Assembly | Overall efficiency | ~70% |
| Resistive Load | Resistance | 176 Ω |

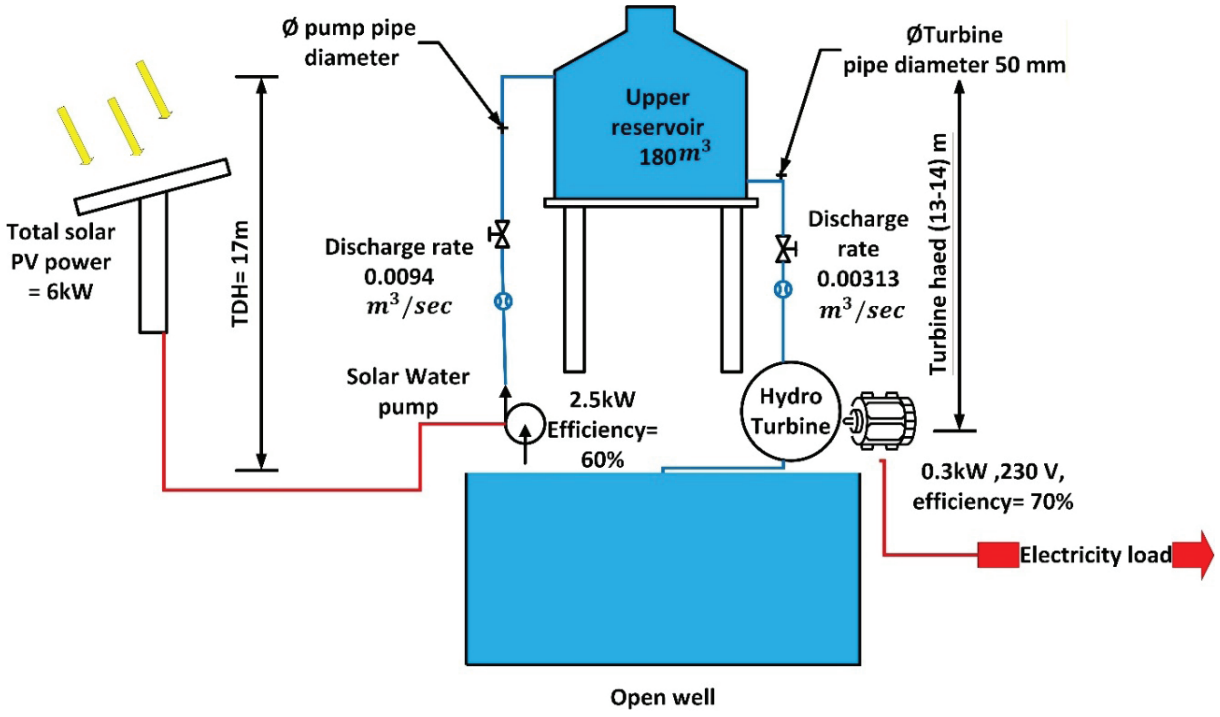



Figure 1: Schematic diagram of the hybrid system components and specifications.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

### B. Numerical model

The mathematical framework developed by Pali and Vadhera [28] formed the foundation for the present numerical analysis. The numerical model represents a PV–PWS system consisting of four main components as described in Section II.A: the PV array, the SWP, a hydro turbine coupled with a generator, and the two interconnected reservoirs. Accordingly, the complete numerical model of the PV–PWS system integrates the mathematical representations of these components and their interdependencies to accurately simulate the system's electro-hydraulic performance.

#### 1. Solar photovoltaic system

The output power of the solar photovoltaic (SPV) system primarily depends on the incident solar irradiance, as well as on factors such as ambient temperature, wind speed, and the electrical characteristics of the PV module. However, for the present hybrid system, the influence of parameters other than irradiance can be neglected for the direct estimation of PV power, since the hydropower output mainly depends on the water flow and Head [28]. Accordingly, the SPV output power at any available solar irradiance can be expressed in simplified form as:

$$P_{SPV(I)} = \eta_{SPV} A_{SPV} I(t) \quad (1)$$

Where $P_{SPV(I)}$ is the solar PV system power (W), $\eta_{SPV}$ is the solar PV efficiency, $A_{SPV}$ is the PV system area in ($m^2$), and $I(t)$ is the solar irradiance in ($W/m^2$).

In the present study, the cell temperature $T_c$ was computed using the empirical NOCT-based [29] correlation given by

$$T_c(t) = T_a(t) + \left(\frac{NOCT - 20}{0.8}\right)\frac{I}{I_{STC}} \quad (2)$$

Where $T_a(t)$ represents the ambient temperature (°C), $I_{STC}$ =1000 $W/m^2$ denotes the standard irradiance under Standard Test Conditions (STC), and $NOCT$ is the Cell Temperature at Nominal Operating (°C). The calculated cell temperature was incorporated as an auxiliary input parameter to find out the indirect influence of thermal variations on the overall energy conversion chain. Consequently, the machine learning model was trained to predict the hydro turbine power and Head, thereby bridging the photovoltaic–hydraulic interaction through data-driven modeling.

#### 2. Solar water pump

In the present PV–PWS system, the total electrical Power generated by the SPV array is directly supplied to the SWP, which lifts water from the open well and stores it in the water reservoir. Thus, the power output from the SPV system, , $P_{WP}$ is equal to the power input to the SWP. Hence,

$$P_{WP} = P_{SPV} \quad (3)$$

The discharge rate of the water into the water reservoir, $Q_{WP}$ ($m^3$/sec), depends on the hydraulic Head and efficiency of the water pump. Hence, it can be expressed as

$$Q_{WP} = \frac{\eta_{WP} P_{SPV}}{\rho_w g TDH} \quad (4)$$

Where $\eta_{WP}$ is the water pump efficiency, $\rho_w$ is the water density in kg/$m^3$, $g$ is the gravity acceleration = 9.81 $m^2$/sec, and $TDH$ is the total dynamic Head in m.

#### 3. Hydro turbine and generator

The hydraulic subsystem of the PV–PWS configuration consists of a hydro turbine coupled with a generator. The total output power of the system is equivalent to the electrical Power produced by the turbine–generator set and can be expressed as:

$$P_t = \eta_{Tg} \rho_w g H_t Q_t \quad (5)$$

Where $P_t$ is the system output power in W, $\eta_{Tg}$ is the overall efficiency of the turbine-generator set, $Q_t$ is the water discharge rate of the turbine in $m^3$/sec, and $H_t$ is the hydro turbine water head in m.

The available water head $H_t$ at any given time depends on the balance between the inflow rate from the solar water pump and the outflow rate through the hydro turbine. It can be determined using the following relation:

$$H_t = H_i + (Q_{WP} - Q_t)\left(\frac{time}{A_{Res}}\right) \quad (6)$$

Where $H_i$ is the initial water head in m at the beginning of the time interval, and $A_{Res}$ is the water

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

reservoir area in m$^2$. When $Q_{WP} > Q_t$, the water level in the reservoir rises, increasing the effective Head; conversely, when $Q_{WP} < Q_t$, the Head gradually decreases. Under normal operating conditions, the reservoir is designed with sufficient capacity to maintain a nearly constant head throughout day–night operation, ensuring stable hydro turbine performance.

## 4. Model solution procedure

Fig. 2 illustrates the proposed PV–PWS working operation and control algorithm. The algorithm was developed and implemented in MATLAB. The process starts with initializing the system parameters, including the PV and pump efficiencies, surface area, water density, total dynamic Head (TDH), solar irradiance, ambient temperature, and the simulation time step (Δt). The simulation begins at $t = 0$ and proceeds incrementally until $t = 1000$ s.

At each time step, the photovoltaic Power, water pump power, water flow rate, and instantaneous Head are calculated. The control logic continuously checks the water level in the upper reservoir:

- If $H_t \geq Hmax$, the Head is fixed at $Hmax$.
- If $H_t$ is below the upper limit but within the operational range, its value is updated based on the computed results.
- If the Head drops below the allowable minimum, $H_t$ is reset to zero, indicating that the reservoir is empty. Once the water head is updated, the system calculates the hydro turbine and cell temperature. The loop continues until the final simulation time is reached. Finally, all computed results are exported to an Excel sheet for further analysis and performance evaluation.

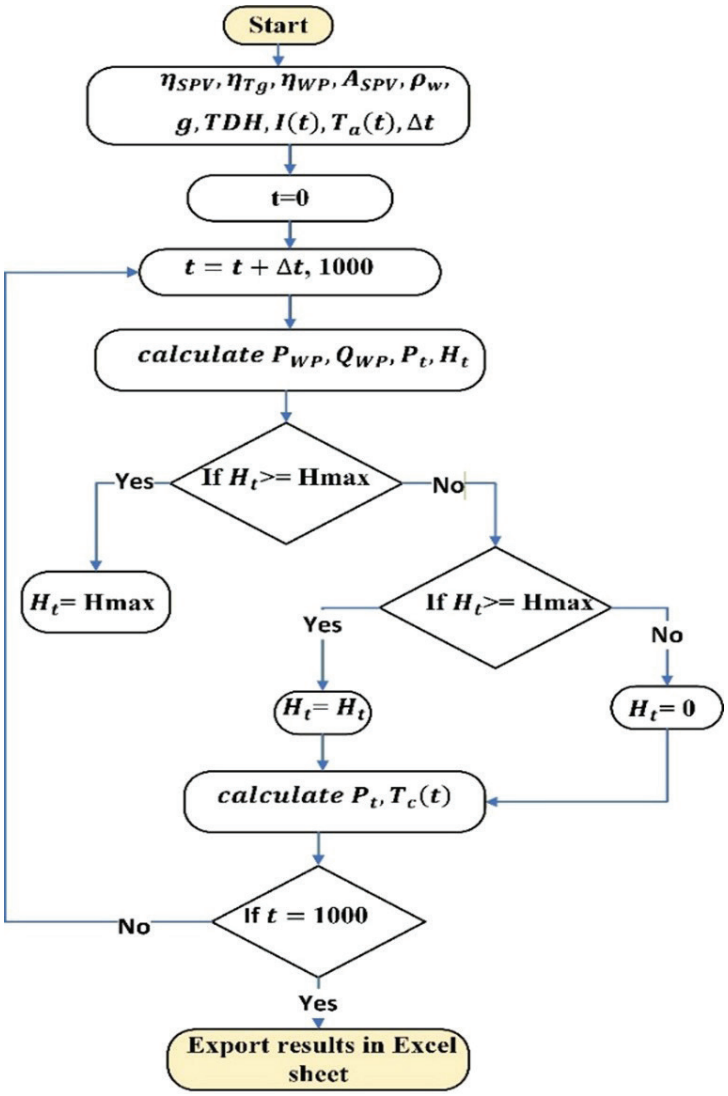



Figure 2: Flowchart of the iterative control and computation steps for the PV–PWS system.

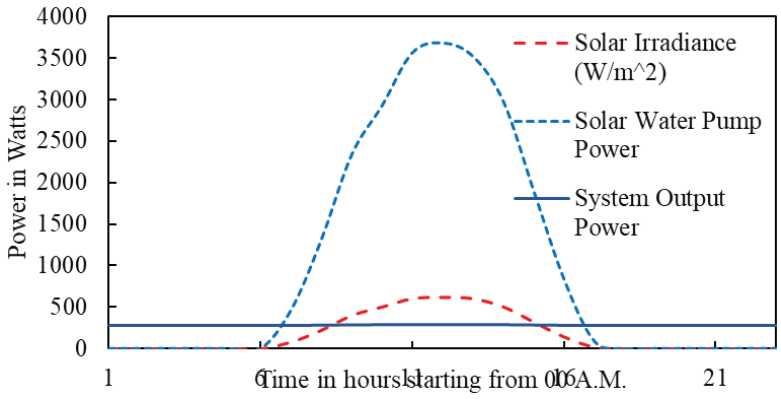
Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

## 5. Model validation

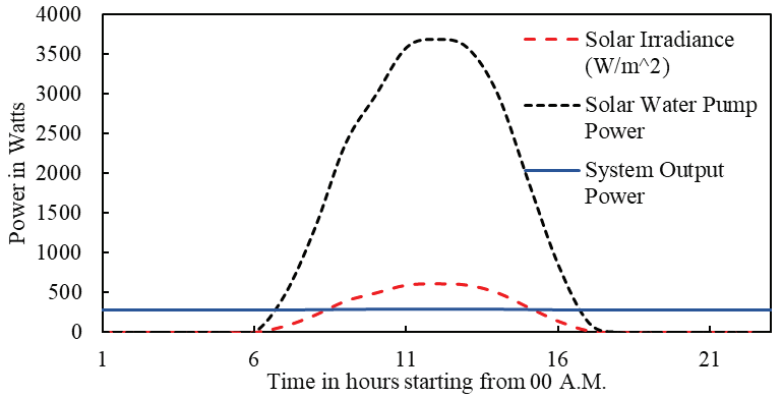
Initially, the upper reservoir is filled with water, and the system operation starts at 00:00 during the night. Fig. 3 shows the available solar irradiance, the Power supplied by the SWP, and the system output power over 24 hours. It can be observed that despite the significant variations in solar irradiance throughout the day and its complete absence at night, the system output power remains almost constant and close to the designed value of 0.3 kW. In contrast, the SWP power follows the solar irradiance profile, increasing during peak sunlight hours and dropping to zero during nighttime. However, to verify the reliability of the implemented model, the obtained simulation results were compared with the results of Pali and Vadhera [28], showing excellent agreement and confirming the accuracy of the adopted modeling approach. Fig. 3 depicts a visual comparison highlighting the validation process and matching performance outcomes.

After successful validation, the study was expanded to the climatic conditions of Egypt, where the hybrid PV–PWS was simulated using representative local meteorological data. Egypt has favorable solar potential, and variable environmental conditions make it an appropriate case for evaluating the real-world electro-hydraulic performance and adaptability of the system under different operating scenarios.

Moreover, the meteorological data for Egypt were obtained from the NASA Langley Research Center POWER Data Access Viewer [30], which provides high-resolution global datasets of solar irradiance, ambient temperature, and wind speed. These data were used as model inputs to represent realistic daily and seasonal variations in the region's climatic conditions.




(a) Pali and Vadhera reported data[28]




(b) Present simulation results

Figure 3: Comparison between (a) reported data, and (b) simulated results for model validation.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

# III. Methodology

This section describes the complete ML pipeline applied in this study, from data generation and preprocessing through model training, optimization, and evaluation. The overall implementation workflow is illustrated in Fig. 4.

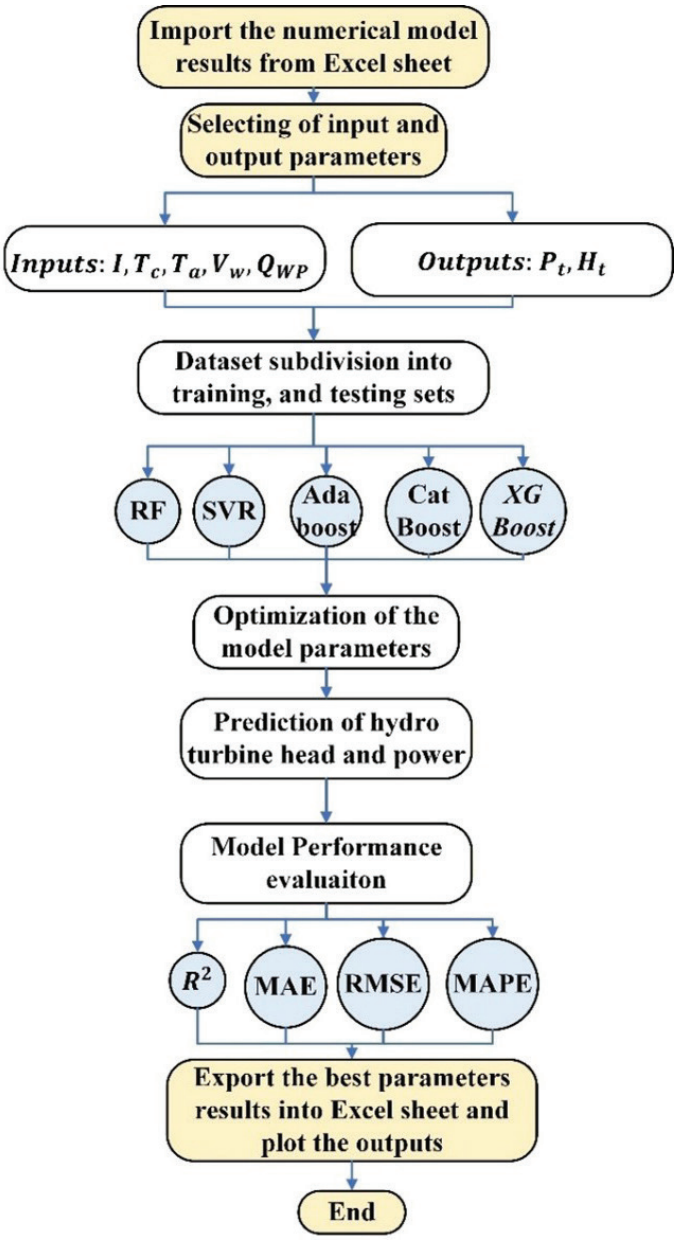



Figure 4: Integrated methodological framework for performance prediction of a hybrid PV–pumped hydro system based on mathematical modeling and machine learning.

## A. Dataset generation and feature selection

Five parameters, namely solar radiation $(I)$, ambient temperature $(T_a)$, wind speed $(V_w)$, cell temperature $(T_c)$, and pump flow rate $(Q_{WP})$, were selected as independent input variables based on their physical significance in governing system behavior, as confirmed by the correlation analysis in Section IV.A. The target outputs were hydro turbine output power $(P_t)$ and hydro turbine water head $(H_t)$. A total of 1000 samples were generated via the validated MATLAB numerical model (Section II.B) under representative Egyptian climatic conditions, providing a physically consistent dataset for model training and evaluation.

## B. Data preprocessing

Rigorous data preprocessing was performed to ensure dataset quality and model readiness, encompassing the following steps:

(i) **Data cleaning:** The dataset was inspected for missing, null, and duplicate entries. No missing values were identified, as the dataset was generated from deterministic numerical simulations. Consistency checks were applied to verify the physical plausibility of all records.

(ii) **Feature scaling:** All input and output variables were standardized using z-score normalization (zero mean, unit standard deviation) to prevent features with larger magnitudes from dominating model training. The normalization is given by[31]:

$$Z = \frac{x - \mu}{\sigma} \tag{7}$$

where $x$ is the original value, $\mu$ and $\sigma$ are the mean and standard deviation of the attribute, respectively. This step is particularly important for SVR, which is sensitive to feature scale.

(iii) **Train/test split:** The dataset was randomly divided into training (80%, 800 samples) and testing (20%, 200 samples) subsets, ensuring that the test set contains only unseen data to provide an unbiased evaluation of model generalization.

## C. Data distribution and class balance

The dataset comprises 1000 samples generated from the validated MATLAB model. The 80:20 train–test split yields **800 training samples** (80%) and **200 test samples** (20%). Since this is a regression task, class-balance concerns do not apply in the classical sense; however, the distribution of target variables

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

was examined to ensure adequate coverage of the operating range. The hydro turbine power Pt ranges from approximately 0 to 320 W, and turbine head $H_t$ ranges from 0 to 14 m, with the distributions broadly reflecting the diurnal and seasonal variation in Egyptian climatic conditions. The training set covers the full output range, and no overrepresentation of specific operating regimes was identified. This is further supported by the cross-validation results reported in Section III.E.

### *D. Machine learning models*

Five ML regression algorithms were implemented to predict both the hydro turbine output power $(P_t)$ and the water head $(H_t)$. Covering linear, nonlinear, and ensemble categories:

#### 1. Random Forest Model

The Random Forest (RF) algorithm is an ensemble learning method that constructs multiple decision trees and averages their predictions to improve accuracy and robustness. Each tree is trained on randomly sampled data and feature subsets, enhancing model diversity and reducing overfitting through bootstrapping and random subspace selection. The overall prediction is given by[19]:

$$\hat{y}(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{8}$$

Where $N$ is the number of trees and $f_i(x)$ denotes the output of the $i^{th}$ tree. Due to its ability to capture nonlinear relationships and handle complex data interactions, RF was employed in this study to predict the hydro turbine output power $(P_t)$ and water head $(H_t)$ under varying climatic and operational conditions. The RF algorithm has been widely applied in energy system modeling due to its robustness and strong predictive capability.

#### 2. Support Vector Regression (SVR)

Support Vector Regression is a supervised learning algorithm that maps input data into a higher-dimensional feature space to construct an optimal hyperplane minimizing prediction error. It applies structural risk minimization and margin maximization principles for better generalization, relying only on support vectors to define the model [32].

#### 3. Adaboost

Adaptive Boosting (AdaBoost) is an ensemble learning algorithm that combines multiple weak learners to form a strong predictive model. It works by sequentially training weak models—typically decision trees—while assigning higher weights to data samples that were mispredicted in previous iterations. This adaptive weighting focuses the model on harder-to-predict cases, thereby improving the overall accuracy. The final prediction of the AdaBoost regression model can be expressed as [33]:

$$\hat{y}(x) = \sum_{t=1}^{T} \alpha_t f_t(x) \tag{9}$$

Where $T$ is the number of weak learners, $\alpha_t$ is the weight assigned to the $t^{th}$ learner, and $f_t(x)$ denotes the prediction of that learner. AdaBoost enhances model robustness and reduces bias without significantly increasing variance, making it effective for regression and classification problems.

#### 4. CatBoost

CatBoost is a gradient boosting algorithm developed by Yandex that efficiently handles both classification and regression problems. It introduces significant innovations in processing categorical features through ordered boosting and target statistics, reducing prediction bias and overfitting. Unlike traditional gradient boosting methods, CatBoost employs symmetric tree structures and random permutations of training data to calculate the average label value of preceding instances with the same category, ensuring balanced and unbiased learning. These design choices enhance model robustness, computational efficiency, and generalization capability. In this study, CatBoost is utilized for predicting the hydro turbine output power and water head, leveraging its strong performance on nonlinear, multivariate datasets[34].

#### 5. XGBoost

XGBoost is an optimized gradient boosting framework that constructs decision trees sequentially, where each new tree aims to minimize the residual errors of the previous ensemble using gradient descent optimization. It is recognized for its high scalability, computational efficiency,

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

and strong predictive performance, particularly when handling sparse or incomplete datasets. The inclusion of regularization terms helps prevent overfitting and improves the model's generalization ability. Furthermore, XGBoost supports parallel computation, enabling faster training and efficient processing of large datasets. By iteratively combining multiple weak learners into an additive model, XGBoost achieves high predictive accuracy for complex nonlinear relationships. The model's prediction at iteration $t$ can be expressed as[23]:

$$\hat{y}^{(t)} = \sum_{k=1}^{t} f_k(x) \quad (10)$$

Here, $f_k(x)$ represents the $k^{th}$ regression tree, and $\Psi$ denotes the space of all possible trees.

### *E. Hyperparameter optimization and cross-validation*

To ensure optimal predictive performance, hyperparameter tuning was performed for all used machine learning models using the Optuna optimization framework. Optuna's Bayesian optimization approach was employed to efficiently explore the hyperparameter space and minimize the selected evaluation metric, the mean absolute error (MAE). Each model was trained and validated using five-fold cross-validation to prevent overfitting and ensure robust generalization capability. Overall, the hyperparameter optimization through Optuna significantly improved each model's predictive accuracy while maintaining computational efficiency.

The hyperparameter search ranges and the corresponding optimal values for the five machine learning models are summarized in Table 3. These values represent the best-performing configurations identified during the Optuna optimization process.

Table 3: Hyperparameter search ranges and optimal values of the five ML models

| Model | Hyperparameters | Tested Range | Selected Value |
|---|---|---|---|
| RF | Number of estimators | 100 – 1000 | 300 |
| | Maximum depth | 3 – 20 | 14 |
| | Minimum samples. split | 2 – 10 | 4 |
| | Minimum samples. leaf | 1 – 5 | 1 |
| | Maximum features | sqrt, log2, None | sqrt |
| | Bootstrap | True, False | True |
| SVR | Regularization parameter | 0.01 – 1000 | 16.14 |
| | Epsilon | 0.01 – 1.0 | 0.53 |
| | Kernel type | Linear, RBF, poly, sigmoid | rbf |
| | Gamma | scale, auto | Scale |
| Adaboost | Number of estimators | 100 – 1000 | 500 |
| | Learning rate | 0.001 – 1.0 | 0.9971 |
| | Maximum depth | 3 – 10 | 7 |
| CatBoost | Tree depth | 3 – 10 | 5 |
| | Learning rate | 0.01 – 0.3 | 0.1110 |
| | L2 regularization | 1 – 10 | 9 |
| | Number of iterations | 300 – 1000 | 800 |
| XGB | Number of estimators | 100 – 1000 | 520 |
| | Maximum depth | 3 – 10 | 5 |
| | Learning rate | 0.001 – 0.3 | 0.0163 |
| | Subsample ratio | 0.6 – 1.0 | 0.9999 |
| | Colsample by tree | 0.6 – 1.0 | 0.9432 |

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

### 1. Evaluation metrics

The predictive performance of the developed models was quantitatively evaluated using four statistical indicators; these metrics are defined as follows [35, 36]:

- **The coefficient of determination ($R^2$):** Measures the proportion of variance in the target variable explained by the model. A value of 1 indicates a perfect fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - \hat{O}_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (11)$$

- **The mean absolute error (MAE):** The average magnitude of prediction errors, in the same units as the target variable. It is robust to outliers compared to RMSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - \hat{O}_i| \quad (12)$$

- **The root mean square error (RMSE):** The square root of the mean squared deviation between predicted and actual values. RMSE penalizes larger errors more heavily and is expressed in the same units as the target variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - \hat{O}_i)^2} \quad (13)$$

- **The mean absolute percentage error (MAPE):** Expresses prediction error as a percentage of the actual value, facilitating comparison across variables with different magnitudes.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{O_i - \hat{O}_i}{O_i}\right| \quad (14)$$

In all equations, $n$ is the number of samples, $O_i$ is the actual value, $\hat{O}_i$ is the predicted value, and $\bar{O}$ is the mean of actual values.

## IV. Simulation work

This section presents the complete simulation and ML prediction results. It begins with a correlation analysis of the input–output relationships in the dataset (Section IV.A), followed by individual model regression results for each of the five ML algorithms (Sections IV.B–IV.F), and concludes with SHAP-based feature importance analysis for the best-performing model (Section IV.G). The optimal statistical performance metrics across all models are compared in Fig 17.

### A. Correlation Analysis

Fig. 5 represents the correlation matrix, which showcases the inherent correlations between the input and output variables. This matrix reveals the mutual relationships among these factors, helping to assess the significance of each variable. As observed in the figure, the color gradient represents the correlation coefficients, with values ranging from blue (negative correlation) to red (positive correlation). Higher absolute values indicate a stronger correlation, while positive and negative signs represent positive and negative correlations, respectively. Values close to zero indicate a negligible influence or a negligible linear relationship. The most significant correlation in the entire matrix is observed between hydro turbine power and hydro turbine water head, with a near-perfect correlation coefficient of 0.99. A strong positive correlation also appears between cell temperature and both solar radiation (0.93) and ambient temperature (0.86). Furthermore, ambient temperature shows a strong positive correlation with hydro turbine power (0.66). Moderate positive correlations are found between solar radiation and pump flow rate (0.56) and between cell temperature and hydro turbine power (0.51). The matrix shows weak negative correlations, specifically between pump flow rate and both hydro turbine water head (-0.17) and hydro turbine power (-0.17). These trends are physically consistent with expected phenomena. For example, the near-perfect relationship (0.99) between the water head and Power is intuitive, as hydraulic Power is a direct function of Head and flow rate. Similarly, the strong correlation (0.93) between solar radiation and cell temperature is entirely logical. These relationships validate the dataset and support its use for further data-driven analysis.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

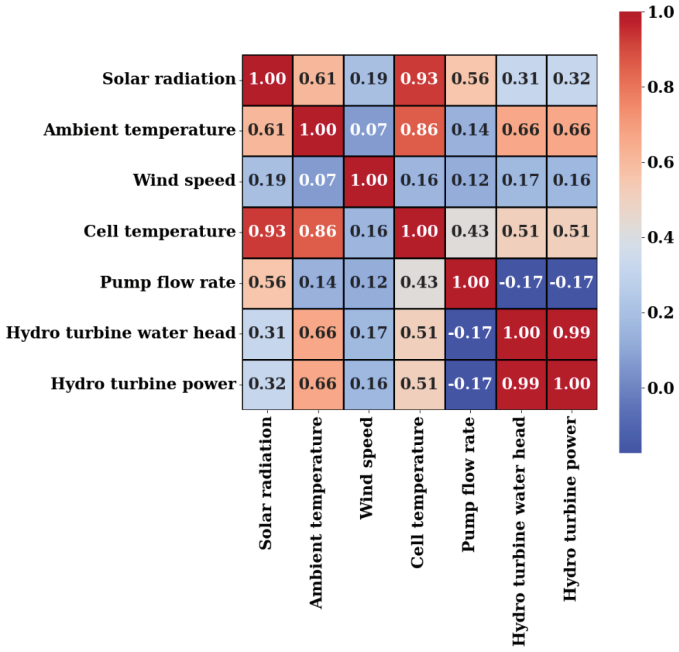



Figure 5: Pearson correlation matrix of input features and output variables.

## B. *Random Forest Model*

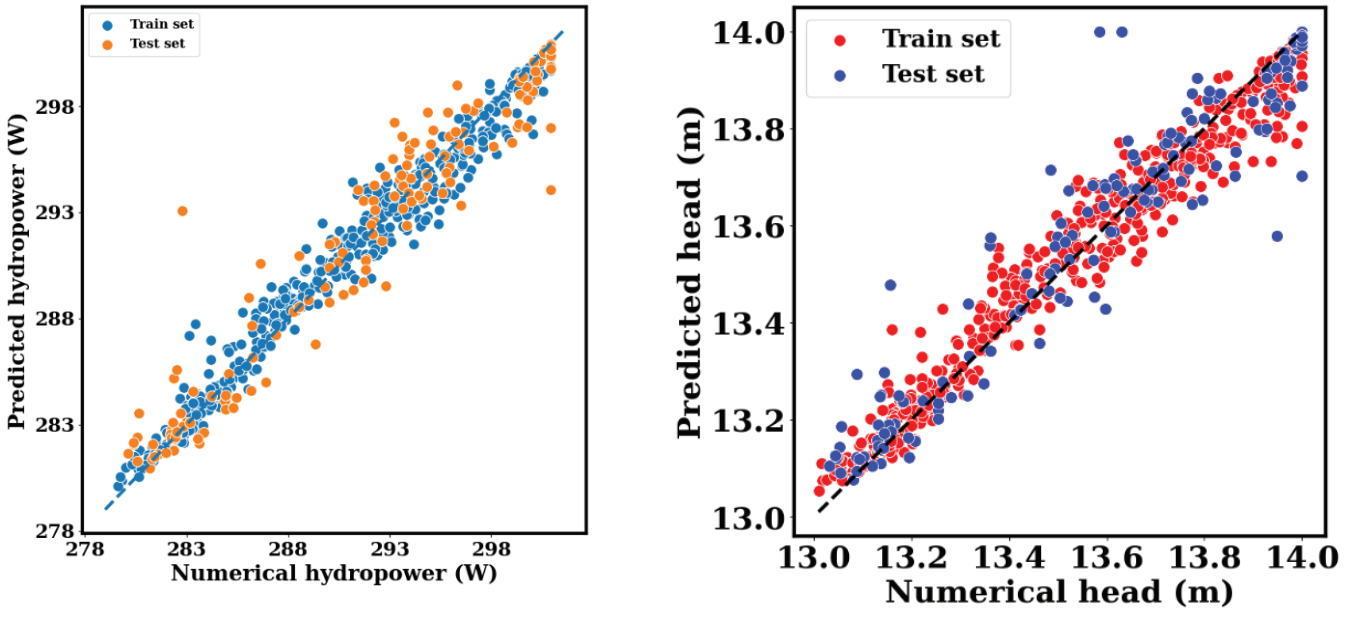
The plots of regression for $(P_t)$ and $(H_t)$ are provided in Fig. 6 for the train, test, and total datasets.

The $R^2$ is 0.9524 for $(P_t)$ and 0.9323 for $(H_t)$. Meanwhile, the associated errors are 0.9078, 1.5210, and 0.31% for $(P_t)$, as well as 0.0509, 0.0849, and 0.38% for $(H_t)$, which is illustrated in Fig. 7.

Furthermore, as shown in Table 4, the RF model has strong predictive accuracy, and it consistently performs well with low MAE and MSE. The model's outstanding $R^2$ values indicate strong explanatory Power.

Table 4: Random Forest regressor model evaluation

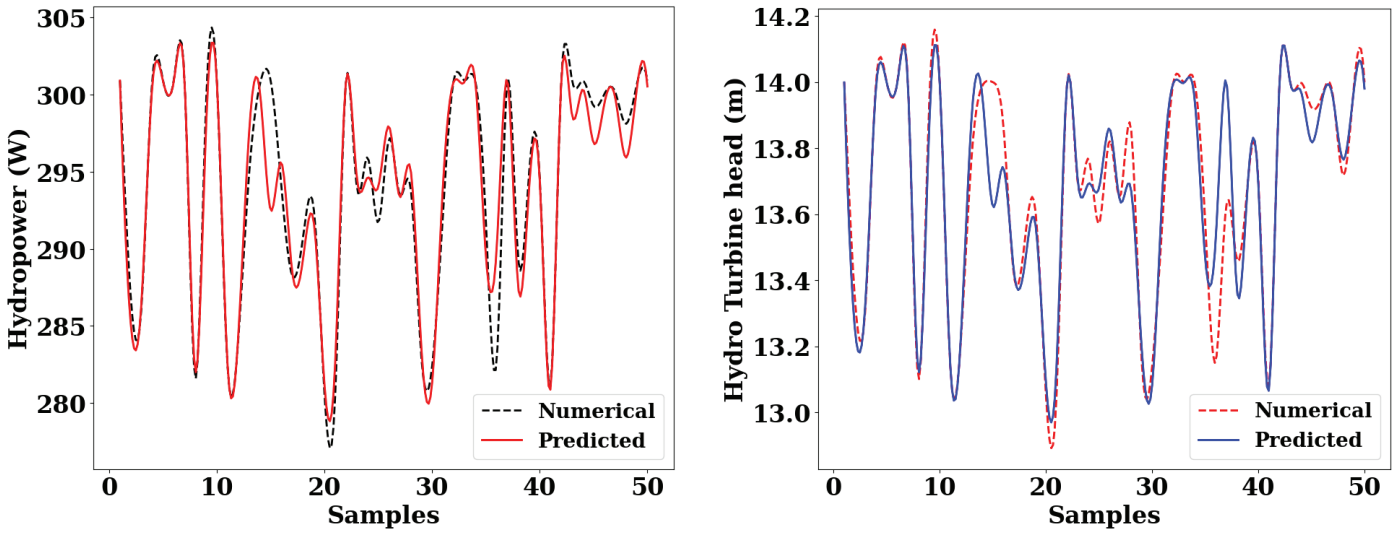
| Random forest | $R^2$ | | MAE | | RMSE | | MAPE | |
|---|---|---|---|---|---|---|---|---|
| | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ |
| Train | 0.9889 | 0.9792 | 0.4492 | 0.0305 | 0.7660 | 0.0484 | 0.15% | 0.22% |
| Test | 0.9524 | 0.9323 | 0.9078 | 0.0509 | 1.5210 | 0.0849 | 0.31% | 0.38% |
| Total | 0.9821 | 0.9703 | 0.5409 | 0.0346 | 0.9654 | 0.0576 | 0.19% | 0.25% |




(a) (b)

Figure 6: Random Forest regression scattered plot for (a) hydro turbine output power $(P_t)$, and (b) hydro turbine water head $(H_t)$.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993




(a) (b)

Figure 7: Random Forest prediction versus numerical dataset for (a) hydro turbine output power $(P_t)$, and (b) hydro turbine water head $(H_t)$.

## C. *Support Vector Regression (SVR)*

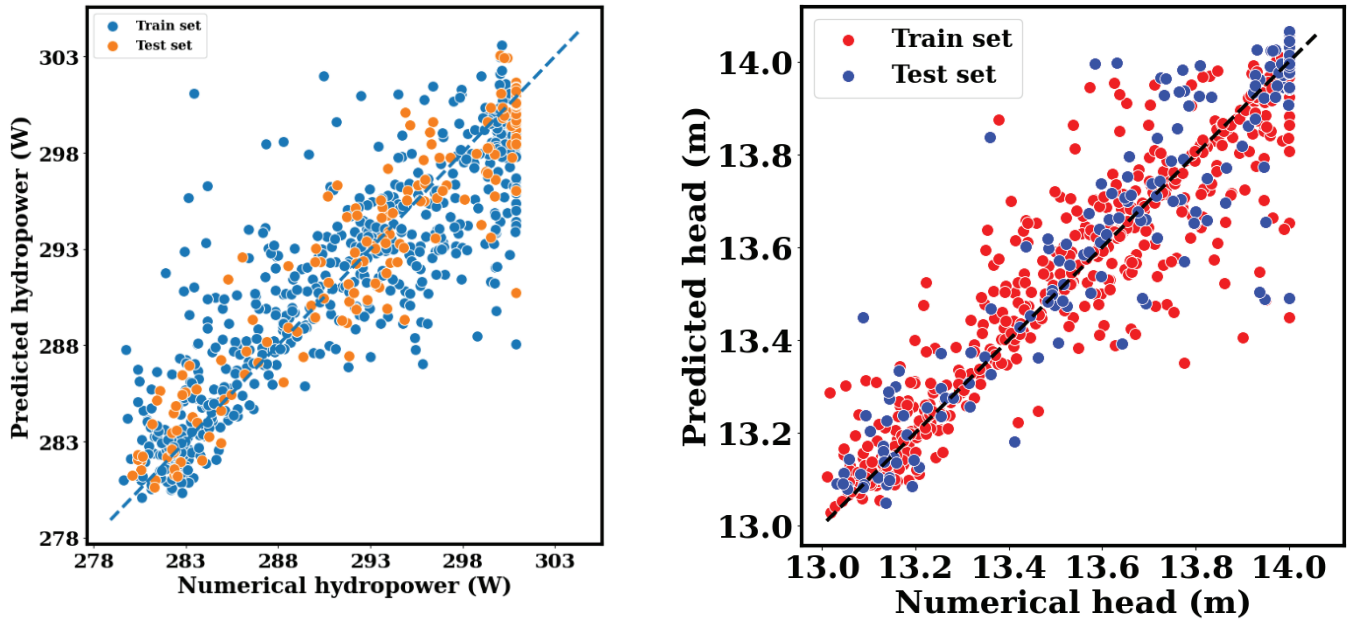
For $P_t$, $H_t$, the regression plots are provided in Fig. 8. The estimated $R^2$ for the two values was 0.9106 and 0.876, respectively. Besides, a comparison between the predicted and numerical data is carried out in Fig. 9. The $P_t$ prediction is shown in Fig. 9 (a), which exhibited MAE, RMSE, and MAPE of 1.4416, 2.0854, and 0.49%, respectively. Similarly, the $(H_t)$, numerical and predicted data, is displayed in Fig. 9 (b) with MAE, RMSE, and MAPE of 0.0691, 0.115, and 0.51%, respectively.

The SVR model achieves competitive MAE and RMSE values depicted in Table 5; however, its performance generally falls behind that of the RF model—particularly for $P_t$, where both $R^2$ and error metrics are less favorable. Nevertheless, the SVR demonstrates moments of strength, such as achieving relatively higher $R^2$ values for $H_t$ in the training phase, suggesting that it can perform well when applied to smoother or less complex datasets. Overall, the RF model consistently provides higher accuracy and better generalization, whereas the SVR tends to be more data-dependent and less resilient to variability in training conditions.

Table 5: The SVR regressor model evaluation

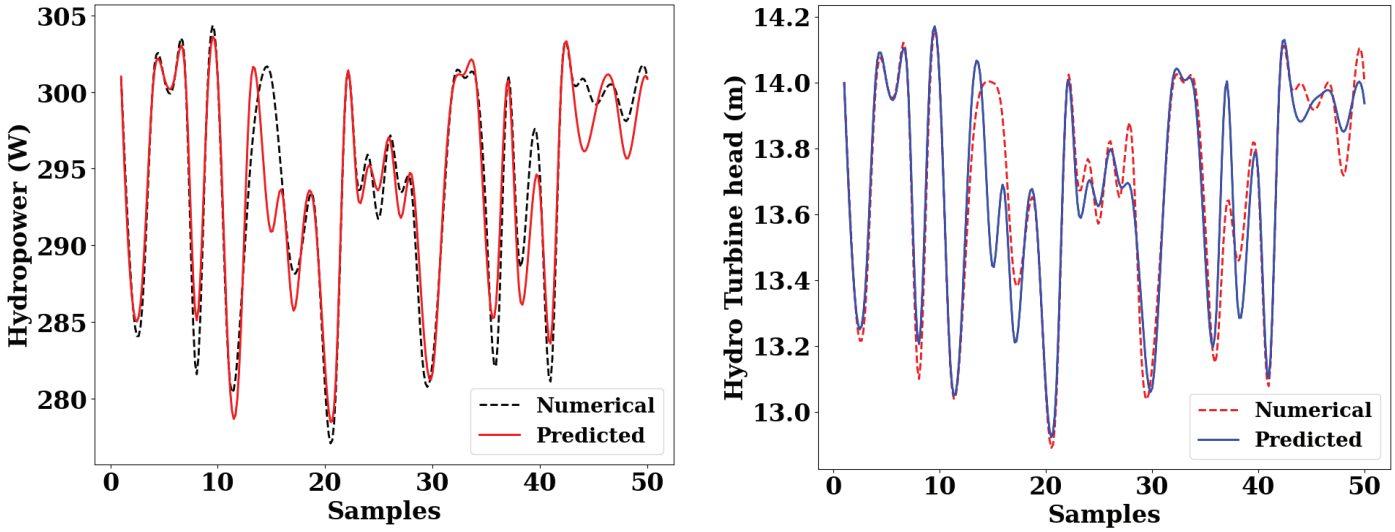
| SVR | $R^2$ | | MAE | | RMSE | | MAPE | |
|---|---|---|---|---|---|---|---|---|
| | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ |
| Train | 0.8506 | 0.929 | 1.7716 | 0.0481 | 2.8115 | 0.0894 | 0.61% | 0.35% |
| Test | 0.9106 | 0.876 | 1.4416 | 0.0691 | 2.0854 | 0.1150 | 0.49% | 0.51% |
| Total | 0.8618 | 0.9189 | 1.7056 | 0.0523 | 2.682 | 0.0951 | 0.58% | 0.38% |

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993




(a) (b)

Figure 8: SVR regression scattered plot for (a) hydro turbine output power ($P_t$), and (b) hydro turbine water head ($H_t$).




(a) (b)

Figure 9: SVR prediction versus numerical dataset for (a) hydro turbine output power ($P_t$), and (b) hydro turbine water head ($H_t$).

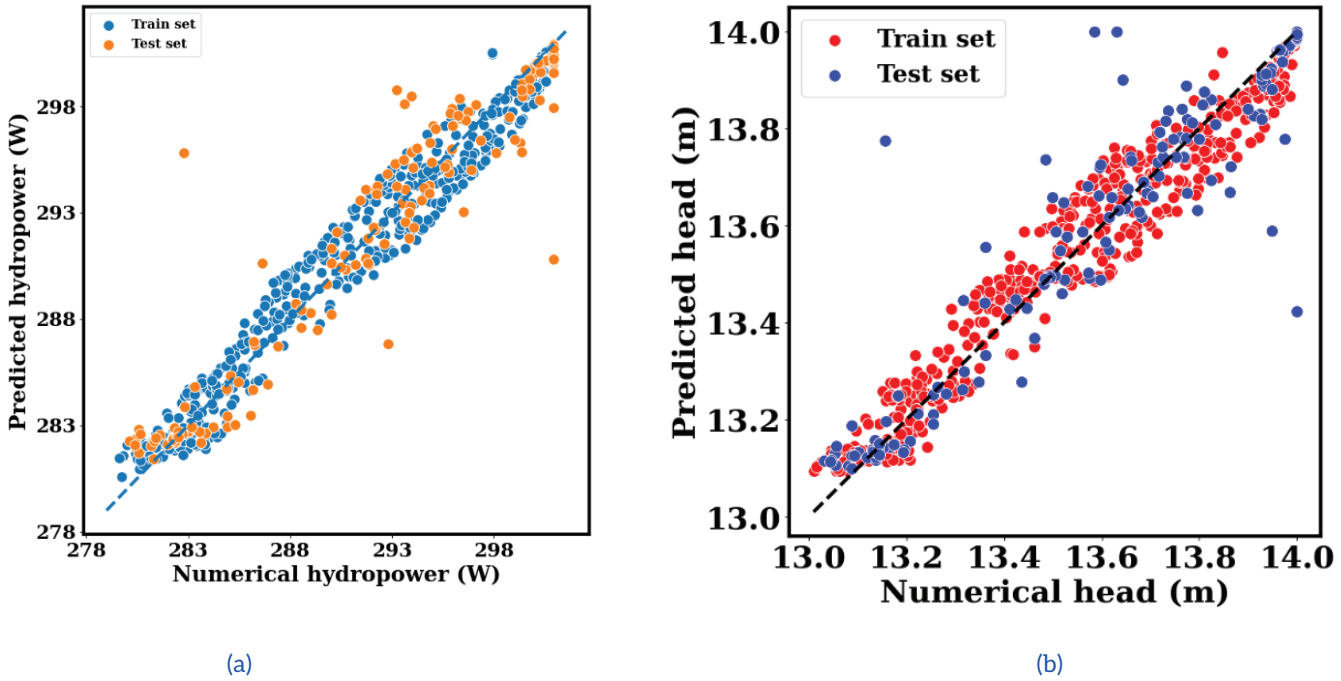
## D. *Adaboost*

For $P_t$, $H_t$, the regression plots are given in Fig. 10. The estimated $R^2$ for the two values was 0.93 and 0.912, respectively. A comparison between the numerical and predicted data is carried out in Fig. 11. The prediction of $P_t$ is shown in Fig. 11 (a), which exhibited MAE, RMSE, and MAPE of 1.0361, 1.791, and 0.35%, respectively. Similarly, the ($H_t$), numerical and predicted data, is displayed in Fig. 11 (b) with MAE, RMSE, and MAPE of 0.0479, 0.0969, and 0.35%, respectively. As shown in Table 6, the Adaboost regressor performs competitively, with both MAE and MSE exhibiting a downward trend. It outperforms SVR in terms of resilience and predictive capability, while the high $R^2$ values demonstrate the strong explanatory Power of the AdaBoost model.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

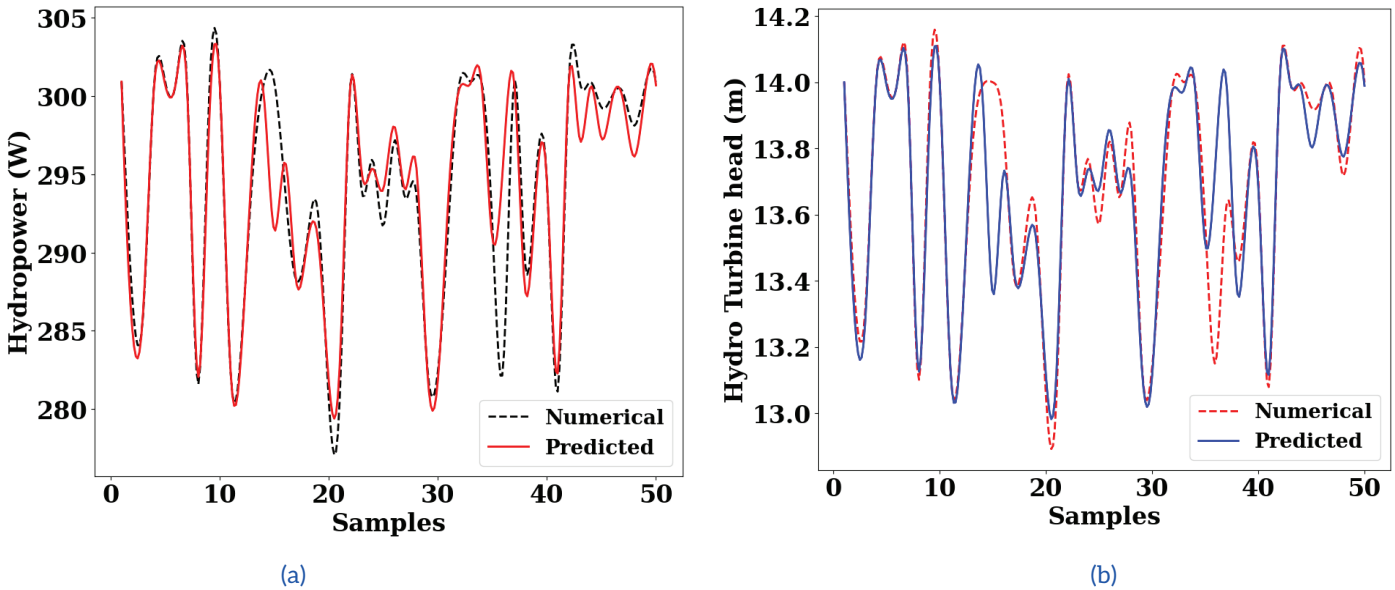
Table 6: The Adaboost regressor model evaluation

| Adaboost | $R^2$ | | MAE | | RMSE | | MAPE | |
|---|---|---|---|---|---|---|---|---|
| | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ |
| Train | 0.9813 | 0.9765 | 0.7418 | 0.0347 | 0.9941 | 0.0514 | 0.25% | 0.26% |
| Test | 0.9341 | 0.9120 | 1.0361 | 0.0479 | 1.7910 | 0.0969 | 0.35% | 0.35% |
| Total | 0.9725 | 0.9642 | 0.8006 | 0.0373 | 1.1967 | 0.0632 | 0.27% | 0.28% |




(a) (b)

Figure 10: Adaboost regression scattered plot for (a) hydro turbine output power ($P_t$), and (b) hydro turbine water head ($H_t$).




(a) (b)

Figure 11: Adaboost prediction versus numerical dataset for (a) hydro turbine output power ($P_t$), and (b) hydro turbine water head ($H_t$).

## E. CatBoost

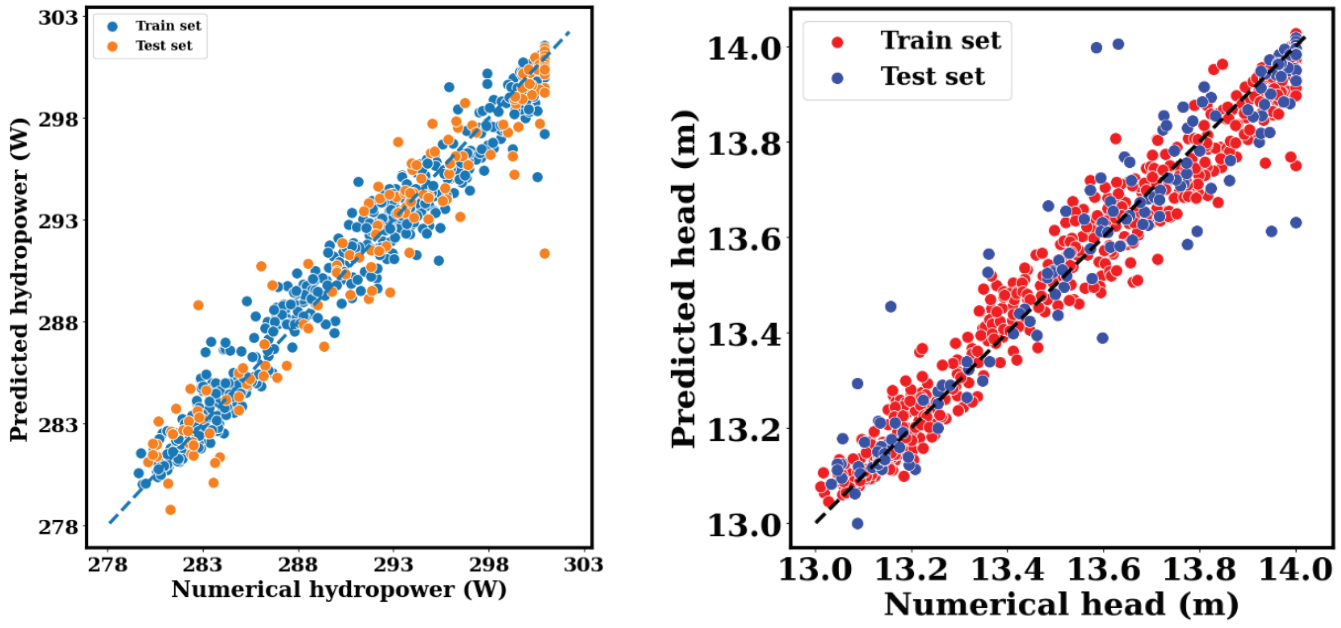
For a more in-depth analysis of the CatBoost, the regression plots for $P_t$ and $H_t$ are illustrated in Fig. 12. The CatBoost model boasts $R^2$ values for the two outputs of 0.9544 and 0.9354, respectively.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

However, Fig. 13 compares the predicted and numerical data. Fig. 13 (a) displays the prediction of $P_t$ , revealing MAE, RMSE, and MAPE values of 0.9426, 1.4893, and 0.32%, respectively. Fig. 13 (b) indicates the $(H_t)$, numerical, and predicted data, with MAE, RMSE, and MAPE of 0.0484, 0.083, and 0.36%, respectively. Moreover, the CatBoost model's ability to deliver satisfactory performance with big datasets could potentially explain the superior performance observed in this study, as shown in Table 7.

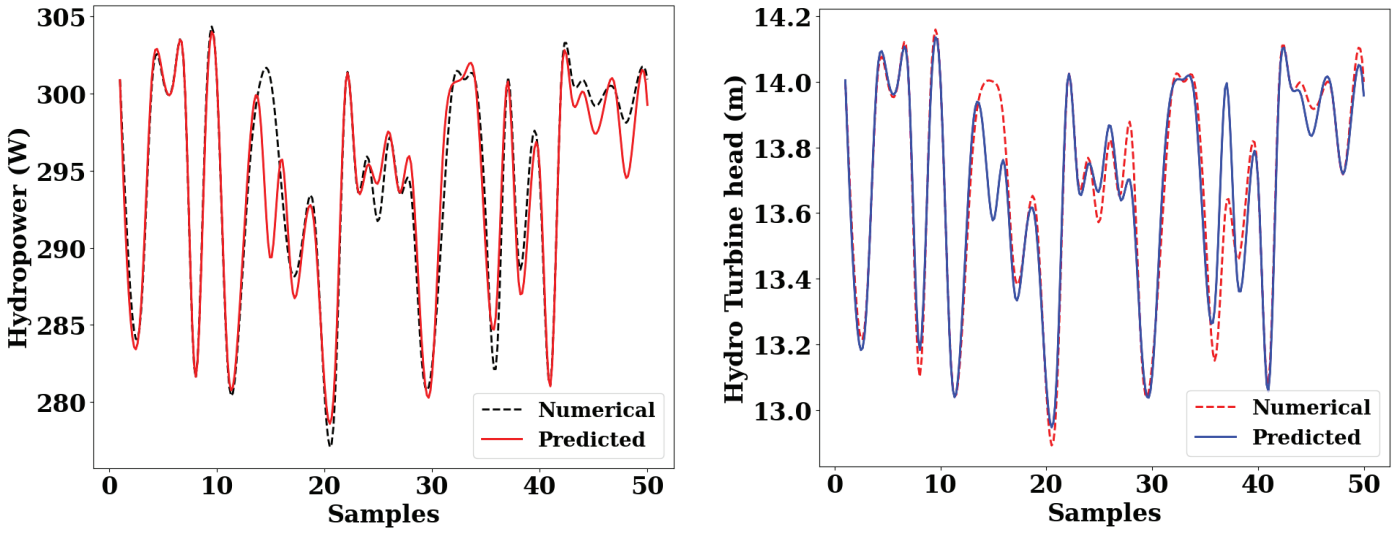
Table 7: The Catboost regressor model evaluation

| Catboost | $R^2$ | | MAE | | RMSE | | MAPE | |
|---|---|---|---|---|---|---|---|---|
| | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ |
| Train | 0.9826 | 0.9795 | 0.6582 | 0.0325 | 0.9600 | 0.0481 | 0.23% | 0.24% |
| Test | 0.9544 | 0.9354 | 0.9426 | 0.0484 | 1.4893 | 0.0830 | 0.32% | 0.36% |
| Total | 0.9773 | 0.9711 | 0.7151 | 0.0357 | 1.0867 | 0.0568 | 0.25% | 0.26% |




(a) (b)

Figure 12: Catboost regression scattered plot for (a) hydro turbine output power $(P_t)$, and (b) hydro turbine water head $(H_t)$.




(a) (b)

Figure 13: Catboost prediction versus numerical dataset for (a) hydro turbine output power $(P_t)$, and (b) hydro turbine water head $(H_t)$.

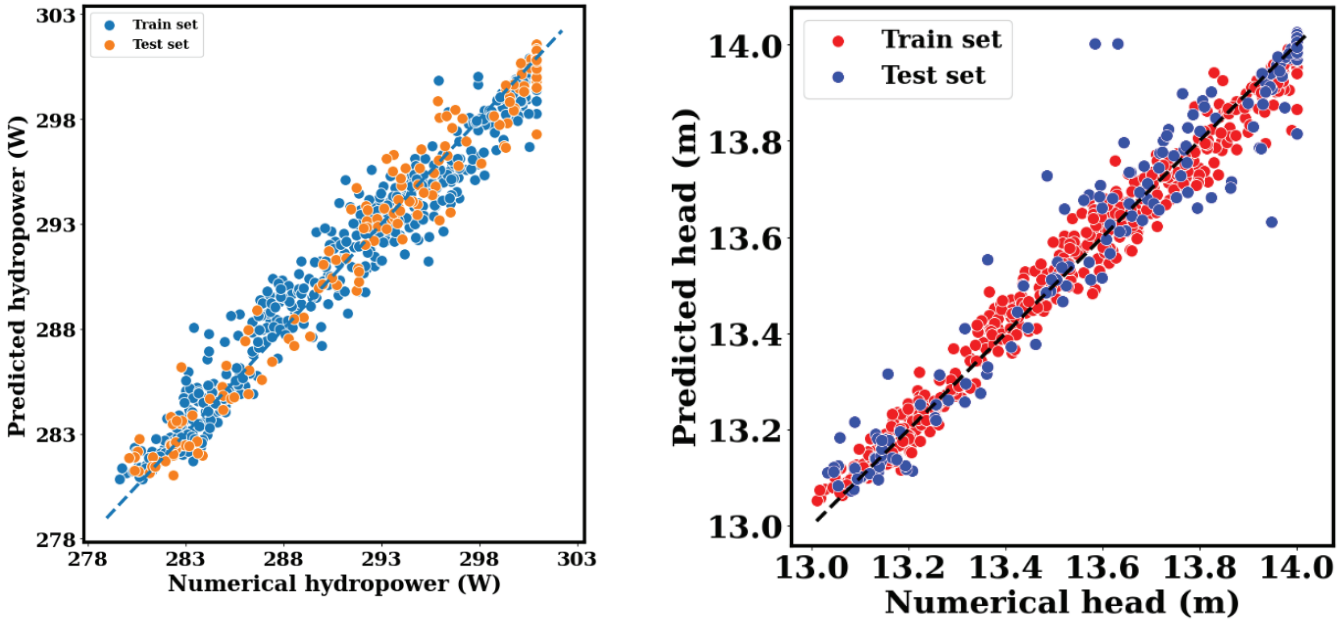
Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

### F. XGBoost

For $P_t$, $H_t$, the regression plots are shown in Fig. 14. The estimated $R^2$ for the two values was 0.9768 and 0.949, respectively. Fig. 15 provides a comparison between the predicted and numerical data. The prediction of $P_t$ is depicted in Fig. 15 (a), which results in values of MAE, RMSE, and MAPE of 0.7309, 1.0626, and 0.25%, respectively. Similarly, the $(H_t)$, numerical and predicted data, is presented in Fig. 15 (b) with MAE, RMSE, and MAPE of 0.0436, 0.0738, and 0.32%, respectively. The performance accuracies attained for $P_t$ and $H_t$ in the training, testing, and total datasets are reported in Table 8.

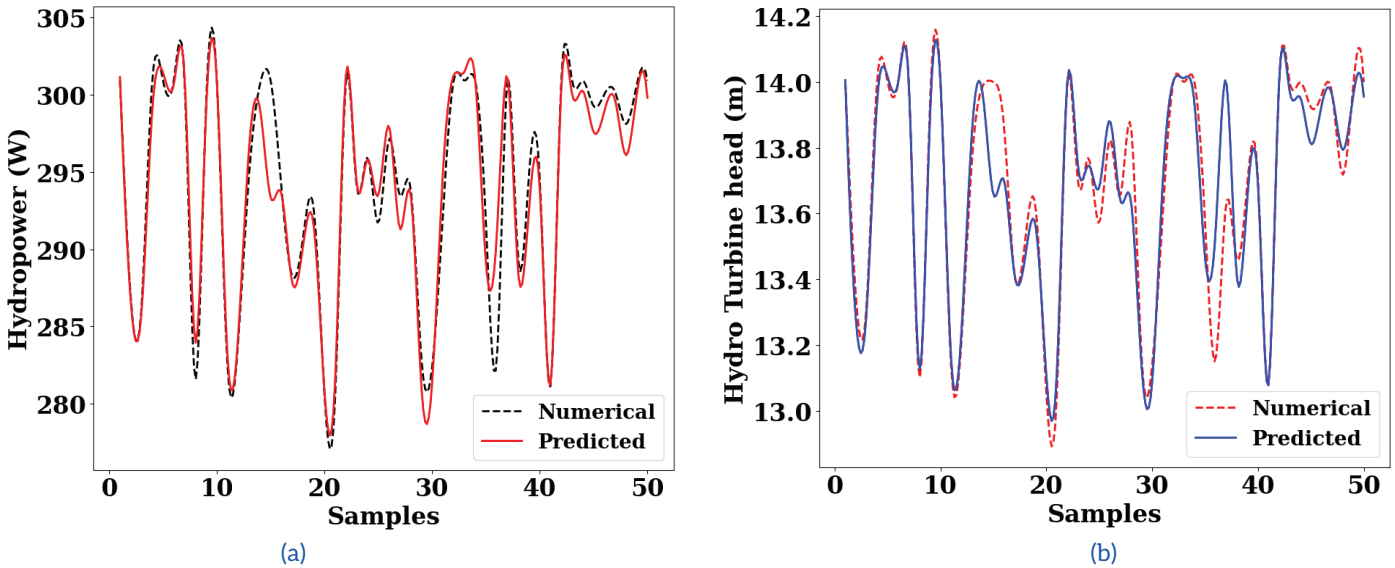
Table 8: The XG Boost regressor model evaluation

| XG boost | $R^2$ | | MAE | | RMSE | | MAPE | |
|---|---|---|---|---|---|---|---|---|
| | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ | $P_t$ | $H_t$ |
| Train | 0.9805 | 0.9901 | 0.6687 | 0.0224 | 1.0163 | 0.0335 | 0.23% | 0.17% |
| Test | 0.9768 | 0.9490 | 0.7309 | 0.0436 | 1.0626 | 0.0738 | 0.25% | 0.32% |
| Total | 0.9798 | 0.9822 | 0.6811 | 0.0267 | 1.0257 | 0.0445 | 0.23% | 0.20% |




(a) (b)

Figure 14: XGboost regression scattered plot for (a) hydro turbine output power $(P_t)$, and (b) hydro turbine water head $(H_t)$.




(a) (b)

Figure 15: XGboost prediction versus numerical dataset for (a) hydro turbine output power $(P_t)$, and (b) hydro turbine water head $(H_t)$.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

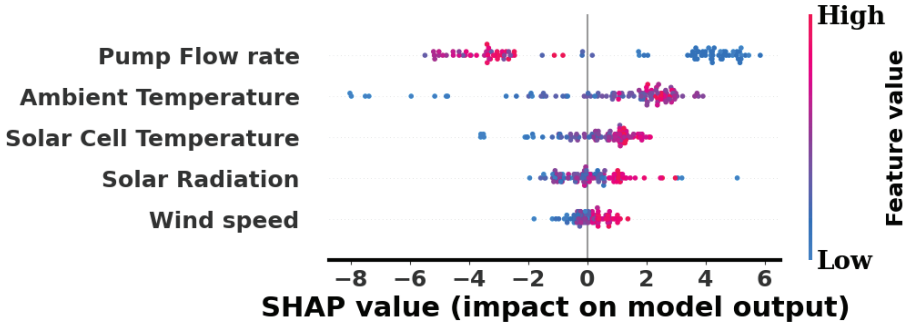
### G. *SHAP features importance analysis*

SHAP analysis was conducted to interpret the contribution of each input parameter to the predicted outputs. This interpretability approach provides a transparent and quantitative assessment of the influence exerted by each feature on the model's predictions, thereby making it easier to comprehend the underlying relationships between system variables. Since the XGBoost model exhibited superior accuracy in forecasting both the $P_t$ and $H_t$, the SHAP analysis was performed exclusively for this model.

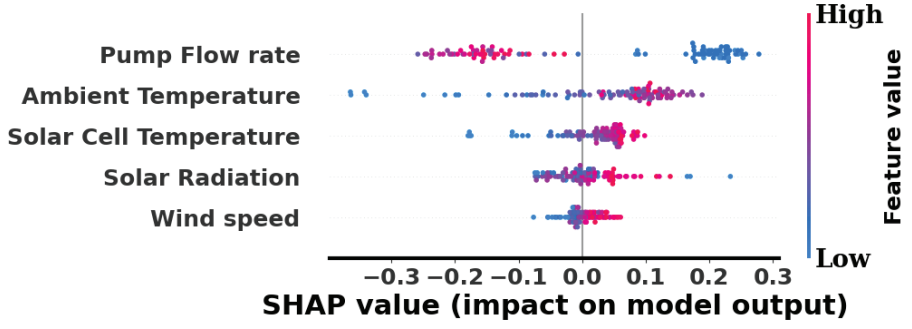
As illustrated in Fig. 16, the SHAP plots for the two outputs depict the relative importance of each input feature and offer further details about the interpretability of the model. The results indicate that the most important features change depending on what is predicted:

- For $P_t$, the pump flow rate was the most impactful feature, exhibiting a consistently positive influence on the model's output. In contrast, solar radiation demonstrated a mixed effect (both positive and negative SHAP values), highlighting its complex relationship with this particular output.
- For $H_t$, Solar radiation was the dominant feature with the strongest positive influence on the prediction, followed by the pump flow rate.
- Solar cell temperature exhibited a clear negative net influence on both predicted outputs, indicating that higher cell temperatures are associated with a decrease in both turbine power and water head. Conversely, ambient temperature and wind speed had relatively minor effects on the model's predictions.

The variation in feature importance ranking between the two outputs underscores the model's capability to discern the distinct physical relationships governing each output. Overall, the SHAP analysis revealed that all five models exhibited comparable feature-importance patterns, reinforcing the reliability and stability of the identified correlations between the operating parameters and the PV–PWS system performance. Although only the SHAP results of the best-performing model (XGBoost) are presented here, the remaining models demonstrated consistent feature relevance, confirming the robustness of the developed predictive framework.




(a)




(b)

Figure 16: SHAP plot for (a) hydro turbine output power *($P_t$)*, and (b) hydro turbine water head *($H_t$)*.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

### *H. Comparative summary of all models*

XGBoost consistently achieved the highest predictive performance across all evaluated metrics, reaching an $R^2$ of 0.9768 for hydro turbine power prediction, indicating excellent variance explanation. This high accuracy is further supported by the low RMSE and MAE values, confirming the model's ability to produce reliable predictions with minimal error. However, a deeper analysis of Fig 17 reveals several important insights beyond this overall superiority.

Figure 17 depicts a structured comparative evaluation of five machine learning models — SVR, AdaBoost, CatBoost, RF, and XGBoost — across four performance metrics: (a) $R^2$, (b) MAE, (c) RMSE, and (d) MAPE, evaluated for both hydraulic Power and pump head. Notably, within each subfigure, the models are deliberately arranged from worst to best (left to right), enabling an intuitive visual interpretation of performance ranking without requiring detailed numerical inspection. This design choice enhances clarity and facilitates rapid comparative assessment.

The most immediately striking observation is the significant disparity in error magnitudes between Power and Head. While Power MAE ranges from 0.731 to 1.443, Head MAE spans only 0.044 to 0.069. This reflects a fundamental physical characteristic of the system: Power exhibits a much wider dynamic range and higher variability than Head, making direct numerical comparisons between the two variables potentially misleading without normalization.

Across all metrics, XGBoost achieves the lowest MAE for both Power (0.731) and Head (0.044), establishing it as the most accurate model. In contrast, SVR performs worst across all metrics, which is consistent with its sensitivity to feature scaling and kernel selection in complex, high-variance regression problems.

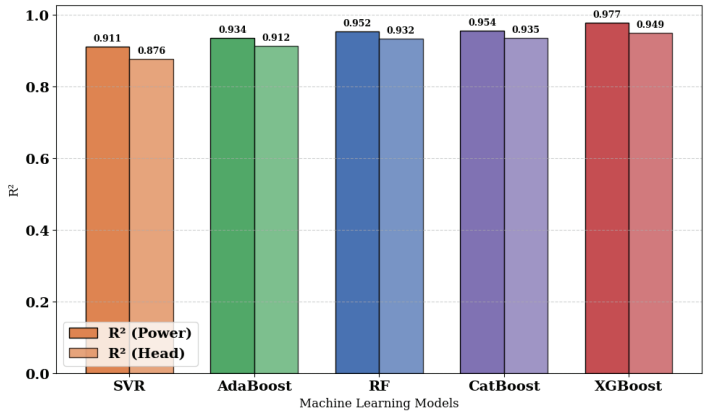
A more nuanced insight emerges when comparing RMSE and MAE between CatBoost and RF. Although RF achieves a lower MAE for Power (0.908 vs. 0.943), CatBoost outperforms RF in RMSE (1.489 vs. 1.521). This indicates that RF produces smaller average errors but is more susceptible to large prediction outliers. From an engineering perspective, such deviations are critical, as large unexpected errors in power prediction can propagate into operational inefficiencies in hydraulic system control.

MAPE provides additional interpretability by expressing error as a percentage, there by eliminating unit-scaling effects. While RF achieves the lowest Power MAPE (0.31%), it records the highest Head MAPE among competitive models (0.38%). In contrast, XGBoost demonstrates a more balanced performance profile, achieving 0.25% for Power and 0.32% for Head, indicating not only high accuracy but also consistent error distribution across both outputs.

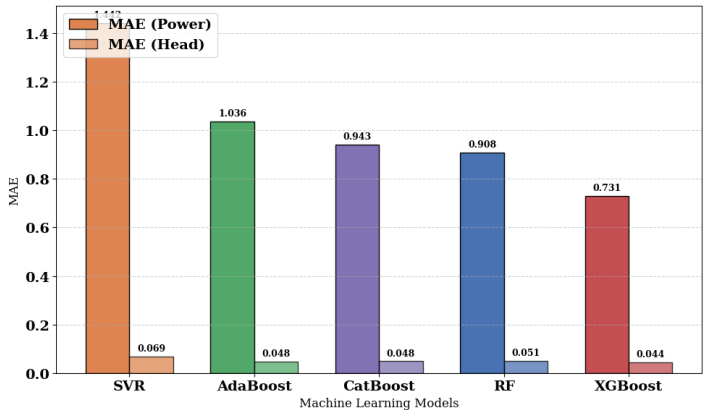
AdaBoost exhibits identical MAPE values (0.35%) for both variables, suggesting symmetric learning behavior, albeit at a lower overall accuracy compared to XGBoost.

Overall, the consistent superiority of XGBoost across all evaluation metrics and both output variables confirms it as the optimal model for this application. Its performance can be attributed to the strength of the gradient boosting framework, which effectively captures complex nonlinear relationships while controlling overfitting through regularization.

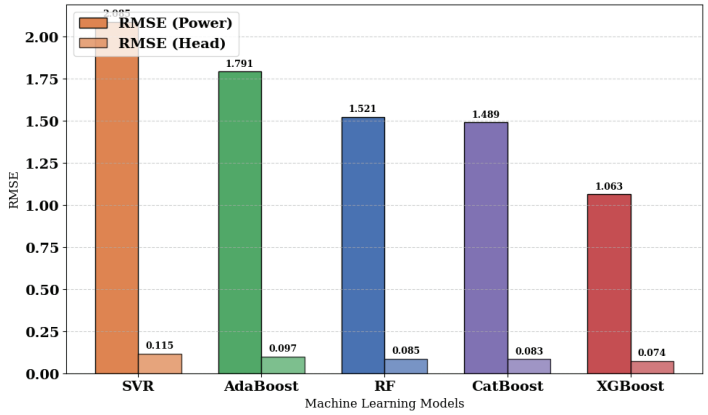
Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993



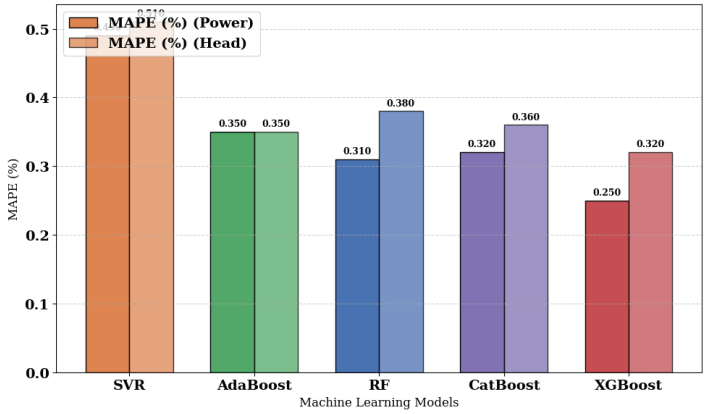

(a)




(b)




(c)




(d)

Figure 17: Optimal statistical performance metrics for the predictive models.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

# V. Discussion

The simulation and ML results demonstrate the effectiveness of ensemble learning methods—particularly XGBoost—for predicting the electro-hydraulic performance of hybrid PV–PWS systems. A consistent performance hierarchy was observed across both target variables ($P_t$ and $H_t$), where XGBoost outperformed all models, followed closely by CatBoost and Random Forest with nearly equivalent performance, while AdaBoost and SVR demonstrated comparatively lower predictive accuracy. This ordering reflects well-established findings in the ML literature on structured datasets: boosting algorithms that iteratively correct residual errors (XGBoost, CatBoost) tend to outperform bagging (RF) and individual nonlinear methods (SVR) when sufficient training data is available, and regularization prevents overfitting [21, 23].

The near-perfect correlation (r = 0.99) between $P_t$ and $H_t$ observed in the correlation matrix is physically expected—hydraulic Power is a direct product of Head and flow rate. This strong coupling also explains why both outputs respond similarly to the same dominant features (pump flow rate and solar irradiance). The negative SHAP contribution of cell temperature to both outputs aligns with established PV physics: elevated cell temperatures reduce panel conversion efficiency, which propagates through the system to reduce pump discharge and available Head.

## A. Limitations of the proposed work

Several limitations of the present study should be acknowledged:

- **Simulation-based dataset:** The 1000-sample dataset was generated from a deterministic numerical model validated against one published reference. While physically consistent, the dataset does not capture real-world measurement noise, sensor uncertainty, or equipment degradation, which could affect model generalization when deployed with real operational data.
- **Single climate region:** The system was simulated using meteorological data representative of Egyptian climatic conditions. Generalization to other climate regions (e.g., high-latitude or arid regions with different irradiance and temperature profiles) has not been evaluated and would require retraining.
- **Dataset size:** Although 1000 samples are adequate for the five ML models selected, deep learning alternatives (LSTM, CNN) would likely require substantially larger datasets to realize their full potential.
- **Static system configuration:** The numerical model assumes fixed system parameters (pump efficiency, turbine efficiency, reservoir geometry). In practice, component aging, fouling, and operational variability would introduce additional uncertainty not captured in the current framework.
- **Real-world deployment:** Transitioning from offline model development to real-time embedded deployment would introduce additional constraints related to computational latency, hardware compatibility, and the need for continuous model updating as operating conditions evolve.

## B. Future work directions

Based on the findings and limitations identified, future work should focus on:

- **Experimental validation:** Collecting real operational data from a physical PV–PWS installation to validate and recalibrate the ML models under real-world measurement noise and component variability.
- **Expanded climate coverage:** Retraining and evaluating the framework under diverse climatic datasets (e.g., NASA POWER data for multiple regions) to assess geographic generalizability.
- **Hybrid ML-optimization integration:** Combining the ML prediction framework with metaheuristic optimization algorithms (e.g., Grey Wolf Optimizer, PSO) for real-time energy management and operational scheduling.
- **Deep learning exploration:** Investigating LSTM or transformer-based architectures for temporal sequence modeling of system dynamics, particularly for multi-step ahead forecasting.
- **Fault detection extension:** Extending the ML framework to anomaly detection and predictive maintenance applications using unsupervised or semi-supervised methods.

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

## VI. Conclusions

This study developed and evaluated a machine learning framework for predicting the electro-hydraulic performance of a hybrid PV–PWS system, specifically targeting hydro turbine output power ($P_t$) and turbine water head ($H_t$). Five ML algorithms—RF, SVR, AdaBoost, CatBoost, and XGBoost—were trained on a 1000-sample MATLAB-generated dataset representing Egyptian climatic conditions, with Optuna-based hyperparameter optimization and 5-fold cross-validation to ensure robust and unbiased evaluation.

XGBoost consistently outperformed all other models, achieving $R^2$ = 0.9768 and 0.9490 for $P_t$ and $H_t$, respectively, with MAPE values below 0.35% for both outputs. SHAP analysis identified pump flow rate and solar irradiance as the dominant features, with PV cell temperature exhibiting a negative net influence on both outputs—consistent with known PV efficiency degradation at elevated temperatures.

The proposed framework demonstrates the viability of ensemble ML as a computationally efficient alternative to full numerical simulation for real-time prediction and operational planning of hybrid PV–PWS systems. Future work should focus on experimental validation with real-world operational data, extension to multiple climate regions, and integration with optimization algorithms for adaptive energy management.

### CRediT authorship Contribution statement

**Said M. A. Ibrahim:** Conceptualization, Suggesting Research Point, Supervision, Validation, Writing - Review & Editing.

**Alhasan Azouz:** Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft.

**Hamdy H. El-Ghetany:** Supervision, Validation.

### Statements and Declarations

#### *Ethical considerations*
Not applicable. This study does not involve human participants, human data, animals, or biological materials. Therefore, ethical approval was not required.

#### *Consent to participate*
Not applicable. This study does not involve human participants.

#### *Consent for publication*
Not applicable. This study does not contain any individual person's data in any form.

#### *Declaration of conflicting interest*
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### *Funding*
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

#### *Data availability*
The data that support the findings of this study are available from the corresponding author upon reasonable request.

## References

[1] M. Rawa *et al.*, "Economical-technical-environmental operation of power networks with wind-solar-hydropower generation using analytic hierarchy process and improved grey wolf algorithm," *Ain Shams Engineering Journal*, vol. 12, no. 3, pp. 2717–2734, Sep. 2021, doi: https://doi.org/10.1016/j.asej.2021.02.004.

[2] B. Oryani, H. Kamyab, A. Moridian, Z. Azizi, S. Rezania, and S. Chelliapan, "Does structural change boost the energy demand in a fossil fuel-driven economy? New evidence from Iran," *Energy*, vol. 254, p. 124391, Sep. 2022, doi: https://doi.org/10.1016/j.energy.2022.124391.

[3] H. Kamyab, A. Naderipour, M. Jahannoush, A. Abdullah, and M. H. Marzbali, "Potential effect of SARS-CoV-2 on solar energy generation: Environmental dynamics and implications," *Sustainable Energy Technologies and Assessments*, vol. 52, p. 102027, Aug. 2022, doi: https://doi.org/10.1016/j.seta.2022.102027.

[4] O. M. Abo Gabl, M. Y. Morgan, and M. S. El-Sobki (Jr.), "Decentralized economic operation of isolated AC, DC, and hybrid microgrids," *Renewable Energy and Sustainable Development*, vol. 11, no. 2, p. 175, Aug. 2025, doi: https://doi.org/10.21622/resd.2025.11.2.1289.

[5] M. S. Nazir, Z. M. Ali, M. Bilal, H. M. Sohail, and H. M. N. Iqbal, "Environmental impacts and risk factors of renewable energy paradigm—a review," *Environmental Science and Pollution Research*, vol. 27, no. 27, pp. 33516–33526, Jun. 2020, doi: https://doi.org/10.1007/s11356-020-09751-8.

[6] R. R. Elbanna, M. H. ElMessmary, H. Diab, and M. Abdelsalam, "A smart hybrid optimization model for DSSE in renewable energy-powered distribution networks," *Renewable Energy and Sustainable Development*, vol. 11, no. 2, p. 314, Sep. 2025, doi: https://doi.org/10.21622/resd.2025.11.2.1271.

[7] P. E. Campana, I. Papic, S. Jakobsson, and J. Yan, "Photovoltaic water pumping systems for irrigation: principles and advances," *Elsevier eBooks*, pp. 113–157, Jan. 2022, doi: https://doi.org/10.1016/b978-0-323-89866-9.00007-9.

[8] H. A. VAIDYA, M. K. RATHOD, and S. CHANNIWALA, "Performance enhancement of box solar cooker with photovoltaic panel," *Journal of Thermal Engineering*, vol. 11, no. 6, pp. 1658–1670, 2025, doi: https://doi.org/10.14744/thermal.0001032.

[9] A. K. Tiwari, V. R. Kalamkar, R. R. Pande, S. K. Sharma, V. C. Sontake, and A. Jha, "Effect of head and PV array configurations on solar water pumping system," *Materials Today: Proceedings*, vol. 46, pp. 5475–5481, 2021, doi: https://doi.org/10.1016/j.matpr.2020.09.200.

[10] A. Maleki, F. Heydari, and A. J. Moghadam, "Hybrid Optimization Method for Optimal Site Selection and Sizing of a Hybrid Photovoltaic Water Pumping/Diesel/Battery System," *Heliyon*, vol. 11, no. 1, p. e40692, Nov. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e40692.

[11] E. Sarmas, E. Spiliotis, V. Marinakis, G. Tzanes, J. K. Kaldellis, and H. Doukas, "ML-based energy management of water pumping systems for the application of peak shaving in small-scale islands," *Sustainable Cities and Society*, vol. 82, pp. 103873–103873, Jul. 2022, doi: https://doi.org/10.1016/j.scs.2022.103873.

[12] A. PAŞAOĞLU and A. HABIBNEZHAD, "Machine learning enhanced hybrid energy storage management system for renewable integration and grid stability optimization in smart microgrids," *Journal of Thermal Engineering*, vol. 11, no. 4, pp. 1040–1050, 2025, doi: https://doi.org/10.14744/thermal.0000962.

[13] C. Wang, C. Li, Y. Feng, and S. Wang, "Predicting hydropower generation: A comparative analysis of Machine learning models and optimization algorithms for enhanced forecasting accuracy and operational efficiency," *Ain Shams Engineering Journal*, vol. 16, no. 3, p. 103299, Mar. 2025, doi: https://doi.org/10.1016/j.asej.2025.103299.

[14] A. Rathore, Almas, and S. Sundaram, "Energy, exergy and performance analysis of a 380 kWP roof-top PV plant assisted with data-driven models for energy generation," *Journal of Thermal Engineering*, vol. 10, no. 5, pp. 1164–1183, Jan. 2024, doi: https://doi.org/10.14744/thermal.0000859.

[15] X. Zhang, Z. Zhang, Y. Liu, Z. Xu, and X. Qu, "A review of machine learning approaches for electric vehicle energy consumption modelling in urban transportation," *Renewable Energy*, vol. 234, p. 121243, Aug. 2024, doi: https://doi.org/10.1016/j.renene.2024.121243.

[16] R. A. Ibrahim and N. E. Zakzouk, "A machine learning framework for predicting fuel consumption and CO2 emissions in hybrid and combustion vehicles: comparative analysis and performance evaluation," *PLOS One*, vol. 21, no. 2, p. e0342418, Feb. 2026, doi: https://doi.org/10.1371/journal.pone.0342418.

[17] A. Kote, A. Khilari, C. Kadam, D. Hud, and A. Ramteke, "CO2 Emission Prediction Using Machine Learning," in *ICT for Intelligent Systems*, Singapore: Springer Nature Singapore, 2024, pp. 317–326. doi: https://doi.org/10.1007/978-981-97-6681-9_28.

[18] Z. Said *et al.*, "Application of novel framework based on ensemble boosted regression trees and Gaussian process regression in modelling thermal performance of small-scale Organic Rankine Cycle (ORC) using hybrid nanofluid," *Journal of Cleaner Production*, vol. 360, p. 132194, Aug. 2022, doi: https://doi.org/10.1016/j.jclepro.2022.132194.

[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: https://doi.org/10.1023/A:1010933404324.

[20] P. S. Kumar, C. Chandrika, P. K. Rao, P. K. Rao, and S. K. Oruganti, "Interpretable hybrid machine learning models for renewable-powered smart grid stability prediction," *Renewable Energy and Sustainable Development*, vol. 11, no. 2, p. 397, Oct. 2025, doi: https://doi.org/10.21622/resd.2025.11.2.1509.

[21] J. Hoxha, Muhammed Yasin Çodur, Enea Mustafaraj, H. Kanj, and Ali El Masri, "Prediction of transportation energy demand in Türkiye using stacking ensemble models: Methodology and comparative analysis," *Applied energy*, vol. 350, pp. 121765–121765, Nov. 2023, doi: https://doi.org/10.1016/j.apenergy.2023.121765.

[22] M. A. Hassan, H. Salem, N. Bailek, and O. Kisi, "Random Forest Ensemble-Based Predictions of On-Road Vehicular Emissions and Fuel Consumption in Developing Urban Areas," *Sustainability*, vol. 15, no. 2, p. 1503, Jan. 2023, doi: https://doi.org/10.3390/su15021503.

[23] Z. Jinbo, L. Yufu, and M. Haitao, "Handling missing data of using the XGBoost-based multiple imputation by chained equations regression method," *Frontiers in Artificial Intelligence*, vol. 8, Apr. 2025, doi: https://doi.org/10.3389/frai.2025.1553220.

[24] A. A. Ajala, O. L. Adeoye, O. M. Salami, and A. Y. Jimoh, "An examination of daily CO2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models," *Environmental Science and Pollution Research*, vol. 32, no. 5, Jan. 2025, doi: https://doi.org/10.1007/s11356-024-35764-8.

[25] S. Haddad, M. Benghanem, A. Mellit, and K. O. Daffallah, "ANNs-based modeling and prediction of hourly flow rate of a photovoltaic water pumping system: Experimental validation," *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 635–643, Mar. 2015, doi: https://doi.org/10.1016/j.rser.2014.11.083.

[26] M. Sapitang, W. M. Ridwan, K. Faizal Kushiar, A. Najah Ahmed, and A. El-Shafie, "Machine Learning Application in Reservoir Water Level Forecasting for Sustainable Hydropower Generation Strategy," *Sustainability*, vol. 12, no. 15, p. 6121, Jul. 2020, doi: https://doi.org/10.3390/su12156121.

[27] M. Dehghani *et al.*, "Prediction of Hydropower Generation Using Grey Wolf Optimization Adaptive Neuro-Fuzzy Inference System," *Energies*, vol. 12, no. 2, p. 289, Jan. 2019, doi: https://doi.org/10.3390/en12020289.

[28] B. S. Pali and S. Vadhera, "A novel solar photovoltaic system with pumped-water storage for continuous power at constant voltage," *Energy Conversion and Management*, vol. 181, pp. 133–142, Feb. 2019, doi: https://doi.org/10.1016/j.enconman.2018.12.004.

[29] J. A. Duffie and W. A. Beckman, *Solar engineering of thermal processes*. Wiley New York, 1980.

[30] NASA, "NASA Langley Research Center POWER Data Access Viewer," *Nasa.gov*, 2024. https://power.larc.nasa.gov/data-access-viewer/

Volume 12, Issue 1, June 2026 - ISSN 2356-8569
http://dx.doi.org/10.21622/RESD.2026.12.1.1993

[31] Oussama Khouili, M. Hanine, and M. Louzazni, "Harnessing Principal Component Analysis and Artificial Neural Networks for Accurate Solar Radiation Prediction," *International Journal of Energy Research*, vol. 2025, no. 1, p. p. 5846114, Jan. 2025, doi: https://doi.org/10.1155/er/5846114.

[32] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: https://doi.org/10.1023/b:stco.0000035301.49549.88.

[33] N. Nguyen and D. Ngo, "Comparative analysis of boosting algorithms for predicting personal default," *Cogent Economics & Finance*, vol. 13, no. 1, Feb. 2025, doi: https://doi.org/10.1080/23322039.2025.2465971.

[34] M. Hamid, F. Hajjej, A. S. Alluhaidan, and N. Waleed, "Fine tuned CatBoost machine learning approach for early detection of cardiovascular disease through predictive modeling," *Scientific Reports*, vol. 15, no. 1, Aug. 2025, doi: https://doi.org/10.1038/s41598-025-13790-x.

[35] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[36] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: https://doi.org/10.5194/gmd-7-1247-2014.