

Hybrid analysis of photovoltaic energy data containing excessive structural zeros

Received: 6 January 2026

Accepted: 16 June 2026

Published online: 22 June 2026

Cite this article as: Altinişik Y., Aydın D., Esen V. *et al.* Hybrid analysis of photovoltaic energy data containing excessive structural zeros. *Sci Rep* (2026). <https://doi.org/10.1038/s41598-026-58719-0>

Yasin Altinişik, Demet Aydın, Vedat Esen, Taner Dindar & Ali Samet Sarkin

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

Hybrid analysis of photovoltaic energy data containing excessive structural zeros

Yasin ALTINIŞIK^{1,2}, Demet AYDIN¹, Vedat ESEN^{3,*}, Taner DİNDAR⁴, Ali Samet SARKIN⁵

¹ Department of Data Science and Analytics, Faculty of Science and Literature, Sinop University, Sinop, Türkiye (Orcid: 0000-0001-9375-2276), (Orcid: 0000-0002-3491-8392)

² Department of Research Methods and Statistics, Institute of Psychology, University of Hamburg, Hamburg, Germany (Orcid: 0000-0001-9375-2276)

^{3,*} vedatesen@topkapi.edu.tr Department of Electric Electronics Engineering, İstanbul Topkapi University, 34087 İstanbul, Türkiye (Orcid: 0000-0001-6230-6070)

⁴ Department of Electronic and Automation, Ankara University, Ankara, Türkiye (Orcid: 0000-0002-4112-2114)

⁵ Department of Electrical and Energy, Osmaniye Korkut Ata University, 80750, Osmaniye, Türkiye (Orcid: 0000-0001-6261-0531)

Abstract

Photovoltaic (PV) power generation plays a critical role in the global transition toward sustainable energy systems. However, accurate PV power forecasting remains challenging due to the non-stationary nature of PV power time series and the presence of structural zeros. Most existing studies rely on parametric models or treat zero inflation as a secondary issue. In this study, a hybrid hurdle modeling framework is implemented to explicitly account for both structural zeros and continuous positive power values through a two-part structure. The zero component is modeled using Bayesian logistic regression (BLR), random forest classifier (RFC), and support vector classifier (SVC). The positive values are analyzed using the parametric models Gamma regression (GR), Log-normal regression (LNR), and Weibull regression (WR) and the non-parametric machine learning (ML) techniques random forest regression (RFR), extreme gradient boosting (XGBoost), and support vector regression (SVR). The proposed framework is validated using the target variable active power (AP, kW) based on a real-world data from a 110 kWe (129.6 kWp) PV plant located in Çaycuma, Zonguldak, Türkiye. Four metrics were used to compare the predictive performances of the models: mean squared error (MSE), mean absolute (scaled) error (MAE and MASE), and the coefficient of determination (R^2). Among the 18 distinct hurdle models, the lowest error values and the highest R^2 were obtained when the positive component was modeled by either RFR, such as BLR-RFR (MSE = 67.425 kW, MAE = 3.974 kW, MASE = 0.615, R^2 = 0.899) or XGBoost, such as RFC-XGBoost (MSE = 69.252 kW, MAE = 4.075 kW, MASE = 0.630, R^2 = 0.898). The findings indicate that hybrid modeling approaches where the zero component is modeled using BLR, RFC, or SVC, and the positive component is modeled with either RFR or XGBoost provide a reliable framework for PV power forecasting, when predicting continuous data characterized by excessive structural zeros.

Keywords: Hybrid Modeling, Hurdle Models, Structural Zeros, Machine Learning, Photovoltaic Power Forecasting.

Introduction

The use of renewable energy sources is on an increasing trend all over the world due to the depletion of fossil fuels and increasing concerns about global warming [1]. Photovoltaic (PV) systems, which convert sunlight directly into electrical energy, are considered the most effective solution in the energy production sector among renewable energy sources because they are clean, cheap and useful [2]. As countries embrace efforts to minimize their carbon footprint, they continue to rapidly add capacity to generate electricity from renewable energy. Solar and wind energy account for 96% of this capacity increase. Solar PV alone accounts for two-thirds of all new renewable electricity technologies. Considering this information, PV is expected to have the largest share in electricity production, which is indispensable for societies in the coming years [3]. Therefore, power plants using PV systems must have stable and reliable power generation. However, electricity production from solar energy varies due to solar radiation, meteorological conditions and seasonal fluctuations. These fluctuations can occur unexpectedly and have a significant impact on how much electricity the PV power plant produces. Additionally, the variability of solar energy production makes it difficult to integrate energy sources and maintain a stable electricity grid. Integration of large-scale PV plants into the electricity grid is an important and difficult issue for electricity operators. Electricity operators have to balance power production and consumption to prevent energy waste and not be penalized due to instability in the electricity grid. They need sophisticated forecasting methods that can precisely predict short-term variations in power output in order to guarantee a steady energy supply in spite of these quick changes [4], [5], [6]. Recently, with new technologies, the field of PV power forecasting has developed strongly and many forecasting models have been put forward. The forecasting approaches of these models derive algorithms that take this unpredictability into account to maximize the performance of PV systems [7].

The motivations driving this research are the desire to solve the problems encountered in PV power forecasting, such as weather dependency, spatial variability, data limitations, and scaling difficulties. Overcoming these challenges requires advanced models, high-quality data, and robust computational methods. Forecasting PV power with high accuracy is crucial for grid stability, proper energy management, and grid integration with renewable energy. In addition, the success of the forecasts allows balancing electricity supply and demand and minimizing grid outages caused by solar energy variability. It helps in optimizations in energy storage systems, corrects the planning of different power sources, and improves market operations by providing reliable forecasts for energy trading. In

addition, PV forecasting reduces operational costs, supports regulatory compliance, and maximizes the use of renewable resources.

Numerous studies have been conducted on PV power forecasting techniques, with physics-based approaches based on PV power models widely recognized as one of the leading methods. In addition to these approaches, statistical methods have also attracted considerable attention. Within the statistical forecasting framework, techniques such as time series analysis, machine learning (ML), and deep learning (DL) are commonly used. In recent years, hybrid approaches have gained increasing importance for further improving forecasting accuracy and reliability. These hybrid methods aim to enhance the predictive performance of both physical and statistical models [8], [9], [10], [11].

Some of the statistical approaches estimate future PV power output using time-series data as input information [12], [13], [14]. Time series models such as autoregressive moving average (ARMA) [15], [16], autoregressive integrated moving average (ARIMA) [17], [18] and seasonal autoregressive moving average (SARIMA) [19], [20] have proven effective for short term forecasting in large scale PV power plants [21]. Due to their ability to handle nonlinear and heterogeneous inputs, ML models such as artificial neural networks (ANNs) [22], [23] random forest regression (RFR) [24], support vector regression (SVR) [25], Gaussian process regression (GPR) [26], [27], and extreme learning machines (ELMs) [28], [29] are widely used in PV power estimation [4], [30]. Moreover, methodologies such as RFR, ensemble regression trees, and GPR provide enhanced stability and reliability in predictive performance [31], [32], [33].

Substantial limitations have been detected in some recent hybrid ML studies for PV power forecasting. Asiedu et al. [34] used data from a grid-connected solar power plant to develop a hybrid prediction system based on RFR, ANNs, and extreme gradient boosting (XGBoost) [35]. The authors developed three hybrid models, namely XGBoost-RFR, ANNs-RFR, and ANNs-XGBoost-RFR, and showed that these models outperform single-stage ML techniques such as linear, Ridge, and Lasso regressions, along with RFR, XGBoost, and ANNs. Nevertheless, the study is subject to several significant limitations. The PV power forecasting was carried out over four time frames, ranging from a single day to one month. Due to this focus on short time horizons, the hybrid models utilized in the analysis remain limited in terms of capturing seasonal variations in PV power forecasting. Additionally, the authors state in the conclusion that the models utilized only three features (irradiation, temperature, and time), thereby significantly restricting the generalizability of their findings on PV power forecasting. Saxena et al. [36] developed a hybrid model by combining the K-Nearest Neighbors (KNN) [37], [38] model with SVR. The authors demonstrated that KNN-SVR outperforms the traditional long short-term memory (LSTM) technique, achieving an accuracy level of 98%. However, while the research examined the influence of factors

including temperature, sunshine duration, solar radiation, and time on solar PV power forecasting, it fails to consider the effects of environmental elements such as wind direction, wind speed, and humidity. Additionally, the study focused exclusively on the performance of the KNN-SVR hybrid model, without considering alternative hybrid configurations. Furthermore, the inherent nature of the employed hybrid model does not have the ability to detect seasonal impacts. In a similar vein, various techniques documented in the literature do not focus on modeling seasonal effects. Esen et al. [39] employed regression-based models for short-term PV power forecasting owing to their robustness, interpretability, and competitive performance on limited datasets. Ren et al. [40] reported that a quantile regression approach based on DL effectively captures uncertainty in short-term PV energy predictions, providing reliable probabilistic predictions while maintaining competitive point prediction accuracy. In their study combining GPR with Bayesian hyperparameter adjustment, Islam et al. [41] obtained accurate and reliable short-term PV power predictions by explicitly modeling uncertainty and maintaining robustness across limited datasets.

The current study demonstrates several notable strengths over the contemporary hybrid ML techniques for PV power forecasting discussed above. When modeling PV power, we account for seasonal variations by incorporating seasonal components into our model. As will be discussed later, in addition to these seasonal components, the model contains various environmental factors influencing PV power generation. Moreover, datasets for PV power forecasting are typically characterized by excessive structural zeros which, unlike sampling zeros, have no potential to take on a different value. However, recent studies often treat zero-inflation as a nuisance factor, despite its potential to exert detrimental effects on parameter estimates. In this regard, particular attention has been paid to the modeling of the structural zeros in the data.

The novelty of the study stems from the fact that it is one of the initial investigations to explicitly integrate structural zeros into the modeling framework. The recommended framework involves hurdle models that serve as specific instances of two-stage models. The application of hurdle models in the domain of PV power forecasting is quite limited. Gillingham and Tsvetanov [42] examine the demand for solar PV power systems using a hurdle model to analyze the related data containing excess zeros. Kim et al. [43] enhance the hurdle modeling methodology for urban PV power forecasting by establishing a two-stage ML approach. Neither of the two aforementioned studies above aligns perfectly with the objectives of this study. Unlike the goals of these studies, our research seeks to model seasonal and environmental impacts on PV power forecasting by addressing the structural zeros in the data. Another innovative feature of our research is that the hybrid modeling approach we utilize to address structural zeros remains effective in scenarios where

maximum likelihood estimates cannot be obtained due to non-convergence and/or separation problem [44]. The study employs techniques that mitigate these problems when dealing with structural zeros in the data. In summary, the current study aims to simultaneously achieve the following objectives: (i) Capturing seasonal effects in PV power forecasting, (ii) Evaluating the effects of environmental factors (and their interactions) in PV power forecasting, (iii) Accounting for structural zeros inherently present in the data, (IV) Dealing with non-convergence and separation problems when incorporating structural zeros into the analysis procedure. To accomplish these tasks simultaneously, we utilize a comprehensive modeling framework containing 18 different two-stage hurdle models.

Many single-stage parametric models such as Gamma regression (GR) [45], Log-normal regression (LNR) [45], and Weibull regression (WR) [46] are not designed to handle data containing (excessive) structural zeros. This circumstance evidently showcases the need for two-stage hybrid modeling in parametric models, when the model under consideration cannot deal with structural zeros. Additionally, our findings in this study indicate that the two-stage hybrid models we implemented can attain results that are either comparable to or superior to those of ML techniques like RFR [24], XGBoost [35], and support vector regression (SVR) [47]. Therefore, our framework enhances the reliability of the analysis, especially in cases where zero values are commonly observed. Although the empirical application focuses on PV power forecasting, the proposed framework can be applied to various renewable energy studies and other fields where structural zeros are common.

Our hybrid modeling approach is composed of two primary components: zero component and positive component. Hybrid modeling exhibits flexibility in that it permits the incorporation of various techniques for modeling structural zeros and positive values in the data independently. The zero component is modeled by Bayesian logistic regression (BLR) [48], random forest classifier (RFC) [24], and support vector classifier (SVC) [47]. These approaches are used to model the occurrences of structural zeros due to their suitability for binary outcomes, providing computational efficiency and stable estimation in our hybrid modeling framework. The solutions provided by these approaches are important especially when the classical maximum likelihood estimation fails to converge or encounters the separation problem [44]. The positive component is examined through the parametric models GR, LNR, and WR, as well as the non-parametric ML techniques RFR, XGBoost, and SVR. The evaluation process illustrates that integrating BLR, RFC, and SVC with these models or ML techniques yields impressive predictive performance. Conducting a thorough comparative assessment of different forecasting accuracy measures allows for a structured evaluation of the utilized hybrid modeling strategy. In this manner, the content of the study is enhanced through the inclusion of a diverse set of characteristics into the analysis of PV power forecasting data.

The proposed hybrid hurdle modeling framework provides several practical implications for decision-making in photovoltaic energy management. By explicitly modeling structural zeros (e.g., nighttime non-production) separately from positive generation values, the approach enables more accurate production forecasting under zero-inflation conditions. Improved forecasting performance can support short-term operational planning, grid balancing strategies, and energy trading decisions. Furthermore, the enhanced modeling of variability and extreme production values contribute to more reliable risk assessment and capacity planning. Overall, the framework offers a data-driven tool that can assist energy managers and system operators in optimizing resource allocation and reducing operational uncertainty.

Methodology

This study analyzes continuous data with excess structural zeros using a hybrid modeling framework. Figure 1 depicts the framework structure associated with hybrid modeling. The hybrid modeling strategy is implemented through hurdle models, a specific class of two-stage models. These models separately account for structural zeros and positive values in the PV power forecasting data by employing two distinct sub-models. Following the completion of the data preparation phase, the final data are divided into a training set comprising 80% of the data and a testing set consisting of 20% of the data. In the training set, BLR, RFC, and SVC are used to model structural zeros, while the positive observations are analyzed using both parametric models (i.e., GR, LNR, and WR) and non-parametric ML techniques (i.e., RFR, XGBoost, SVR). In the testing set, the fitted values from the models used in Stages 1 and 2 are combined to evaluate the predictive performance of each hurdle model employed in the research. We commence by introducing the parametric models and non-parametric ML techniques used in Stages 1 and 2. Subsequently, we elaborate on data preparation and the application of hybrid modeling for the training set and testing set of the PV power forecasting data, respectively.

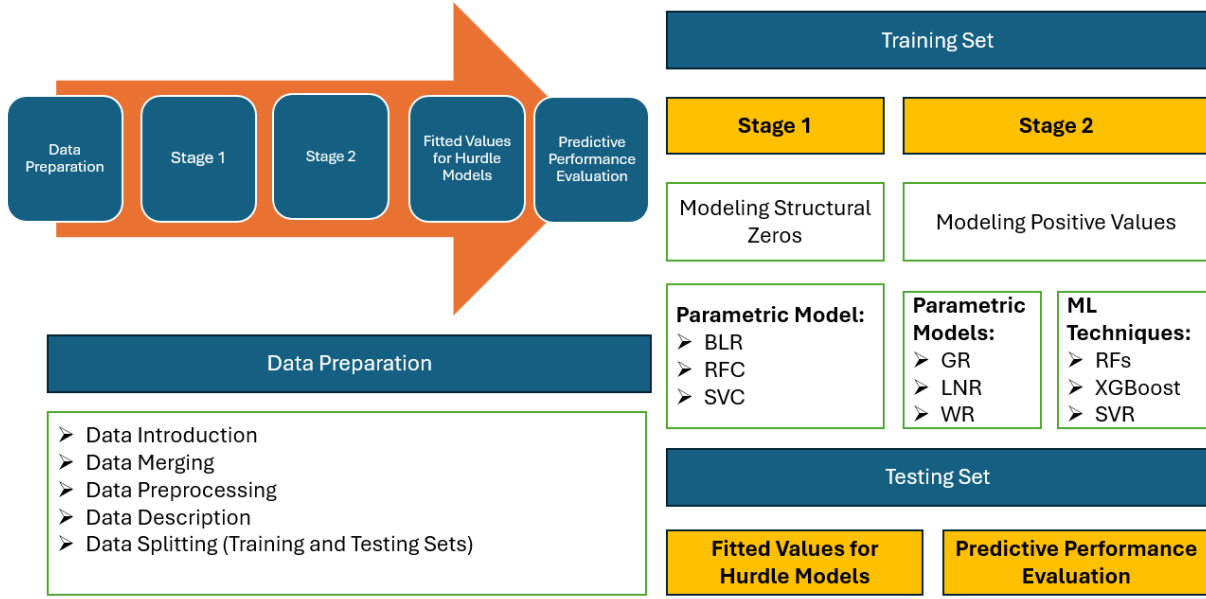


Fig. 1. Structure pertaining to the hybrid modeling framework.

Modeling Structural Zeros

At the first stage, we analyze structural zeros in the PV power forecasting utilizing BLR, RFC, and SVC.

Bayesian Logistic Regression (BLR)

The classical logistic regression (LR) is widely utilized in modeling binary variables in both statistics and ML disciplines. This technique is applicable when the target variable can solely take on the values of 0 (failure) and 1 (success). It relies on the principle of connecting success probabilities to linear predictors by utilizing the logit link function. The LR model is described as:

$$\text{logit}(\pi_i) = X_i^T \boldsymbol{\theta}, \quad (1)$$

where π_i is the (conditional) probability of success for the i th observation ($i = 1, 2, \dots, n$), X_i is the i th row of the design matrix containing the values of the attributes in the model, and $\boldsymbol{\theta}$ is the vector of model parameters. The logit link function in Equation 1 relates the success probability for the i th observation with the corresponding linear predictor, that is, $X_i^T \boldsymbol{\theta}$. Here, the logit function is defined as the logarithm of the ratio between the probability of success and the probability of failure (odds), that is, $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$.

As will be elaborated later, when modeling the structural zeros within the training set of the PV power forecasting data, the classical LR cannot be utilized in our context due to issues like non-convergence and separation problem [44]. Bayesian logistic regression (BLR) [48]

provides an effective prediction environment in situations where the classical LR encounters such issues. BLR operates on the principle of generating a posterior distribution and utilizing insights gained from this distribution to make predictions. The posterior distribution in the log scale is defined as follows:

$$\log P(\theta|y) = L(\theta|y) + \log P(\theta), \quad (2)$$

where $L(\theta|y)$ signifies the information derived from the data, while $\log P(\theta)$ denotes the prior information. We obtain the estimates of the model parameters in BLR using the *stan_glm* function with *family = binomial(link = "logit")* option, available in the *rstanarm* package [49] of *R* statistical software. This function implements a weakly informative $N(0, 2.5)$ prior distribution for regression coefficients as the default option.

Random Forest Classifier (RFC)

The RFC [24] is a widely used non-parametric ensembling technique in the field of ML. In the realm of classification, RFC aims to collate the information obtained from multiple decision trees, and thus, makes more robust predictions compared to a single decision tree. In this regard, RFC represents an advanced version of the decision tree algorithm designed to address more intricate issues. We implement the RFC algorithm using the *randomForest* function of the *randomForest* package [50] in *R*.

Decision trees are mostly suitable for solving simple problems. The primary reason for this is that intricate data types with many attributes lead the decision tree to branch more frequently (and become increasingly complex) in order to model the target variable, a phenomenon referred to as overfitting [51], [52], [53]. RFC is less susceptible to overfitting compared to a single decision tree. In RFC, decision trees (e.g., $n_{tree} = 1000$) are generated by drawing independent samples from the training set through non-parametric bootstrapping (with replacement) [54]. In each bootstrap sample, the process of selecting random subsets of features occurs separately for each node within the corresponding decision tree. Based on the data gathered from these bootstrap samples, the classification that receives the highest vote is identified as the final output. Thanks to its non-parametric characteristics and ability to account for non-linear relationships between variables, RFC provides a modeling framework that is remarkably sturdy against issues such as non-convergence and separation, especially when compared to conventional LR models.

Support Vector Classifier (SVC)

Support vector machines are applicable for both classification and regression tasks. Within the domain of classification, support vector machines are designated as SVC [45]. It conducts classification in the presence of linear or non-linear relationships between variables utilizing different kernel functions (e.g., linear, polynomial, Gaussian, and

sigmoid). In the linear SVC technique, it is assumed that the relationship between the attributes (or features) and the target variable can be elucidated using a linear function, whereas this assumption is relaxed in the non-linear SVC technique. This study employs the Gaussian kernel, which is the most commonly utilized kernel function within SVC, also known as the radial basis function (RBF) kernel [55], [56]. We employ SVC with the Gaussian kernel function using the *svm* function of the *e1071* package [57] in *R* with the *kernel = "radial"* option. The default setting for the kernel parameter was employed, which is the inverse of the total number of attributes (P) in the data. Note that for classification purposes, the target variable must be defined as a factor. Additionally, the "*probability = TRUE*" option must be enabled in *svm*, as it provides the class probabilities necessary to calculate the fitted values in hurdle modeling.

Modeling Positive Values Using ML Techniques

At the secondary stage, we investigate positive values in the PV power forecasting data using both parametric models and non-parametric ML techniques. We employ three non-parametric ML techniques, namely RFR, XGBoost, and SVR.

Random Forest Regression (RFR)

Random forests can also be used for both classification and regression purposes. When random forests are employed for regression tasks, they are referred to as RFR [24]. Therefore, similar to RFC, RFR is an ensembling technique based on bootstrap samples for decision trees. One of the key distinctions between RFC and RFR lies in the approach used to extract information from bootstrap samples. In RFR, the predictions regarding the values of the target variable in the testing set are established by averaging the fitted values obtained based on the decision trees. The *randomForest* function of the *randomForest* package [50] in *R* is also utilized for prediction in RFR. Note that this function automatically executes either classification or regression tasks, depending on the nature of the target variable. Specifically, it conducts classification when the target variable is categorized as a factor, and it performs regression when the variable is numeric.

Extreme Gradient Boosting (XGBoost)

The XGBoost [35] algorithm refers to a boosting technique that successively and dependently trains multiple weak models to develop a strong final model. The study examines decision trees also within the framework of boosting. This approach allows us to compare RFR (using the bagging algorithm) and XGBoost (using the boosting algorithm) on the same basis. The main objective of the bagging technique is to achieve more reliable predictions by minimizing the variance of the outcomes produced by independent decision trees. The boosting approach centers on the sequential construction of weak (biased)

decision trees that are interdependent, aiming to achieve a collection of decision trees that is unbiased (or less biased). In this regard, employing RFR and XGBoost within the same framework allows us for the distinct evaluation of prediction performance through both variance and bias reductions. We implement XGBoost using the *xgb.Dmatrix* function in the *xgboost* package [58] in *R*.

Support Vector Regression (SVR)

In the realm of regression modeling, support vector machines are denoted as SVR [45]. SVR is used to reveal linear or non-linear relationships between variables with the assistance of the same kernel functions elaborated for SVC above. We also utilize the Gaussian kernel within the context of SVR, which is the default option. The *svm* function of the *e1071* package [57] in *R* is also used for regression purposes. To be able to use this function for regression, the target variable must be a numeric vector. As will be elaborated later, we utilize an epsilon-insensitive approach within the SVR framework, where a medium size cost parameter is chosen for the errors. Note that the “*probability = TRUE*” option must be omitted when using the *svm* function for regression purposes.

Modeling Positive Values Using Parametric Models

At the secondary stage, we also utilize parametric models to address positive values in the PV power forecasting data. We implement three different parametric models which are GR, LNR, and WR.

Gamma Regression (GR)

The GR [45] model serves as a special form of generalized linear models (GLMs), which is defined as:

$$\log(\mu_i) = Z_i^T \eta, \quad (3)$$

where μ_i is the expected value of the positive value for the i th observation ($i = 1, 2, \dots, n$), Z_i is the i th row of the design matrix, and η is the vector of regression coefficients [45]. Notably, the design matrices $\mathbf{X}$ in Equation 1 and $\mathbf{Z}$ in Equation 3 do not necessarily contain the values of the same predictors in the data. The elements of the mean vector are related to the elements of the parameter vector(s) based on the distribution of the target variable. That is, $\mu_i = \alpha_i \beta_i$ ($i = 1, 2, \dots, n$) for GR with α_i and β_i are the i th values of the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ containing the values of the shape and scale parameters, respectively. We perform the parameter estimation for GR utilizing the *glm* function in the *stats* package [59] with *family = Gamma(link = "log")* option.

Log-normal Regression (LNR) and Weibull Regression (WR)

The LNR [45] and WR [46] models are the special instances of survival models, which are defined as:

$$\log(T_i) = Z_i^T \eta + \phi \varepsilon_i, \quad (4)$$

where T_i is the i th value of the target variable (survival time) and ϕ is the scale parameter [60], [61]. Here, ε_i represents the i th value of the random error which follows a standard normal distribution for the LNR model and a standard extreme value (Gumbel Type I) distribution for the WR model. Similar to GLMs, the components of the mean vector are associated with the components of parameter vector(s) in accordance with the distribution of the target variable. For the LNR model, $\mu_i = \exp(\gamma_i + \frac{\sigma_i^2}{2})$ with $\gamma_i = Z_i^T \eta$ and σ_i are the i th values of the parameter vectors containing the mean and standard deviations for the logarithm of the survival times. For the WR model, $\mu_i = \xi_i \Gamma(1 + \frac{1}{\nu})$ with ν (constant) and ξ_i are the shape and scale parameters for the i th observation, respectively.

Table 1 displays the i th value of the parameter vector (λ_i), the probability density function (pdf), and the mean vector (μ_i) for each of the models used in the second stage of the hurdle models. For each of these distributions, the dependent variable is defined only for positive values. The parameter estimation for the LNR and WR models was carried out through the *survreg* function in the *survival* package [62], [63] in R, utilizing the *dist* = “lognormal” and *dist* = “weibull” options, respectively.

Model	λ_i	$P(y=y_i; \lambda_i)$	μ_i
Gamma	α_i, β_i	$\frac{1}{\Gamma(\alpha_i)\beta_i^{\alpha_i}} y_i^{\alpha_i-1} e^{-\frac{y_i}{\beta_i}}$	$\alpha_i \beta_i$
Log-normal	γ_i, σ_i	$\frac{1}{y_i \sigma \sqrt{2\pi}} \exp(-\frac{(\log(y_i) - \gamma_i)^2}{2\sigma^2})$	$\exp(\gamma_i + \frac{\sigma_i^2}{2})$
Weibull	ν (constant), ξ_i	$\frac{\nu}{\xi_i} (\frac{y_i}{\xi_i})^{\nu-1} \exp(-(\frac{y_i}{\xi_i})^{\nu-1})$	$\xi_i \Gamma(1 + \frac{1}{\nu})$

Table 1. The i th value of the parameter vector (λ_i), probability density function (pdf), and the mean vector (μ_i) for the models in the second stage of the parametric hurdle models.

The parametric hurdle models rely on the distribution described below:

$$P(y = y_i; \pi, \lambda) = \begin{cases} \pi_i, & y_i = 0, \\ (1 - \pi_i)P(y = y_i; \lambda), & y_i > 0, \end{cases} \quad (5)$$

where π_i is the conditional probability of success, stated otherwise, the probability of obtaining a structural zero in the first stage in Equation 1, λ is the parameter vector in the

second stage, and $(1 - \pi_i) \times P(y = y_i; \lambda)$ is the probability of obtaining the continuous value (y_i) in the second stage [64], [65], [66]. Table 2 displays the parametric hurdle models employed in the study, specifically, the hurdle Gamma regression (HGR) [64], [65] hurdle Log-normal regression (HLNR) [66], [67], and hurdle Weibull regression (HWR) [65], [66] models, which are derived from the distribution in Equation 5.

Model	λ_i	Stage 1 $y_i = 0$	Stage 2 $y_i > 0$
Hurdle Gamma , [64], [65]	α_i, β_i	π_i	$(1 - \pi_i) \frac{1}{\Gamma(\alpha_i)\beta_i^{\alpha_i}} y_i^{\alpha_i-1} e^{-\frac{y_i}{\beta_i}}$
Hurdle Log-normal , [65], [66]	γ_i, σ_i	π_i	$(1 - \pi_i) \frac{1}{y_i \sigma \sqrt{2\pi}} \exp(-\frac{(\log(y_i) - \gamma_i)^2}{2\sigma^2})$
Hurdle Weibull [65], [66]	ν (constant), ξ_i	π_i	$(1 - \pi_i) \frac{\nu_i}{\xi_i} (\frac{y_i}{\xi_i})^{\nu_i-1} \exp(-(\frac{y_i}{\xi_i})^{\nu_i-1})$

Table 2. The parametric hurdle models based on the distribution in Equation 5.

Data Preparation

As depicted in Figure 1, our data preparation process comprises five different subprocesses: Data introduction, data merging, data preprocessing, data description, and data splitting. In the following subsections, we provide a brief elaboration on each of these subprocesses, respectively.

Data Introduction

This study conducted a hybrid modeling that involves two different datasets: the UNALLAR Solar Power Plant (SPP) and a meteorological data obtained from the Open-Meteo website.

The UNALLAR SPP has an installed capacity of 110 kWe (129.6 kWp) and is located in the Çaycuma district of Zonguldak province. Çaycuma is located at the coordinates of 41° 25' 35.004" N and 32° 4' 32.016 E. The data contain the measurements regarding the target variable, which is the total active power (AP, kW) representing the alternating current active power transferred to the grid at the inverter output. This measure plays an important role in calculating the system performance ratio, energy production efficiency, and inverter conversion losses. Its values are typically 2-5% lower than the total direct current (DC, kW) power.

The meteorological data contain 10 different attributes. That is, temperature-2m (°C, T), relative humidity-2m (% , RH), wind speed-10m (km/h, WS), direct radiation (W/m², DR), direct normal irradiance (W/m², DNI), apparent temperature (°C, AT), terrestrial radiation (W/m², TR), wind direction-100m (°, WD), dew point-2m (°C, DP) and cloud-cover (% , CC).

Data Merging

The UNALLAR plant's monitoring system recorded the measurements for AP each 15 minutes from January 4, 2021 to January 5, 2025. The values of the attributes specified above for the meteorological data were recorded on an hourly basis from January 1, 2021 to January 1, 2025. In order to work on both datasets mentioned above simultaneously, the ultimate data we possess contain hourly records spanning from January 4, 2021 to January 1, 2025.

Data Preprocessing

Two preprocessing steps were applied prior to the analysis when dealing with the structural zeros in the data. First, in 86 cases where direct radiation (W/m^2 , DR), direct normal irradiance (W/m^2 , DNI), and terrestrial radiation (W/m^2 , TR) were zero but active power (AP, kW) took small positive values, AP was corrected to zero in accordance with the structural zero mechanism in the data. When $\text{DR} = 0$, $\text{DNI} = 0$, and $\text{TR} = 0$, the corresponding solar AP value is expected to be structurally zero. This is because such a scenario signifies an absence of solar energy, which can occur during nighttime, system outages, equipment failures, or measurement errors. Note that, since $\text{TR} = 0$ denotes absolute zero (-273.15°C), this condition cannot occur in practice. Second, 829 observations (2.4% of the data) in which $\text{DR} > 0$, $\text{DNI} > 0$, and $\text{TR} > 0$, while $\text{AP} = 0$ were removed, as they were inconsistent with the structural zero mechanism in the data. Because the lack of PV power generation ($\text{AP} = 0$) during daylight hours ($\text{DR} > 0$ and $\text{DNI} > 0$) indicates that there may have been system outages, equipment failures, or measurement errors. It should be noted that these occurrences do not signify structural zeros. Conversely, these instances indicate situations in which $\text{AP} = 0$ is observed despite the expectation of $\text{AP} > 0$. After these adjustments, the final data set consisted of 32,943 hourly observations. It is important to note that a decrease in the accuracy of the classification methods (BLR, RFC, and SVC) is observed when these preprocessing steps are omitted. The implementation of these preprocessing steps is also crucial for aligning the data with the concept of structural zeros. This also acts as tangible evidence demonstrating the importance of DR, DNI, and TR in predicting solar PV power. Figure 2 displays the algorithmic description of the preprocessing rules when dealing with the structural zeros in the PV power forecasting data.

It should be emphasized that there is no standardization involved in the two-stage hurdle modeling procedure. Note that the GR, LNR, and WR parametric models necessitate the target variable AP to be positive. In this context, the traditional standardization approach is not suitable for these models, as it produces negative values. Additionally, since RFR and XGBoost are based on decision trees, they generally do not require standardization. Although a standardization procedure for the geometry-based SVR technique (and RFR and

XGBoost) was contemplated, it has been observed that this approach significantly distorts the analysis results based on the four evaluation metrics employed in this study. All results regarding the analysis of the unstandardized data can be duplicated by utilizing `set.seed(111)` in *R* statistical software.

Data Description

Descriptive statistics of the processed data are presented in Table 3. Skewness coefficients demonstrate that all variables exhibit departures from zero, indicating that none of the distributions can be characterized as symmetric. The attributes T ($S = 0.07$), WS ($S = 0.99$), DR ($S = 1.72$), DNI ($S = 1.04$), AT ($S = 0.09$), TR ($S = 0.84$), and the target variable AP ($S = 1.54$) display positively (right) skewed distributions, typically associated with elongated or heavy right tails. Conversely, RH ($S = -0.59$), WD ($S = -0.53$), DP ($S = -0.23$), and CC ($S = -0.05$) exhibit negatively (left) skewed distributions, in which the tail extends toward smaller values and most of the observations are concentrated at higher values. The attributes WS ($K = 4.74$), DR ($K = 4.93$), and AP ($K = 4.22$) have kurtosis values above three, reflecting leptokurtic distributions with sharper peaks and heavier tails, suggesting a higher propensity for extreme observations. All the other variables in the data exhibit kurtosis values below three, indicating platykurtic distributions characterized by flatter peaks and lighter tails, consistent with more stable data and fewer extreme values.

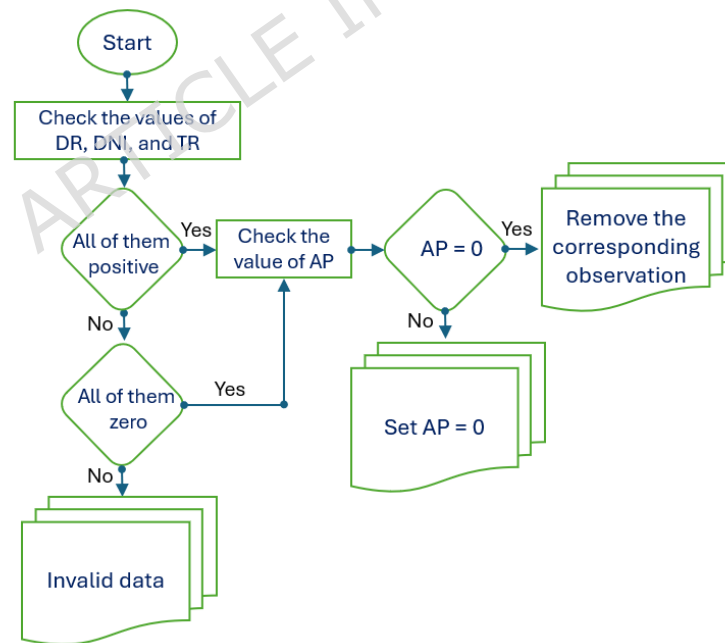


Fig. 2. Algorithmic description of the preprocessing rules when dealing with the structural zeros in the PV power forecasting data.

	Mean	Standard deviation	Median	Skewness (S)	Kurtosis (K)	SE	Coefficient of variation (CV)
T	14.98	7.85	14.80	0.07	2.39	0.04	0.52
RH	77.12	16.40	80.00	-0.59	2.36	0.09	0.21
WS	7.43	4.13	6.60	0.99	4.74	0.02	0.56
DR	119.40	199.89	0.00	1.72	4.93	1.10	1.67
DNI	205.98	288.52	0.00	1.04	2.49	1.59	1.40
AT	14.56	9.80	14.00	0.09	2.26	0.05	0.67
TR	332.46	408.81	57.20	0.84	2.24	2.25	1.23
WD	209.76	102.26	209.00	-0.53	2.53	0.56	0.49
DP	10.56	6.66	10.60	-0.23	2.35	0.04	0.63
CC	52.20	42.64	53.00	-0.05	1.24	0.23	0.82
AP	16.33	25.35	0.26	1.54	4.22	0.14	1.55

Table 3. Descriptives of the processed data.

The coefficient of variation (CV) is a standardized measure used to compare the variability of distributions expressed in different units. Values of $CV < 1$ indicate low variability, whereas $CV > 1$ suggests high variability. Moreover, higher CV values further imply that the observations are more widely dispersed around the mean. The greatest values of CV were observed in DR ($CV = 1.67$), AP ($CV = 1.55$), DNI ($CV = 1.40$), and CC ($CV = 0.82$), respectively, while RH ($CV = 0.21$) displayed the lowest degree of variability.

Figure 3 displays the pairwise relationships between the attributes and target variable. This figure indicates that only RH exhibits a negative correlation with AP, while all the other variables in the data demonstrate positive correlations with AP. Figure 4 depicts the Pearson correlation coefficients between the attributes and the target variable AP. This figure shows that, in line with the plot in Figure 3, AP is positively correlated with DR ($r = 0.90$, $p < 0.001$), DNI ($r = 0.84$, $p < 0.001$), and TR ($r = 0.86$, $p < 0.001$) and negatively correlated with RH ($r = -0.66$, $p < 0.001$). There is a weak but statistically significant positive correlation between AP and WD ($r = 0.03$, $p < 0.001$). Note that since there is an almost perfect correlation between T and AT ($r = 0.99$, $p < 0.001$), we utilize T (but not AT) to avoid a potential overfitting problem in our hurdle modeling.

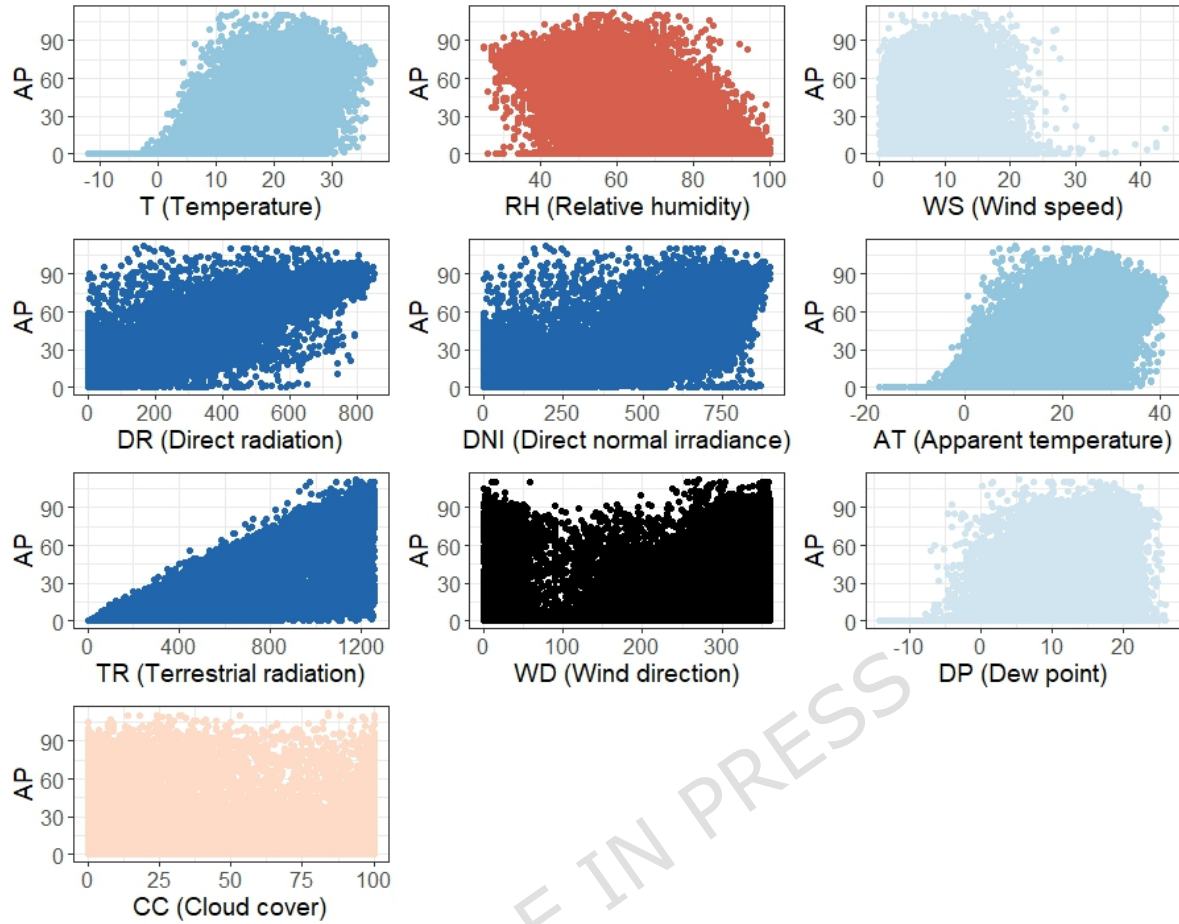


Fig. 3. The pairwise relationships between the independent variables and the dependent variable active power (AP) in the processed data.

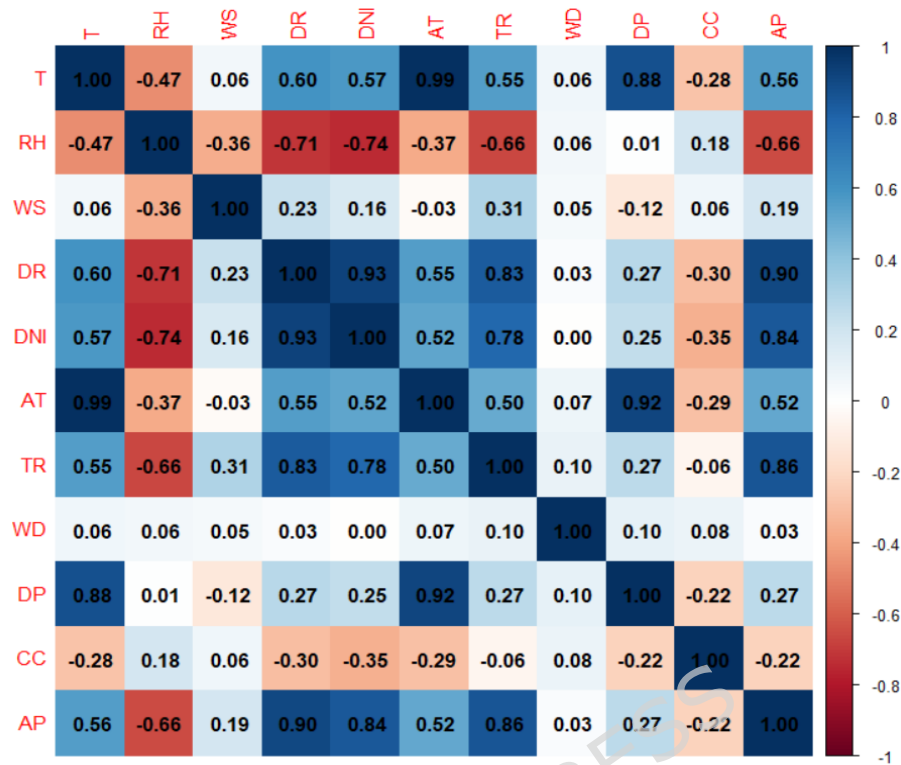


Fig. 4. The correlation plot between the variables in the processed data.

Data Splitting

To prevent look-ahead bias or data leakage issues, we do not utilize random splitting when forming the training and testing sets. Instead, we utilize the initial 80% of the data for the training set, reserving the remaining 20% for the testing set. The training sample contains 26,354 observations, while the testing set consists of 6,589 observations. The training set observations are utilized to develop the hurdle models addressing both the modeling of structural zeros and the positive values present in the processed data. The observations in the testing set are employed to obtain the combined fitted values for the hurdle models, which are subsequently used to evaluate the predictive performance of these models.

Application of Hybrid Modeling in the Training Set

We utilize the model in Equation 1 to train the zero component of the hurdle model. The training performance of the zero component is evaluated based on its ability to identify the structural zeros present in the training set in terms of prediction probabilities and confusion matrix. We employ each of the non-parametric ML techniques (i.e., RFR, XGBoost, and SVR) and the parametric models (i.e., GR, LNR, and WR) at the secondary phase of the hurdle modeling in the training set. The performance of the component for positive values is examined by analyzing the relationship between the fitted values and the target variable AP.

Stage 1: Modeling Structural Zeros in the Training Set

The initial phase of hurdle modeling involves incorporating structural zeros into the modeling process. The appropriate integrations of BLR, RFC, and SVC into the analytical process is essential for examining the training set, given that 49.43% of the values for the target variable AP are structural zeros (i.e., 13,027 structural zeros out of a total of 26,354 cases). The LR model in Equation 1 fails to converge when analyzing the training set. The model also suffers from the separation problem [44] which causes detrimental effects on the standard errors of the regression coefficients for DR (SE = 45.83) and DNI (SE = 43.01). We address the issues of non-convergence and separation simultaneously by utilizing BLR, RFC, and SVC, respectively.

The BLR model provides the (conditional) probability of success for each observation in the data depending on the positions of the structural zeros (i.e., 0: not a structural zero and 1: structural zero). Consequently, the model is expected to produce high success probabilities for the cases of structural zeros, specifically for ones. In the *rstanarm* package [49], the BLR model adopts a chain-based strategy (by default 4 chains) involving to produce posterior samples through the Bayesian Markov Chain Monte Carlo (MCMC) algorithm. Table 4 displays the training times for the warm-up and sample procedures associated with each chain.

	Warm-up	Sample
Chain 1	301.500sec	336.884sec
Chain 2	316.605sec	306.546sec
Chain 3	286.203sec	324.676sec
Chain 4	312.686sec	293.138sec

Table 4. Training times for the warm-up and sample procedures associated with each of the four chains in BLR.

We implement the Bayesian MCMC using $N(0, 2.5)$ prior distribution for regression coefficients which is the default option in *rstanarm* package [49]. Altinisik [67] performs a comparative analysis that involves the application of various prior distributions to address the separation problem.

The BLR used to model the structural zeros in the training set is defined as:

$$\text{logit}(\pi_i) = \theta_0 + \theta_1 T_i + \theta_2 RH_i + \theta_3 WS_i + \theta_4 DR_i + \theta_5 DNI_i + \theta_6 TR_i + \theta_7 WD_i + \theta_8 DP_i + \theta_9 CC_i + \theta_{10} t_i + \theta_{11} \sin\left(\frac{2\pi t_i}{P}\right) + \theta_{12} \cos\left(\frac{2\pi t_i}{P}\right), (6)$$

where θ_0 is the intercept term, $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8$, and θ_9 are the regression coefficients of variables T, RH, WS, DR, DNI, TR, WD, DP, and CC, respectively, for $i = 1, 2,$

..., n . Here, t_i represents the linear time-trend component (i.e., $t_i = i$ for $i = 1, 2, \dots, n$) and θ_{10} is the corresponding regression coefficient that indicates how the target variable AP varies over time, assuming that the other variables in the model are constant. Similarly, $\sin(\frac{2\pi t_i}{P})$ and $\cos(\frac{2\pi t_i}{P})$ are the seasonal components of the model, where $P = 4$ is utilized to address seasonality, given that the data were collected on an annual basis. Therefore, θ_{11} and θ_{12} are used to capture periodic fluctuations in AP over time.

The relationship between the observed structural zeros and the fitted values (predicted probabilities) of the BLR model above is depicted in Figure 5. The orange dots in the figure indicate the occurrences of structural zeros, namely, ones. The green dots represent the predicted probabilities of observing structural zeros, which are obtained by fitting the model in Equation 6. The figure reveals that the BLR model predominantly provides high predicted values for the structural zeros. This means that the zero component of our hurdle model is proficient in modeling the structural zeros in the training set.

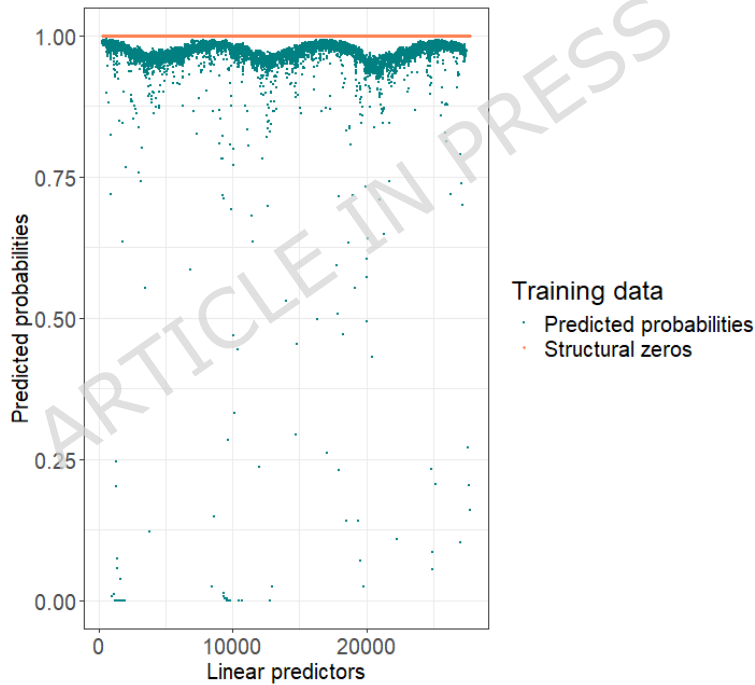


Fig. 5. The relationship between the structural zeros in the data and predicted probabilities of the BLR model.

For both RFC and SVC, the target variable AP along with the identical independent variables from Equation 6 were utilized. Unlike BLR, RFC and SVC do not utilize regression-based methodologies. While RFC performs classification using entropy-based information theory, SVC conducts classification by examining the degree to which multivariate observations deviate from a decision boundary in a geometric space. Therefore, RFC and SVC directly

produce class labels without operating on a probability basis. The duration of training for RFC (i.e., 20.40sec) and SVC (i.e., 35.21sec) is significantly less than that of BLR (see Table 4).

Table 5 displays the confusion matrices used to compare the performances of BLR, RFC, and SVC based on their correct classification rates (CCR). Since BLR is a regression-based approach, the relevant confusion matrix is constructed by designating predicted values greater than 0.5 as one, and those that are 0.5 or lower as zero. There is no cut-off value involved in the construction of the confusion matrices for RFC and SVC. Table 5 shows that RFC exhibits the highest level of accuracy according to CCR, specifically, $CCR = \frac{13,219+12,928}{13,219+99+108+12,928} = 0.992$. The lower and upper bounds of the 95% confidence interval of the CCR for RFC are obtained as 0.991 and 0.993, respectively. The Kappa coefficient for RFC is calculated as $K = 0.984$, signifying a strong model for predicting the structural zeros within the training set. Similar to RFC, high levels of accuracy are obtained for BLR and SVC, with $CCR = 0.980$ for both. For both BLR and SVC, the corresponding 95% confidence intervals are obtained as [0.978, 0.982]. High values of the Kappa coefficient are also observed for BLR and SVC, with $K = 0.960$ for both.

	Predicted values	Status	
		0: Not a structural zero	1: Structural zero
BLR	0	12,369	64
	1	458	12,963
RFC	0	13,219	99
	1	108	12,928
SVC	0	12,853	56
	1	474	12,971

Table 5. The confusion matrices between the actual and predicted values based on BLR, RFC, and SVC for the training set.

In summary, the results indicate that each of the BLR, RFC, and SVC techniques exhibits satisfactory performance in terms of modeling the structural zeros present in the PV power forecasting data. The primary disadvantage of the BLR technique is considered to be its significantly longer execution time compared to RFC and SVC. Additionally, as mentioned, the confusion matrix for BLR was established under the weakly informative $N(0, 2.5)$ prior distribution and utilized the default cut-off value of 0.5. The sensitivity of the results derived from BLR concerning the choice of prior distribution and cut-off value was not examined, as this falls outside the scope of this study. The fact that the RFC technique provides the highest accuracy in terms of CCR also deters us from undertaking such a comprehensive sensitivity analysis. Note that RFC and SVC utilize identical values of tuned hyperparameters as RFR and SVR, which are elaborated in the next subsection.

Stage 2: Modeling Positive Values in the Training Set

As previously stated, we employ the non-parametric ML techniques RFR, XGBoost, and SVR and the parametric models GR, LNR, and WR for modeling positive values in the data. The same attributes in BLR, RFC, and SVC from the preliminary phase are employed in the execution of RFR, XGBoost, and SVR for modeling positive values. Figure 6 displays the relationship between the fitted values obtained using RFR and the values of the target variable AP. There appears to be a positive linear relationship between the fitted values and AP, suggesting that RFR is effective in predicting the values of AP. Figure 7 shows the relationship between the linear predictors and residuals (the differences between the values of AP and the corresponding fitted values) using RFR. This figure shows that the residuals fluctuate around zero, which implies that there is no systematic error associated with the use of RFR for predicting AP.

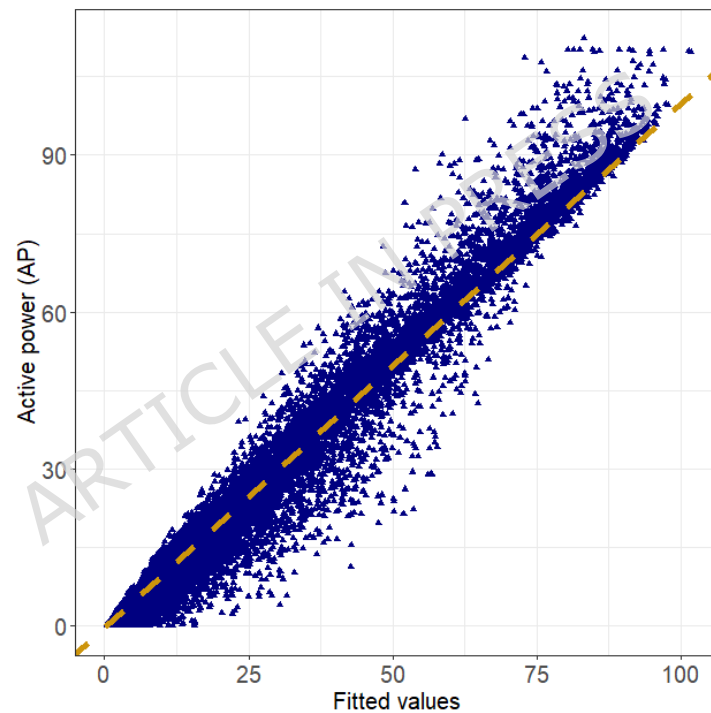


Fig. 6. The relationship between the fitted values derived from RFR and the values of AP.

Figure 8 depicts the relationship between the linear predictors and the values of AP using RFR. This illustration suggests that RFR is effective in capturing the seasonal impact of the attributes on AP. Figures 6, 7, and 8 are also obtained by using XGBoost, SVR, GR, LNR, and WR, which are not included here but can be provided upon request.

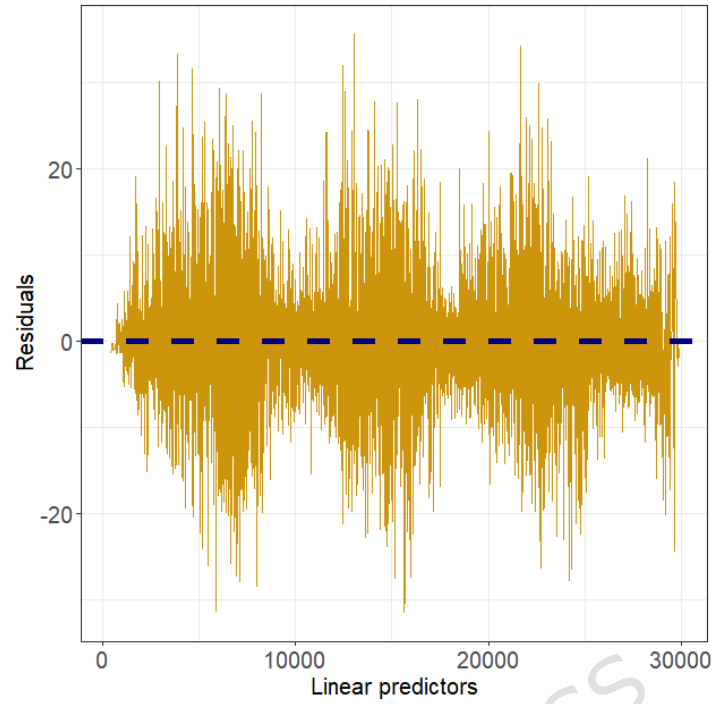


Fig. 7. The relationship between the linear predictors and residuals in RFR.

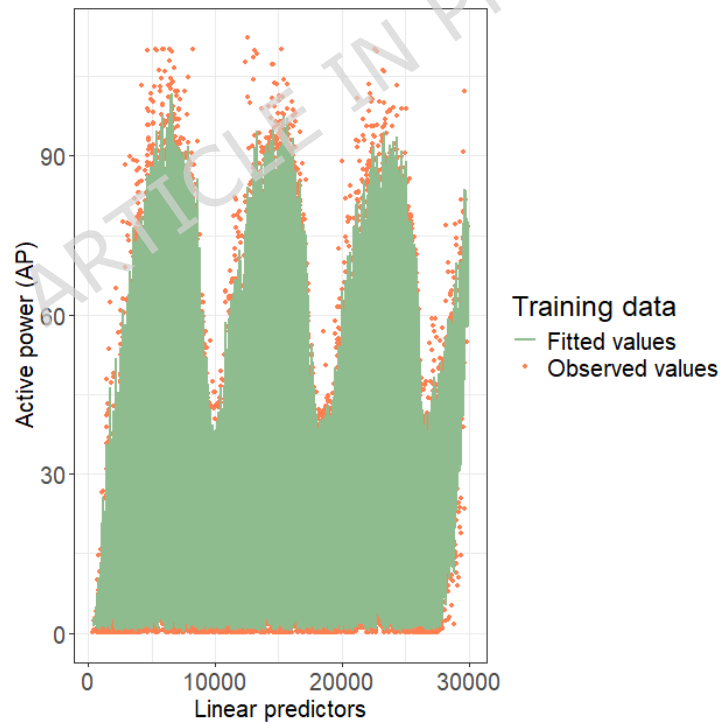


Fig. 8. The relationship between the linear predictors and the values of AP using RFR.

Table 6 shows the information pertaining to RFR, XGBoost, and SVR, including their corresponding tuned hyperparameters. The table additionally displays the system time

necessary to execute the *R* code for each ML technique. It appears that XGBoost and SVR provide analysis results that are considerably quicker than those produced by RFR. In RFR, the mean squared residuals (MSRs) are derived from the out-of-bag (OOB) observations, which refers to the data excluded from the bootstrap sample of the training set for each decision tree (with *ntree* = 1000). Figure 9 shows how the MSRs vary in relation to the number of decision trees in RFR. This graph illustrates that the MSRs stabilize significantly, particularly after achieving a tree count of 250-300. Thus, the number of trees utilized (i.e., *ntree* = 1000) is quite sufficient. The minimum value of the MSRs recorded is 144.738 (refer to the blue dashed line in Figure 9), which serves as an estimate for the MSR in the testing set. In addition, the rapid decrease in the MSRs demonstrates that RFR executes an effective learning in the training set.

Random Forest Regression (RFR)	
☐	type: Regression
☐	R package: randomForest
☐	ntree: 1000 (The number of decision trees)
☐	mtry: 3 (The number of attributes to be randomly selected at the split points)
☐	nodesize: 5 (The smallest number of observations that can be contained in a node)
☐	system time: user (130.36sec), system (0.33sec), elapsed (130.90sec)
Extreme Gradient Boosting (XGBoost)	
☐	booster: gbtrees (Decision tree-based approach)
☐	R package: xgboost
☐	nrounds: 300 (The number of decision trees)
☐	eta: 0.01 (The proportion of errors rectified in the preceding tree in the system)
☐	max_depth: 5 (Maximum depth of decision trees)
☐	subsample: 0.7 (The randomly selected portion of the training set in each decision tree within the system)
☐	colsample_bytree: 0.7 (The randomly selected portion of the attributes in the training set)
☐	system time: user (4.33sec), system (3.67sec), elapsed (1.86sec)
Support Vector Regression (SVR)	
☐	type: eps-regression (Epsilon insensitive based approach)
☐	R package: e1071
☐	kernel: radial (Gaussian kernel function is used when determining the values in the kernel matrix)

- cost: 1 (A medium size of the cost term is chosen for the errors, which serves as the default option)
- epsilon: 0.5 (The chosen maximum magnitude of the errors that are ignored)
- system time: user (4.93sec), system (0.01sec), elapsed (4.94sec)

Table 6. Details regarding RFR, XGBoost, and SVR along with their respective tuned hyperparameters.

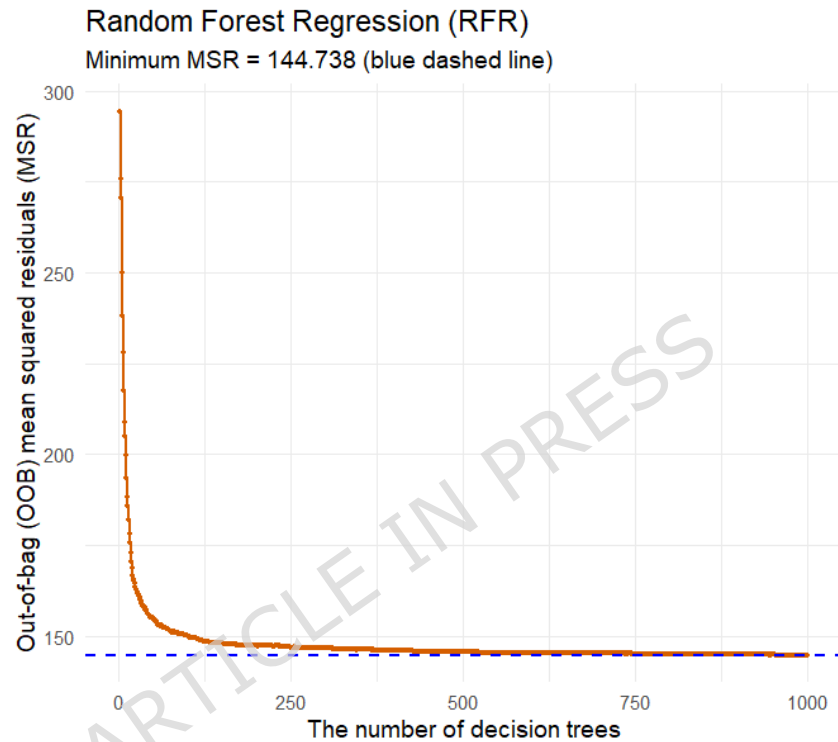


Fig. 9. Variation of out-of-bag (OOB) mean squared residuals (MSRs) in relation to the number of trees in RFR.

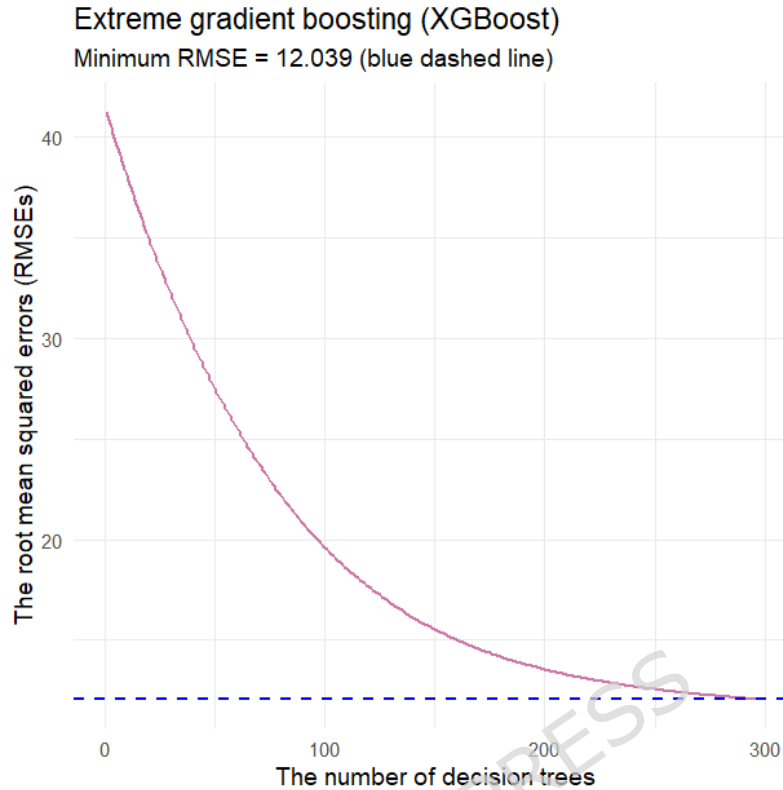


Fig. 10. The values of root mean squared errors (RMSEs) in relation to the number of trees in XGBoost.

Figure 10 illustrates a similar plot containing the values of the root mean squared errors (RMSEs) for each decision tree in the context of XGBoost (with nrounds = 300). This figure demonstrates that the RMSEs stabilize across the updated decision trees, indicating that a total of 300 decision trees is adequate. The minimum value of RMSEs for XGBoost is obtained as 12.039 (refer to the blue dashed line in Figure 10). Table 7 shows the values of the importance measures for the attributes in the training set using RFR, XGBoost, and SVR, respectively. The technique we employ in RFR randomly shuffles the values of each attribute and utilize the percentage increase in the out-of-bag MSE as a measure of the importance of attributes in the training set. In a similar manner, the gain in XGBoost indicates the extent to which the attributes reduce the value of the loss function, which is the MSE in our context. Permutation feature importance is utilized in SVR to assess the contribution of each attribute to the predictive performance of the model in the training set. Table 7 shows that the metrics employed in RFR, XGBoost, and SVR exhibit consistent results, indicating that DR, DNI, and TR are the most important attributes in predicting the positive values of AP. These features play a significant role not only in modeling the positive values of AP but also in modeling the structural zeros that are associated with AP. In this regard, the importance

of these attributes for PV power forecasting was elaborated in the subsection “Data Preprocessing”.

It is important to note that the similarity matrix employed in SVR and the decision trees utilized in RFR and XGBoost possess the ability to capture the interactions among the attributes. Consequently, no interaction terms are incorporated in the non-parametric techniques. However, the parametric models GR, LNR, and WR lack this capability. Therefore, in addition to the attributes in Equation 6, we use a set of interaction terms in our parametric models. By undertaking this approach, our objective is to enhance the performances of our parametric models within both the training and testing sets.

Attributes	RFR: Percentage increase in MSE	XGBoost: Gain	SVR: Permutation feature importance
T	75.869	0.015	1.090
RH	37.101	0.020	1.065
WS	21.153	0.016	1.082
DR	429.523	0.547	1.444
DNI	168.336	0.124	1.133
TR	334.781	0.210	1.839
WD	16.573	0.012	1.040
DP	27.877	0.006	1.071
CC	37.929	0.013	1.059
t	18.742	0.008	1.029
$\sin(\frac{2\pi t}{P})$	24.151	0.007	1.040
$\cos(\frac{2\pi t}{P})$	55.456	0.020	1.114

Table 7. The values of importance measures for the attributes in the context of RFR, XGBoost, and SVR.

In this regard, six interaction terms have been incorporated into each parametric model, which are $T \times DNI$, $RH \times DNI$, $T \times DR$, $RH \times DR$, $T \times WS$, and $RH \times WS$. Table 8 displays the maximum likelihood estimates (with the corresponding values of the test statistics between brackets), the (minus) log likelihood (-LL), the Akaike information criterion (AIC) and Bayesian information criterion (BIC) values obtained by GR, LNR, and WR, respectively. Each of the parametric models in Table 8 has a training time of less than one second. According to the findings shown in Table 8, the parametric model that most accurately represents the training set for positive values is identified as WR. Because this model produces smaller minus log likelihood, AIC, and BIC values compared to the other two models. Nevertheless, this conclusion is derived based on classical statistical inference on only positive values. Performance comparisons of both the parametric and non-parametric

hurdle models for predicting the values of the target variable AP are carried out using a series of metrics on the testing set, which will be discussed in the next subsection.

	Gamma regression		Log-normal regression		Weibull Regression	
Effects	β (t-value)	SE	β (z-value)	SE	β (z-value)	SE
(Intercept)	0.911** (2.606)	0.350	-0.727 (-1.432)	0.508	1.434*** (4.146)	0.346
T	-0.001 (-0.018)	0.017	0.020 (0.815)	0.025	-0.012 (-0.688)	0.017
RH	0.001 (0.105)	0.004	0.009 (1.784)	0.005	-0.003 (-0.766)	0.004
WS	0.093*** (10.568)	0.009	0.133*** (10.472)	0.013	0.080*** (9.581)	0.008
DR	-0.001** (-2.919)	0.001	-0.002** (-2.824)	0.001	-0.001 (-2.313)	0.001
DNI	0.001 (1.798)	0.001	0.001 (1.254)	0.001	0.001 (1.635)	0.001
TR	0.003*** (91.233)	0.001	0.003*** (73.909)	0.001	0.003*** (77.560)	0.001
WD	-0.001*** (-5.331)	0.001	-0.001*** (-4.584)	0.001	-0.001*** (-4.405)	0.001
DP	-0.005 (-0.259)	0.018	-0.013 (-0.529)	0.025	0.007 (0.382)	0.017
CC	0.001 (0.785)	0.001	-0.001 (-0.798)	0.001	0.001 (1.516)	0.001
t	-0.001*** (-3.809)	0.001	0.001 (0.419)	0.001	-0.001*** (-4.599)	0.001
$\sin(\frac{2\pi t}{P})$	-0.057*** (-5.682)	0.010	-0.023 (-1.567)	0.014	-0.055*** (-5.604)	0.010
$\cos(\frac{2\pi t}{P})$	-0.162*** (-9.930)	0.016	-0.079*** (-3.353)	0.024	-0.167*** (-10.496)	0.016
T $\times$ DNI	-0.001** (-2.770)	0.001	0.001** (3.068)	0.001	-0.001*** (-3.588)	0.001
RH $\times$ DNI	0.001*** (15.524)	0.001	0.001*** (8.829)	0.001	0.001*** (11.701)	0.001
T $\times$ DR	0.001*** (3.729)	0.001	0.001 (0.123)	0.001	0.001*** (3.867)	0.001
RH $\times$ DR	-0.001*** (-6.587)	0.001	-0.001*** (-4.519)	0.001	-0.001*** (-6.569)	0.001
T $\times$ WS	-0.001*** (-7.298)	0.001	-0.002*** (-6.541)	0.001	-0.001*** (-7.077)	0.001
RH $\times$ WS	-0.001*** (-14.229)	0.001	-0.002*** (-13.316)	0.001	-0.001*** (-13.218)	0.001
Statistics						
-LL	52,113.0		54,396.3		51,662.7	
AIC	104,265.9		108,832.7		103,365.4	

BIC	104,415.9	108,982.6	103,515.4
-----	-----------	-----------	-----------

Table 8. The maximum likelihood estimates (with the corresponding values of the test statistics between brackets), the (minus) log likelihood, and AIC and BIC values obtained by GR, LNR, and WR. Note: * : p -value < 0.05, **: p -value < 0.01, and ***: p -value < 0.001.

Fitted Values and Performance Evaluation in the Testing Set

We obtain the fitted values, and consequently, evaluate the predictive performance of each hurdle model under consideration using the testing set ($n = 6,589$). As mentioned, the zero component of the hurdle models was modeled using BLR, RFC, and SVC. The positive component was examined using the parametric models GR, LNR, and WR and the non-parametric ML techniques RFR, XGBoost, and SVR. Notably, this yields $3 \times 6 = 18$ distinct hurdle models. In the following two subsections, we provide a detailed explanation of how to derive the fitted values for each of the 18 hurdle models and evaluate their predictive performance, respectively.

Obtaining Fitted Values for Hurdle Models

To derive the fitted values for the hurdle models, we initially acquire those for BLR, RFC, and SVC when modeling the structural zeros in the testing set, that is, $(1 - \hat{\pi}_i)$ for $i = 1, 2, \dots, 6,589$. Then, we obtain the fitted values using RFR, XGBoost, SVR, GR, LNR, and WR when modeling positive values in the testing set, that is, $\hat{y}_i = E(y_i | y_i > 0)$ for $i = 1, 2, \dots, 6,589$. These two types of fitted values are integrated within the framework of hurdle models by multiplying the corresponding values, that is, $(1 - \hat{\pi}_i) \times \hat{y}_i$ for $i = 1, 2, \dots, 6,589$. We perform comparisons of the hurdle models regarding their ability to predict the values of the target variable AP by utilizing four different metrics, each of which employs the integrated fitted values above.

Predictive Performance Evaluation

The effectiveness of the BLR, RFC, and SVC techniques employed to model structural zeros was evaluated using the confusion matrices on the testing set. Table 9 displays the confusion matrices between the actual and predicted values based on BLR, RFC, and SVC for the testing set. This table indicates that each of the BLR, RFC, and SVC techniques demonstrates satisfactory performance in the testing set according to the CCR. Similar to the training set results in Table 5, RFC produces slightly higher accuracy (CCR = 0.992) compared to BLR (CCR = 0.981) and SVC (CCR = 0.970). The Kappa coefficient for RFC is determined to be $K = 0.985$, indicating a robust model for forecasting the structural zeros in the testing set.

	Predicted	Status
--	-----------	--------

	values	0: Not a structural zero	1: Structural zero
BLR	0	3,398	14
	1	109	3,068
RFC	0	3,478	21
	1	29	3,061
SVC	0	3,319	9
	1	188	3,073

Table 9. The confusion matrices between the actual and predicted values based on BLR, RFC, and SVC for the testing set.

The list of the four metrics used to assess the predictive performance of the 18 hurdle models is as follows: mean squared error (MSE), mean absolute error (MAE), mean absolute scaled error (MASE), and the coefficient of determination (R^2). The evaluation metrics are defined as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

$$\text{MASE} = \frac{\text{MAE}}{\frac{1}{n-1} \sum_{i=2}^n |y_i - y_{i-1}|},$$

$$R^2 = \frac{[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}, \quad (7)$$

where y_i and $\hat{y}_i$ denote the observed and fitted values for the testing data with the corresponding means $\bar{y}$ and $\bar{\hat{y}}$, respectively. Since the AP is measured in kW, MAE and RMSE are expressed in kW, whereas MASE and R^2 are unit-free due to scaling.

Table 10 displays the values of these metrics for a set of single-stage models (RFR, XGBoost, SVR, and LSTM) and each hurdle model under consideration. Note that the single-stage GR, LNR, and WR cannot be utilized in this context, as these models are designed to operate exclusively with positive values (and not with zeros). The findings concerning RFR versus its two-stage hybrid counterparts (BLR-RFR, RFC-RFR, and SVC-RFR) and XGboost versus its counterparts (BLR-XGBoost, RFC-XGBoost, and SVC-XGBoost) were notably similar, whereas the two-stage BLR-SVR, RFC-SVR, and SVC-SVR exhibited enhanced performance over SVR across all predictive performance metrics. Since LSTM is sensitive to the scale of the data, it was employed on the (min-max) standardized data. It should be noted that standardization also alters the scale of the some of the performance metrics above. Therefore, the comparison of LSTM with the other techniques was conducted following a back-transformation procedure. A sigmoid function was employed within LSTM to restrict outputs to non-negative values in line with the characteristics of the data. Since the training

data contain a sufficient number of observations, a look-back window of 100 timesteps was employed. Further details regarding the single-stage LSTM procedure will be provided upon request.

Based on the results presented in Table 10, the best predictive performances were achieved when the positive component was modeled using either RFR or XGBoost. In this sense, for example, BLR-RFR (MSE = 67.425 kW, MAE = 3.974 kW, MASE = 0.615, $R^2 = 0.899$) and BLR-XGBoost (MSE = 68.815 kW, MAE = 4.052 kW, MASE = 0.627, $R^2 = 0.899$) performed better than BLR-SVR (MSE = 96.191 kW, MAE = 5.864 kW, MASE = 0.907, $R^2 = 0.872$). It appears that the hurdle models with the ML techniques perform better than that with the parametric models. All the hurdle models exhibit satisfying performance in terms of explaining the variance of the target variable AP. It is important to note that even the least effective model RFC-LNR manages to account for more than 86% (with $R^2 = 0.861$) of the variance in AP. This percentage reaches almost 90% (with $R^2 = 0.899$) for BLR-RFR and RFC-RFR.

	MSE	MAE	MASE	R^2
RFR	70.733	3.865	0.223	0.889
XGBoost	69.595	4.091	0.633	0.898
SVR	147.021	10.486	1.622	0.870
LSTM	73.959	4.442	0.685	0.893
BLR-RFR	67.425	3.974	0.615	0.899
BLR-XGBoost	68.815	4.052	0.627	0.899
BLR-SVR	96.191	5.864	0.907	0.872
BLR-GR	80.247	4.415	0.683	0.882
BLR-LNR	98.194	4.997	0.773	0.862
BLR-WR	82.526	4.444	0.687	0.881
RFC-RFR	68.416	4.094	0.633	0.899
RFC-XGBoost	72.028	4.281	0.662	0.898
RFC-SVR	102.899	6.443	0.996	0.872
RFC-GR	76.685	4.461	0.690	0.882
RFC-LNR	95.715	4.967	0.768	0.861
RFC-WR	79.565	4.470	0.691	0.881
SVC-RFR	67.542	3.969	0.614	0.898
SVC-XGBoost	69.252	4.075	0.630	0.898
SVC-SVR	96.511	5.805	0.898	0.871
SVC-GR	80.038	4.401	0.681	0.882
SVC-LNR	97.656	4.981	0.770	0.862
SVC-WR	82.357	4.428	0.685	0.881

Table 10. The values of the predictive performance metrics for the 18 hurdle models under consideration.

Discussion, Conclusion, and Limitation

This study presents an analysis of PV power forecasting data with excessive structural zeros using a hybrid modeling strategy. The analysis divided the data into two parts: a training set (the first 80% of the data) and a testing set (the remaining 20% of the data). The study examined both the training and testing sets using two-stage hurdle models. In the first stage, BLR, RFC, and SVC were employed to model the structural zeros in the data. In the second stage, the remaining positive values were analyzed using the parametric models GR, LNR, and WR, as well as the non-parametric ML techniques RFR, XGBoost, and SVR.

The training and testing performances of BLR, RFC, and SVC were evaluated according to their ability to identify structural zeros. In the training set, although each of these models exhibited a high level of accuracy in terms of the correct classification rate (CCR), RFC (CCR = 0.992) performed slightly better than BLR and SVC (both with CCR = 0.980). The execution times for RFC (20.40sec) and SVC (35.21sec) were significantly shorter than that for BLR (41.30min) (see Table 4). In the testing set, similar results were achieved with respect to the level of accuracy. Namely, RFC (CCR = 0.992) provided a slightly higher accuracy when compared to BLR (CCR = 0.981) and SVC (CCR = 0.970). The Kappa coefficients (K) calculated from the confusion matrices of the actual and predicted values in Table 5 (training set) and Table 9 (testing set) indicated that each technique is highly effective for predicting the structural zeros in the PV power forecasting data. In this sense, for example, RFC produced high Kappa coefficient values, specifically, $K = 0.984$ and $K = 0.985$ for the training and testing sets, respectively. When analyzing the positive values in the training set, the metrics employed in RFR, XGBoost, and SVR exhibited consistent results, indicating that DR, DNI, and TR are the most important attributes in predicting the values of AP (see Table 7). The WR model was chosen as the best model among the parametric models when modeling the positive values in the training set, as it produced the smallest -LL (51,662.7), AIC (103,365.4), and BIC (103,515.4) values compared to GR and LNR.

The hurdle modeling strategy achieved the highest predictive accuracy when the positive component was modeled using either RFR or XGBoost (see Table 10). In this regard, BLR-RFR (MSE = 67.425 kW, MAE = 3.974 kW, MASE = 0.615, $R^2 = 0.899$) and BLR-XGBoost (MSE = 68.815 kW, MAE = 4.052 kW, MASE = 0.627, $R^2 = 0.899$) demonstrated superior performance compared to BLR-SVR (MSE = 96.191 kW, MAE = 5.864 kW, MASE = 0.907, $R^2 = 0.872$). Additionally, the hurdle models using the ML techniques RFR, XGBoost, and SVR for modeling positive values in the data outperformed using the parametric alternatives GR, LNR, and WR. All hurdle models exhibited satisfactory performance in explaining the variance of the target variable AP. It is important to note that even the least effective RFC-LNR model managed to account 86% of the variance in AP. This rate reached to almost 90% when the positive component was modeled using RFR or XGBoost regardless of the technique used for modeling the structural zeros in the data. This rate almost reached 90%

when the positive component was modeled using either RFR or XGBoost, regardless of the technique employed to model the structural zeros in the data.

The effectiveness of RFR and XGBoost in the single-stage settings highlights their potential for two-stage hurdle modeling. This is primarily attributed to their capacities to capture intricate non-linear relationships among the variables in the presence of outliers when compared to their parametric alternatives. Note that, the structural zeros are *not* modeled separately from the positive values in the single-stage settings. Therefore, these structural zeros render many positive values of the target variable AP as outliers when using the single-stage techniques. In the literature, both RFR and XGBoost are considered as robust techniques when dealing with non-linearity [68], [69], [70], [71] and outliers [72], [73], [74] in the data. Conversely, the parametric models GR, LNR, and WR are sensitive to outliers due to their functional characteristics [75], [76], [77], [78]. Additionally, these models are not defined for non-positive values, and do not capture the non-linear relationships between the variables in the data. Consequently, hurdle models incorporating RFR and XGBoost can be asserted to produce superior modeling performance compared to those utilizing GR, LNR, and WR. It is important to note that while both RFR and XGBoost employ decision trees for modeling the target variable AP there exist notable differences in their implementations. While RFR models the data by capturing the underlying relationships between the variables through independent decision trees, XGBoost constructs decision trees in a sequential manner, treating them as interdependent models. Moreover, the predictive performance of SVR is directly related to the capacity of the kernel function employed for capturing the intricate relationships among the variables in the data. The hurdle models containing SVR for modeling positive component with a Gaussian kernel exhibited inferior predictive performance relative to RFR and XGBoost (see Table 10).

Despite its promising results, the proposed framework has several limitations that should be acknowledged. Below, we address four main limitations that we deem significant. One of the key limitations of our investigation is that the analysis is based on the data from a single photovoltaic plant, which may limit the generalizability of the findings to other geographical regions, climatic conditions, or system configurations. Another important limitation is that hybrid models, which are part of the two-stage model family, usually demand more observations than the traditional parametric models or ML techniques. Therefore, employing these models with small samples may lead to overfitting problem. Third limitation is that hybrid models exhibit greater memory consumption and longer training time when compared to the classical single-stage models. The convergence of these models requires more time compared to classical techniques, particularly when dealing with large datasets. Last but not least, the three classification techniques implemented in this study (BLR, RFC, and SVC) produced comparable and satisfactory predictive

performances when modeling the structural zeros in the data. As mentioned, we set the tuned hyperparameters for RFC and SVC identically to those of RFR and SVR, respectively (see Table 6). In this regard, the use of different hyperparameters may yield different results. Additionally, since BLR produced results comparable to those of RFC and SVC, and furthermore, its execution time was considerably longer than that of the other techniques, the robustness of the findings in relation to the choice of the cut-off value or prior selection was not investigated. In this regard, undertaking sensitivity analysis for both BLR and the ML techniques RFC, SVC, RFR, and XGBoost for PV power forecasting might be seen as a potential subject for future research.

Funding

No funding was received to assist with preparing this manuscript.

References

- [1] Li, M., Wang, W., He, Y. & Wang, Q. Deep learning model for short-term photovoltaic power forecasting based on variational mode decomposition and similar day clustering. *Computers and Electrical Engineering*, **115**, 109116; (2024).
- [2] Bayrak, F., Abu-Hamdeh, N., Alnefaie, K. A. & Öztö, H. F. A review on exergy analysis of solar electricity production. *Renewable and Sustainable Energy Reviews*, **74**, 755–770 (2017).
- [3] IEA (INTERNATIONAL ENERGY AGENCY), *Renewables 2023 analysis and forecast to 2028*, (2024).
- [4] AlKandari, M. & Ahmad, I. Solar power generation forecasting using ensemble approach based on deep learning and statistical methods. *Applied Computing and Informatics*, **20(3–4)**, 231–250 (2024).
- [5] Eseye, A. T., Zhang, J. & Zheng, D. Short-term photovoltaic solar power forecasting using a hybrid Wavelet-PSO-SVM model based on SCADA and Meteorological information. *Renew Energy*, **118**, 357–367 (2018).
- [6] Pretto, S., Ogliari, E., Niccolai, A. & Nespoli, A. A new probabilistic ensemble method for an enhanced day-ahead power forecast. *IEEE J Photovolt*, **12(2)**, 581–588 (2022).
- [7] Xu, Y., Li, Q., Chew, B. S. H., Ding, H. & Zhao, G. An integrated missing-data tolerant model for probabilistic PV power generation forecasting. *IEEE Transactions on Power Systems*, **37(6)**, 4447–4459 (2022).

- [8] Harrou, F., Kadri, F. & Sun, Y. Forecasting of photovoltaic solar power production using LSTM approach. *Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems, London*, **3** (2020).
- [9] Hou, Z., Zhang, Y., Liu, Q. & Ye, X. A hybrid machine learning forecasting model for photovoltaic power. *Energy Reports*, **11**, 5125–5138 (2024).
- [10] El Hendouzi, A. & Bourouhou, A. Solar photovoltaic power forecasting. *Journal of Electrical and Computer Engineering*, **2020(1)**, 8819925 (2020).
- [11] Tuncar, E. A., Sağlam, Ş. & Oral, B. A review of short-term wind power generation forecasting methods in recent technological trends. *Energy Reports*, **12**, 197–209 (2024).
- [12] De Giorgi, M. G., Congedo, P. M. & Malvoni, M. Photovoltaic power forecasting using statistical methods: impact of weather data. *IET Science, Measurement & Technology*, **8(3)**, 90–97 (2014).
- [13] Zhou, Y., Zhou, N., Gong, L. & Jiang, M. Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine. *Energy*, **204**, 117894 (2020).
- [14] Graditi, G., Ferlito, S. & Adinolfi, G. Comparison of Photovoltaic plant power production prediction methods using a large measured dataset. *Renewable energy*, **90**, 513–519 (2016).
- [15] Tian, B., Piao, Z., Guo, D. & Wang, H. Wind power ultra short-term model based on improved EEMD-SE-ARMA. *Power System Protection and Control*, **45(4)**, 72–79 (2017).
- [16] Yang, Z., Ce, L. & Lian, L. Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods. *Applied Energy*, **190**, 291–305 (2017).
- [17] Atique, S. et al. Forecasting of total daily solar energy generation using ARIMA: A case study. In 2019 IEEE 9th annual computing and communication workshop and conference (CCWC). *IEEE*, (2019).
- [18] Fara, L., Diaconu, A., Craciunescu, D. & Fara, S. Forecasting of energy production for photovoltaic systems based on ARIMA and ANN advanced models. *International Journal of Photoenergy*, **2021(1)**, 6777488 (2021).
- [19] Vagropoulos, S. I., Chouliaras, G. I., Kardakos, E. G., Simoglou, C. K. & Bakirtzis, A. G. Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. In 2016 IEEE international energy conference (ENERGYCON). *IEEE*, (2016).

- [20] Paulescu, M., Badescu, V. & Brabec, M. (2013). Tools for PV (photovoltaic) plant operators: Nowcasting of passing clouds. *Energy*, **54**, 104–112.
- [21] Sharadga, H., Hajimirza, S. & Balog, R. S. Time series forecasting of solar power generation for large-scale photovoltaic plants. *Renewable Energy*, **150**, 797–807 (2020).
- [22] Khatib, T., Mohamed, A., Sopian, K. & Mahmoud, M. Assessment of artificial neural networks for hourly solar radiation prediction. *International Journal of Photoenergy*, **2012(1)**, 946890 (2012).
- [23] Saberian, A., Hizam, H., Radzi, M. A. M., Ab Kadir, M. Z. A., & Mirzaei, M. Modelling and prediction of photovoltaic power output using artificial neural networks. *International Journal of Photoenergy*, **2014(1)**, 469701 (2014).
- [24] Breiman, L. Random forests. *Machine Learning*, **45(1)**, 5–32 (2001).
- [25] Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning*, **20(3)**, 273–297 (1995).
- [26] Islam, M. S. S. et al. Optimizing short-term photovoltaic power forecasting: A novel approach with gaussian process regression and bayesian hyperparameter tuning. *Processes*, **12(3)**, 546 (2024).
- [27] Kanwal, S., Khan, B., Ali, S. M., Mehmood, C. A. & Rauf, M. Q. Support vector machine and Gaussian process regression based modeling for photovoltaic power prediction. In 2018 International Conference on Frontiers of Information Technology (FIT). *IEEE*, (2018).
- [28] Teo, T. T., Logenthiran, T. & Woo, W. L. Forecasting of photovoltaic power using extreme learning machine. In 2015 IEEE Innovative Smart Grid Technologies-Asia (ISGT ASIA). *IEEE*, (2015).
- [29] Al-Dahidi, S., Ayadi, O., Adeeb, J., Alrbai, M. & Qawasmeh, B. R. Extreme learning machines for solar photovoltaic power predictions. *Energies*, **11(10)**, 2725 (2018).
- [30] Al-Dahidi, S., Alrbai, M., Alahmer, H., Rinchi, B. & Alahmer, A. Enhancing solar photovoltaic energy production prediction using diverse machine learning models tuned with the chimp optimization algorithm. *Scientific Reports*, **14(1)**, 18583 (2024).
- [31] Amiri, A. F., Chouder, A., Oudira, H., Silvestre, S. & Kichou, S. Improving photovoltaic power prediction: Insights through computational modeling and feature selection. *Energies*, **17(13)**, 3078 (2024).
- [32] Nespoli, A., Leva, S., Mussetta, M. & Ogliari, E. G. C. A selective ensemble approach for accuracy improvement and computational load reduction in ANN-based PV power forecasting. *IEEE Access*, **10**, 32900–32911 (2022).

- [33] Tahir, M. F., Yousaf, M. Z., Tzes, A., El Moursi, M. S. & El-Fouly, T. H. M. Enhanced solar photovoltaic power prediction using diverse machine learning algorithms with hyperparameter optimization. *Renewable and Sustainable Energy Reviews*, **200**, 114581 (2024).
- [34] Asiedu, S. T., Nyarko, F. K. A., Boahen, S., Effah, F. B. & Asaaga, B. A. Machine learning forecasting of solar PV production using single and hybrid models over different time horizons. *Heliyon*, **10(7)**, e28898 (2024).
- [35] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794 (2016).
- [36] Saxena, N., et al. Hybrid KNN-SVM machine learning approach for solar power forecasting. *Environmental Challenges*. **14**, 100838 (2024).
- [37] Mas' ud, A. A. Comparison of three machine learning models for the prediction of hourly PV output power in Saudi Arabia. *Ain Shams Engineering Journal*, **13(4)**, 101648 (2022).
- [38] Ramli, N. A., Hamid, M. F. A., Azhan, N. H. & Ishak, M. A. A. S. Solar power generation prediction by using k-nearest neighbor method. In AIP Conference Proceedings, **2129(1)**, 020116 (2019).
- [39] Esen, V., Coban, B., Kavus, B. Y., Karaca, T. K., Dindar, T. & Sarkin, A. S. An interpretable statistical approach to photovoltaic power forecasting using factor analysis and ridge regression. *Scientific Reports*, **15(1)**, 38947 (2025).
- [40] Ren, X., Liu, Y., Zhang, F. & Li, L. A deep learning quantile regression photovoltaic power-forecasting method under a priori knowledge injection. *Energies*, **17(16)**, 4026 (2024).
- [41] Islam, M. S. S. et al. Optimizing short-term photovoltaic power forecasting: A novel approach with gaussian process regression and bayesian hyperparameter tuning. *Processes*, **12(3)**, 546 (2024).
- [42] Gillingham, K. & Tsvetanov, T. Hurdles and steps: Estimating demand for solar photovoltaics. *Quantitative Economics*, **10(1)**, (2019).
- [43] Kim, H., Jang, H. & Kim, J. Two-stage machine learning for urban PV generation forecasting: model development and SHAP-based interpretation of key climatic drivers. *International Journal of Thermophysics*, **47(5)**, (2026).
- [44] Albert, A. & Anderson, J. A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71(1)**, 1–10 (1984).

- [45] McCullagh, P. & Nelder, J. Generalized linear models (Chapman & Hall, 1989).
- [46] Caron, R., Sinha, D., Dey, D. K. & Polpo, A. Categorical data analysis using a skewed Weibull regression model. *Entropy*, **20(3)**, 176 (2018).
- [47] Vapnik, V. N. Constructing learning algorithms. In *The nature of statistical learning theory*. (Springer, 1995).
- [48] Held, L., & Holmes, C. C. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, **1(1)**, 145–168.
- [49] Goodrich, B., Gabry, J., Ali, I. & Brilleman, S. rstanarm: Bayesian applied regression modeling via Stan. R package version 2.32. 1. 2024 (2025).
- [50] Liaw, A., & Wiener, M. Classification and regression by randomForest. *R News*, **2(3)**, 18–22 (2002).
- [51] Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. Classification and regression trees. (Chapman and Hall/CRC, 2017).
- [52] Schonlau, M. & Zou, R. Y. The random forest algorithm for statistical learning. *The Stata Journal*, **20(1)**, 3–29 (2020).
- [53] Barreñada, L., Dhiman, P., Timmerman, D., Boulesteix, A. L. & Van Calster, B. Understanding overfitting in random forest for probability estimation: a visualization and simulation study. *Diagnostic and Prognostic Research*, **8(1)**, 14 (2024).
- [54] Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution* (Springer, 1992).
- [55] Schölkopf, B. & Smola, A. J. Learning with kernels: support vector machines, regularization, optimization, and beyond (MIT Press, 2002).
- [56] Ding, X., Liu, J., Yang, F. & Cao, J. Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute*, **358(18)**, 10121–10140 (2021).
- [57] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F., e1071: Misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-16, (2024).
- [58] Chen, T. et al., Xgboost: extreme gradient boosting. R package version 0.4-2, **1(4)**, 1–4 (2015).
- [59] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (2025).
- [60] Lawless, J. F. Statistical models and methods for lifetime data. (John Wiley & Sons, 2011).

- [61] Cavalcante, T., Ospina, R., Leiva, V., Cabezas, X. & Martin-Barreiro, C. Weibull regression and machine learning survival models: Methodology, comparison, and application to biomedical data related to cardiac surgery. *Biology*, **12(3)**, 442 (2023).
- [62] Therneau, T. M. & Grambsch, P. M. Modeling Survival Data: The Cox Model (2000, Springer).
- [63] Therneau, T. M. A package for survival analysis in R, R package version 3.8–3 (2024).
- [64] Zeileis, A., Kleiber, C. & Jackman, S. Regression models for count data in R. *Journal of Statistical Software*, **27**, 1–25 (2008).
- [65] Cragg, J. G. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 829–844 (1971).
- [66] Deb, P., & Trivedi, P. K. The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics*, **21(4)**, 601–625 (2002).
- [67] Altinisik, Y. A comparative study on the performance of frequentist and Bayesian estimation methods under separation in logistic regression. *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, **69(2)**, 1083–1103 (2020).
- [68] Zhu, C. et al., Short-term power prediction of photovoltaic power station based on LSTM-XGBoost model. *Solar Energy*, **300**, 113819 (2025).
- [69] Li, X. et al., Exploring interactive and nonlinear effects of key factors on intercity travel mode choice using XGBoost, *Applied Geography*, **166**, 103264 (2024).
- [70] Parmar, C. B., A comparative analysis of K-nearest neighbors and random forest regression for predicting solar panel power generation, *International Research Journal of Engineering and Technology (IRJET)*, **12(10)**, 651-655 (2025).
- [71] Auret, L., & Aldrich, C., Interpretation of nonlinear relationships between process variables by use of random forests, *Minerals Engineering*, **35**, 27-42 (2012).
- [72] Prakoso, D. P., Irfan, M. & Siddique, Q. A comparative analysis of linear regression and XGBoost algorithms for predicting GPU prices using technical specifications. *International Journal of Informatics and Information Systems*, **7(4)**, 189–199 (2024).
- [73] Jogunuri, S. Random forest machine learning algorithm based seasonal multi-step ahead short-term solar photovoltaic power output forecasting. *IET Renewable Power Generation*, **19**, 1-16 (2025).
- [74] Atiea, M. A., Abdelghaffar, A. A., Aribia, H. B., Nouredine, F., & Shaheen, A. M., Photovoltaic power generation forecasting with Bayesian optimization and stacked ensemble learning. *Results in Engineering*, **26**, 104950 (2025).

- [75] Alshangiti, A. M. et al., New robust estimator for handling outliers and multicollinearity in gamma regression model with application to breast cancer data. *Scientific Reports*, **15**, 38436 (2025).
- [76] Motulsky, H. J., Head, T., & Clarke, P. B. S., Analyzing lognormal data: A nonmathematical practical guide. *Pharmacological Reviews*, **77(3)**, 100049 (2025).
- [77] Aydi, W., & Alduais, F. S., Estimating Weibull parameters using least squares and multilayer perceptron vs. Bayes estimation. *Computers, Materials and Continua*, **71(2)**, 4033-4050 (2021).
- [78] Safari, M. A. M., Masseran, N., Majid, M. H. A., & Tajuddin, R. R. M., Robust estimation of the three parameter Weibull distribution for addressing outliers in reliability analysis. *Scientific Reports*, **15**, 11516 (2025).

ARTICLE IN PRESS