Ain Shams Engineering Journal 15 (2024) 102716

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: https://www.sciencedirect.com

Full Length Article

On reliability enhancement of solar PV arrays using hybrid SVR for soiling forecasting based on WT and EMD decomposition methods

Abhijeet Redekar [a], Harsh S. Dhiman [b,*], Dipankar Deb [a], S.M. Muyeen [c]

[a] Department of Electrical Engineering, Institute of Infrastructure Technology Research and Management, Ahmedabad, 380026, India
[b] Department of Artificial Intelligence & Machine Learning, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, 412115, India
[c] Department of Electrical Engineering, Qatar University, Doha, Qatar

ARTICLE INFO

Keywords:
Soiling forecasting
Wavelet transform
Support vector regression
Empirical mode decomposition

ABSTRACT

Solar farms have PV arrays in arid and semi-arid regions where ensuring the system's reliability is paramount and face uncertain events like dust storms. The deposition of random dust patterns over panel arrays is called uneven soiling, which diminishes the power generation of such farms. This paper finds the most suitable hybrid algorithm model, the wavelet transform-based support vector regression variants (WT-SVR) algorithm, and the empirical model decomposition-based support vector regression variants (EMD-SVR) to predict the extent of soiling levels and uncertain events on PV arrays. The soiling dataset is taken from NREL's Soiling Station Number 3 in Imperial County, Calipatria, California, from December 30, 2014, to December 31, 2015. This research tested four SVR variants on soiling data, viz., εSVR, LSSVR, TSVR, and εTSVR, then compared with the benchmark random forest. The hyperparameters for each model are meticulously tuned to enhance the robustness of the trained algorithms. Results reveal that the WT-TSVR model outperforms the WT-SVR model in terms of wavelet transform decomposition by a margin of 91.6%. Similarly, the EMD-TSVR model showcases an 85.7% enhancement in performance over the EMD-SVR model based on empirical mode decomposition. All SVR variants outperform the benchmark model (RF). Furthermore, EMD models exhibit enhanced efficiency in forecasting random events compared to WT, which is attributed to their reduced computational time. This model applies to multi-cleaning agent robots, aligning with recommendations from the state-of-the-art literature.

1. Introduction

Photovoltaic (PV) power is emerging as a crucial renewable energy source within the interconnected power grid network. Efficient economic dispatch of a PV power plant necessitates the anticipation of power efficiency, a metric influenced by solar irradiation and various factors like the geographical location of the plant, tilt angle, incident angle, dust accumulation, wind speed, and surface temperature of PV panels, among others [1]. Notably, the dust accumulation, or the soiling rate, is particularly high in arid regions like Gulf countries (e.g., Qatar, Saudi Arabia, Kuwait), significantly impacting power generation. In these regions, soiling is an inevitable challenge that diminishes the transmittance of solar rays from the glazing of solar panels to the cells. This reduction occurs due to the accumulation of diverse local particles on the panel's surface. The nature and extent of soiling vary across different locations, and non-uniform (uneven) soiling occurs due to spatial irregularities, random events in arid regions, dust storms, and other factors [2]. The impact of soiling losses on electrical efficiency is noteworthy, especially in large-scale solar farms [3]. As an illustration, the efficiency declined from 7.2% to 5.6% over a 108-day observation period in a commercial site in Santa Clara, California [4]. Similarly, Dhahran, Saudi Arabia experienced a power loss ranging from 45.9% to 49% over eight months [5], the Colorado desert exhibited a 2.6% loss in six months [6], and instances of power losses surpassing 50% in various studies [7–11].

Due to gravity, lightweight dust particles uniformly settle on the panel's surface in typical wind conditions, as noted in the study by Majeed et al. (2020). However, the dispersion of larger particles is influenced by factors like wind speed and storms, resulting in an irregular or non-uniform distribution of dust on the glazing of the photovoltaic (PV) panel. [12]. Research indicates that the impact of non-uniform soiling on photovoltaic energy generation is more significant in proportion to the effect of uniform soiling [13]. Wind erosion is the primary factor behind desertification, uncertainty, and temporal changes in arid

* Corresponding author.
E-mail address: harsh.dhiman@sitpune.edu.in (H.S. Dhiman).

https://doi.org/10.1016/j.asej.2024.102716
Received 26 August 2023; Received in revised form 17 January 2024; Accepted 16 February 2024

Available online 6 March 2024

Abbreviations

PV	Photo-Voltaic	ARIMA	Auto-regressive moving integrated moving average
DSI	Dust Storm Index	MAE	Mean squared error
GRA	Granger-Ramanathan averaging	RMSE	Root mean square error
DT	Decision Tree	SVR	Support vector regression
LR	Linear Regression	SSR/SST	Ratio of sum of squared residuals to sum of squared deviation
RF	Random Forest		
ANN	Artificial neural network	SSR	Sum of squared residuals
ARMA	Auto-regressive moving average	SST	Sum of squared deviation of testing samples

regions. Consequently, accurately forecasting wind erosion events and the Dust Storm Index (DSI) is crucial for effectively managing panel soiling. Ballestrín et al. emphasized the significance of achieving accurate solar power forecasting results, essential for the seamless integration of solar energy into the smart grid [14]. Researchers have discovered various cleaning methods for solar panels, such as air blowing [15], spiral brush cleaning [16], microfiber cloth cleaning [17], electrostatic cleaning [18], and dew cleaning [19]. These cleaning agents typically involve a robotic structure moving across the entire solar panel row. However, these rigid cleaning practices result in unnecessary cleaning of already clean portions, increasing cleaning energy and time consumed. Therefore, there is a motivation to implement a machine learning-based prediction algorithm that considers forecasted random events. This approach optimizes the cleaning process, making it quicker and more energy-efficient for robotic structures.

Many researchers developed prediction models for optimization of power, dust storm index, energy, etc. Gostein et al. observed that nonuniform soiling detected by the soiling ratio using maximum power metric is insufficient to differentiate random events [20]. Ebrahimi et al. found that combining the Granger-Ramanathan averaging (GRA) model with individual machine learning models provides a more accurate dust event forecast in arid areas [21]. Zhang et al. discussed a deep learning technique for probabilistic estimation of soiling loss [22]. Benhmed et al. developed a power prediction model using feature selection methods. The decision tree (DT) model is more accurate than the linear regression (LR) with and without feature selection but more accurate with all features [1]. However, the short-term solar energy prediction model developed by Shahid et al. and the obtained random forest (RF) and ridge regression model outperformed [23]. Khandakar et al. took environmental parameters of the site in Qatar, like solar irradiance, wind speed, relative humidity, ambient temperature, panel surface temperature, and accumulated dust for feature selection, and compared among artificial neural network (ANN), linear regression model (LR), M5P decision tree (DT), and Gaussian Process Regression (GPR) for output power prediction, in terms of RSME [24]. A summary of forecasting models for the soiling and targeted parameters is recorded in Table 1.

Dhiman et al. successfully devised hybrid machine intelligent models, specifically support vector regressor (SVR) variants like TSVR and ε-TSVR, incorporating wavelet transform (WT) decomposition. This approach was applied for short-term wind speed forecasting and identification of ramp events in wind farms [25]. Given that soiling on solar panels depends on weather-related stochasticity, like rainfall or dust storms that rapidly alter soiling levels akin to ramp events in wind farms, the current work addresses the uncertainties of random events like distributed dust or soiling patterns. The study seeks to detect these patterns and quantify the amount of dust present utilizing SVR machine learning algorithms. Consequently, SVR variants, with WT and empirical mode decomposition (EMD) methods, are applied to analyze the soiling dataset. Also, the forecasting models of Random Forest (RF) and Ridge consistently yield the lowest Mean Absolute Error (MAE) [23]. Consequently, the RF model is the benchmark for comparison in this research study.

Table 1
Forecasting models to predict soiling related parameters in terms of MAE.

Paper	Forecasting model	Target	Out-performer
[1]	LR (9.433), M5P-DT	PV Power	M5P-DT (8.632)
[21]	MARS (1.17), SVR (1.08) Lasso (1.16), Cubist (1.02) K-NN (1.13), ANN (1.18) GP (1.12), XGB (1.04) RF (0.96), GRA (0.84)	Dust storm index	GRA (0.84)
[23]	LR(67e5), DT (39e5) Ridge (22e5), RF (22e5) Lasso (40e5), ANN (37e5)	Energy	Ridge (22e5) RF(22e5)
[24]	ANN (2.1436), M5P DT (7.69) LR (8.95), and GPR (6.69)	PV Power	ANN (2.1436)

This research aims to develop an optimal hybrid prediction model for a dataset related to soiling. The study evaluates four Support Vector Regression (SVR) variants, namely ε-SVR, LSSVR, TSVR, and εTSVR, using two signal decomposition techniques—wavelet transform (WT) and empirical mode decomposition (EMD). The comparison is conducted within SVR variants, with wavelet transform (WT-SVR variants), and with empirical mode decomposition (EMD-SVR variants). The hybrid prediction model, incorporating these techniques, is then tested on soiling data. The performance of the WT-εTSVR hybrid model compared with εSVR, LSSVR, TSVR, and εTSVR, first optimally tunes the regularization and ε tube width tuning parameters to achieve the best model performance while retaining the regularization value for subsequent models. Subsequently, the model utilizes long-term historical data to predict uncertain events, such as dust storms over solar farms. These solar farms have fully automatic solar panel cleaning robots with appropriate cleaning agents for addressing these unpredictable events. This manuscript significantly contributes by employing machine learning and signal preprocessing for soiling quantification. This quantification serves as input for an algorithm that predicts soiling events, enabling the automatic cleaning robot at the solar PV farm site to initiate cleaning actions energy-efficiently and selectively clean heavily soiled sections of solar panels instead of the entire photovoltaic (PV) plant.

The organization of our work is given as follows. Section 2 presents the support vector regressor variants model formulation, and Section 3 describes the hybrid models framework process. Then, Section 4 describes the state of the art of soiling categorization. Section 5 is followed by the results and discussions. Section 6 concludes the work with a recommendation.

2. Support vector regression and its variants

Regression is one of the applications of Support Vector Machines (SVMs), and SVR is the extension of Support Vector Classification (SVC) [26]. The principle of structural risk minimization develops support vector regression (SVR) (SRM) using statistical learning for identifying non-linearity in the data and provides a best prediction algorithm [27–29]. The objective of SVR is to fit a line (linear or nonlinear) over

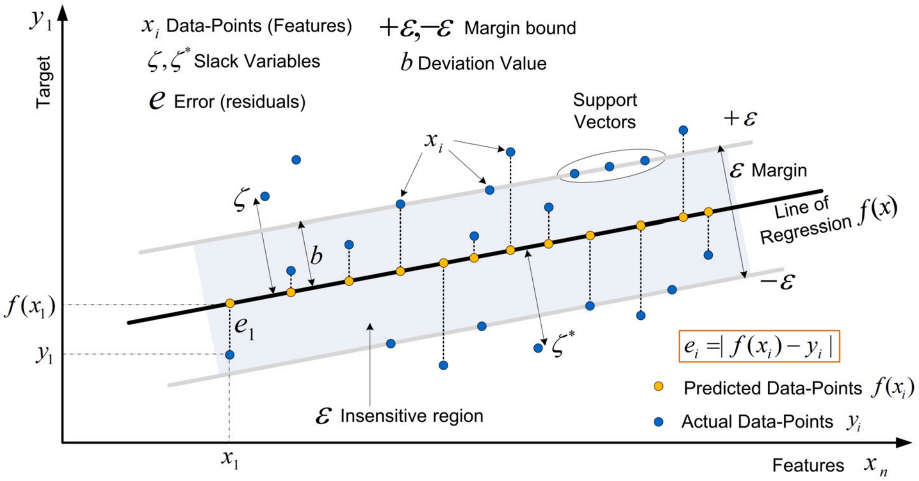

Fig. 1. Regression line and its hyperparameters.

the data in n-dimensional space, called a regressor. These are decision boundaries that predict the output. The closest data points on either side of the hyperplane created two support vectors. The support vectors influence the regressor that helps to build the model [30].

2.1. εSVR

Consider a training set $T = \{(x_1, y_1), ..., (x_i, y_i)\}$, $i = 1, 2, ..., l$. $y_i \in \mathbb{R}$ is target vector, $x_i \in \mathbb{R}^n$ for n feature vectors, l is number of training instances. The basic SVR, as understood with Fig. 1, aims to find a regressor function for prediction data as $f(x) = w^T x + b$, with $w \in X$, $b \in \mathbb{R}$, where w is the weight coefficient for each input vector x_i and b is the deviation value (bias term) [31].

The weight parameter is such that $(w^T x_i)$ shrinks towards the target vector y_i. The regressor $f(x)$ is as flat as possible with maximum deviation ε. For model flatness, minimization of the square of the norm of weight vector w is achievable through a convex optimization problem [32]:

$$min\left(\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\left(\zeta_i + \zeta_i^*\right)\right) \tag{1}$$

subject to

$$y_i - w^T x_i \le \varepsilon + \zeta_i^*, \quad w^T x_i - y_i \le \varepsilon + \zeta_i, \quad \zeta_i, \zeta_i^* \ge 0, \qquad i = 1, \dots, n,$$

where C is a tuneable regularization parameter providing more weight to minimize the flatness or the error, with slack variables ζ_i and ζ_i^* which give a soft margin to determine tolerable points outside the ε tube. However, kernel functions are also required in the case of non-linear features to transform the data to a higher dimensional space via suitable mapping function ϕ: $\mathbb{R} \to Z$.

The inner product $\langle w^T, \phi(x)\rangle$ in the target space is represented using the kernel function that satisfies Mercer's theorem such that $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle$ are the elements of the kernel matrix k [33]. This SVR optimization problem can be solved by finding Lagrangian multipliers or dual variables $\lambda, \lambda^*, \alpha, \alpha^*$ which are non-negative real numbers:

$$min\left(\frac{1}{2}||w||^2 + C\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)^T k(x_i, x_j)(\alpha_j - \alpha_j^*) + \varepsilon\sum_{i=1}^{n}(\alpha + \alpha^*) - \sum_{i=1}^{n} y_i(\alpha - \alpha^*)\right), \tag{2}$$

subject to

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \qquad 0 \le \alpha, \qquad \alpha^* \le C.$$

The performance of the regressor $f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x, x_i) + b$, depends on kernel function, regularization parameter (C) and tube margin ε.

2.2. LSSVR

In the Least square support vector regression (LSSVR), the square of the error term ε is minimized. The regressor function is $f(x) = w^T \phi(x) + b$, with $x_i \in \mathbb{R}^n$, $y \in \mathbb{R}$. The objective function is given as

$$min\left(\frac{1}{2}||w||^2 + \frac{1}{2}\gamma\sum_{i=1}^{n}\varepsilon_i^2\right), \tag{3}$$

subject to

$$y_i = \phi(x_i) + b + \varepsilon_i, \qquad i = 1, \dots, n,$$

where γ is the margin parameter and ε_i is the error term corresponding to each x_i. For optimization, equality constraints are chosen. It takes less computational time than classical εSVR.

2.3. TSVR

The twin support vector regressor (TSVR) with two non-parallel hyperplane functions $f_1(x) = w_1^T x_1 + b_1$ and $f_2(x) = w_2^T x_2 + b_2$ aims to find the ε-insensitive down-bound and up-bound regressor respectively [34], as shown in Fig. 2.

The objective functions are

$$\begin{gathered} min\left(\frac{1}{2}\sum_{i=1}^{n}(y_i - e\varepsilon_1 - (x_i w_1 + eb_1))^T(y_i - e\varepsilon_1 - (x_i w_1 + eb_1)) + C_1 e^T\sum_{i=1}^{n}\zeta_i\right) \\ min\left(\frac{1}{2}\sum_{i=1}^{n}(y_i - e\varepsilon_2 - (x_i w_1 + eb_2))^T(y_i - e\varepsilon_2 - (x_i w_2 + eb_2)) + C_2 e^T\sum_{i=1}^{n}\eta_i\right), \end{gathered} \tag{4}$$

subject to

$$y_i - (x_i w_1 + eb_1) \ge e\varepsilon_1 - \zeta_i, \quad y_i - (x_i w_2 + eb_2) \ge e\varepsilon_2 - \eta_i,$$

where $C_1, C_2 > 0$ and $\varepsilon_1, \varepsilon_2 \ge 0$ are the TSVR hyperparameters and ζ_i, η_i are the slack vectors introduced as a soft margin to the error ε in an optimization problem. This TSVR is much faster than the standard SVR.

The final regressor:

$$f(x) = \frac{1}{2}((w_1 + w_2)^T x + (b_1 + b_2)). \tag{5}$$

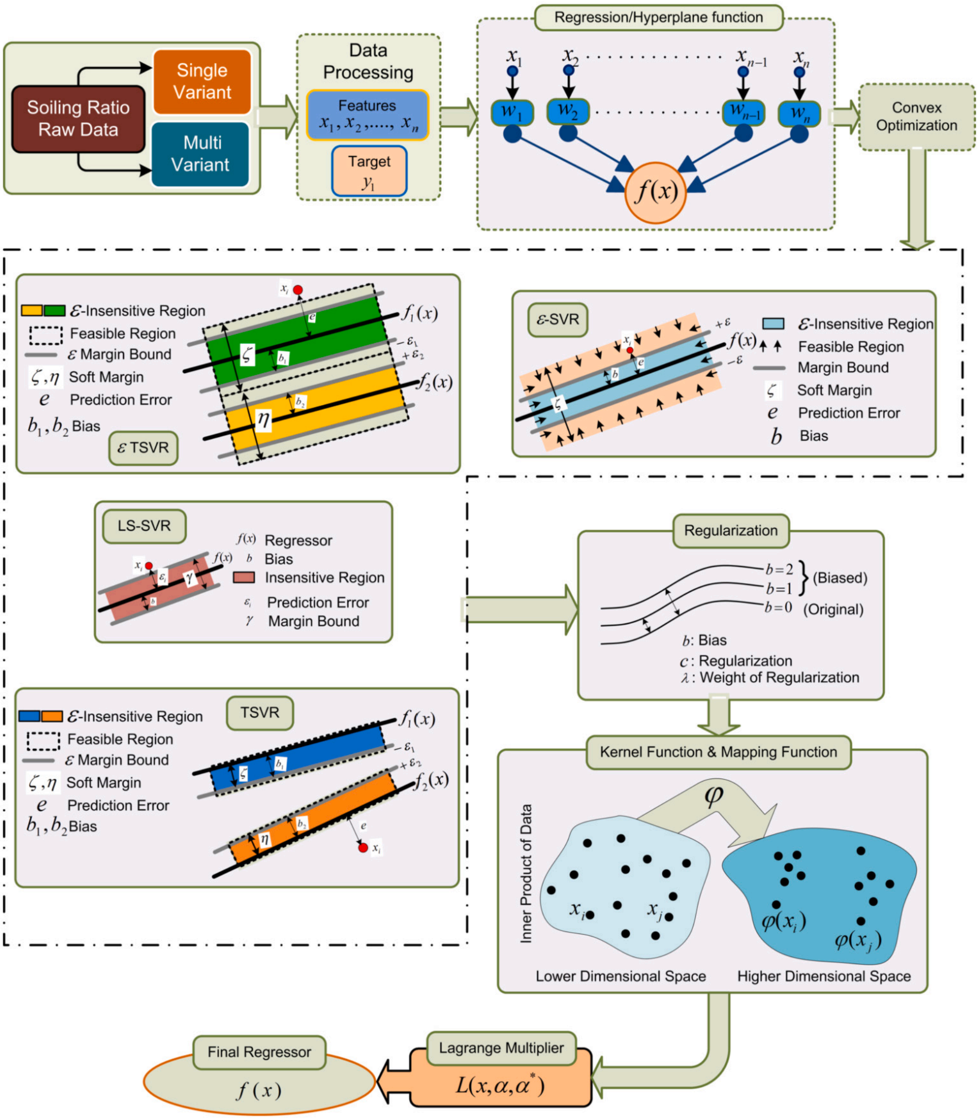

Fig. 2. Model framework development of SVR variants.

Table 2
Hyper-parameters of support vector regression variants with rbf kernel.

Model	Tunable parameters with range
ε-SVR	ε: error insensitivity parameter σ: kernel bandwidth , C: regularization
LS-SVR	σ: kernel bandwidth γ: regularization
TSVR	C_1 and C_2: regularization, ε_1, ε_2 σ: kernel bandwidth
ε-TSVR	C_1 C_2C_3 C_4: Regularization, ε_1, ε_2 σ: kernel bandwidth

2.4. εTSVR

The εTSVR is extended from TSVR that determines the pair of ε-insensitive functions by solving two convex optimization problems [35].

εTSVR considers an added regularization term that solves the ill-conditioning problem, the objective function as

$$min\left(\frac{1}{2}C_3(w_1^T w_1 + b_1^2) + \frac{1}{2}\zeta^{*T}\zeta + C_1 e^T\zeta\right),$$

$$min\left(\frac{1}{2}C_4(w_2^T w_2 + b_2^2) + \frac{1}{2}\zeta^{*T}\zeta + C_2 e^T\zeta\right), \tag{6}$$

subject to

$$Y - (Xw_1 + eb_1) = \zeta^*, \quad Y - (Xw_1 + eb_1) \geq -e\varepsilon_1 - \zeta, \quad \zeta \geq 0,$$
$$Y - (Xw_2 + eb_2) = \zeta^*, \quad Y - (Xw_2 + eb_2) \geq -e\varepsilon_2 - \eta, \quad \eta \geq 0.$$

In the optimization problem C_1, C_2, ε_1, ε_2 are the hyperparameters that determine the regression performance (See Table 2.). The final regressor is the mean of two functions $f_1(x)$ and $f_2(x)$ given as

$$f(x) = \frac{1}{2}[f_1(x) + f_2(x)] + \frac{1}{2}[(w_1 + w_2)^T x + (b_1 + b_2)]. \tag{7}$$

2.5. Performance metrics

The prediction accuracy in terms of error is assessed by computing the following metrics mathematically expressed as

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2\right]^{1/2}, \quad MAE = \left[\frac{1}{n}\sum_{i=1}^{n}|\hat{x}_i - x_i|\right]$$

$$SSR/SST = \frac{\sum_{i=1}^{n}(\hat{x}_i - \bar{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \quad SSE/SST = \frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}$$

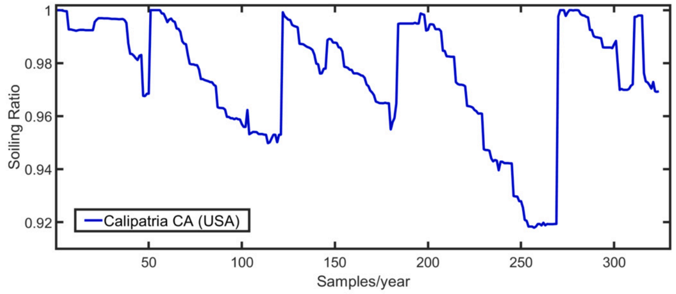

Fig. 3. Soiling station waveform of selected dataset.

$$IOA = 1 - \sum_{i=1}^{n}(\hat{x}_i - x_i)^2 / \sum_{i=1}^{n}(|\hat{x}_i - \bar{x}_i| + |x_1 + \bar{x}_i|)^{1/2}$$

$$U1 = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n}(\hat{x}_i - x_i)^2} / \left(\sqrt{\frac{1}{n} \times \sum_{i=1}^{n} x_i^2} + \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} \hat{x}_i^2}\right)$$

$$U2 = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n}((x_{i+1} - \hat{x}_{i+1})/x_i)^2} / \sqrt{\frac{1}{n} \times \sum_{i=1}^{n}((x_{i+1} - \hat{x}_i)/x_i)^2}$$

where, $\hat{x}_i$ is predicted, x_i actual, and $\bar{x}$ is the mean value notation of testing samples.

3. Hybrid models framework process

The decomposition of the soiling signal involves utilizing both Wavelet Transform (WT) and Empirical Mode Decomposition (EMD) methods. The subsequent section elucidates the framework necessary for the hybrid model.

3.1. Soiling datasets

To evaluate the SVR-based hybrid forecasting models, soiling data obtained from NREL (National Renewable Energy Laboratory) Soiling Station Number 3 in Imperial County, Calipatria, CA, spanning from December 30, 2014, to December 31, 2015, with a total of 324 recorded samples utilized [36]. In Fig. 3, the variations in soiling ratio for selected datasets are illustrated. The soiling ratio exhibits a saw-tooth wave-type pattern. Instances of natural events, such as rainfall, or manual cleaning result in a sudden upward shift in the daily soiling ratio, also referred to as the performance index. This is followed by a linear decline each day due to soiling until the occurrence of the next cleaning event.

3.2. Decomposition of mother signal

Soiling forecasting employs hybrid models that integrate wavelet transform or Empirical Mode Decomposition of the data signal along with various versions of Support Vector Regression (SVR) prediction models. In Fig. 4 (a), the conceptual block diagram illustrates the hybrid model incorporating wavelet transform.

Soiling raw data undergoes decomposition into various frequency signals to be utilized in SVR forecasting models. The selected filter configuration is Daubechies 'db4,' employing a 5-level decomposition. Wavelet transform formulated as discrete (DWT) and continuous wavelet transform (CWT) are defined as

$$B(u,v) = 2^{-u/2} \sum_{t=0}^{N-1} d(x) \ \ \phi\left(\frac{t - v.2^u}{2^u}\right),$$

$$B(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} d(x) \ \ \phi\left(\frac{x-b}{a}\right), \tag{8}$$

Table 3
Statistical information of decomposed signals by EMD.

Component	Mean	Std	Min	Max
Original	0.9753	0.0222	0.9178	1.0000
IMF1	-0.0006	0.0062	-0.0282	0.0273
IMF2	-0.0001	0.0095	-0.0270	0.0321
IMF3	-0.0008	0.0186	-0.0598	0.0400
IMF4	0.0004	0.0070	-0.0134	0.0141
IMF5	-0.0016	0.0065	-0.0100	0.0122
Res	0.9782	0.0046	0.9708	0.9856

where $d(x)$ is the soiling time series data of length N, $\phi(.)$ is the mother wavelet signal. The two signals obtained at each decomposition stage are approximate (A5) and detail signals (D1, D2, D3, D4, and D5) together form a matrix of input features for low-frequency and high-frequency components, respectively. The soiling ratio is the output or target used in the forecasting algorithm. The decompositions of selected data along with the mother signal (S0) are shown in Fig. 4 (b) in red color, approximate signal (a5) in blue, and detailed signals (D1 to D5) in green color.

The EMD method decomposes the mother soiling signal into a number of Intrinsic Mode Functions (IMFs) and residue signals in time series [37]. The IMFs and residue are determined as following steps: (a) From the time series signal $x(t)$ forms its maxima $x_u(t)$ and minima $x_l(t)$ (b) determine to mean of maxima and minima as $x_m(t) = \frac{x_u(t)+x_l(t)}{2}$ (c) Obtain detailed component as $x_d(t) = x(t) - x_m(t)$ (d) The decomposition process will continue until any one of the following criteria satisfies, (i) $x_m = 0$, and (ii) number of zero crossings and extrema should differ by one or zero (e) Repeat above steps until obtaining the residue. EMD hybrid model framework is shown in Fig. 5 (a) and the decomposed results of the soiling signal are shown in (b).

In Fig. 5 (b), the IFM components are arranged in descending order of frequency, encompassing local features and trends of the soiling signal at various time scales. Notably, the high-frequency component, i.e., IMF 1, reflects abrupt shifts in the soiling signal primarily attributed to cleaning events. Conversely, the low-frequency components from IMF 2 to IMF 5 delineate the rate and trend of soiling between two cleaning events or uncertain occurrences such as dust storms. Additionally, Table 3 provides statistical details for all EMD decomposed components, including mean value (Mean), standard deviation (Std), minimum value (Min), and maximum value (Max).

The standard deviation is a metric to gauge the statistical distribution and dispersion between two points within a given dataset. Upon comparing the standard deviation (std) values of all Intrinsic Mode Functions (IMFs) with the original data, it becomes evident that the decomposed signals exhibit greater smoothness and stability than the original data. Consequently, all IMFs and a residue signal are chosen as features for training the EMD-SVR variants model, as they provide detailed information about the soiling signal and its cumulative trend, respectively. Similarly, all detailed and approximation signals are employed as features for training the model in WT-SVR variants.

4. State of the art: soiling categorization

Categorizing dust accumulation is crucial for the robot system to choose the right cleaning agent and activate the actuator. Table 4 presents the statistical data on soiling. In contrast, Fig. 6 illustrates the categorization of soiling ratio data and dust levels across different large-scale PV generating stations. The figure reveals four distinct cumulative dust levels, demarcated by natural cleaning events indicated by vertical blue lines. Anticipating soiling occurrences based on dust levels enables the robot to prepare and deploy appropriate cleaning agents with functional actuators. In scenarios of nonuniform soiling, it becomes essential to identify dust levels and their locations for optimizing the cleaning process. Events like dust storms introduce uncertainties, leading to nonuniform dust profiles over the panels.

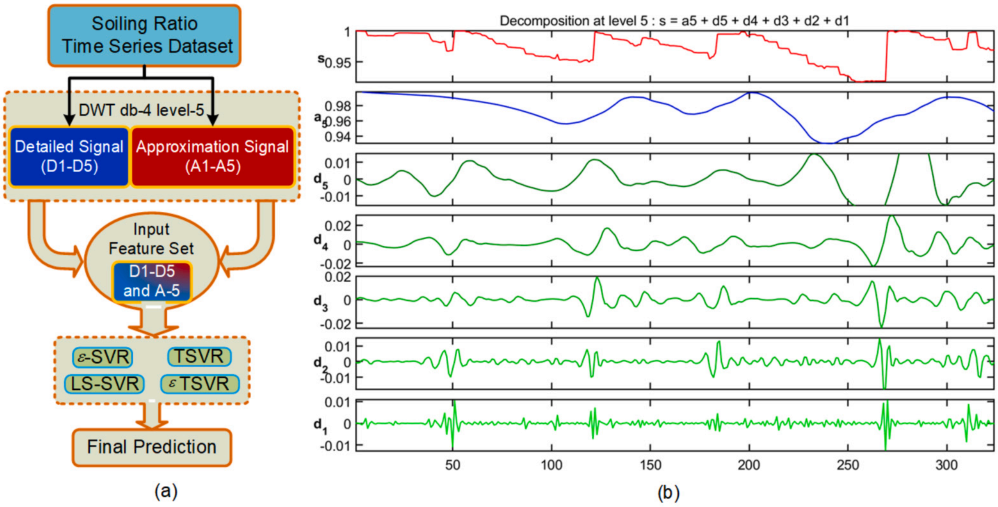

Fig. 4. Hybrid model using wavelet transform and SVR models (a) framework of hybrid model (b) Transformed wave of selected dataset.

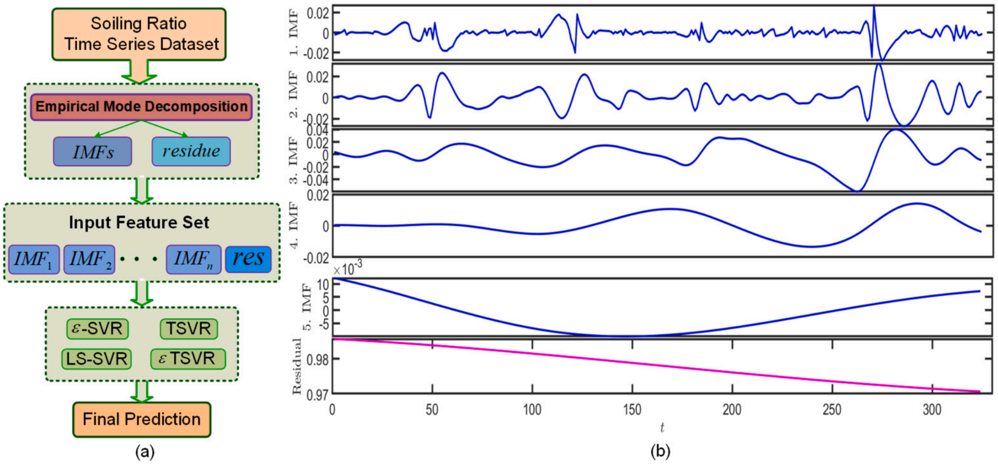

Fig. 5. Decomposed Intrinsic modes functions using empirical mode decomposition methods (a) framework of EMD hybrid model, (b) decomposed soiling signal.

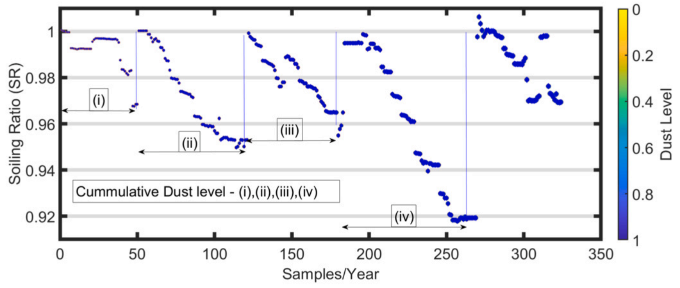

Fig. 6. Soiling categorization by accumulation level.

Table 4
Statistics of selected soiling data.

Soiling stations (Dataset)	Max	Min	Mean	Std Dev
Calipatria CA (A)	324	1.0	162.5	93.67497

4.1. Adaptation in optimized cleaning

The primary challenge faced by photovoltaic panels in desert regions lies in the accumulation of surrounding dust and periodic dust storms, resulting in unpredictable dust deposition. The uneven nature of dust accumulation renders traditional cleaning robots inefficient, leading to extended cleaning times and higher energy consumption. To address this issue, short-term forecasting methods are employed alongside fully automatic cleaning robots equipped with various cleaning agents, ensuring effective panel cleaning even during stormy conditions with improved energy efficiency and reduced cleaning time. The proposed robot integrates diverse cleaning actuators and agents, including synthetic jet actuators, spherical actuators, solenoid actuators, microfiber brushes, steam generation kits, and more. Short-term forecasting hybrid models utilize meteorological features such as seasons, geographic plant location, weather conditions, and wind speed to predict dust levels for the next 2 to 3 hours. The robot then adapts to the predicted dust level, preparing for optimal panel cleaning. For instance, if the prediction model anticipates a dust storm within the next hour with a level of 0.80, relying solely on a brush would be insufficient. Consequently, the robot activates an array of synthetic jet actuators. This adaptive approach allows the robot to choose the most suitable actuator for efficient cleaning based on the predicted soiling levels. Fig. 7 illustrates fully automatic robots, each selected according to the predicted soiling levels.

5. Results and discussions

The results are represented into three categories: (i) without any decomposition method, and only SVR variants are used, (ii) wavelet transforms decomposition plus SVR variants, and (iii) empirical mode

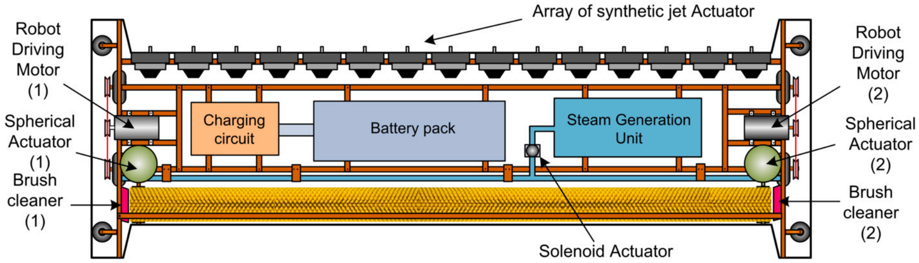

Fig. 7. Different actuators and cleaning agents of robot.

Table 5
Performance metrics of SVR variants only.

Model	RMSE	MAPE	MAE	SSR/SST	SSE/SST
	IOA	Acc	U_1	U_2	CPU time (s)
εSVR	0.0096	0.3620	0.0035	0.9142	0.0925
	0.9996	99.6380	0.0050	1.1060	0.6015
LS-SVR	0.0096	0.3807	0.0037	0.8933	0.0920
	0.9996	99.6193	0.0050	1.3796	0.0837
TSVR	**0.0097**	0.3386	**0.0033**	0.9307	0.0938
	0.9996	99.6614	0.0050	1.2443	**0.5134**
εTSVR	**0.0100**	0.3389	**0.0033**	1.0365	0.0992
	0.9996	99.6611	0.0052	0.2101	**0.0931**
RF	0.0147	1.0487	0.0099	4.3e-04	0.2151
	0.9989	98.9513	0.0076	0000	1.6801

Table 6
Performance metrics for SVR variants using wavelet transform decomposition.

Model	RMSE	MAPE	MAE	SSR/SST	SSE/SST
	IOA	Acc	U_1	U_2	CPU time (s)
WT-εSVR	0.0014	01202	0.0012	0.9290	0.0019
	1.0000	99.8798	0.0007	1.2576	0.5359
WT-LSSVR	0.0006	0.0508	0.0005	0.9667	0.0004
	1.0000	99.9492	0.0003	0.5831	0.0547
WT-TSVR	**0.0001**	**0.0107**	**0.0001**	**0.9921**	0.0000
	1.0000	99.9893	0.0001	0.0019	0.3220
WT-εTSVR	0.0009	0.0517	0.0005	0.9795	0.0008
	1.0000	99.9483	0.0005	4.7077	**0.1301**
WT-RF	0.0238	2.0552	0.0196	1.8558	0.5656
	0.9954	97.9448	0.0123	0000	3.7535

decomposition plus SVR variants. The Mean Absolute Error (MAE) is the main focus of observation, and the relative percentage improvement of the two decomposed methods is shown in terms of MAE. Around 75% of the data is used for training and the rest as testing for all three categories. The Radial basis (kernel) function (RBF) $k(x,x_i) = e^{\left(-\frac{||x-x_i||^2}{2\sigma^2}\right)}$ helps build the regression models with bandwidth σ. The hyper-parameters C_1, C_2, C_3, and C_4 are tuned manually and by a grid search. The Dataset of soiling taken from NREL's Soiling Station Number 3 in Imperial County, Calipatria, California, from December 30, 2014, to December 31, 2015, is chosen to test the categories' models and consists of 324 samples, out of which 75% (243) are used for training, and the remaining (81) for testing purposes. The forecasting models of Random Forest (RF) are selected as the benchmark because they exhibit the minimum Mean Absolute Error (MAE) [23]. Consequently, all SVR variants are compared among themselves and with the RF model. All performance metrics for soiling forecasting depicted for SVR variants with wavelet transform decomposition models and with empirical mode decomposition, respectively are shown in the Tables 5, 6, and 7.

5.1. Prediction results

The performance indices for SVR variants in terms of forecasting values are documented in Table 5. In terms of Mean Absolute Error (MAE), all four variants of Support Vector Regression (SVR) surpass the Random Forest (RF) model. The identification of the optimal SVR model involves a comparison among these four SVR variants. For the given soiling data, MAE-wise εTSVR and TSVR model outperformed εSVR and LSSVR by 5.71% and 10.81% respectively. Among TSVR and εTSVR, TSVR outperformed εTSVR by 3% in terms of RSME and except CPU time. Overall, TSVR is better than other models quantitatively also, the accuracy of TSVR is presented graphically in Fig. 8 (a). The hybrid model forecasted performance indices for WT-SVR variants are collected in Table 6. The benchmark model parameters are identified using the wavelet transform to decompose the soiling signal in unison with the Random Forest method.

Table 7
Performance metrics for SVR variants using empirical mode decomposition.

Model	RMSE	MAPE	MAE	SSR/SST	SSE/SST
	IOA	Acc	U_1	U_2	CPU time (s)
EMD-εSVR	0.0009	0.0763	0.0007	**0.9788**	0.0009
	1.0000	99.9237	0.0005	1.4421	0.5201
EMD-LSSVR	0.0006	0.0302	0.0003	0.9837	0.0002
	1.0000	99.9698	0.0002	2.7780	0.0851
EMD-TSVR	**0.0002**	**0.0158**	**0.0001**	0.9876	**0.0001**
	1.0000	99.9842	0.0001	0.0611	0.3948
EMD-εTSVR	0.0003	0.0246	0.0002	0.9909	0.0001
	1.0000	99.9754	0.0002	10.8911	0.0707
EMD-RF	0.0247	2.1599	0.0205	2.1442	0.6083
	0.9945	97.8401	0.0127	0000	4.8047

The WT-TSVR model demonstrated superior performance across all metrics except computational speed. The WT-εTSVR variant exhibits faster computation times in comparison. WT-TSVR surpassed WT-ε SVR, WT-LSSVR, and WT-εTSVR by 91.67%, 80%, and 80% in terms of MAE, respectively. Fig. 8(b) illustrates the prediction accuracy of the WT-TSVR model, particularly evident at the sharp edges of the actual data.

Similar performance indices for empirical mode decomposition are shown in Table 7. A combination of the empirical mode decomposition and the Random Forest model helps identify the parameters for the benchmark model. Within the EMD-SVR variants, EMD-TSVR outperforms all other models across all performance metrics except for the required computation time. In MAE comparison, EMD-TSVR surpasses EMD-εSVR, EMD-LSSVR, and EMD-εTSVR by 85.71%, 66.67%, and 50%, respectively. Fig. 8(c) illustrates that EMD-εSVR and, in some instances, EMD-εTSVR exhibit lower accuracy or fail to predict the original data, particularly noticeable at sharp curves.

The main essence of this research is to compare decomposed SVR variants, specifically WT-SVR and EMD-SVR variants, with conventional SVR models. The absolute MAE errors for all three categories are de-

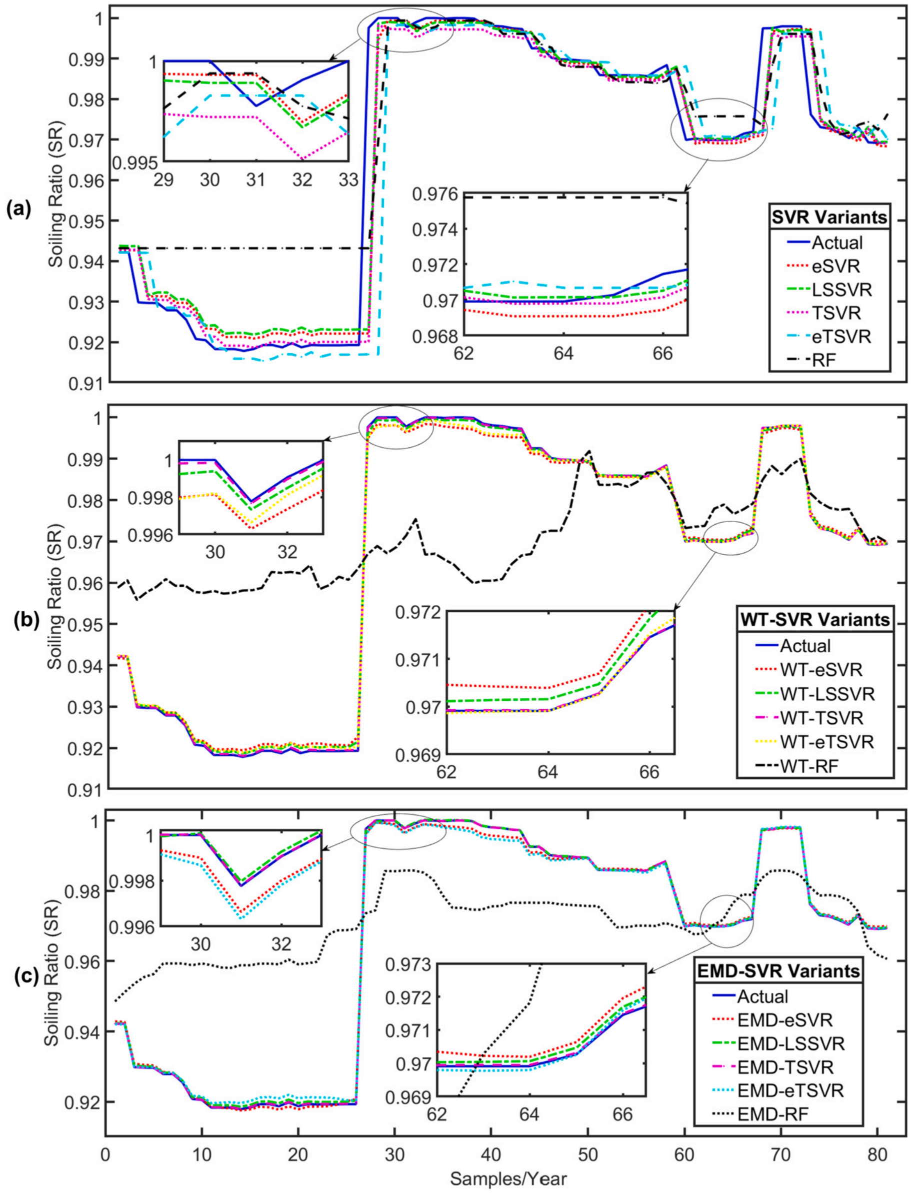

Fig. 8. Forecasting charts (a) for SVR variants, (b) for WT-SVR variants, and (3) for EMD-SVR variants.

picted in Fig. 9(a). Among these, the EMD-SVR variant exhibits superior prediction results compared to the others.

In terms of MAE, EMD-εSVR outperforms SVR variants and WT-variants by 80% and 41.67%, respectively. In the case of EMD-LSSVR, it is better by 91.89% and 40%. In the case of EMD-ε TSVR, it is better by 93.93% and 60%. And finally, EMD-TSVR gives the best result, i.e., 96.97%. The WT-TSVR and EMD-TSVR performance are the same. Next, the relative percentage improvement in MAE error of WT variants and EMD variants concerning MAE error in SVR variants are calculated and shown in Table 8 and Fig. 9 (b).

Within each model, the EMD decomposition prediction method demonstrates a greater reduction in error compared to the WT decomposition prediction method. At the end of this result discussion, the TSVR model, WT-TSVR model, and EMD-TSVR model are best among the rest of the SVR variants, respectively. The EMD-TSVR method gives better results than the WT-TSVR method, and the superiority of the EMD-SVR variants is checked by comparing TSVR with WT and EMD decomposition methods. In identifying random events, the EMD-TSVR surpasses the other two models in accuracy, as seen in the enlarged view shown in Fig. 10.

Table 8
Relative percentage error improvement of mean absolute error with respect to SVR variants.

Model	WT (%)	EMD (%)
ε-SVR	65.71	80.00
LSSVR	85.71	91.43
TSVR	97.14	97.14
ε-TSVR	85.71	94.29

6. Conclusions

Random events and dust storms detrimentally impact PV energy generation, and cleaning robots require additional energy for detecting and cleaning soiling. For energy-efficient cleaning, accurate input data regarding random events and uneven soiling is crucial for the robot. In this research work, a machine learning model for soiling detection in solar farms is developed using four SVR variants (namely, εSVR, LSSVR, TSVR, and εTSVR) and hybrid SVR models incorporating the wavelet

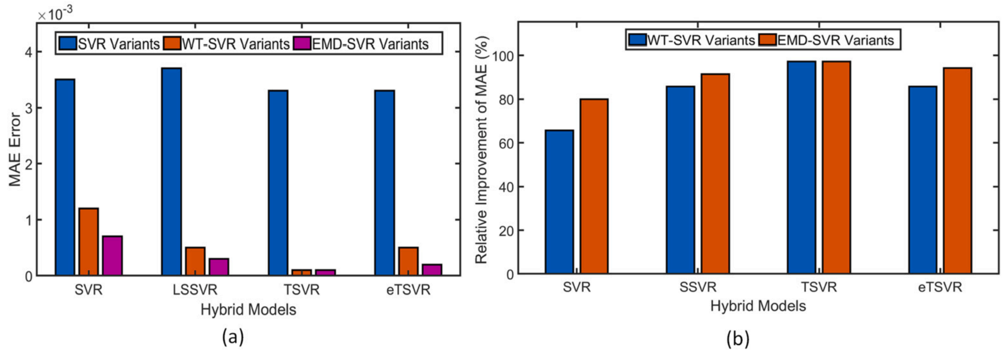

Fig. 9. MAE error of TSVR, WT-TSVR, and EMD-TSVR variants (a) absolute error, (b) relative percentage error.

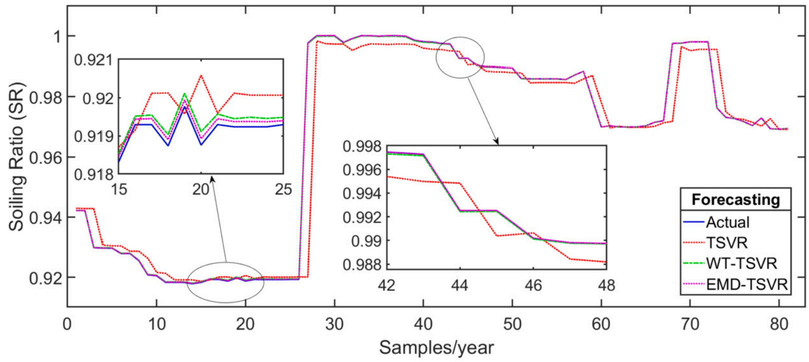

Fig. 10. Forecasting results of TSVR, WT-TSVR and EMD-TSVR.

transform (WT) and empirical mode decomposition (EMD) techniques. The comparison is conducted across three categories: within SVR variants, with wavelet transform (WT-SVR variants), and with empirical mode decomposition (EMD-SVR variants). The random forest algorithm serves as the benchmark model. The soiling signal is decomposed using a DB-4 wavelet filter at level 5. The optimal selection of SVR hyperparameters, such as regularization constants (c) and RBF bandwidth (s), is crucial in enhancing the soiling forecasting accuracy of these hybrid models. Model performance is assessed based on Mean Absolute Error (MAE), and the findings are as follows:

- Across all three categories, both conventional SVR variants and the hybrid SVR variants with WT and EMD outperform the benchmark Random Forest (RF) model for the soiling dataset.
- TSVR exhibits the lowest Mean Absolute Error (MAE) among the SVR variants. However, εTSVR requires the least computational time among the four models due to a smaller-sized optimization problem. The preference should be as per the most critical requirements.
- Within the three categories, the two hybrid models, WT and EMD, surpass SVR variants in MAE error and computational time. Notably, the EMD decomposed model is more efficient than WT, exhibiting an even further reduced relative percentage error. Additionally, EMD models demonstrate superior efficiency in forecasting random events compared to WT, which is attributed to their lower computational time requirements.
- When evaluating the overall performance indices across all regressors, the conventional TSVR and its EMD hybrid models emerge as preferred choices for short-term forecasting.

A significant constraint of short-term forecasting models is their dependency on extensive datasets for identifying random events, a factor crucial for distinguishing dust storms, uncertainties of natural cleanings, and the nonlinearity of the soiling rates. The future work of this research is to implement the hybrid model on a controller for a multi-cleaning agent robot, as recommended in the state-of-the-art literature, for validation purposes.

Declaration of competing interest

Authors can confirm that there is no conflict of interest with respect to the current manuscript.

References

[1] Benhmed K, Touati F, Al-Hitmi M, Chowdhury NA, Gonzales AS, Qiblawey Y, et al. Pv power prediction in Qatar based on machine learning approach. In: 2018 6th international renewable and sustainable energy conference (IRSEC). IEEE; 2018. p. 1–4.

[2] Bessa JG, Micheli L, Almonacid F, Fernández EF. Monitoring photovoltaic soiling: assessment, challenges, and perspectives of current and potential strategies. iScience 2021;24(3):102165.

[3] Dahlioui D, Laarabi B, Barhdadi A. Investigation of soiling impact on pv modules performance in semi-arid and hyper-arid climates in Morocco. Energy Sustain Dev 2019;51:32–9.

[4] Mejia F, Kleissl J, Bosch J. The effect of dust on solar photovoltaic systems. Energy Proc 2014;49:2370–6.

[5] Adinoyi MJ, Said SA. Effect of dust accumulation on the power outputs of solar photovoltaic modules. Renew Energy 2013;60:633–6.

[6] Caron JR, Littmann B. Direct monitoring of energy lost due to soiling on first solar modules in California. IEEE J Photovolt 2012;3(1):336–40.

[7] Micheli L, Deceglie MG, Muller M. Mapping photovoltaic soiling using spatial interpolation techniques. IEEE J Photovolt 2018;9(1):272–7.

[8] Sarver T, Al-Qaraghuli A, Kazmerski LL. A comprehensive review of the impact of dust on the use of solar energy: history, investigations, results, literature, and mitigation approaches. Renew Sustain Energy Rev 2013;22:698–733.

[9] Costa SC, Diniz ASA, Kazmerski LL. Dust and soiling issues and impacts relating to solar energy systems: literature review update for 2012–2015. Renew Sustain Energy Rev 2016;63:33–61.

[10] Chiteka K, Arora R, Sridhara S. A method to predict solar photovoltaic soiling using artificial neural networks and multiple linear regression models. Energy Syst 2020;11(4):981–1002.

[11] Terhag F, Wolfertstetter F, Wilbert S, Hirsch T, Schaudt O. Optimization of cleaning strategies based on ANN algorithms assessing the benefit of soiling rate forecasts. AIP conference proceedings, vol. 2126. AIP Publishing; 2019.

[12] Sahana L, Kumaar N, Waldl HP, Das PK, Ramanathan K, Balaraman K, et al. Impact of soiling on energy yield of solar pv power plant and developing soiling correction factor for solar pv power forecasting. Eur J Energy Res 2021;1(2):21–9.

[13] Cui Y-Q, Xiao J-H, Xiang J-L, Sun J-H. Characterization of soiling bands on the bottom edges of PV modules. Front Energy Res 2021;9:665411.

[14] Ballestrín J, Polo J, Martín-Chivelet N, Barbero J, Carra E, Alonso-Montesinos J, et al. Soiling forecasting of solar plants: a combined heuristic approach and autoregressive model. Energy 2022;239:122442.

[15] King M, Li D, Dooner M, Ghosh S, Roy JN, Chakraborty C, et al. Mathematical modelling of a system for solar PV efficiency improvement using compressed air for panel cleaning and cooling. Energies 2021;14(14):4072.

[16] Al Shehri A, Parrott B, Carrasco P, Al Saiari H, Taie I. Impact of dust deposition and brush-based dry cleaning on glass transmittance for pv modules applications. Sol Energy 2016;135:317–24.

[17] Younis A, Onsa M. A brief summary of cleaning operations and their effect on the photovoltaic performance in Africa and the middle East. Energy Rep 2022;8:2334–47.

[18] Kawamoto H. Electrostatic cleaning equipment for dust removal from soiled solar panels. J Electrost 2019;98:11–6.

[19] Belihi S, Dahlioui D, Laarabi B, Barhdadi A. On the use of dew for cleaning pv panels in Morocco: literature survey and experimental results. In: 2019 7th international renewable and sustainable energy conference (IRSEC). IEEE; 2019. p. 1–4.

[20] Gostein M, Düster T, Thuman C. Accurately measuring pv soiling losses with soiling station employing module power measurements. In: 2015 IEEE 42nd photovoltaic specialist conference (PVSC). IEEE; 2015. p. 1–4.

[21] Ebrahimi-Khusfi Z, Taghizadeh-Mehrjardi R, Mirakbari M. Evaluation of machine learning models for predicting the temporal variations of dust storm index in arid regions of Iran. Atmos Pollut Res 2021;12(1):134–47.

[22] Zhang W, Liu S, Gandhi O, Rodríguez-Gallegos CD, Quan H, Srinivasan D. Deep-learning-based probabilistic estimation of solar pv soiling loss. IEEE Trans Sustain Energy 2021;12(4):2436–44.

[23] Shahid F, Zameer A, Afzal M, Hassan M. Short term solar energy prediction by machine learning algorithms. arXiv preprint. arXiv:2012.00688, 2020.

[24] Khandakar A, Chowdhury MEH, Khoda Kazi M, Benhmed K, Touati F, Al-Hitmi M, et al. Machine learning based photovoltaics (PV) power prediction using different environmental parameters of Qatar. Energies 2019;12(14):2782.

[25] Dhiman HS, Deb D, Guerrero JM. Hybrid machine intelligent SVR variants for wind forecasting and ramp events. Renew Sustain Energy Rev 2019;108:369–79.

[26] developers scikit-learn. Support vector machines. https://scikit-learn.org/stable/modules/svm.html. [Accessed 13 November 2022].

[27] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.

[28] Yang H, Huang K, King I, Lyu MR. Localized support vector regression for time series prediction. Neurocomputing 2009;72(10–12):2659–69.

[29] Dhiman HS, Deb D, Balas VE. Supervised machine learning in wind forecasting and ramp event prediction. Academic Press; 2020.

[30] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14(3):199–222.

[31] Ma Z, Ye C, Ma W. Support vector regression for predicting building energy consumption in southern China. Energy Proc 2019;158:3433–8.

[32] Awad M, Khanna R. Support vector regression. In: Efficient learning machines. Springer; 2015. p. 67–80.

[33] Hsia J-Y, Lin C-J. Parameter selection for linear support vector regression. IEEE Trans Neural Netw Learn Syst 2020;31(12):5639–44.

[34] Peng X. Tsvr: an efficient twin support vector machine for regression. Neural Netw 2010;23(3):365–72.

[35] Shao Y, Gong Z. The existence of ε-approximate solutions to fuzzy functional differential equations. In: 2012 9th international conference on fuzzy systems and knowledge discovery. IEEE; 2012. p. 175–8.

[36] Leonardo Micheli MGD, Ruth Daniel, Muller M. Time series analysis of photovoltaic soiling station data: version 1.0. https://www.nrel.gov/research/publications.html. [Accessed 3 June 2022].

[37] Liu M, Sun X, Wang Q, Deng S. Short-term load forecasting using EMD with feature selection and TCN-based deep learning model. Energies 2022;15(19):7170.

Abhijeet Redekar, received his B.E. in Electrical Engineering from Shivaji University, Kolhapur, Maharashtra, India in 2010 and M.E. degree in Control system from Savitribai Phule Pune University (SPPU), formerly known as Pune University, Pune, Maharashtra, India in 2013. He is PhD scholar in Institute of Infrastructure, Technology, Research And Management (IITRAM), Ahmedabad, Gujarat, India and his research interests include control theory, adaptive and nonlinear controller design for electrical actuators used in the renewable applications, and FACTS controller.

Harsh S. Dhiman, Senior Member, IEEE received the B.Tech. degree in Electrical Engineering from Nirma University, Ahmedabad, India, in 2014, the master's degree in Electrical Power Engineering from The Maharaja Sayajirao University of Baroda, Vadodara, India, in 2016, and the Ph.D. degree from the Department of Electrical Engineering, Institute of Infrastructure Technology Research and Management, Ahmedabad, India, in June 2020. He is currently an Assistant Professor with the Department of Artificial Intelligence & Machine Learning, Symbiosis Institute of Technology, Pune, India. He has authored or coauthored 13 SCI-indexed journal articles and two books with Springer and Elsevier. His core research interests include condition monitoring and predictive maintenance of wind turbines. Apart from wind energy, he also remains interested in prognostics and diagnostics of fuel cell technology. He is an active reviewer with renowned journals such as IEEE Transactions in Instrumentation and Measurement, Expert Systems with Applications, and Journal of Fuzzy and Intelligent Systems.

Dipankar Deb is a Senior Member of IEEE and has served a couple of years at IIT Guwahati as an Assistant Professor (AGP 8000) during 2010-2012. He has over 6 years of Industrial experience both in New York (USA) and GE Global Research (Bengaluru) India. From July 2015 to Jan 2019, he has served as an Associate professor, and from Jan 24, 2019, onward he is a Professor in Electrical Engineering at Institute of Infrastructure Technology Research and Management (IITRAM) Ahmedabad. He holds 6 US patents and 3 Indian Utility Patents, and 1 Indian Design registration, and has published 52 SCI indexed Journal articles and 40+ International conference papers. He has also authored and edited 12 books with reputed publishers like Springer and Elsevier. He is a Book Series Editor with Studies in Infrastructure and Control (Springer) and CRC Press on Control Theory and Applications, and has served as an Associate Editor for IEEE Access (2019-2022). He has also worked extensively in areas such as Adaptive Control, Active flow control, Renewable Energy, Cognitive Robotics and Machine Learning. He is listed in the top 2% of Researchers worldwide for 2020 & 2021 as published by Stanford University.

S. M. Muyeen, (S'03–M'08–SM'12) is a full professor in the Electrical Engineering Department of Qatar University. He received his B.Sc. Eng. Degree from Rajshahi University of Engineering and Technology (RUET), Bangladesh, formerly known as Rajshahi Institute of Technology, in 2000 and M. Eng. and Ph.D. Degrees from Kitami Institute of Technology, Japan, in 2005 and 2008, respectively, all in Electrical and Electronic Engineering. His research interests are power system stability and control, electrical machine, FACTS, energy storage system (ESS), Renewable Energy, and HVDC system. He has been a Keynote Speaker and an Invited Speaker at many international conferences, workshops, and universities. He has published more than 350+ articles in different journals and international conferences. He has published seven books as an author or editor. He is serving as Editor/Associate Editor for many prestigious Journals from IEEE, IET, and other publishers, including IEEE Transactions on Energy Conversion, IEEE Power Engineering Letters, IET Renewable Power Generation and IET Generation, Transmission &amp; Distribution, etc. Dr. Muyeen is a senior member of IEEE, Chartered Professional Engineers, Australia, and a Fellow of Engineers Australia.