



Data preprocessing techniques for statistical photovoltaic forecasting models

Tolulope Olumuyiwa Falope^{a,*}, Liyun Lao^a, Da Huo^b

^a Energy and Power, School of Water, Energy & Environment, Cranfield University, MK43 0AL, UK

^b Queen Mary University of London, 327 Mile End Road, London, E1 4NS

ARTICLE INFO

Keywords:

Datapreprocessing
Renewable energy sources
Statistical photovoltaic method
Machine learning
Photovoltaic forecasting

ABSTRACT

Accurate photovoltaic power forecasting depends on high-quality data, yet raw meteorological and photovoltaic datasets often contain noise, missing values, and redundant features. Data preprocessing is therefore critical, but existing studies typically apply isolated techniques without a structured framework or systematic evaluation of combined strategies. This paper addresses these gaps by introducing a functional classification of data preprocessing methods and empirically testing sixteen (16) widely used techniques—individually and in combination—on raw photovoltaic and meteorological data. Two scenarios are considered: (i) comparing optimal data preprocessing combinations to a base-case approach using minimal cleaning, and (ii) benchmarking against an established three-step photovoltaic forecasting model. Forecasting performance is assessed using Root Mean Square Error across more than twenty (20) regression algorithms in MATLAB's Regression Learner App. The experimental findings are further validated through cross-testing with an independent dataset, and an iterative ranking procedure is introduced to examine preprocessing order and cumulative impact. Results demonstrate that integrated data preprocessing strategies significantly outperform both baseline and conventional preprocessing pipelines, highlighting the importance of method synergy. This study provides a reproducible framework for data preprocessing classification and evaluation, offering practical guidance for researchers seeking to optimize photovoltaic forecasting models.

1. Introduction

Accurate photovoltaic (PV) forecasting relies heavily on data-driven models, which in turn depend on large, high-quality datasets [1]. Raw meteorological and PV data often contain noise, redundancy, and missing values, making data preprocessing (DP) essential for improving model performance [2–4]. While DP techniques—such as outlier handling, normalization, and feature engineering—are widely acknowledged for enhancing forecasting accuracy, their application in PV forecasting remains fragmented. Existing studies typically focus on isolated methods or narrow subsets of techniques, offering limited guidance on how different DP strategies interact or complement each other [5–8].

Recent research demonstrates DP's influence on PV forecasting accuracy. For example, Ranjan et al. [9] classify outlier detection techniques into statistical, density, and distance-based approaches. Bramm et al. [10] report that preprocessing—including outlier removal, daylight filtering, and missing data handling—reduced Root Mean

Square Error (RMSE) by 75.9% in linear regression. Phan et al. [11] demonstrate that normalization, scaling, and feature engineering significantly improve Gated Recurrent Unit (GRU)-based models compared to unprocessed inputs. However, these studies rarely explore combined strategies across multiple functional categories. A summary of prior findings is presented in Table 1.

Further review of literature shows a lack of a systematic framework for categorizing DP methods in PV forecasting. Asma and Nagaraja [15] enumerate DP categories and techniques separately but do not explicitly map techniques to categories. By contrast, Fan et al. [16] associate techniques with five DP categories; however, their framework targets general data analysis rather than PV forecasting. Other studies provide only partial coverage: Lee et al. [17] consider four DP categories; Chiranjeevi et al. [18] emphasize outlier removal, normalization, and correlation analysis; Nguyen and Ha [19] focus primarily on normalization (e.g., Adaptive Min–Max Scaling, log transforms); and Shin et al. [20] present a mathematical preprocessing model without a technique-level categorization. Collectively, these gaps underscore the need for a domain-specific, comprehensive, and systematically organized DP

* Corresponding author.

E-mail addresses: t.o.falope@cranfield.ac.uk (T.O. Falope), d.huo@qmul.ac.uk (D. Huo).

<https://doi.org/10.1016/j.measurement.2026.121997>

Received 22 September 2025; Received in revised form 28 March 2026; Accepted 20 May 2026

Available online 25 May 2026

0263-2241/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Nomenclature		MM	Moving Mean
Abbreviations		MSE	Mean Squared Error
		NC	Negative Correlation
		NMSE	Normalized Mean Squared Error
		NN	Neural Networks
		PCA	Principal Component Analysis
		PCC	Pearson Correlation Coefficient
		PV	Photovoltaic
		PVC	Photovoltaic Correlation
		RLA	Regression Learner App
		RM	Rescaling Method
		RRVFLN	Robust Minimum Variance Random Vector Functional Link Network
		RMSE	Root Mean Square Error
		SD	Seasonal Decomposition
		SR	Solar Radiation
		SRC	Solar Radiation Correlation
		SVM	Support Vector Machines
		VMD	Variational Mode Decomposition
		WD	Wavelet Decomposition
		WNN	Wide Neural Network
		XGBoost	Extreme Gradient Boosting
		ZS	Z-score Method
ANN		Artificial Neural Network	
ASD		Atomic Sparse Decomposition	
BiLSTM		Bidirectional Long Short-Term Memory	
DF		Daylight-free	
DFT		Discrete Fourier Transform	
DO		Delete Outliers	
DP		Data Preprocessing	
DS		Data Smoothing	
DSM		Decimal Score Method	
EMD		Empirical Mode Decomposition	
ENN		Elman Neural Network	
EWT		Empirical Wavelet Transform	
FO		Fill Outliers	
FRNN		Full Recurrent Neural Network	
FWPD		Full Wavelet Packet Decomposition	
GPR		Gaussian Process Regression	
GRU		Gated Recurrent Unit	
LIM		Linear Interpolation Method	
LWDM		List-wise Deletion Method	
MAE		Mean Absolute Error	
MGSM		Medium gaussian Support Vector Machine	

Table 1
Study comparisons on data preprocessing methods.

DP Methods	Studies	Ranjan et al. [9]	Bramm et al. [10]	Phan et al. [11]	Carrera and Kim [12]	Alshdaifat et al. [13]	Malvoni et al. [14]	This study
LWDM			✓			✓		✓
LIM						✓		✓
DO			✓	✓				✓
FO	✓							✓
DS	✓							✓
ZS						✓		✓
RM				✓		✓		✓
LLDF	✓			✓	✓			✓
PCC				✓				✓
PCA				✓			✓	✓
DTM				✓				✓
DF			✓		✓			✓
SD								✓
Resub							✓	✓
Cross					✓	✓		✓
Holdout				✓				✓

Note: Cross = Cross-validation, DF = Daylight-free, DO = Delete Outliers, DS = Data Smoothing, DTM = Day/Time/Month, FO = Fill Outliers, Holdout = Holdout Validation, LIM = Linear Interpolation Method, LWDM = List-wise Deletion Method, PCA = Principal Component Analysis, PCC = Pearsons Correlation Coefficient, Resub = Resubstitution validation, RM = Rescaling Method, ZS = Z-Score.

framework for PV forecasting. The main contributions of this study are:

1. A systematic categorization of DP techniques is introduced based on their functional roles in statistical PV forecasting. This framework enables researchers to quickly identify method categories, streamline preprocessing strategies, and scope additional techniques for future studies.
2. Unlike prior studies that assess DP methods in isolation, this work evaluates sixteen (16) widely used techniques—both individually and in combination—on raw meteorological and PV datasets. Through RMSE-based benchmarking across multiple regression algorithms and a sequential testing approach, the study derives an optimal ordering of DP methods, demonstrating how both individual techniques and their sequence materially influence forecasting accuracy.

3. Building on this sequential framework, a dedicated DP ranking method is introduced to validate the optimal sequence by selecting techniques that deliver the greatest incremental reductions in RMSE. Cross-scenario testing and multiple robustness analyses further confirm the method’s ability to identify the relative importance and interactive effects of DP choices, reinforcing the reliability of the derived sequence.
4. This study investigates whether integrating multiple DP methods yields measurable improvements over both a base-case scenario (minimal cleaning) and an established statistical three-step PV forecasting model. This analysis offers actionable insights for designing optimized preprocessing pipelines.

This paper is organized as follows. Section 2 outlines the rationale for the selected data preprocessing methodology. Section 3 introduces a

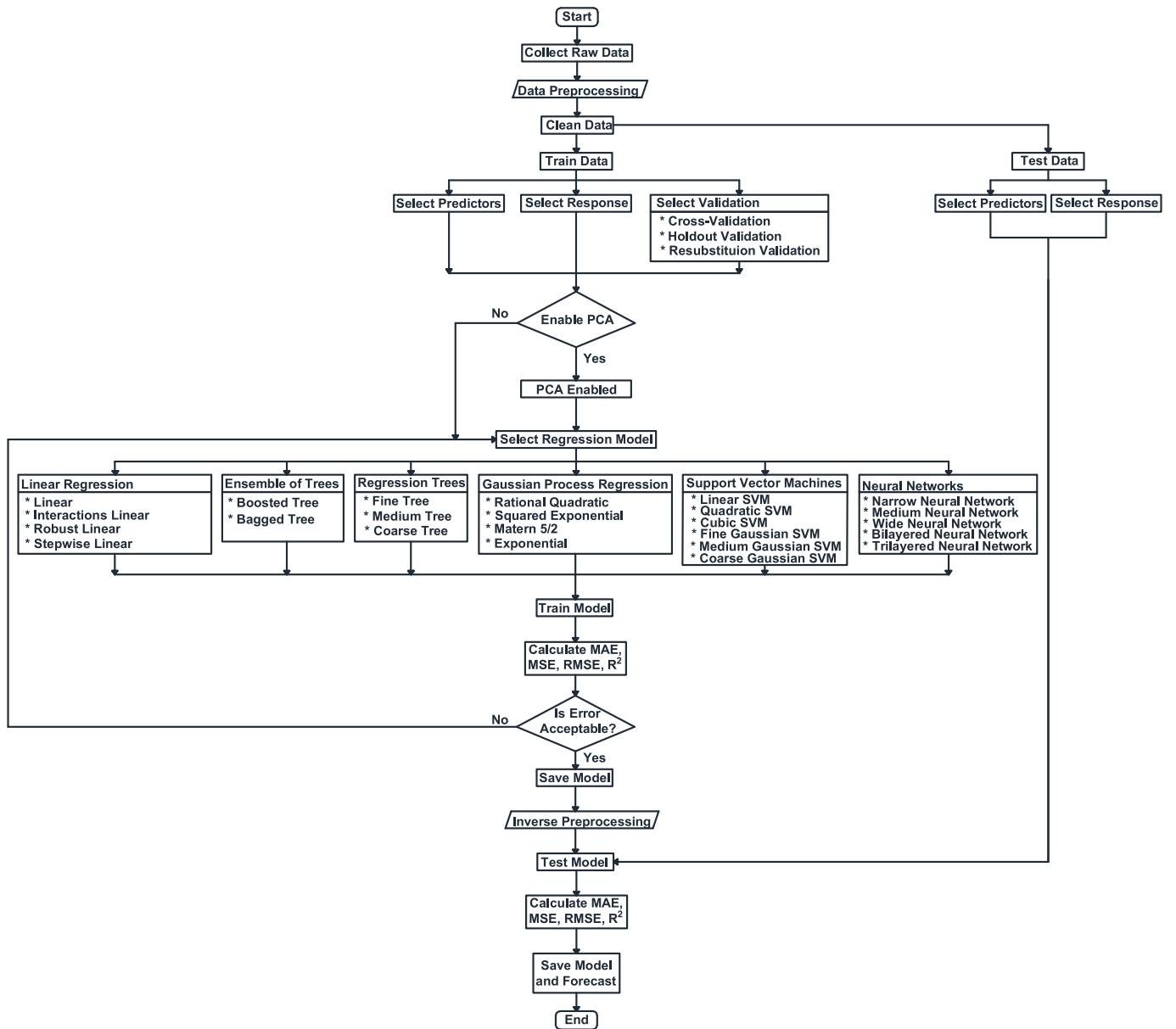


Fig. 1. Flow Chart of the Regression Learner App.

systematic classification of DP methods. Section 4 details the methodology underlying the three-step forecasting model. Section 5 outlines the evaluation framework, including validation procedures and error metrics. Section 6 presents the case study, describing the dataset and analysis workflow. Section 7 reports and discusses the experimental results. Finally, Section 8 concludes the paper and provides recommendations for future research.

2. Methodology

The methodology adopted in this study was designed to address two critical gaps identified in existing PV forecasting research: (i) the absence of a structured framework for categorizing DP methods, and (ii) the lack of empirical evidence on how integrated DP strategies influence forecasting accuracy.

Framework development: To address these gaps, a DP framework was developed and organized into a functional classification hierarchy. This hierarchy reflects the operational dependencies among DP methods: some methods commute (order-independent, e.g., data decomposition

vs. low-level data fusion) [21], whereas others exhibit strong order dependence (e.g., gap-filling before normalization) [22,23]. DP methods were grouped into stages that represent a natural processing order, informed by prior literature and best practices in PV forecasting [24]. This ensured that the experimental design mirrored real-world workflows with demonstrated relevance.

Experimental design: Once the framework was established, trials were conducted sequentially from left to right across the hierarchy, beginning with foundational techniques such as missing-data handling and progressing toward advanced methods like feature selection and dimensionality reduction [17]. For this design, two and a half years of historical data (2018–2020), comprising nine weather parameters from the Kisanhub weather station (KH075) [25] located at Cranfield University and twelve parameters from the UK Met Office in Bedford was used. Corresponding PV generation data were obtained from Cranfield's online energy management platform [26]. Iterative trials were performed, progressively adding DP techniques to assess incremental improvements in RMSE. The optimal configuration was identified by comparing RMSE reductions across all trials.

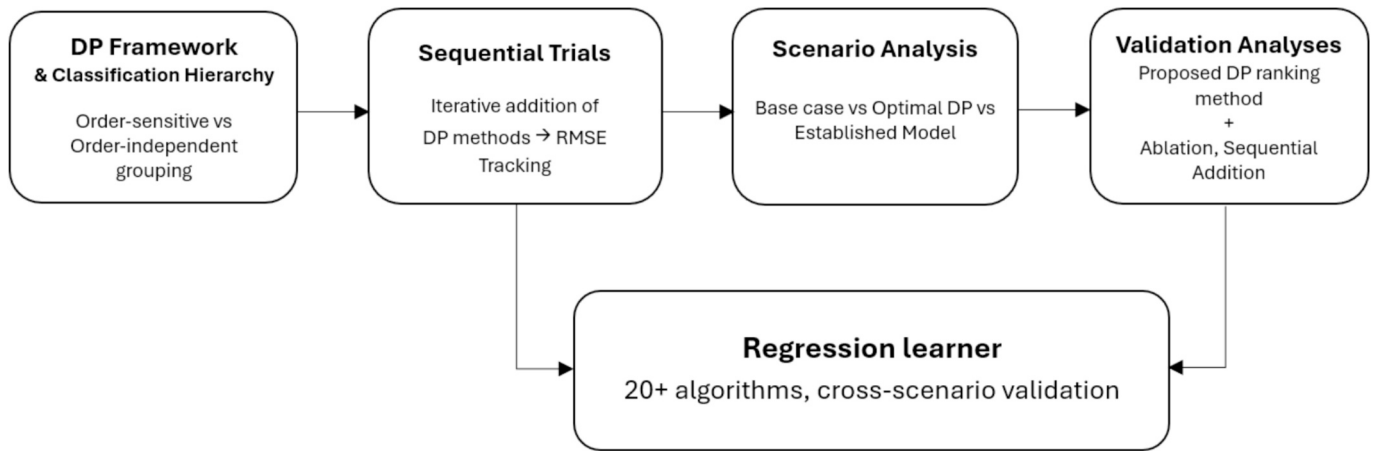


Fig. 2. Methodology for DP evaluation in PV forecasting.

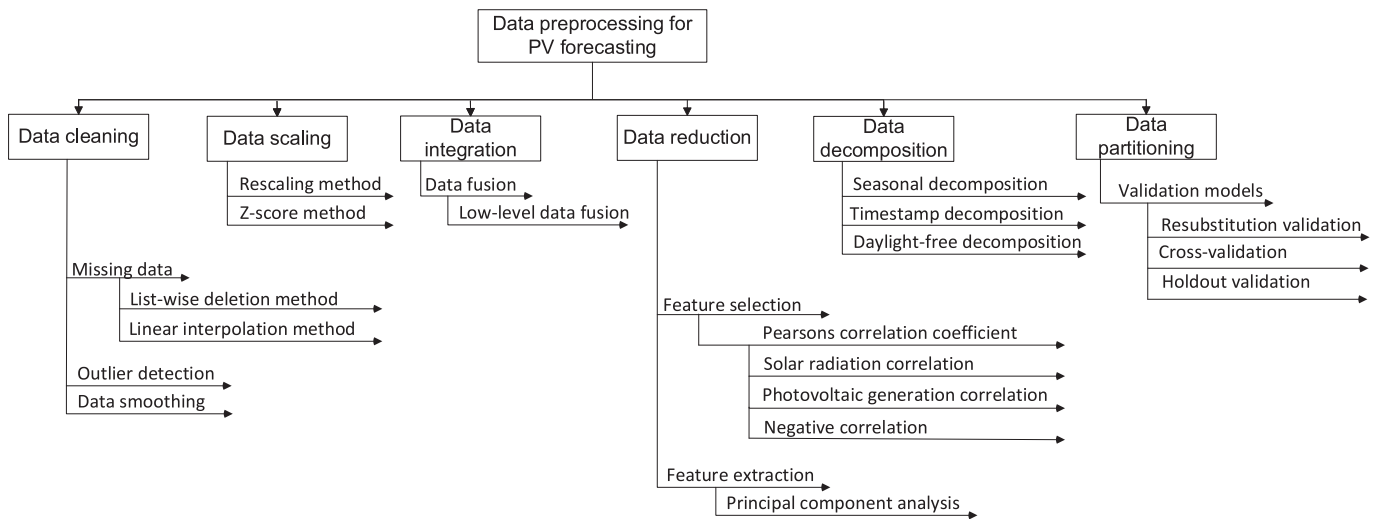


Fig. 3. Data preprocessing framework for statistical PV forecasting models.

Two comparative scenarios were considered: (a) benchmarking optimal DP combinations against a base-case configuration using minimal cleaning, and (b) comparing these combinations to an established three-step PV forecasting model; a model which relies on integrating several DP methods [27].

Performance evaluation was conducted using MATLAB's Regression Learner App (RLA) across more than twenty regression algorithms to model the relationship between photovoltaic (PV) generation and multiple weather-related predictors. The RLA, as seen in Fig. 1, was used to identify the most accurate model based on the lowest Root Mean Square Error (RMSE). RMSE was chosen as the evaluation metric because it is well-suited for relatively clean datasets, penalizes larger deviations effectively, and provides an interpretable measure of prediction accuracy.

Order selection and validation: To validate the optimal sequence of DP methods obtained in the experimental design, an iterative DP ranking approach was developed. Cross-scenario validation was employed, using a one-year (2020 – 2021) historical dataset from the same data sources for training and testing to ensure generalizability [28]. Initially, Linear Interpolation Method (LIM) and List-wise Deletion Method (LWDM) were applied individually to the raw dataset, and its corresponding RMSE was computed. The method yielding the lowest RMSE was selected as the first component in the sequence. Subsequently, the remaining DP methods were evaluated by applying them to the newly pre-processed dataset, and the next method producing the greatest

incremental improvement (lowest RMSE) was chosen. This process was repeated iteratively, adding one method at a time based on performance gains, an optimal ordering of DP methods was established. The resulting sequence represents the empirically derived optimal order of DP techniques for the forecasting pipeline.

To assess the robustness of the final ordering, three complementary validation techniques—ablation analysis and sequential addition analysis—were conducted. These analyses quantified the relative contribution and operational impact of each selected DP method, thereby confirming the validity of the proposed ranking sequence. Fig. 2 illustrates the methodology's structure.

3. Data preprocessing framework

Fig. 3 illustrates the proposed DP framework along with 16 selected DP techniques classified into six categories: (1) data cleaning, (2) data reduction, (3) data decomposition, (4) data scaling, (5) data integration and (6) data partitioning.

3.1. Data cleaning

Data cleaning refers to the process of addressing missing or erroneous data within a dataset. This includes detecting and correcting outliers, as well as removing incomplete or inconsistent records [29]. In this study, two data cleaning methods are investigated: handling missing

data and outlier detection.

3.1.1. Missing data

Missing data is a common challenge in large datasets, particularly when measurements originate from devices susceptible to malfunction, communication failures, or human error. Unaccounted values can introduce bias into forecasting models [30], thereby compromising predictive accuracy. Additionally, distorted or corrupted data may occur during or after collection, necessitating strategies to mitigate their impact on model development [31]. Addressing missing data is typically the first step in DP and serves as the baseline scenario in this study.

This work utilized three datasets—historical PV generation, onsite weather measurements, and local weather station data—all exhibiting missing values completely at random (MCAR). To manage these gaps, two approaches were implemented: LWDM and LIM.

LWDM is among the simplest and most widely used techniques for handling missing data [31]. Under this method, if any predictor variable is missing for a given timestamp, all associated observations for that timestamp are removed [32]. In other words, LWDM excludes entire rows when one or more variables are missing [31,33]. For example, Ziyabari et al. [34] applied LWDM to omit missing and unnecessary data from a 15-year dataset split into training (10 years), testing (3 years), and validation (2 years). Using a multi-branch residual network, they achieved a low RMSE of 0.073, outperforming other architectures.

In this study, LWDM was applied when synthesizing the three raw data sources into a unified dataset. Specifically, if a day's PV meter data was missing, the corresponding timestamps in both onsite and local weather datasets were also removed. While this approach reduces sample size, its impact is minimal when the dataset is sufficiently large, thereby preserving data integrity and avoiding over-approximation.

Conversely, LIM is a widely adopted imputation technique [35], which substitutes missing values to create a complete dataset [29]. Kim et al. [36], demonstrated LIM's effectiveness in estimating missing environmental and behavioural data. The LIM formula [37] is presented in Equation (1).

$$y = y_1 + (x - x_1) \frac{(y_2 - y_1)}{(x_2 - x_1)} \quad (1)$$

Where x and y are the current points, y_1, x_1 are the previous points, and y_2, x_2 are the next points.

3.1.2. Outlier detection and data smoothing

Outlier detection involves identifying observations that deviate substantially from much of the dataset or conflict with expected patterns [38]. Common approaches to fill outliers (FO) or delete outliers (DO) include quartile-based methods, the generalized extreme studentized deviate test, and moving average techniques, among others [39]. Data smoothing (DS), on the other hand, aims to reduce noise and irregular fluctuations while preserving the underlying structure of the data. This is typically achieved through statistical or mathematical transformations, with the choice of method determined by the characteristics of the dataset. A widely used technique that serves both outlier detection and DS is the Moving Average (MA), also referred to as the Moving Mean. The MA method computes the average of a subset of consecutive observations within a sliding window, thereby attenuating short-term variations and highlighting longer-term trends.

3.2. Data reduction

Dimensionality reduction refers to the process of transforming a high-dimensional dataset into a lower-dimensional representation while preserving as much of the original structure and information as possible [40]. This process typically involves techniques such as feature selection and feature extraction, which aim to reduce redundancy and improve computational efficiency. Principal Component Analysis (PCA) is among

the most widely adopted methods due to its ability to capture the maximum variance in the data through orthogonal components, thereby preserving essential patterns while reducing complexity. Its robustness, interpretability, and effectiveness across diverse statistical and machine learning applications make PCA a standard approach for dimensionality reduction.

3.2.1. Feature selection

Feature selection, also referred to as variable or attribute selection, is the process of identifying a subset of relevant and non-redundant features that effectively represent the original dataset while reducing dimensionality [41]. This reduction enhances model performance by improving accuracy, computational efficiency, generalizability, and interpretability [42].

In the proposed three-step forecasting framework, Pearson Correlation Coefficient (PCC) analysis is employed to select weather parameters that exhibit a positive moderate-to-strong correlation with solar radiation (SR). PCC was chosen because it provides a simple yet effective measure of the linear relationship between two continuous variables, which is particularly relevant for PV forecasting where weather parameters often exhibit linear or near-linear associations with solar radiation and power output. Unlike more complex methods such as mutual information or wrapper-based approaches, PCC is computationally efficient, interpretable, and well-suited for large datasets. Its ability to quantify both the strength and direction of correlation allows for straightforward identification of features that contribute meaningfully to model accuracy without introducing unnecessary complexity. Additionally, PCC is widely used in energy forecasting literature, making it a robust and accepted choice for preliminary feature selection in statistical models.

The PCC formula [43], presented in Equation (2), quantifies the strength of the linear relationship between variables, enabling the identification of parameters most strongly associated with PV generation. A higher correlation coefficient indicates a stronger predictive relationship, thereby contributing to improved forecasting accuracy [44].

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \quad (2)$$

Where r is the correlation coefficient, x is one of variables in the sample that is to be compared, y represents the other variable in the sample, $\bar{x}$ and $\bar{y}$ are the mean values of the respective variables.

Similar to the proposed model, several studies [45–49] have employed PCC for feature selection. For instance, Zeng et al. [50] examined correlations between PV output and more than ten meteorological parameters, whereas Mansood et al. [48] focused on five parameters. In this study, both solar radiation correlation (SRC) and PV correlation (PVC) are investigated to determine which approach is more effective in selecting relevant features.

Additionally, the inclusion of negatively correlated features is explored. Negative correlation (NC) represents an inverse relationship between two variables, where an increase in one corresponds to a decrease in the other [51]. NC plays a significant role in various predictive models; for example, heating energy demand typically rises as temperature falls. In a review of solar irradiation forecasting methods, Patel et al. [52] identified an inverse relationship with humidity, classifying it as a key parameter for model training—an example of NC. Similarly, Kong et al. [49] applied an R-value threshold of 0.25 for selecting seven weather parameters, where humidity, with an R-value of -0.5681 , was included as an input feature.

3.2.2. Feature extraction

PCA is a widely used technique for feature extraction and dimensionality reduction [53]. It reduces the number of variables in a dataset by transforming them into a new set of uncorrelated variables, known as

principal components, which are linear combinations of the original features. PCA identifies underlying patterns or driving factors shared among variables, removes redundancy, and captures the maximum variance in fewer dimensions. This transformation simplifies the dataset, making it easier for forecasting algorithms to detect patterns and relationships. Importantly, all principal components are orthogonal, ensuring no redundant information, and the complete set of components retains the full variance of the original dataset.

PCA has been extensively applied in solar forecasting studies to enhance model accuracy and computational efficiency [14,54–58]. For example, Kazem et al. [54] demonstrated that PCA improved PV power forecasting performance, achieving a normalized mean square error (NMSE) of 0.0780 compared to 0.1778 for a Full Recurrent Neural Network (FRNN). Similarly, Munawar et al. [58] employed PCA for feature selection in an XGBoost-based model, which produced the lowest RMSE and outperformed feature importance-based selection. Mallika et al. [57] combined PCA with artificial neural networks (ANN) for ionospheric forecasting, achieving a correlation of 0.95 compared to 0.41 for standard ANN models. Lan et al. [55] integrated PCA with Discrete Fourier Transform (DFT) and Elman Neural Networks (ENN) to improve SR frequency tracking.

Further, Malvoni et al. [14] reported that PCA combined with Wavelet Decomposition (WD) reduced prediction error by up to 1% compared to other preprocessing techniques. Zhang and Wei [56] applied PCA to reduce the dimensionality of decomposed historical SR data for daily forecasting, demonstrating superior performance when PCA was included in a hybrid model with Wavelet Transformation, Extreme Learning Machine, and Bat Algorithm. Collectively, these studies confirm PCA's effectiveness in reducing dimensionality, improving computational efficiency, and enhancing forecasting accuracy in PV and related energy prediction models.

PCA in MATLAB uses the singular value decomposition algorithm. Let X be a real $m \times n$ matrix of rank n , and I_n is the identity matrix. A singular value decomposition of X [59] is such that:

$$X = U\Sigma Z^T \quad (3)$$

Where: U is a $m \times r$ matrix called the left singular values of X and $U \times U^T = I_r$.

Σ is an $r \times r$ matrix that is diagonal with single values of X .

Z is a $n \times r$ matrix called the right singular value of X and $Z \times Z^T = I_r$.

Using singular decomposition,

$$T = XZ \quad (4)$$

$$T = U\Sigma Z^T Z \quad (5)$$

$$T = U\Sigma \quad (6)$$

The output of PCA is the pair of matrices U and Σ . Matrix T is now comparable to the original matrix X such that each column of T is a left singular vector of X multiplied by the corresponding singular value.

3.3. Data decomposition

Data decomposition is a technique used to partition the original time series into multiple relatively stable subseries [60]. The objective is to isolate components with distinct characteristics, thereby simplifying pattern recognition for forecasting algorithms. For instance, decomposition can reveal seasonality and long-term trends within the data. Forecasting models are then developed for each subseries, and the final prediction is obtained by aggregating the individual forecasts. Common decomposition methods include WD, Atomic Sparse Decomposition (ASD), Empirical Mode Decomposition (EMD), and Variational Mode Decomposition (VMD) [61].

Decomposition approaches are generally classified as additive or multiplicative. Seasonal decomposition (SD), an example of the additive approach, separates a time series into trend, seasonal, and residual

components, capturing seasonal variations in the observed data [62]. Timestamp-based decomposition can also group data by days, months, or specific time intervals. The rationale for seasonal grouping is that days within the same season exhibit similar weather patterns, enabling machine learning models to detect patterns more effectively. Several studies have explored this concept. Fraihat et al. [63] examined the influence of 24 weather parameters on SR forecasting in Jordan, concluding that parameter impacts vary by season. Singla et al. [64] proposed a hybrid model combining Full Wavelet Packet Decomposition (FWPD) and Bidirectional Long Short-Term Memory (BiLSTM), using seasonal divisions defined by the Indian Meteorological Department. Results indicated lower forecasting errors in summer and winter compared to autumn and monsoon periods. Similarly, Ali-Ou-Salah et al. [65] introduced a hybrid approach integrating ANN with seasonal clustering for ultra-short-term SR forecasting. Another study combined Empirical Wavelet Transform (EWT) with Robust Minimum Variance Random Vector Functional Link Networks (RRVFLN), acknowledging the influence of seasonal and geographical factors on solar power generation.

Decomposition can also be applied to daily PV generation patterns, which typically follow a bell-shaped curve aligned with solar intensity. Certain hours—such as early morning and nighttime—yield zero PV output due to the absence of sunlight, referred to as daylight-free (DF) hours. Several studies have adopted the practice of excluding DF hours to improve model accuracy [64,66–70].

3.4. Data scaling

Data scaling—also referred to as data transformation or normalization—is the process of adjusting the scale of variables in a dataset to a common range. This step ensures that features with larger numerical values do not disproportionately influence the model, allowing all variables to contribute equally to the learning process [71]. Common scaling techniques include Min–Max scaling, Z-score normalization (ZS), robust scaling, and logarithmic transformation.

The Min–Max method, also known as the Rescaling method (RM), rescales data to a predefined interval, typically [1], thereby mitigating the risk of dominance by features with differing magnitudes. The RM formula [72] is given in Equation (7).

$$X_{rescaled} = a + \left[\frac{X - X_{min}}{X_{max} - X_{min}} \right] (b - a) \quad (7)$$

Where $X_{rescaled}$ is the normalized data between the intervals $[a, b]$, X represents the values of the x -variable in the sample, X_{min} is the minimum while X_{max} is the maximum value of X in the sample.

Z-score normalization (ZS), also known as the standard score method, is another widely used approach for data normalization. Several studies [63,66,73–76] have identified ZS as a preferred technique due to its ability to standardize features by centering them around zero and scaling based on variability. The ZS formula [77] presented in Equation (8), computes the normalized value using the sample mean (μ) and standard deviation (σ), ensuring that transformed data have a mean of zero and a standard deviation of one. This approach is particularly effective when features exhibit different units or scales, as it preserves relative differences while mitigating bias introduced by magnitude disparities.

$$Z = \frac{X - \mu}{\sigma} \quad (8)$$

3.5. Data integration

Data integration refers to the process of consolidating data from multiple sources into a unified dataset to provide a comprehensive and consistent view [78], creating data that is accurate, consistent, and formatted uniformly. Data fusion, a subset of data integration, involves

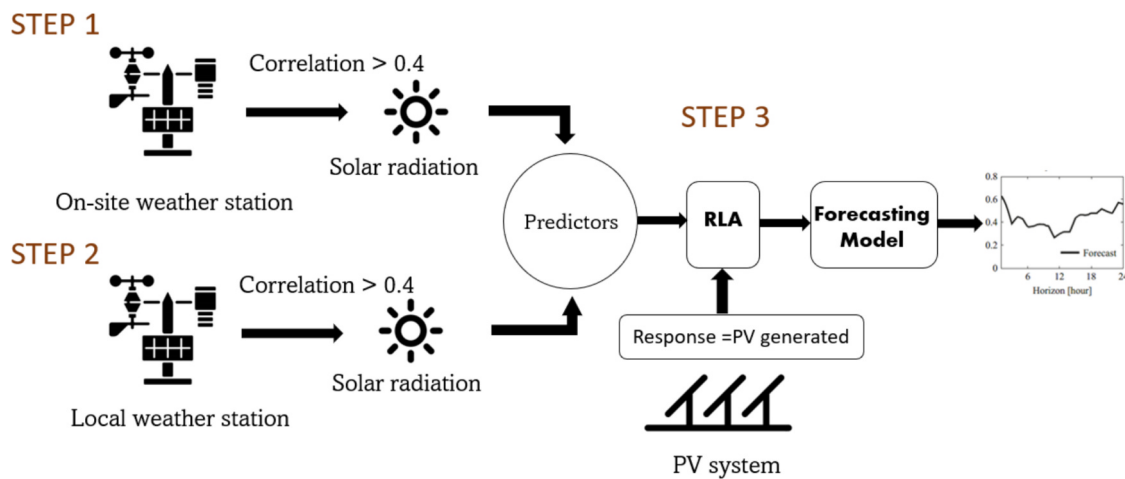


Fig. 4. Three-Step Forecasting Architecture.

merging information from diverse sources to extract meaningful insights [79]. This approach enables the creation of a richer and more complete representation of the data than would be possible using a single source. Data fusion techniques are generally categorized into low, intermediate, and high-level fusion, depending on the stage of processing at which integration occurs [80].

Low-Level Data Fusion (LLDF) is a straightforward method that concatenates raw data from multiple sources into a single sample for subsequent analysis [81]. Teglia et al. [82] proposed an LLDF strategy for processing second-order liquid chromatographic data, successfully addressing issues such as baseline contributions, peak distortions, and overlaps, while improving prediction accuracy and reducing uncertainty. Similarly, Lanza et al. [83] demonstrated that LLDF provided superior discriminative capability for milk classification compared to individual datasets (fatty acid-nonpolar and nuclear magnetic resonance polar), resulting in improved predictive performance of the Canonical Discriminant Analysis algorithm.

3.6. Data partitioning

Data partitioning, also referred to as data splitting, is the process of dividing a dataset into subsets for distinct purposes, such as training machine learning models, validating performance, and testing generalization capability. The most common approach involves splitting the data into training and testing sets, where the training subset is used to fit the model and the testing subset evaluates its predictive accuracy [84,85].

In this study, 75% of the available data was allocated for training, while the remaining 25% was reserved for testing. This ratio was selected to achieve an optimal balance between model learning and evaluation. Allocating 75% of the data for training ensures that the model has sufficient observations to capture underlying patterns and seasonal variability, which is critical for time-series forecasting. Reserving 25% for testing provides a robust sample for assessing generalization performance without introducing bias from overfitting.

4. Three-step forecasting model

The existing three-step statistical PV forecasting model [86] utilizes weather parameters that exhibit positive correlation with SR, sourced from both on-site and local weather stations, to predict PV generation. SR was selected as the primary reference parameter because it demonstrates the strongest correlation with PV output among all meteorological variables [87]. The raw data undergoes preprocessing using LLDF, PCC, LWDM and RM. For model validation, resubstitution was employed as the evaluation method. The architecture of the existing forecasting

framework is presented in Fig. 4.

This study aims to enhance the accuracy of the existing three-step PV forecasting model by evaluating the impact of additional DP techniques beyond the four methods previously employed. By systematically incorporating and testing alternative DP strategies, the proposed approach seeks to improve data quality and optimize model performance.

5. Model evaluation and validation

This involves the tools and techniques used in assessing and validating the performance of the model.

5.1. Validation methods

Validation methods play a critical role in model development by determining how data is partitioned for training, testing, and performance evaluation. These methods help prevent overfitting or underfitting by assessing the forecasting algorithm on unseen data. Typically, validation involves splitting the dataset into training and testing subsets. MATLAB provides three primary validation options: cross-validation, holdout validation, and resubstitution validation.

In cross-validation, the dataset is divided into k folds of approximately equal size. Each fold is used once for validation while the remaining folds are used for training, and this process is repeated until all folds have served as the validation set. The average validation error across all folds is then computed as the model's performance metric. Holdout validation, by contrast, partitions the dataset into training and validation subsets based on a predefined percentage, offering the advantage of simplicity since training and validation occur only once. However, caution is advised when applying this method to small datasets due to potential bias. Resubstitution validation uses the entire dataset for both training and error computation, which provides no safeguard against overfitting.

Several studies have applied these methods in forecasting and predictive modelling. Zhang et al. [88] employed both resubstitution and cross-validation, noting that cross-validation effectively corrected resubstitution bias in confirmatory clinical trials with finite samples. Ku et al. [89] used holdout validation with 30% of the data for testing when training gradient boosting and Gaussian Process Regression (GPR) models for respiratory disease prediction. Cross-validation has also been widely adopted in solar forecasting studies [54,58]; for example, Mallika et al. [57] applied k -fold cross-validation ($k = 5$ for feature selection and $k = 3$ for hyperparameter tuning) in ionospheric space weather forecasting. In a comparative study of classification tools [90], all three validation methods were evaluated, with holdout ratios ranging from

Table 2
Performance of data preprocessing methods.

Data Cleaning														
Trial	LWDM	LIM	DO	FO	DS	RMSE	Algorithm							
1	✓					38.52	MGSVM							
2	✓		✓			35.79	CGSVM							
3	✓			✓		38.40	MGSVM							
4	✓				✓	11.39	WNN							
5	✓		✓		✓	12.03	WNN							
6		✓				37.56	MGSVM							
7		✓	✓			37.15	MGSVM							
8		✓		✓		38.09	MGSVM							
9		✓			✓	11.53	WNN							
10		✓	✓		✓	12.74	WNN							
Data Scaling														
	LWDM	LIM	DS	ZS	RM	RMSE	Algorithm							
11	✓		✓	✓		0.2044	MNN							
12	✓		✓		✓	0.0403	WNN							
13		✓	✓	✓		0.1745	WNN							
14		✓	✓		✓	0.0375	WNN							
Data Integration														
	LIM	DO	FO	DS	RM	Cranf	Bed	LLDF	RMSE	Algorithm				
15	✓			✓	✓	✓			0.0422	FGSVM				
16	✓			✓	✓		✓		0.0530	FGSVM				
17	✓			✓	✓			✓	0.0375	WNN				
18	✓	✓		✓	✓			✓	0.0379	WNN				
19	✓		✓	✓	✓			✓	0.0295	MGSVM				
Data/Dimentionality Reduction – Feature Selection by Correlation (SRC,PVC, NC)														
	LWDM	LIM	FO	DS	RM	Cranf	Bed	LLDF	SRC	PVC	NC	RMSE	Algorithm	
20	✓				✓	✓			✓			0.1037	BNN	
21	✓		✓		✓	✓			✓			0.1042	NNN	
22	✓		✓	✓	✓	✓			✓			0.0604	WNN	
23	✓				✓		✓		✓			0.1049	TNN	
24	✓		✓		✓		✓		✓			0.1073	NNN	
25	✓		✓	✓	✓		✓		✓			0.0688	FGSVM	
26	✓				✓			✓	✓			0.0945	MNN	
27	✓		✓		✓			✓	✓			0.0928	BNN	
28	✓		✓	✓	✓			✓	✓			0.0549	FGSVM	
29		✓	✓	✓	✓			✓	✓			0.0547	WNN	
30		✓	✓	✓	✓			✓		✓		0.0417	WNN	
31		✓	✓	✓	✓			✓	✓		✓	0.0362	FGSVM	
32		✓	✓	✓	✓			✓		✓	✓	0.0365	FGSVM	
Data Decomposition – Day, Time, Month, Daylight-free														
	LIM	FO	DS	RM	LLDF	SRC	PVC	NC	DTM	DF	RMSE	Algorithm		
33	✓	✓	✓	✓	✓	✓		✓	✓		0.0264	WNN		
34	✓	✓	✓	✓	✓		✓	✓	✓		0.0251	WNN		
35	✓	✓	✓	✓	✓		✓	✓	✓	✓	0.1984	WNN		
Data Decomposition – Seasons														
	LIM	FO	DS	RM	LLDF	PVC	NC	DTM	SD	RMSE	Algorithm			
36	✓	✓	✓	✓	✓	✓	✓	✓	Summer	0.0128	RQGPR			
37	✓	✓	✓	✓	✓	✓	✓	✓	Spring	0.0144	MGPR			
38	✓	✓	✓	✓	✓	✓	✓	✓	Autumn	0.0090	SEGPR			
39	✓	✓	✓	✓	✓	✓	✓	✓	Winter	0.0079	MGPR			
Data/Dimentionality Reduction – PCA														
	LIM	FO	DS	RM	LLDF	PVC	NC	DTM	PCA	RMSE	Algorithm			
40	✓	✓	✓	✓	✓	✓	✓	✓	PCA 11	0.0265	WNN			
41	✓	✓	✓	✓	✓	✓	✓	✓	PCA 9	0.0264	WNN			

(continued on next page)

Table 2 (continued)

Data Cleaning									
42	✓	✓	✓	✓	✓	✓	✓	PCA 8	WNN
43	✓	✓	✓	✓	✓	✓	✓	PCA 7	WNN
44	✓	✓	✓	✓	✓	✓	✓	PCA 5	FGSVM
Data Partitioning									
45	Validation		PCA	RMSE	Algorithm	PCA	RMSE	Algorithm	
46	Cross: 5	9	0.0280	0.0253	WNN	8	0.0253	WNN	
47	Cross: 10	9	0.0272	0.0254	WNN	8	0.0248	WNN	
48	Cross: 15	9	0.0265	0.0248	WNN	8	0.0268	WNN	
49	Cross: 20	9	0.0268	0.0253	WNN	8	0.0253	WNN	
50	Holdout: 10	9	0.0250	0.0248	WNN	8	0.0256	WNN	
51	Holdout: 15	9	0.0250	0.0256	WNN	8	0.0266	WNN	
52	Holdout: 20	9	0.0253	0.0266	WNN	8			
53	Holdout: 25	9							
Data Decomposition – SD									
Seasons			PCA	Validation	RMSE	Algorithm	Validation	RMSE	Algorithm
53	Spring	8	Cross: 15	Cross: 15	0.0142	MGPR	Holdout: 15	0.0115	MGPR
54	Summer	8	Cross: 15	Cross: 15	0.0117	RQGR	Holdout: 15	0.0116	RQGR
55	Autumn	8	Cross: 15	Cross: 15	0.0078	MGPR	Holdout: 15	0.0077	MGPR
56	Winter	8	Cross: 15	Cross: 15	0.0076	MGPR	Holdout: 15	0.0077	MGPR

Note: Bed = Bedford, BNN = Bilayered Neural Network, Granf. = Granfield, Cross. # = Cross-validation: Number of folds, CSVM = Coarse Gaussian Support Vector Machine, DF = Daylight-free, DO = Delete Outliers, DS = Data Smoothing, DTM = Day/Time/Month, FGSVM = Fine Gaussian Support Vector Machine, FO = Fill Outliers, Holdout = Holdout Validation: Percent hold out, LIM = Linear Interpolation Method, LLDF = Low-level Data Fusion, LWDM = List-wise Deletion Method, MGPR = Matern 5/2 Gaussian Process Regression, MGSVM = Medium Gaussian Support Vector Machine, MNN = Medium Neural Network, NC = Negative Correlation, NNN = Narrow Neural Network, PCA = Principal Component Analysis, PVC = Photovoltaic Correlation, RM = Rescaling Method, RMSE = Root Mean Square Error, RQGPR = Rational Quadratic Gaussian Process Regression, SD = Seasonal Decomposition, SEGPR = Squared Exponential Gaussian Process Regression, SRC = Solar Radiation Correlation, TNN = Trilayered Neural Network, WNN = Wide Neural Network, ZS = Z-Score.

60% to 80% and $k = 10$ folds for cross-validation.

5.2. Performance metrics

Several metrics are available to evaluate the performance of forecasting models, each designed to quantify the discrepancy between observed values and predicted outputs. These error measures provide insight into model accuracy and reliability by comparing actual data against forecasted results. In this study, the following error formula [91,92] are employed to assess model performance:

- I. Mean Square Error (MSE): Is the average of the squares of the errors or the sum of the average difference squared between predicted and actual values.

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{x}_i - x_i)^2 = RMSE^2 \quad (9)$$

- II. Mean Absolute Error (MAE): Measures the average absolute difference of the forecasted values from the actual values.

$$MAE = \frac{1}{N} \sum_{i=1}^N |(\hat{\mathbf{x}}_i - \mathbf{x}_i)| \quad (10)$$

- III. **Root Mean Square Error (RMSE):** Penalizes big forecast errors and is preferably used where small errors are acceptable [68]. It is calculated by taking the square root of the MSE. The lower the RMSE, the higher the prediction model accuracy.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{x}_i - x_i)^2} \quad (11)$$

- IV. R-Squared: The coefficient of determination or R-squared indicates the proportionate amount of variation in the dependent variable response explained by the independent variable. In other words, R-squared is a goodness of fit measure that shows how well the data fits the regression model [52].

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{x}_i - x_i)^2}{\sum_{i=1}^N (\hat{x}_i - \hat{x}_{average})^2} \quad (12)$$

Where N is the number of data points, $\hat{x}_i$ are the predicted values and x_i are the observed values.

Using a combination of MSE, MAE, RMSE, and R^2 provides a comprehensive evaluation of forecasting model performance. MSE is particularly useful due to its sensitivity to large errors, as squaring amplifies their impact. MAE, on the other hand, offers robustness by assigning equal weight to all errors and is easily interpretable since it is expressed in the same units as the original data. RMSE combines the interpretability of MAE with the error sensitivity of MSE, making it a widely accepted and versatile metric. Additionally, R^2 facilitates model comparison on the same dataset, with higher values indicating better fit. Collectively, these metrics capture both the magnitude and distribution of errors, ensuring a rigorous and reliable assessment of forecasting models for practical applications.

5.3. Ranking validation methods

- I. **Ablation Analysis (Drop-One Approach):** Ablation analysis evaluates the importance of individual components of a model or system by systematically removing one element at a time and measuring the change in a performance metric like accuracy or error rate [93].

$$\Delta_i = M_{optimal} - M_{(i)} \quad (13)$$

Table 3

Correlation matrix of weather Parameters in Cranfield and Bedford weather stations to ((a) Solar radiation (b) Photovoltaic generation).

SRC		PVC	
Solar Radiation	1	PV	1
Wind Speed	0.0175	Solar Radiation	0.9444
Relative Humidity	-0.5961	Wind Speed	0.0251
Pressure	0.1973	Relative Humidity	-0.5026
Dew Point	0.2025	Pressure	0.2144
Heat Index	0.5247	Dew Point	0.1511
Temperature	0.5464	Heat Index	0.4243
Wind Chill	0.5412	Temperature	0.4427
Hourly Pressure at Mean Sea Level	0.1902	Wind Chill	0.4383
Hourly Dewpoint Temperature	0.2014	Hourly Pressure at Mean Sea Level	0.2085
Hourly Rainfall Total	-0.1544	Hourly Dewpoint Temperature	0.1331
Hourly Relative Humidity	-0.6979	Hourly Rainfall Total	-0.1634
Hourly Mean Wind Direction	-0.0611	Hourly Relative Humidity	-0.6249
Hourly Mean Windspeed	-0.0127	Hourly Mean Wind Direction	-0.0458
Hourly Maximum Gust	0.0287	Hourly Mean Windspeed	-0.1091
Hourly Temperature	0.5437	Hourly Maximum Gust	0.0214
Hourly Wet Bulb Temperature	0.4092	Hourly Temperature	0.4443
Hourly Global Radiation	0.9504	Hourly Wet Bulb Temperature	0.3215
Hourly Total Cloud Cover	-0.2850	Hourly Global Radiation	0.8962
Hourly Visibility	0.2397	Hourly Total Cloud Cover	-0.3406
		Hourly Visibility	0.2017

Where: M_{optimal} is the performance metric of the complete or optimal model, $M_{(i)}$ is the performance metric after removing component, and Δ_i is the impact of component i .

II. Sequential Addition Analysis: Sequential addition analysis measures the incremental contribution of components by adding them one at a time to a baseline model and recording performance improvements at each step [94]. This approach highlights cumulative benefits and marginal gains, helping justify model complexity.

$$\Delta_i = M_{(1,\dots,i)} - M_{(1,\dots,i-1)} \quad (14)$$

Where $M_{(1,\dots,i)}$ and $M_{(1,\dots,i-1)}$ are the performances after and before adding component i respectively, and Δ_i is the marginal gain from component i .

6. Results and discussion

Table 2 presents the results for the systematic testing of DP methods, based on the lowest RMSE values and the best-performing algorithm for each trial. RMSE was selected as the evaluation metric between trials because it performs well on relatively clean datasets, effectively penalizing larger deviations while providing an interpretable measure of prediction accuracy. For clarity, the minimum RMSE within each DP method subgroup has been highlighted.

This work set out to quantify how integrated DP strategies influence PV forecasting accuracy by coupling a structured DP framework with an iterative order selection procedure. Baseline cleaning via LWDM and LIM yielded high RMSE values of 38.52 and 37.56, respectively, confirming limited benefit from minimal cleaning alone. These methods are considered the baseline scenarios, as outlined in Section 3.1.1, and serve as benchmarks for evaluating other DP techniques.

In contrast, DS consistently improved accuracy (e.g., RMSE $\approx$ 11.4 with LWDM, 11.5 with LIM; Trial 22: 0.0604), highlighting its ability to suppress short-term noise.

Normalization further reduced error (Trials 11–56 < Trials 1–10),

with RM outperforming ZS. LIM + DS + RM provided the best normalization-stage result (RMSE 0.0375), narrowly beating LWDM + DS + RM (0.0403), indicating that preserving continuity via interpolation is preferable to discarding observations, especially when dataset size is limited.

Merging data from relevant stations with LLDF (Trials 26–32) improved performance over single-source inputs (Trials 20–25), underscoring the value of spatial diversity; within these setups, FO surpassed DO (RMSE 0.0295 vs 0.0379).

Subsequent tests explored data reduction through correlation-based feature selection. In the original three-step forecasting architecture, features were selected based on moderate-to-strong positive correlation ($r > 0.4$) with SR. In this study, we investigate the effect of replacing SR with PV generation as the reference variable and incorporating weather parameters exhibiting strong negative correlation. PCC analysis was applied to data cleaned using LIM, as defined in Equation (2). The results of this analysis are presented in Table 3.

As shown in Table 3, correlation-based feature reduction cut features to 9 (SRC) and 8 (PVC). Using PV as the reference (PVC) performed similarly to SR-based selection (SRC: RMSE 0.0417 vs PVC: 0.054), but adding negatively correlated predictors (e.g., relative humidity) improved both (SRC: 0.0362; PVC: 0.0365) highlighting the importance of capturing inverse relationships in PV generation dynamics. Across Trials 20–32, DS, LLDF, and LIM remained consistently beneficial.

Augmenting predictors with timestamp features (DTM) improved both PVC and SRC (PVC + DTM: 0.0251; SRC + DTM: 0.0264) in Trials 33–34. Removing zero-production (DF) hours degraded accuracy (RMSE 0.1984), suggesting those periods encode essential temporal structure (e.g., sunrise and sunset) that aid training and validation.

SD was evaluated by segmenting the dataset into Spring (March 20–June 20), Summer (June 21–September 21), Autumn (September 22–December 20), and Winter (December 21–March 19) [95], producing the lowest errors overall (Trials 36–38). This likely reflects that homogeneous training/testing windows simplify the mapping between inputs and outputs.

The effect of PCA was investigated by varying the number of retained components. With PVC + DTM, the feature count increased from 8 to 11. Using Equations (3)–(6), Trials 40–44 evaluated PCA configurations from 11 down to 5 components. Trial 42 (PCA-8) achieved the lowest RMSE = 0.0253, with PCA-9 close behind at 0.0264. One possible explanation is that, among the 11 features, certain variables (relative humidity, temperature, SR) were represented twice; compressing these duplications into 8 components may preserve salient variance while removing redundancy. Importantly, there was no clear advantage over Trial 34 (without PCA), potentially due to limited sample size or nonlinear structure in the data [96].

Validation comparisons showed k-fold ($k = 15$) and holdout (15%) achieved similar accuracy (both RMSE 0.0248). While resubstitution tends to create bias and overfitting in smaller data samples, cross-validation is preferable for small samples and holdout for larger ones. SD was re-examined under cross-validation vs. holdout (Trials 53–56). Holdout consistently outperformed cross-validation, producing the lowest RMSE values across all 56 trials. The optimal sequence of DP methods (LLDF + LIM + FO + DS + RM + PVC + NC + DTM + PCA-8 + Holdout:15%) achieved an RMSE = 0.0248. The full flow chart is depicted in Fig. 5 and highlighted as the critical path for implementation.

Table 2 indicates a consistent reduction in RMSE as additional DP methods are applied. To contextualize these gains, we contrasted the optimal DP sequence with two baselines: (a) a minimal cleaning base-case scenario, and (b) the established three step PV model. Comparing LWDM (RMSE = 38.52) and LIM (RMSE = 37.56), the base case scenarios, to the optimal DP combination (RMSE = 0.0248) demonstrates near-complete improvement in forecasting accuracy, confirming that integrated DP approaches significantly enhance performance. In the second scenario, the impact of this configuration is further assessed by

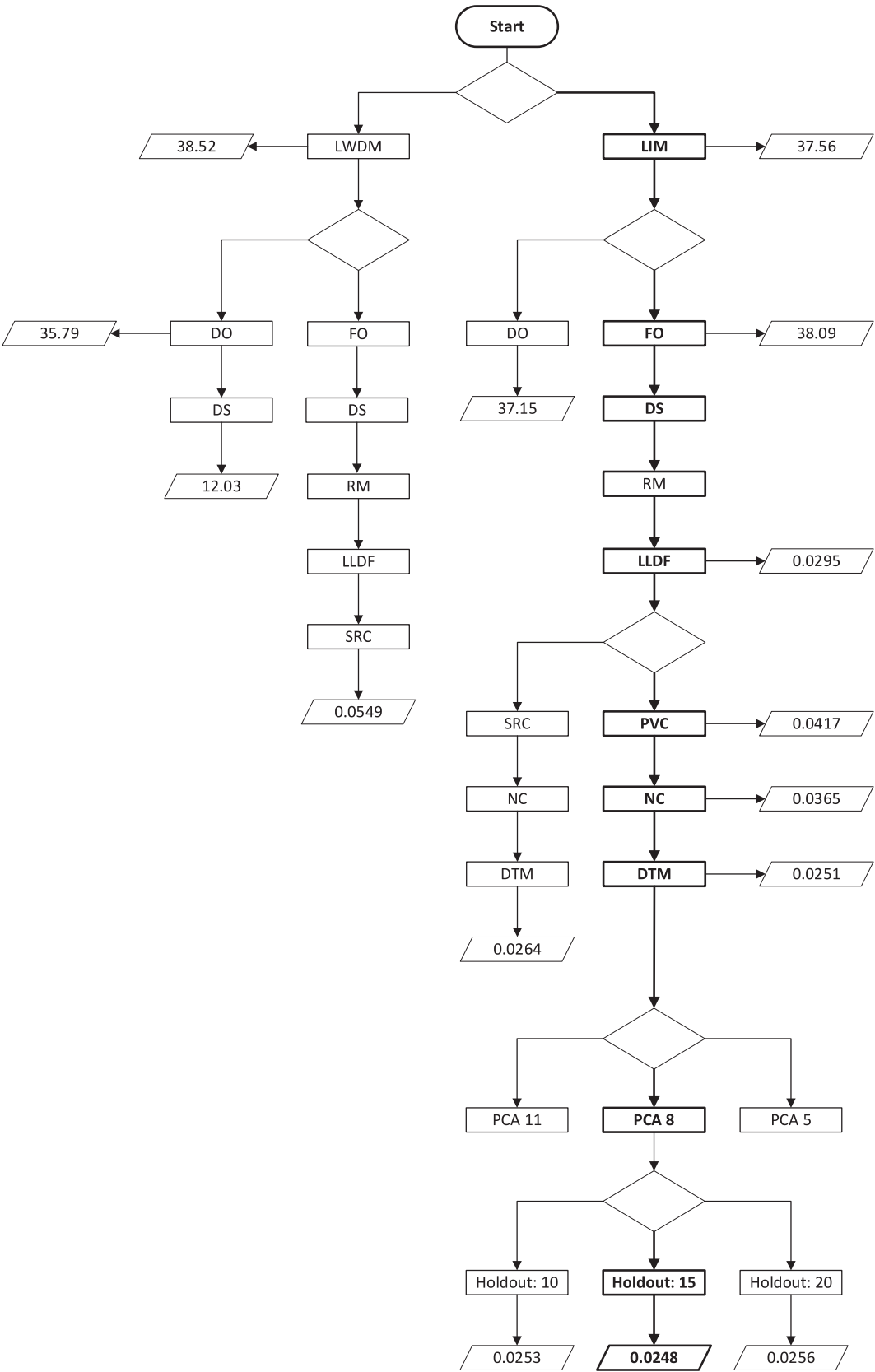


Fig. 5. Flow chart summary of data preprocessing iterations.

Table 4

Comparison of original Three-Step to Optimized model.

Three-Step Model	Training RMSE	MSE	MAE	R ²	Testing RMSE	MSE	MAE	R ²
MNN	0.0914	0.0084	0.0491	0.86	0.0989	0.0098	0.0531	0.84
Optimized Model	Training RMSE	MSE	MAE	R ²	Testing RMSE	MSE	MAE	R ²
WNN	0.0254	0.0006	0.0181	0.98	0.0248	0.0006	0.0176	0.98

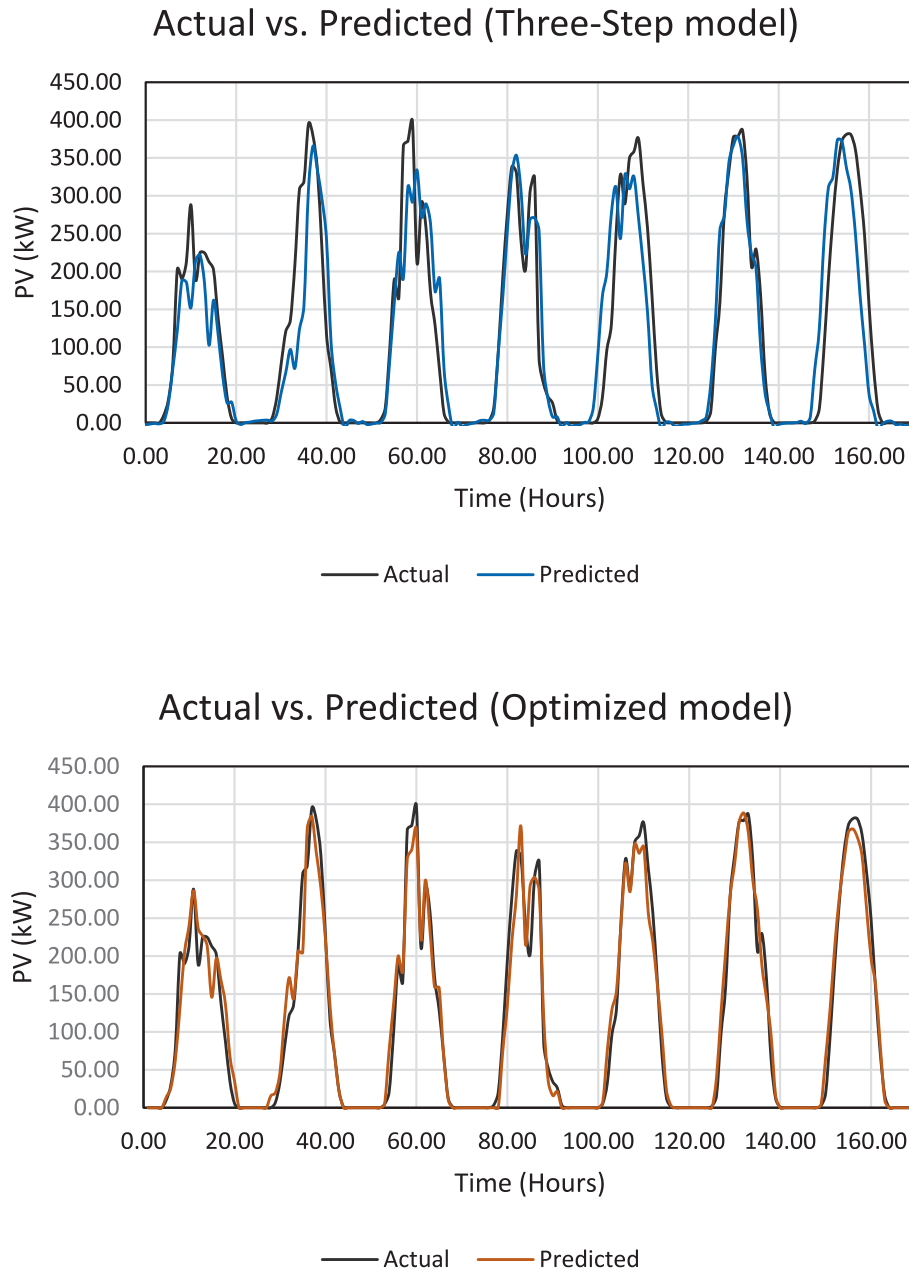


Fig. 6. One-week prediction comparison between Three-step and Optimized model.

contrasting the three-step PV forecasting architecture (LLDF + LWDM + RM + SRC) with the optimized model (LLDF + LIM + FO + DS + RM + PVC + NC + DTM + PCA8 + Holdout:15). Results are presented in Table 4.

The application of DP methods reduces RMSE by 75%, from 0.0989 to 0.0248, across the full test sample, indicating a substantial improvement in model accuracy. Furthermore, the optimized model

achieves an R^2 value of 0.98 compared to 0.84 for the original three-step model, demonstrating that predicted outputs closely align with actual PV generation. To illustrate this performance gain, Fig. 6 presents a 7-day comparison of actual versus predicted values for both models.

The optimized model provides a closer approximation of actual PV generation, including peak values, compared to the original architecture, underscoring the importance of accurate data in forecasting.

Table 5

Results for datapreprocessing ranking approach.

DP Ranking (with Holdout:15)			LIM	DTM	LLDF	RM	DS	FO	PCA	SRC	NC
Data Type	DP Methods		1	2	3	4	5	6	7	8	9
Raw Data	LIM	RMSE	38.45								
	LWDM		38.96		36.06						
	LLDF			34.99	33.21						
	DTM			34.70							
Preprocessed Data	DF	RMSE		47.72	42.27	38.78	0.0951				
	FO			39.31	36.24	33.58	0.0794	0.0156			
	DO			39.34	34.72	32.36	0.0788	0.0166			
	DS			13.18	6.331	6.712	0.0328				
	RM			0.0936	0.0886	0.0786					
	ZS			0.3954	0.3694	0.3390					
	PVC			41.60	38.49	33.82	0.0782	0.0317	0.0315	0.0402	
	PVC + NC			41.74	39.52	32.39	0.0755	0.0218	0.0225		
	SRC			41.60	39.42	32.44	0.0846	0.0250	0.0272	0.0144	
	SRC + NC			41.74	38.89	31.91	0.0811	0.0174	0.0196		
	NC			12.34	81.46	44.66	0.1105	0.0700	0.0683	0.0782	0.0285
	PCA			43.27	55.54	55.75	0.0783	0.0157	0.0133		

Note: DF = Daylight-free, DO = Delete Outliers, DS = Data Smoothing, DTM = Day/Time/Month, FO = Fill Outliers, LIM = Linear Interpolation Method, LWDM = List-wise Deletion Method, NC = Negative Correlation, PCA = Principal Component Analysis, PVC = Photovoltaic Correlation, RM = Rescaling Method, RMSE = Root Mean Square Error, SRC = Solar Radiation Correlation, ZS = Z-Score, Holdout:15 = Holdout Validation: 15%t hold out.

Table 6

Ablation (Drop-one) analysis.

DP Method	RMSE with method OFF	Δ (Contribution) = ALL_ON – OFF	Normalized (%)
LIM	0.1074	−0.0941	0.0016
DTM	0.0441	−0.0308	0.0005
LLDF	0.0512	−0.0379	0.0007
RM	57.03	−57.019	0.9960
DS	0.0748	−0.0615	0.0011
FO	0.0157	−0.0024	0.00004
PCA	0.0156	−0.0023	0.00004
		Δ Sum	−57.25
DP Method RMSE with ALL methods ON			
All Methods	0.0133		

For the model-validation phase, the optimized model was then applied to a separate one-year training and testing dataset to enable cross-scenario validation, resulting in an RMSE of 0.02195. This indicates that the optimized model transfers well across datasets. To validate the optimized model, a dedicated DP ranking procedure was then applied using a performance-guided, sequential selection process. First, the LIM and LWDM methods were evaluated on the raw incomplete dataset, and the method yielding the lower RMSE was fixed as the initial step in the sequence. Next, the remaining DP techniques were applied to the newly preprocessed dataset, and the method producing the lowest RMSE was added to the sequence. This iterative procedure

continued until all DP methods had been assessed and an ordered ranking was established. A summary of results is illustrated in Table 5.

In Table 5, LIM outperforms LWDM on raw data and with data enrichment steps LLDF and DTM. The strongest performance gains emerge once RM is applied, after which DS and FO provide reliable incremental improvements. PCA yields small but consistent final-stage refinement, whereas correlation-based methods (PVC, SRC, NC) offer modest, order-sensitive gains when used after the core cleaning is completed. The lowest RMSE achieved in the pipeline was 0.0133 and will mark the last DP in the sequence. Therefore, the optimized model based on DP ranking is: LIM + DTM + LLDF + RM + DS + FO + PCA + Holdout: 15. The optimized model with DP ranking (RMSE = 0.0133) outperforms the original optimized model (RMSE = 0.0220).

To evaluate the marginal and sequential contributions of the DP methods, the study applies the ranking-based validation procedures described in Section 5.3 to the validation dataset. This evaluation is implemented through an ablation framework consisting of a 10-run design: one ALL-ON baseline and nine single-factor-OFF ablation runs. Results from the ablation analysis in Table 6 show that the ALL-ON configuration achieved an RMSE of 0.0133. Turning RM OFF increased RMSE to 57.03, a deterioration of −57.02 (defined as ALL-ON minus OFF), accounting for ~99.60% of the total absolute ablation magnitude. This single step dominates all other error changes and establishes range normalization as a non-negotiable precondition for achieving low error in this pipeline. FO and PCA exhibit comparatively small ablation deltas: −0.0024 and −0.0023, respectively (each < 0.005% of the total ablation magnitude). This suggests that, once data are appropriately scaled and basic temporal/interpolation/smoothing logic is applied, outlier filling and dimensionality reduction provide only incremental improvements for this dataset and model.

The order-specific “Sequential Addition” experiment highlighted in

Table 7

Sequential addition analysis.

Step	Method added at step	RMSE after step	Prev RMSE	Δ (Prev – Current)	% change vs prev	Cumulative improvement vs Step 3	Normalized Contribution
1	LIM	33.45				−0.2350	
2	DTM	34.70	33.45	−1.256	−0.0376	−1.491	−0.0376
3	LLDF	33.21	34.70	1.491	0.0430	0.0000	0.0446
4	RM	0.0786	33.21	33.13	0.9976	33.13	0.9910
5	DS	0.0328	0.0786	0.0457	0.5820	33.18	0.0014
6	FO	0.0156	0.0328	0.0173	0.5262	33.19	0.0005
7	PCA	0.0133	0.0156	0.0023	0.1451	33.20	0.0001

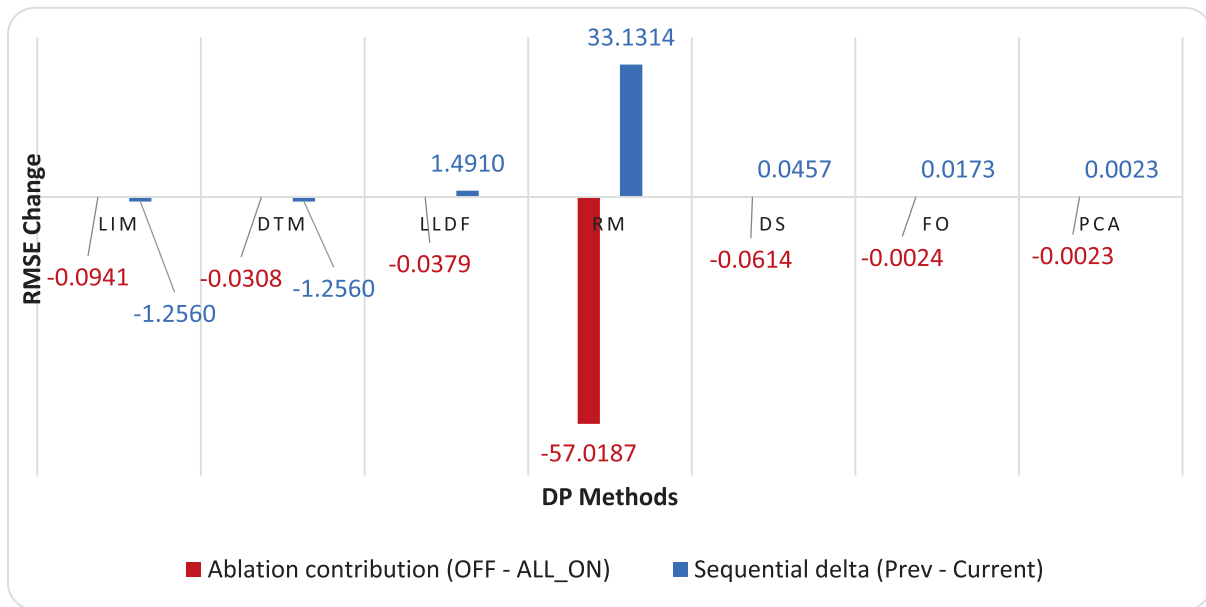


Fig. 7. Ablation vs sequential contributions.

Table 7 is consistent with the ablation pattern. From a high-error baseline, the error decreases marginally in Trials 1–3, but RM produces the major inflection, cutting RMSE from 33.21 to 0.0786 ($\Delta = 33.13$). Subsequent steps yield diminishing returns: DS from 0.0786 to 0.0328, FO from 0.0328 to 0.0156, and PCA from 0.0156 to 0.0133. The culminating ALL-ON RMSE (0.0133) matches the ablation baseline, reinforcing internal consistency. Fig. 7 shows the combined ablation and sequential analysis.

In summary, the consistent pattern across phases is to impute rather than delete (LIM over LWDM), encode time early (DTM) to expose daily and seasonal patterns, fuse sources (LLDF) when available, and normalize before downstream steps (RM), followed by smoothing (DS), outlier filling (FO), and optional PCA as an end-stage redundancy reducer. Finally, although combining multiple DP methods can greatly improve forecasting accuracy, their performance depends on factors such as parameter tuning, data quality, and model-specific biases. The sensitivity of DP choices is evident in Appendices A1–A4: using different DP methods during training and testing increased RMSE (A1); removing DS and DO or altering PPC thresholds reduced accuracy (A2); and experiments with smaller training sets showed that better-quality data lead to stronger predictions (A4).

7. Conclusion

This study demonstrates that data preprocessing (DP) plays a decisive role in improving the accuracy of statistical PV forecasting models. By introducing a structured classification framework and systematically evaluating sixteen (16) DP techniques—both individually and in combination—this work addresses a critical gap in existing literature, which often lacks comprehensive categorization and empirical analysis of integrated DP strategies.

Methodologically, the paper contributes two complementary perspectives. First, the experimental phase show that combining multiple DP methods (RMSE = 0.0248) significantly enhances forecasting performance compared to both a base-case minimal cleaning scenario (RMSE = 37.56) and an established three-step model (RMSE = 0.0989), underscoring the importance of method synergy rather than isolated application. Second, the order-and-validation phase improved the performance from the experimental phase by applying a dedicated DP ranking procedure that sequentially identified the best preprocessing steps, achieving an RMSE of 0.0133 on the validation dataset.

Ablation (single-factor-off) and sequential-addition analyses corroborate this ordering and isolate rescaling method as the dominant inflection point for error reduction—both when added (massive RMSE drop) and when removed (catastrophic error rise).

Beyond empirical findings, this study provides a reproducible workflow and practical guidance for researchers seeking to optimize PV forecasting pipelines. The classification framework offers a foundation for scoping additional DP techniques, while the comparative analysis serves as a benchmark for future studies.

This study has two primary limitations. Firstly, although the optimized preprocessing pipeline demonstrated strong performance and transferred well to an independent one-year dataset, the study remains limited by the relatively narrow geographical and temporal scope of the available data. As shown in the Appendix, the model's performance varies noticeably when trained on shorter datasets, indicating sensitivity to data volume and diversity. To fully assess potential overfitting and strengthen the generalizability of the proposed forecasting framework, future research should incorporate external multi-site datasets collected under different climatic and operational conditions. Expanding the dataset will enable more rigorous out-of-sample validation and provide stronger evidence of the model's robustness across broader contexts.

Secondly, the set of DP methods analysed in this study is not exhaustive. Some techniques were excluded due to irrelevance to the statistical PV model or research constraints, leaving room for future studies to incorporate additional preprocessing strategies and assess their impact. Further research should explore the scalability of combined DP strategies across different climates, PV technologies, and hybrid forecasting approaches, including machine learning and deep learning models. Additionally, integrating automated DP selection and optimization algorithms could further streamline preprocessing and enhance forecasting accuracy. Another important direction is the systematic tuning and sensitivity analysis of DP parameters—for example, interpolation settings, smoothing window sizes, outlier-filling thresholds, or principal component analysis (PCA) variance levels—as small changes in these parameters can materially influence forecasting performance and interact with model behaviour in non-linear ways. By bridging the gap between theory and practice, this study lays the groundwork for more robust, data-driven PV forecasting systems.

CRedit authorship contribution statement

Tolulope Olumuyiwa Falope: . **Liyun Lao:** Writing – review & editing, Supervision. **Da Huo:** Writing – review & editing, Supervision.

Declaration of competing 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.

Acknowledgement

This research was supported by the Petroleum Technology Development Fund (PTDF), Nigeria; [PTDF/ED/OSS/PHD/TOF/1945/20].

Appendix A**Table A1**

Performance results of applying separate data preprocessing methods to training and testing data sets.

Data Cleaning									
Test	LWDM	LIM	DO	FO	DS	RMSE	Algorithm		
1	✓					37.13	BOT		
2	✓		✓			37.10	BOT		
3	✓			✓		37.22	BOT		
4	✓				✓	40.36	BAT		
5	✓		✓		✓	40.43	BAT		
6	✓			✓	✓	40.28	BAT		
7		✓				42.16	LR		
8		✓	✓			37.08	BOT		
9		✓		✓		37.19	BOT		
10		✓			✓	40.38	BAT		
11		✓	✓		✓	40.31	BAT		
12		✓		✓	✓	40.34	BAT		
Data Scaling									
	LWDM	LIM	DO	FO	ZS	RM	DS	RMSE	Algorithm
*13	✓				✓			107.5	MT
*14	✓					✓		108.6	MT
*15	✓		✓		✓			107.7	CT
*16	✓		✓			✓		108.6	MT/CT
*17	✓			✓	✓			107.5	MT
*18	✓			✓		✓		108.6	MT
19	✓				✓			0.4010	BAT
20	✓					✓		0.0927	BOT
21	✓		✓		✓			0.4035	BAT
22	✓		✓			✓		0.1257	LR
23	✓			✓	✓			0.3990	BAT
24	✓			✓		✓		0.0927	BOT
25	✓					✓	✓	0.0989	BAT

Note: BAT = Bagged Trees, BOT = Boosted Trees, CT = Coarse Tree, DO = Delete Outliers, DS = Data Smoothing, FO = Fill Outliers, LIM = Linear Interpolation Method, LR = Linear Regression

LWDM = List-wise Deletion Method, MT = Medium Tree, RM = Rescaling Method, RMSE = Root Mean Square Error, ZS = Z-Score.

*# represent trials that used LIM and RM for the test data set while using other data processing methods for the training data set.

Table A2

Performance results of data preprocessing methods with different Pearsons Correlation Coefficients.

PCC > 0.5									
Data Integration and data reduction									
	LIM	SRC	RM	ZS	Cranf	Bed	Cranf + Bed	RMSE	Algorithm
1	✓	✓	✓		✓			0.1039	TNN
2	✓	✓	✓			✓		0.1216	MNN
3	✓	✓	✓				✓	0.0965	BNN
4	✓	✓	✓				✓	0.4084	MNN
5	✓	✓		✓					
Data Decomposition – Day, Time, Month, Daylight-free									
	LIM	SRC	RM	Cranf + Bed	DTM	DF	RMSE	Algorithm	
6	✓	✓	✓	✓	✓		0.0856	BNN	
7	✓	✓	✓	✓	✓	✓	0.1627	MGPR	
8	✓	✓	✓	✓		✓	0.1744	SEGPR	
Data Decomposition – Seasons									
	LIM	SRC	RM	Cranf + Bed	DTM	Season	RMSE	Algorithm	
9	✓	✓	✓	✓		Spring	0.1161	SEGPR	
10	✓	✓	✓	✓	✓	Spring	0.0945	SEGPR	
11	✓	✓	✓	✓	✓	Summer	0.0850	MGPR	
12	✓	✓	✓	✓	✓	Autumn	0.0756	BAT	

(continued on next page)

Table A2 (continued)

13	✓	✓	✓	✓	✓	Winter	0.0811	QSVM		
PCC > 0.4										
Data Integration and Data Reduction										
	LIM	SRC	RM	Cranf + Bed	NC	RMSE	Algorithm			
14	✓	✓	✓	✓	✓	0.0818	TNN			
15	✓	✓	✓	✓	✓	0.0832	MNN			
Data Partitioning										
	LIM	SRC	RM	Cranf + Bed	NC	DTM	Validation	RMSE	Algorithm	
16	✓	✓	✓	✓	✓	✓	Cross: 5	0.0854	BNN	
17	✓	✓	✓	✓	✓	✓	Cross: 10	0.0851	TNN	
18	✓	✓	✓	✓	✓	✓	Cross: 15	0.0840	BNN	
19	✓	✓	✓	✓	✓	✓	Cross: 20	0.0834	BNN	
20	✓	✓	✓	✓	✓	✓	Cross: 40	0.0852	TNN	
21	✓	✓	✓	✓	✓	✓	Holdout: 10	0.0840	MNN	
22	✓	✓	✓	✓	✓	✓	Holdout: 20	0.0836	WNN	
23	✓	✓	✓	✓	✓	✓	Holdout: 30	0.0817	BNN	
24	✓	✓	✓	✓	✓	✓	Holdout: 40	0.0779	BNN	
Data Reduction – PCA										
	LIM	SRC	RM	Cranf + Bed	NC	DTM	PCA	Validation	RMSE	Algorithm
25	✓	✓	✓	✓	✓	✓	10	Resubstituion	0.0860	MNN
26	✓	✓	✓	✓	✓	✓	10	Holdout: 40	0.0840	BNN
27	✓	✓	✓	✓	✓	✓	8	Holdout: 40	0.0818	TNN
28	✓	✓	✓	✓	✓	✓	6	Holdout: 40	0.0873	TNN
29	✓	✓	✓	✓	✓	✓	4	Holdout: 40	0.1011	BNN
Data Decomposition – Seasons										
	LIM	SRC	RM	Cranf + Bed	NC	DTM	Season	Validation	RMSE	Algorithm
30	✓	✓	✓	✓	✓	✓	Spring	Resubstituion	0.0983	EGPR
31	✓	✓	✓	✓	✓	✓	Summer	Resubstituion	0.0879	SEGPR
32	✓	✓	✓	✓	✓	✓	Autumn	Resubstituion	0.0649	NNN
33	✓	✓	✓	✓	✓	✓	Winter	Resubstituion	0.0691	EGPR
34	✓	✓	✓	✓	✓	✓	Spring	Holdout: 40	0.0876	EGPR
35	✓	✓	✓	✓	✓	✓	Summer	Holdout: 40	0.0892	MGPR
36	✓	✓	✓	✓	✓	✓	Autumn	Holdout: 40	0.0688	EGPR
37	✓	✓	✓	✓	✓	✓	Winter	Holdout: 40	0.0667	NNN
Overall Performance										
	LIM	SRC	RM	Cranf + Bed	NC	DTM	PCA	Validation	RMSE	Algorithm
38	✓	✓	✓	✓	✓	✓	8	Resubstituion	0.0813	BAT
39	✓	✓	✓	✓	✓	✓	8	Holdout: 40	0.0837	TNN
	LIM	SRC	RM	Cranf + Bed	DTM	Season	RMSE	Algorithm		
40	✓	✓	✓	✓	✓	Spring	0.1161	SEGPR		
41	✓	✓	✓	✓	✓	Spring	0.0945	SEGPR		

Note: BAT = Bagged Trees, Bed = Bedford, BNN = Bilayered Neural Network, Cranf. = Cranfield, Cross: # = Cross-validation: Number of folds, DF = Daylight-free, DTM = Day-Time-Month, Holdout = Holdout Validation: Percent hold out, LIM = Linear Interpolation Method, LWDM = List-wise Deletion Method, MGPR = Matern 5/2 Gaussian Process Regression, MNN = Medium Neural Network, NC = Negative Correlation, NNN = Narrow Neural Network, PCA = Principal Component Analysis, QSVM = Quadratic Support Vector Machine, RM = Rescaling Method, RMSE = Root Mean Square Error, SEGPR = Squared Exponential Gaussian Process Regression, SRC = Solar Radiation Correlation, Trilayered Neural Network, WNN = Wide Neural Network, ZS = Z-Score.

Table A3

Performance results of data preprocessing methods.

Data Cleaning, Data Scaling, Data reduction, and Data Integration											
	LWDM	LIM	FO	DS	RM	SRC	Cranf	Bed	Cranf + Bed	RMSE	Algorithm
1	✓		✓		✓	✓	✓			0.1042	NNN
2	✓		✓		✓	✓		✓		0.1073	MNN
3	✓		✓		✓	✓			✓	0.0928	BNN
4	✓		✓	✓	✓	✓	✓			0.0604	WNN
5	✓		✓	✓	✓	✓		✓		0.0688	FGSVM
6	✓		✓	✓	✓	✓			✓	0.0549	FGSVM
7		✓	✓		✓	✓	✓			0.1037	BNN
8		✓	✓		✓	✓		✓		0.1049	TNN
9		✓	✓		✓	✓			✓	0.0946	MNN
10		✓	✓	✓	✓	✓			✓	0.0547	WNN
Data Scaling											
	LIM	FO	DS	ZS	RM	SRC	PVC	Cranf + Bed	RMSE	Algorithm	
11	✓	✓	✓	✓		✓		✓	0.2552	WNN	
12	✓	✓	✓		✓	✓		✓	0.0547	WNN	
13	✓	✓	✓		✓		✓	✓	0.0543	MNN	
Data Decomposition											
	LIM	FO	DS	RM	Cranf + Bed	PVC	DTM	NC	DF	RMSE	Algorithm
14	✓	✓	✓	✓	✓	✓	✓			0.0283	FGSVM
15	✓	✓	✓	✓	✓	✓	✓	✓		0.0268	WNN
16	✓	✓	✓	✓	✓	✓	✓	✓	✓	0.1984	WNN
Data Decomposition – Seasons											
	LIM	FO	DS	RM	Cranf + Bed	PVC	DTM	NC	Season	RMSE	Algorithm
17	✓	✓	✓	✓	✓	✓	✓	✓	Summer	0.0128	RQGPR

(continued on next page)

Table A3 (continued)

18	✓	✓	✓	✓	✓	✓	✓	✓	Spring	0.0144	MGPR	
19	✓	✓	✓	✓	✓	✓	✓	✓	Autumn	0.0090	SEGPR	
20	✓	✓	✓	✓	✓	✓	✓	✓	Winter	0.0079	MGPR	
Data/Dimentionality Reduction – PCA												
	LIM	FO	DS	RM	Cranf + Bed	PVC	DTM	NC	PCA	RMSE	Algorithm	
21	✓	✓	✓	✓	✓	✓	✓	✓	PCA 11	0.0265	WNN	
22	✓	✓	✓	✓	✓	✓	✓	✓	PCA 9	0.0264	WNN	
23	✓	✓	✓	✓	✓	✓	✓	✓	PCA 8	0.0133	WNN	
24	✓	✓	✓	✓	✓	✓	✓	✓	PCA 7	0.0276	WNN	
25	✓	✓	✓	✓	✓	✓	✓	✓	PCA 5	0.0522	FGSVM	
Data Partitioning												
	LIM	FO	DS	RM	Cranf + Bed	PVC	DTM	NC	PCA	Validation	RMSE	Algorithm
26	✓	✓	✓	✓	✓	✓	✓	✓	9	Cross: 5	0.0280	MGSVM
27	✓	✓	✓	✓	✓	✓	✓	✓	9	Cross: 10	0.0272	WNN
28	✓	✓	✓	✓	✓	✓	✓	✓	9	Cross: 15	0.0265	WNN
29	✓	✓	✓	✓	✓	✓	✓	✓	9	Cross: 20	0.0268	WNN
30	✓	✓	✓	✓	✓	✓	✓	✓	9	Holdout: 10	0.0269	WNN
31	✓	✓	✓	✓	✓	✓	✓	✓	9	Holdout: 20	0.0265	WNN
32	✓	✓	✓	✓	✓	✓	✓	✓	9	Holdout: 30	0.0272	WNN
33	✓	✓	✓	✓	✓	✓	✓	✓	9	Holdout: 40	0.0263	WNN
Data Decomposition												
	Seasons	Validation	PCA	RMSE	Algorithm	PCA	RMSE	Algorithm				
34	Summer	Holdout: 40	9	0.0112	RQGPR	8	0.0096	RQGPR				
35	Spring	Holdout: 40	9	0.0103	RQGPR	8	0.0108	RQGPR				
36	Autumn	Holdout: 40	9	0.0055	MGPR	8	0.0075	MGPR				
37	Winter	Holdout: 40	9	0.0064	RQGPR	8	0.0072	MGPR				

Note: Bed = Bedford, BNN = Bilayered Neural Network, Cranf. = Cranfield, Cross: # = Cross-validation: Number of folds, DF = Daylight-free, DO = Delete Outliers, DS = Data Smoothing, DTM = Day-Time-Month, FGSVM = Fine Gaussian Support Vector Machine, FO = Fill Outliers, Holdout = Holdout Validation: Percent hold out, LIM = Linear Interpolation Method, LWDM = List-wise Deletion Method, MGPR = Matern 5/2 Gaussian Process Regression, MGSVM = Medium Gaussian Support Vector Machine, MNN = Medium Neural Network, NC = Negative Correlation, NNN = Narrow Neural Network, PCA = Principal Component Analysis, PVC = Photovoltaic Correlation, RM = Rescaling Method, RMSE = Root Mean Square Error, RQGPR = Rational Quadratic Gaussian Process Regression, SEGPR = Squared Exponential Gaussian Process Regression, SRC = Solar Radiation Correlation, TNN = Trilayered Neural Network, WNN = Wide Neural Network, ZS = Z-Score.

Table A4

Performance results of data preprocessing methods using only one year of data.

LWMD Performance											
	LWMD	SRC	PVC	RM	Cranf + Bed	NC	Validation	RMSE	Algorithm		
1	✓	✓		✓	✓		Resubstitution	0.0859	BOT		
2	✓		✓	✓	✓		Resubstitution	0.0864	BOT		
3	✓	✓		✓	✓	✓	Resubstitution	0.0858	BOT		
4	✓		✓	✓	✓	✓	Resubstitution	0.0865	BOT		
5	✓	✓		✓	✓		Cross: 5	0.0859	BOT		
6	✓	✓		✓	✓	✓	Cross: 10	0.0859	BOT		
7	✓	✓		✓	✓	✓	Cross: 15	0.0859	BOT		
8	✓	✓		✓	✓	✓	Cross: 20	0.0859	BOT		
9	✓	✓		✓	✓	✓	Holdout: 10	0.0859	BOT		
10	✓	✓		✓	✓	✓	Holdout: 20	0.0859	BOT		
11	✓	✓		✓	✓	✓	Holdout: 30	0.0859	BOT		
12	✓	✓		✓	✓	✓	Holdout: 40	0.0859	BOT		
Data Reduction – PCA											
	Validation		PCA		RMSE	Algorithm	PCA	RMSE	Algorithm		
13	Resubstitution		95%		0.1807	RL	7	0.0103	RL		
14	Resubstitution		90%		0.1807	RL	5	0.1010	RL		
15	Resubstitution		75%		0.1509	MGSVM	3	0.1023	RL		
16	Resubstitution		50%		0.1509	MGSVM					

Note: Bed = Bedford, BOT = Boosted Trees, Cranf. = Cranfield, Cross: # = Cross-validation: Number of folds, Holdout = Holdout Validation: Percent hold out, LWDM = List-wise Deletion Method, NC = Negative Correlation, PCA = Principal Component Analysis, PVC = Photovoltaic Correlation, RL = Robust Linear, RM = Rescaling Method, RMSE = Root Mean Square Error, SRC = Solar Radiation Correlation

Data availability

Data will be made available on request.

References

- [1] U.K. Das, K.S. Tey, M. Seyedmahmoudian, S. Mekhilef, M.Y.I. Idris, W. van Deventer, et al., Forecasting of photovoltaic power generation and model optimization: a review, *Renew. Sustain. Energy Rev.* 81 (2018) 912–928, <https://doi.org/10.1016/J.RSER.2017.08.017>.
- [2] B. Pan, K. Hirota, Z. Jia, Y. Dai, A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods, *Neurocomputing* 561 (2023) 126866, <https://doi.org/10.1016/J.NEUCOM.2023.126866>.
- [3] G.K. Adari, M. Raja, P. Vijaya, Machine learning in genomics: identification and modeling of anticancer peptides, *Data Science for Genomics* (2023) 25–68, <https://doi.org/10.1016/B978-0-323-98352-5.00007-0>.
- [4] H. Liu, C. Chen, Data processing strategies in wind energy forecasting models and applications: a comprehensive review, *Appl. Energy* 249 (2019) 392–408, <https://doi.org/10.1016/J.APENERGY.2019.04.188>.
- [5] H.T. Duong, T.A. Nguyen-Thi, A review: preprocessing techniques and data augmentation for sentiment analysis, *Comput. Soc. Netw.* 8 (2021) 1–16, <https://doi.org/10.1186/S40649-020-00080-X/FIGURES/4>.
- [6] Y. Jiao, Z. Li, X. Chen, S. Fei, Preprocessing methods for near-infrared spectrum calibration, *J. Chemom.* 34 (2020) e3306.

- [7] S. Albahra, T. Gorbett, S. Robertson, G. D'Aleo, S.V.S. Kumar, S. Ockunzzi, et al., Artificial intelligence and machine learning overview in pathology & laboratory medicine: a general review of data preprocessing and basic supervised concepts, *Semin. Diagn. Pathol.* 40 (2023) 71–87, <https://doi.org/10.1053/J.SEMDP.2023.02.002>.
- [8] García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: methods and prospects. *Big Data Analytics* 2016 1:1 2016;1:1–22. DOI: 10.1186/S41044-016-0014-0.
- [9] K.G. Ranjan, B.R. Prusty, D. Jena, Review of preprocessing methods for univariate volatile time-series in power system applications, *Electr. Pow. Syst. Res.* 191 (2021) 106885, <https://doi.org/10.1016/J.EPSR.2020.106885>.
- [10] Bramm A, Eroshenko S, Khalyasmaa A. Effect of Data Preprocessing on the Forecasting Accuracy of Solar Power Plant. 2021 18th International Scientific Technical Conference Alternating Current Electric Drives, ACED 2021 - Proceedings 2021. DOI: 10.1109/ACED50605.2021.9462288.
- [11] Q.T. Phan, Y.K. Wu, Q.D. Phan, H.Y. Lo, A Novel forecasting Model for Solar Power Generation by a Deep Learning Framework with Data Preprocessing and Postprocessing, *IEEE Trans. Ind. Appl.* 59 (2023) 220–231, <https://doi.org/10.1109/TIA.2022.3212999>.
- [12] Carrera B, Kim K. Comparison Analysis of Machine Learning Techniques for Photovoltaic Prediction Using Weather Sensor Data. *Sensors* 2020, Vol 20, Page 3129 2020;20:3129. DOI: 10.3390/S20113129.
- [13] Alshdaifat E, Alshdaifat D, Alsarhan A, Hussein F, El-Salhi SMFS. The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. *Data* 2021, Vol 6, Page 11 2021;6:11. DOI: 10.3390/DATA6020011.
- [14] M. Malvoni, M.G. de Giorgi, P.M. Congedo, Forecasting of PV Power Generation using weather input data-preprocessing techniques, *Energy Procedia* 126 (2017) 651–658, <https://doi.org/10.1016/J.EGYPRO.2017.08.293>.
- [15] N.A.M. Asma, Machine Learning Models, Data Preprocessing Techniques and Suite of Metrics for Assessing Solar Power forecasting: a Comprehensive Review. *Communications on Applied, Electronics* 8 (2025).
- [16] C. Fan, M. Chen, X. Wang, J. Wang, B. Huang, A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data, *Front. Energy Res.* 9 (2021), <https://doi.org/10.3389/FENERG.2021.652801/PDF>.
- [17] D.S. Lee, C.W. Lai, S.K. Fu, A short- and medium-term forecasting model for roof PV systems with data pre-processing, *Heliyon* 10 (2024) e27752, <https://doi.org/10.1016/j.heliyon.2024.e27752>.
- [18] Chiranjeevi M, Karlamangal S, Moger T, Jena D, Agarwal A. Preprocessing Techniques of Solar Irradiation Data. 2023 IEEE Renewable Energy and Sustainable E-Mobility Conference, RESEM 2023 2023. DOI: 10.1109/RESEM57584.2023.10236100.
- [19] T.A. Nguyen, C.K. Ha, View of effect of Data Preprocessing on Short-Term load forecasting: a Convolutional Neural Network Approach, *Journal of Robotics and Control* 6 (2025) 2349–2361, <https://doi.org/10.18196/jrc.v6i5.28399>.
- [20] W.G. Shin, J.S. Lee, Y.C. Ju, H.M. Hwang, S.W. Ko, Data preprocessing and machine learning method based on ameliorated mathematical models for inferring the power generation of photovoltaic system, *Energy Convers Manag* 333 (2025) 119793, <https://doi.org/10.1016/J.ENCONMAN.2025.119793>.
- [21] E.R.M. Shouman, E.R.M. Shouman, Solar Power Prediction with Artificial Intelligence, *Advances in Solar Photovoltaic Energy Systems* (2024), <https://doi.org/10.5772/INTECHOPEN.1002726>.
- [22] K. Manimekalai, A. Kavitha, Missing value imputation and normalization techniques in myocardial infarction. *ICTACT Journal on, Soft. Comput.* (2018) 1655–1662.
- [23] S. Roy, P. Sharma, K. Nath, D.K. Bhattacharyya, J.K. Kalita, Pre-processing: a data preparation step, *Encyclop Bioinform Comput Biol ABC Bioinform* 1 (2018) 1–5.
- [24] T.A. Alghamdi, N. Javadi, A Survey of Preprocessing Methods used for Analysis of big Data Originated from Smart Grids, *IEEE Access* 10 (2022) 29149–29171, <https://doi.org/10.1109/ACCESS.2022.3157941>.
- [25] Cranfield University. Cranfield University's Living Laboratory and Urban Observatory - Data. Living Laboratory 2021. <https://www.livinglab.ac.uk/data.html> (accessed February 21, 2022).
- [26] Cranfield University. Energy Manager Live 2022. <https://cranfield.energymanagerlive.com/> (accessed February 21, 2022).
- [27] Falope TO, Lao L, Hanak D. A Unique Three-Step Weather Data Approach in Solar Energy Prediction Using Machine Learning. *Sustainable Energy Solutions for a Post-COVID Recovery towards a Better Future: Part VII*, Bangkok: International Conference on Applied Energy; 2021, p. 1–7. DOI: DOI: 10.46855/energy-proceedings-9863.
- [28] Xiao J, Liu S, Peng S, Zhang W, Shan C, Xu Z, et al. Leveraging Pretrained Large Language Models for Power System Source-Load Forecasting. 2025 7th International Conference on Power and Energy Technology, ICPE 2025 2025: 171–5. DOI: 10.1109/ICPET66029.2025.11160348.
- [29] Benhar H, Idri A, L Fernández-Alemán J. Data preprocessing for heart disease classification: A systematic literature review. *Comput Methods Programs Biomed* 2020;195:105635. DOI: 10.1016/J.CMPB.2020.105635.
- [30] M. Pazhoohesh, M. Sadegh Javadi, M. Gheisari, S. Aziz, Dealing with Missing Data in the Smart buildings using innovative Imputation Techniques; dealing with Missing Data in the Smart buildings using innovative Imputation Techniques, in: IECON 2021–47th Annual Conference of the IEEE Industrial Electronics Society, 2021, <https://doi.org/10.1109/IECON48115.2021.9612650>.
- [31] F. Bashir, H.L. Wei, Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm, *Neurocomputing* 276 (2018) 23–30, <https://doi.org/10.1016/J.NEUCOM.2017.03.097>.
- [32] Ayilara OF, Tolulope -, Sajobi T, Barclay - Ruth, Bohm E, Mohammad -, et al. A comparison of methods to address item non-response when testing for differential item functioning in multidimensional patient-reported outcome measures. *Quality of Life Research* 2022 2022;1:1–12. DOI: 10.1007/S1136-022-03129-8.
- [33] J. Sophia Wang, P.M. Aronow, Listwise Deletion in High Dimensions, *Polit. Anal.* (2021), <https://doi.org/10.48550/10.1109/2021.11470>.
- [34] Ziyabari S, Du L, Biswas S. Short-term Solar Irradiance Forecasting Based on Multi-Branch Residual Network. *ECCE 2020 - IEEE Energy Conversion Congress and Exposition 2020:2000–5*. DOI: 10.1109/ECCE44975.2020.9235930.
- [35] M. Bencheikroun, B. Chevallier, V. Zalc, D. Istrate, D. Lenne, N. Vera, The Impact of Missing Data on Heart Rate Variability Features: a Comparative Study of Interpolation Methods for Ambulatory Health monitoring, *IRBM* 44 (2023) 100776, <https://doi.org/10.1016/J.IRBM.2023.100776>.
- [36] J. Kim, Y. Kwak, S.H. Mun, J.H. Huh, Imputation of missing values in residential building monitored data: Energy consumption, behavior, and environment information, *Build. Environ.* 245 (2023) 110919, <https://doi.org/10.1016/J.BUILDENV.2023.110919>.
- [37] C. Scott, M. Ahsan, A. Albarbar, Machine learning for forecasting a photovoltaic (PV) generation system, *Energy* 278 (2023) 127807, <https://doi.org/10.1016/J.ENERGY.2023.127807>.
- [38] J. Li, Z. Liu, Attribute-weighted outlier detection for mixed data based on parallel mutual information, *Expert Syst. Appl.* 236 (2024) 121304, <https://doi.org/10.1016/J.ESWA.2023.121304>.
- [39] H. Alimohammadi, C.S. Nancy, Performance evaluation of outlier detection techniques in production timeseries: a systematic review and meta-analysis, *Expert Syst. Appl.* 191 (2022) 116371, <https://doi.org/10.1016/J.ESWA.2021.116371>.
- [40] D. Olszewski, M. Iwanowski, W. Graniszewski, Dimensionality reduction for detection of anomalies in the IoT traffic data, *Futur. Gener. Comput. Syst.* 151 (2024) 137–151, <https://doi.org/10.1016/J.FUTURE.2023.09.033>.
- [41] S. Solorio-Fernández, J.A. Carrasco-Ochoa, J.F. Martínez-Trinidad, Filter unsupervised spectral feature selection method for mixed data based on a new feature correlation measure, *Neurocomputing* 571 (2024) 127111, <https://doi.org/10.1016/J.NEUCOM.2023.127111>.
- [42] I.K. Ihianle, P. Machado, K. Owa, D.A. Adama, R. Otuka, A. Lotfi, Minimising redundancy, maximising relevance: HRV feature selection for stress classification, *Expert Syst. Appl.* 239 (2024) 122490, <https://doi.org/10.1016/J.ESWA.2023.122490>.
- [43] Microsoft. PEARSON function. Microsoft Support 2021. <https://support.microsoft.com/en-us/office/pearson-function-0c3e30fc-e5af-49c4-808a-3ef66e034c18> (accessed October 15, 2021).
- [44] P. Singla, M. Duhan, S. Saroha, A comprehensive review and analysis of solar forecasting techniques, *Front. Energy* 2021 (2021) 1–37, <https://doi.org/10.1007/S11708-021-0722-7>.
- [45] Monfared M, Fazeli M, Lewis R, Searle J. Day-Ahead Prediction of PV Generation Using Weather Forecast Data: A Case Study in the UK. 2nd International Conference on Electrical, Communication and Computer Engineering, ICECCE 2020 2020. DOI: 10.1109/ICECCE49384.2020.9179454.
- [46] H. Eom, Y. Son, S. Choi, Feature-selective ensemble learning-based long-term regional PV generation forecasting, *IEEE Access* 8 (2020) 54620–54630, <https://doi.org/10.1109/ACCESS.2020.2981819>.
- [47] de Jongh S, Riedel T, Mueller F, Yacoub AE, Suriyah M, Leibfried T. Spatio-Temporal Short Term Photovoltaic Generation Forecasting with Uncertainty Estimates using Machine Learning Methods. *UPEC 2020 - 2020 55th International Universities Power Engineering Conference, Proceedings* 2020. DOI: 10.1109/UPEC49904.2020.9209764.
- [48] Masood N al, Asif MI, Alam AM, Deeba SR, Aziz T. Forecasting of Photovoltaic Power Generation: Techniques and Key Factors. *Proceedings of 2019 IEEE Region 10 Symposium, TENSYP 2019* 2019:457–61. DOI: 10.1109/TENSYP46218.2019.8971337.
- [49] J. Kong, S. Oh, J. Jung, I. Lee, Improving the accuracy of photovoltaic generation forecasting by considering particulate matter variables, *IEEE Power and Energy Society General Meeting (2020; 2020-August.)*, <https://doi.org/10.1109/PESGM41954.2020.9282054>.
- [50] Zeng J, Wang C, Hu X. Photovoltaic power generation prediction based on correlation analysis and GA - BP neural network. 2021 4th International Conference on Pattern Recognition and Artificial Intelligence, PRAI 2021 2021: 306–9. DOI: 10.1109/PRAI53619.2021.9551035.
- [51] L. Qin, W. Huang, Z. Guo, Y. Zhou, B. Zhang, Topology Identification Method of Low-Voltage Distribution Network Based on Improved Pearson Correlation Coefficient Method (2022) 1–6, <https://doi.org/10.1109/CYCEE53554.2021.9676722>.
- [52] M.D. Patel, S. Shah, S. Sinha, R.K. Gupta, S.K. Bharti, V.P. Singh, Solar Irradiation forecasting - Comparative Analysis of Various Methods, in: 2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES), 2021, <https://doi.org/10.1109/TRIBES52498.2021.9751626>.
- [53] Q. Xiong, H. Xiong, Q. Kong, X. Ni, Y. Li, C. Yuan, Machine learning-driven seismic failure mode identification of reinforced concrete shear walls based on PCA feature extraction, *Structures* 44 (2022) 1429–1442, <https://doi.org/10.1016/J.ISTRUC.2022.08.089>.
- [54] H.A. Kazem, J.H. Yousif, M.T. Chaichan, A.H.A. Al-Waeli, K. Sopian, Long-term power forecasting using FRNN and PCA models for calculating output parameters in solar photovoltaic generation, *Heliyon* 8 (2022) e08803, <https://doi.org/10.1016/J.HELIYON.2022.E08803>.
- [55] H. Lan, C. Zhang, Y.Y. Hong, Y. He, S. Wen, Day-ahead spatiotemporal solar irradiation forecasting using frequency-based hybrid principal component analysis

- and neural network, *Appl. Energy* 247 (2019) 389–402, <https://doi.org/10.1016/J.APENERGY.2019.04.056>.
- [56] X. Zhang, Z. Wei, A hybrid model based on principal component analysis, wavelet transform, and extreme learning machine optimized by bat algorithm for daily solar radiation forecasting, *Sustainability (switzerland)* 11 (2019), <https://doi.org/10.3390/SU11154138>.
- [57] I.L. Mallika, D.V. Ratnam, Y. Ostuka, G. Sivavaraprasad, S. Raman, Implementation of Hybrid Ionospheric TEC forecasting Algorithm using PCA-NN Method, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 12 (2019) 371–381, <https://doi.org/10.1109/JSTARS.2018.2877445>.
- [58] U. Munawar, Z. Wang, A Framework of using Machine Learning Approaches for Short-Term Solar Power forecasting, *Journal of Electrical Engineering & Technology* 15 (2020) 561–569, <https://doi.org/10.1007/s42835-020-00346-4>.
- [59] W.J. Machine, L.P. Component, Analysis (2017).
- [60] A. Ahmad, X. Xiao, H. Mo, D. Dong, Tuning data preprocessing techniques for improved wind speed prediction, *Energy Rep.* 11 (2024) 287–303, <https://doi.org/10.1016/J.EGYR.2023.11.056>.
- [61] Y. Chen, X. Hu, L. Zhang, A review of ultra-short-term forecasting of wind power based on data decomposition-forecasting technology combination model, *Energy Rep.* 8 (2022) 14200–14219, <https://doi.org/10.1016/J.EGYR.2022.10.342>.
- [62] G. Xie, S. Wang, K.K. Lai, Short-term forecasting of air passenger by using hybrid seasonal decomposition and least squares support vector regression approaches, *J. Air Transp. Manag.* 37 (2014) 20–26, <https://doi.org/10.1016/J.JAIRTRAMAN.2014.01.009>.
- [63] Fraihat H, Almbaideen AA, Al-Odienat A, Al-Naami B, de Fazio R, Visconti P. Solar Radiation Forecasting by Pearson Correlation Using LSTM Neural Network and ANFIS Method: Application in the West-Central Jordan. *Future Internet* 2022, Vol 14, Page 79 2022;14:79. DOI: 10.3390/FI14030079.
- [64] P. Singla, M. Duhan, S. Saroha, A Hybrid Solar Irradiance forecasting using Full Wavelet Packet Decomposition and Bi-Directional Long Short-Term memory (BiLSTM), *Arab. J. Sci. Eng.* (2022) 1–27, <https://doi.org/10.1007/S13369-022-06655-2/TABLES/18>.
- [65] H. Ali-Ou-Salah, B. Oukarfi, T. Mouhaydine, Short-term solar radiation forecasting using a new seasonal clustering technique and artificial neural network, <https://doi.org/10.1080/15435075.2021.1946819>.
- [66] S.-G. Kim, J.-Y. Jung, M.K. Sim, A Two-step Approach to Solar Power Generation Prediction based on Weather Data using Machine Learning, *Sustainability* (2019), <https://doi.org/10.3390/su11051501>.
- [67] D. Shin, E. Ha, T. Kim, C. Kim, Short-term photovoltaic power generation predicting by input/output structure of weather forecast using deep learning, *Soft. Comput.* 25 (2021) 771–783, <https://doi.org/10.1007/S00500-020-05199-7/FIGURES/9>.
- [68] C.N. Obiora, A.N. Hasan, A. Ali, N. Alajarmeh, Forecasting Hourly Solar Radiation using Artificial Intelligence Techniques, *IEEE Canadian Journal of Electrical and Computer Engineering* 44 (2021).
- [69] M. Rana, I. Koprinska, V.G. Agelidis, Univariate and multivariate methods for very short-term solar photovoltaic power forecasting, *Energy Convers Manag* 121 (2016) 380–390, <https://doi.org/10.1016/J.ENCONMAN.2016.05.025>.
- [70] H. Zang, L. Liu, L. Sun, L. Cheng, Z. Wei, G. Sun, Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations, *Renew. Energy* 160 (2020) 26–41, <https://doi.org/10.1016/J.RENENE.2020.05.150>.
- [71] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, *Appl. Soft Comput.* 97 (2020) 105524, <https://doi.org/10.1016/J.ASOC.2019.105524>.
- [72] Normalize data - MATLAB normalize - MathWorks United Kingdom n.d. <https://uk.mathworks.com/help/matlab/ref/double.normalize.html> (accessed October 30, 2021).
- [73] Y. Chu, M. Li, C.F.M. Coimbra, D. Feng, H. Wang, Intra-hour irradiance forecasting techniques for solar power integration: a review, *Iscience* 24 (2021) 103136, <https://doi.org/10.1016/J.JSCI.2021.103136>.
- [74] K. Santhanakrishnan, V. Senthoooran, A Parallel distributed Cluster Computing Model for Z-Score Computation in respect of Sri Lankan University Admissions, in: 2022 2nd International Conference on Advanced Research in Computing (ICARC), 2022, <https://doi.org/10.1109/ICARC54489.2022.9754093>.
- [75] Herwanto HW, Handayani AN, Wibawa AP, Chandrika KL, Arai K. Comparison of Min-Max, Z-Score and Decimal Scaling Normalization for Zoning Feature Extraction on Javanese Character Recognition. 7th International Conference on Electrical, Electronics and Information Engineering: Technological Breakthrough for Greater New Life, ICEEE 2021 2021. DOI: 10.1109/ICEEEI52663.2021.9616665.
- [76] I.I. Prado-Rujas, A. Garcia-Dopico, E. Serrano, M.S. Perez, A flexible and robust deep learning-based system for solar irradiance forecasting, *IEEE Access* 9 (2021) 12348–12361, <https://doi.org/10.1109/ACCESS.2021.3051839>.
- [77] H.W. Herwanto, A.N. Handayani, A.P. Wibawa, K.L. Chandrika, K. Arai, Comparison of Min-Max, Z-Score and Decimal Scaling Normalization for Zoning Feature Extraction on Javanese Character Recognition, in: 7th International Conference on Electrical, Electronics and Information Engineering (ICEEE), 2021, <https://doi.org/10.1109/ICEEEI52663.2021.9616665>.
- [78] S.A. Alasadi, W.S. Bhaya, Review of data preprocessing techniques in data mining, *J. Eng. Appl. Sci.* 12 (2017) 4102–4107, <https://doi.org/10.3923/JEASCI.2017.4102.4107>.
- [79] S. Hassani, U. Dackermann, M. Mousavi, J. Li, A systematic review of data fusion techniques for optimized structural health monitoring, *Inf. Fusion* 103 (2024) 102136, <https://doi.org/10.1016/J.INFUS.2023.102136>.
- [80] A. Munir, E. Blasch, J. Kwon, J. Kong, A. Aved, Artificial Intelligence and Data Fusion at the Edge, *IEEE Aerosp. Electron. Syst. Mag.* 36 (2021) 62–78, <https://doi.org/10.1109/MAES.2020.3043072>.
- [81] K. Rammelkamp, S. Schröder, S. Kubitz, D.S. Vogt, S. Frohmann, P.B. Hansen, et al., Low-level LIBS and Raman data fusion in the context of in situ Mars exploration, *J. Raman Spectrosc.* 51 (2020) 1682–1701, <https://doi.org/10.1002/JRS.5615>.
- [82] C.M. Teglia, S.M. Azcarate, M.R. Alcaráz, H.C. Goicoechea, M.J. Culzoni, Exploiting the synergistic effect of concurrent data signals: Low-level fusion of liquid chromatographic with dual detection data, *Talanta* 186 (2018) 481–488, <https://doi.org/10.1016/J.TALANTA.2018.04.090>.
- [83] I. Lanza, V. Lolli, S. Segato, A. Caligiani, B. Contiero, A. Lotto, et al., Use of GC–MS and 1H NMR low-level data fusion as an advanced and comprehensive metabolomic approach to discriminate milk from dairy chains based on different types of forage, *Int. Dairy J.* 123 (2021) 105174, <https://doi.org/10.1016/J.IDAIRYJ.2021.105174>.
- [84] Q.H. Nguyen, H.B. Ly, L.S. Ho, N. Al-Ansari, H. Van Le, V.Q. Tran, et al., Influence of data splitting on performance of machine learning models in prediction of shear strength of soil, *Math. Probl. Eng.* (2021; 2021.), <https://doi.org/10.1155/2021/4832864>.
- [85] J.J. Zhu, M. Yang, Z.J. Ren, Machine Learning in Environmental Research: Common pitfalls and best Practices, *Environ. Sci. Technol.* 57 (2023) 17671–17689, <https://doi.org/10.1021/ACS.EST.3C00026/ASSET/IMAGES/LARGE/ES3C00026.0007.JPEG>.
- [86] T.O. Falope, L. Lao, D. Hanak, A Three-step Weather Data Approach in Solar Energy Prediction using Machine Learning, *Renewable Energy Focus* 50 (2024) 100615, <https://doi.org/10.1016/J.REF.2024.100615>.
- [87] Pawar P, Mithulananthan N, Raza MQ. Solar PV Power Forecasting Using Modified SVR with Gauss-Newton Method. *Proceedings - 2020 IEEE 2nd Global Power, Energy and Communication Conference, GPECOM 2020 2020:226–31*. DOI: 10.1109/GPECOM49333.2020.9247935.
- [88] Z. Zhang, M. Li, M. Lin, G. Soon, T. Greene, C. Shen, Subgroup selection in adaptive signature designs of confirmatory clinical trials, *J. R. Stat. Soc. Ser. C. Appl. Stat.* 66 (2017) 345–361, <https://doi.org/10.1111/RSSC.12175>.
- [89] Y. Ku, S. Kwon bin, J.-H. Yoon, S.-K. Mun, M. Chang, Machine Learning Models for predicting the Occurrence of respiratory Diseases using Climatic and Air-Pollution Factors, *Clin Exp Otorhinolaryngol* (2022), <https://doi.org/10.21053/ceo.2021.01536>.
- [90] Mirza HB, Ratnaparkhe VR. Classifier Tools: A Comparative Study. *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018 2019:1543–7*. DOI: 10.1109/ICCONS.2018.8662904.
- [91] M. Ghofrani, M. Alolayan, Time Series and Renewable Energy forecasting, *Time Series Analysis and Applications* (2017), <https://doi.org/10.5772/INTECHOPEN.70845>.
- [92] Meenal R, Binu · D, Ramya · K C, Prawin ·, Michael A, Vinoth Kumar · K, et al. Weather Forecasting for Renewable Energy System: A Review. *Archives of Computational Methods in Engineering* n.d.;1:3. DOI: 10.1007/s11831-021-09695-3.
- [93] Wu L, Won Y-S, Jap D, Perin G, Bhasin S, Picek S. Explain Some Noise: Ablation Analysis for Deep Learning-based Physical Side-channel Analysis n.d.
- [94] R. Trivedi, S. Patra, S. Khadem, A Data-Driven Short-Term PV Generation and load forecasting Approach for Microgrid applications, *IEEE Journal of Emerging and Selected Topics in Industrial Electronics* 3 (2022) 911–919, <https://doi.org/10.1109/JESTIE.2022.3179961>.
- [95] National Physical Laboratory. When do the four seasons officially begin? - NPL 2021. <https://www.npl.co.uk/resources/q-a/when-do-the-four-seasons-begin> (accessed May 24, 2022).
- [96] J. Lever, M. Krzywinski, N. Altman, Points of significance: principal component analysis, *Nat. Methods* 14 (2017) 641–642, <https://doi.org/10.1038/NMETH.4346>.



DSPACE

<https://dspace.org/>

Data preprocessing techniques for statistical photovoltaic forecasting models

Falope, Tolulope Olumuyiwa; Lao, Liyun; Huo, Da

2026-07-14

Elsevier

<https://doi.org/10.1016/j.measurement.2026.121997>

Falope TO, Lao L, Huo D. (2026) Data preprocessing techniques for statistical photovoltaic forecasting models. Measurement, Volume 282, July 2026, Article number 121997