Desalination and Water Treatment
www.deswater.com
doi: 10.5004/dwt.2021.27950

244 (2021) 1–11
December

Predictive supervised machine learning models for double-slope solar stills

Hisham A. Maddah

Department of Chemical Engineering, Faculty of Engineering—Rabigh Branch, King Abdulaziz University, Jeddah 21589, Saudi Arabia, email: hmaddah@kau.edu.sa

Received 13 April 2021; Accepted 31 October 2021

ABSTRACT

There has been a growing trend to develop predictive solar-desalination models *via* machine learning (ML) and artificial intelligence (AI) tools. However, forecasting productivities of solar stills of different designs remains a challenge that can be overcome by establishing regression correlations *via* built-in and pre-existing ML toolboxes. Herein, the author developed accurate supervised predictive ML models for the productivity predictions in a double-slope still based on literature experimental results. Training datasets were constructed from the earlier observations (inputs/outputs) from various designed passive and/or active solar stills which were used to treat brackish water or wastewater with 45% TDS. A semi-proportional relationship between water-glass temperature ($T_w - T_g$) and water distillate was established with a minimum statistical error. The relationship proposed that an increase in both $T_w - T_g$ and basin temperature (T_B) would result in the maximum distillate at time 14:00. The regression models (FGSVM, EBoT, SEGPR) showed that they had the least achieved root mean square error (RMSE) of <138 indicating their reliability to accurately predict the distillate amounts in double-slope designs. The high accuracy of the SEGPR trained model with (R^2 = 1) and the very low RMSE < 8 showed the capability of the model to predict the performance in similar solar-desalination systems. Yet, the FGSVM was found to be more reliable in predicting $T_w - T_g$ whereas that the stepwise linear regression (SLR) better predicted the T_B pattern against the water distillate. This work suggests the importance of both the FGSVM and the SLR models for water outputs predictions which can pave the way towards establishing a unified theoretical tuning-parameters model to maximize the performance and the distillate water in double-slope solar stills.

Keywords: Supervised regression; Machine learning models; Solar still; Double slope; Distillates; Water-glass temperature

1. Introduction

The current problem of securing enough freshwater for the increasing population remains unsolved due to the constant and scarce available freshwater sources which have been estimated to be only around 1% of the total earth's water surface [1,2]. The excessive production of industrial wastewater has never been in today's skyrocketing rates owing to the high demands for technology and consumer products from various water-related manufacturing and industrial processes [3]. Thus, finding alternatives to the expensive membrane technologies *via* the emerging solar-desalination and solar distillation will help in covering the increasing world's demands for freshwater and water supply for industrial uses [3–5].

Solar distillation systems can be categorized as (i) direct: referring to the use of directly absorbed solar (radiation) energy in solar stills, and (ii) indirect: referring to the use of converted solar-to-thermal energy or solar-to-photovoltaic energy like in thermal and membranes systems [6]. The discovery of the interesting water distillation technology named "solar stills" since the late 19th century [7,8], which emerged into the water manufacturing and supply market in the last few decades, was found to be promising as an

alternative to the expensive membranes technology. The possibility of utilizing free and environmental-friendly energy provided by solar radiation for the water distillation is a breakthrough since no wastes are produced and no further energy is needed, reducing potential costs in using solar stills for desalination [7,8]. The water solar-distillation replicates the nature manner of "rain" by the recurring condensation/evaporation process of water [9]. The produced distilled water is usually potable with a very high quality due to the complete removal of total dissolved solids (TDS), inorganic, and organic contaminants [10].

Considering the direct passive solar stills for water distillation is a common practice due to their simple system design and ability to improve the water productivity in single slope solar stills *via* thermal insulation, energy storage medium, and solar radiation absorption [6]. Solar radiation, wind speed, surrounding temperature and air humidity, feed water temperature (or glass/water temperature difference), basin surface area, top-cover inclination angel, glass transparency for solar rays, and the desired feed flowrates (or maximum allowable water levels in the tank) are some of the important controlling factors needed to be taken into account for the optimization of the solar still productivity [9]. Among these parameters, the wall insulation film thickness (1–5 cm), the water depth (2–3.5 cm), the solar intensity, and the ambient, the water, and the vapor temperatures were found to critically impact the performance of direct passive solar stills. These parameters were previously examined experimentally and correlated to the performance *via* mathematical models with good prediction accuracy [11].

Wang et al. [12] observed that the saltwater temperature, the basin temperature, and the solar radiation were among the most important predictors (40.87%, 32.43%, and 18.2%, respectively) for productivity prediction in tubular solar stills. In terms of the design, the single slope solar still was found to be more effective than a pyramid-shaped still, with a 30% higher yield in winter, due to the minimized escape of absorbed radiation energy from the large surface cover [13]. Moreover, Cuce et al. [6] created a novel non-insulated solar distillation unit using a sensible medium for energy storage coupled with a passive booster reflector and found that total water productivity of 2,197.4 ml/week was achievable from the passive design with the sensible medium energy storage. But, the water productivity exclusively depended on the thermal insulation and the cooling for the aperture glazing and external structure which would drive up condensation-evaporation rates by minimizing heat losses and enhancing convective coefficients.

However, the utilization of solar radiation for solar stills and/or power conversion is still not widely industrialized due to the relatively high installation costs and low conversion rates. There should be more innovative research on the application of supervised machine learning (ML) and cross-validation (CV) techniques. This would help in materials selection [14], solar harvesting [15], and the application of solar-desalination *via* developing novel technology methodologies capable of comprehensively analyzing available literature datasets and patents designs. Such created algorithms can facilitate the advancement in the field of solar-desalination technologies for the commercialization of large-scale solar stills which will benefit the community, the companies' technical R&D centers, and the business sectors [16].

ML is an alternative way of dealing with complex non-linear problems [17] such as prediction of the solar still productivity [18], rather than using the conventional numerical analysis and the inaccurate complex regression models [19]. Based on real experimental data, the conventional methods which were utilized for the prediction of the still performance included: (i) numerical models based on solutions of differential equations of heat and mass transfer [20,21], (ii) regression models capable of predicting the relationship between multi-dependent variables (inputs) and the independent output [22,23], and (iii) trained models constructed from the ML and artificial intelligence (AI) toolboxes which were used for the energy and solar-desalination systems. The ML/AI analysis gave accurate predictions for the productivity and thermal efficiency as compared to the data resulting from the conventional regression or linear models [24,25]. Predictions from the non-linear trained ML models were found to be much accurate than those from the multilinear regression (MLR) models, showing the potential of such ML/AI toolboxes.

A major problem in solar stills is the low productivity, where many previous studies [26–30] attempted to increase the still productivity that is a prime design target which would ensure maintaining the solar still design simplicity and operation feasibility (economic-wise). Thus, there has been a growing trend for using ML and AI models for creating complex computational simulators or numerical models capable of solving environmental engineering-related problems. Mashaly et al. [31] established an artificial neural network (ANN) algorithm for the construction of a mathematical model as a useful and valuable tool meant to determine solar still productivity. Passive solar still fed with agricultural drainage water was studied to predict the instantaneous thermal efficiency. The ANN model predicted the experimental results accurately with minimum errors confirmed from the coefficient of determination (R^2 = 0.96), proving that it was possible to apply ANN to establish a general model for estimating the water productivity [31].

However, forecasting the performance or the water output of solar stills with various designs (or based on different surrounding conditions) remains a challenge to be investigated *via* the built-in and freely available ML toolboxes. This is because the solar still productivity depends on many parameters that need to be considered both implicitly and explicitly to ensure model adequacy for predicting the distillation efficiency to treat saline water from various sources [31]. In the simplest case, a solar still system can be modeled based on the common explicit and/or experimentally measured parameters (e.g., the water-glass temperature, the basin temperature, etc.) which can be used as independent variables correlated to the water distillates (as the only dependent variable) using various supervised ML regression learners.

Herein, the author created a framework to develop a supervised model with high prediction accuracy for the water productivity in double-slope solar stills based on the available built-in ML tools (MATLAB) and previous experiments. Collected data were taken from previously conducted experiments in a double-slop solar still which

were utilized for the treatment of (i) brackish water with high contents of sodium carbonates (40% soap solution), or (ii) wastewater of reverse osmosis (RO) plant with 45% TDS; with/without reflectors and/or phase change materials (PCM) [32]. The water-glass temperature difference and the basin temperature were correlated to the still outputs using various ML models to select the reliable models that would have the minimum statistical errors. Input variables included the following: the basin (T_B), the glass (T_g), the water (T_w) temperatures, and the average water-glass temperature difference ($T_w - T_g$) which were correlated to the water distillates. A thorough comparison between the different trained/tested model results was established to evaluate the performance of the developed models. The most reliable and promising models were then selected for further analysis to choose the optimum model that should predict similar results to those results found from experiments (testing) for forecasting the water productivity in double-slop solar stills.

2. Emergence of ML/AI predictive solar-desalination models

Knowledge-based ML can provide fast and precise predictions for solar-desalination systems and cost-effective designs based on previously proposed frameworks. Potential designs like water-in-glass evacuated tubes for solar water heaters are possibly feasible *via* system optimization using ML tools. A good framework consists of both a predictive model and datasets as screening candidates (testing), which require a combination of computational and experimental case studies for efficient designing and optimized performance of the solar-desalination systems [33]. There should be a clear understanding of the developed mathematical modeling methodologies for solar-desalination in the different classes of solar stills before carrying out such ML prediction analysis [34]. However, it is also necessary to have both high availabilities of data and powerful computational algorithms to perform highly accurate predictions. The ANN and gradient boosting regression trees (GBRT) were found to be the most accurate models in predicting solar conversion across five different sites in Sweden [35]. Li et al. [36] developed ML models using >300 ensemble data points to optimize material composition, design strategies, and performance of perovskite-based devices from studying structural stability [14]. The model predicted the maximum theoretical limit showing the promise of ML to provide an insightful understanding of the associated physical phenomena in the areas of energy engineering [36].

Theoretically, simple models of solar-desalination systems can be obtained from a comparison of analytical results with experimental data to express the long-term changes in water productivity [37,38]. Srivastava et al. [39] found from their computational work that there was an evident relationship between water temperatures and distilled output as a function of solar insolation, which impacted water levels and basin temperature. Another established mathematical model matched the experimental results in predicting that the maximum efficiency from a solar still was usually in the early afternoon due to the high solar radiation, where the ambient temperature and/or the solar intensity were both proportionally related to the solar productivity [11].

Sohani et al. [40] employed ANN for design enhancement of a solar desalination system using the obtained experimental data of a one-year operation. Four inputs were considered (ambient temperature, wind speed, sun's radiation, and water depth in the basin) and were correlated with only two outputs (water temperature and distillate) in the modeling analysis from the input-output relationship. Annual error analysis of the created models showed that the error for prediction of the daily water production was in the range of 2.41%–5.84%. High heat transfer rates were maintained using a flat-plate solar collector insulated with glass wool with blackened bottom-side (basin) solar still to maximize the solar absorption. Prediction behavior and accuracy of the created ANN models differed based on the used structure type, namely feedforward, backpropagation, and radial basis function structures. Results showed that both feedforward and backpropagation types had the highest $R^2 > 0.96$ indicating their potential for the estimation of the hourly water production and water temperature. Therefore, adopting such proven models in the future analysis will save time and costs using numerical computational analysis rather than experiments to recommend the optimum design before experimentations.

Moreover, the non-linear ANN and random forest (RF) ML regression methods were previously used for tubular solar still [12] to generate prediction models for the estimation of productivity. The built models were further optimized with the Bayesian optimization algorithm (BOA) from adjusting the hyperparameters. The RF is a supervised ML algorithm that consists of many decision trees combined into one model meant to improve statistical results [12]. Results of the established models were compared with those obtained from the MLR model with average productivity of 4.3 L/(m^2 d). The ANN model achieved optimal predictions with very high accuracy confirmed from determination coefficients ($R^2 > 0.997$) and was found to be much accurate than the MLR models. However, the RF was found to be a more stable model achieving better performance predictions than the ANN+BOA. In both cases, the models accurately predicted hourly production closer to true experimental observations [12]. Such models hold the promise in forecasting productivity from effective designing to achieve the highest distillate outputs [11]. These models were attained through exploiting ML/AI tools for easy simulation and optimization leading to efficient and economical systems for maximum distillate outputs [12].

2.1. Stepwise linear regression

Stepwise linear regression (SLR) works by regressing multiple variables while removing the weakest variables with low impact on the studied (predicted) parameter, following the general formula shown in Eq. (1); where Y is the predicted (dependent) parameter, X_i (i = 1, 2, …, n) is the predictor (independent variable), β_0 is the intercept, β_i (i = 1, 2, …, n) is the coefficient on the ith predictor [31]. A fitted distribution was carried out by an automatic procedure to have only those variables which best explain the correlation requiring a normal distribution behavior [41].

The addition or subtraction of a variable from the set of explanatory variables could be developed by a series of T-tests or F-tests performed *via* starting the test with either all available predictor variables or with no predictor variables to predict estimated outputs errors [41–43]. The ordinary least squares (OLS) method [43] identified the optimal values of β_n from finding the parameters that minimized the sum of the squared errors (MSE), as shown in Eq. (2), where y_i is the actual value and $\hat{y}_i$ is the predicted value [44].

$$\gamma = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{1}$$

$$\min \sum_{i=1}^{m} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{m} \left(y_i - \gamma\right)^2 \tag{2}$$

2.2. Support vector machines

Support vector machines (SVM) analysis was first identified by Vladimir Vapnik and his colleagues in 1995 [45] as a nonparametric statistical regression technique relying on kernel function and parameters selection. Cross-validation learning and gradient descent learning were some of the primary methods which were commonly used for kernel optimization and parameters selection. Considering a mixed kernel function strategy would result in models with decent learning ability for generalization purposes [46]. The predictor parameters and the response should be selected and analyzed carefully, respectively, from the training detests. Such selection would ensure having models with minimum errors and highest accuracy from finding a flat function $f(x)$ with ε as the maximum deviation from y_i for each training point x [47]. In other words, the function should have at most ε-deviation from the target from convex optimization based on three constraints and a tradeoff complexity. Typically, the goal should to find a regression function: f: $R^D \rightarrow R$:

$$y = f\left(x\right) = \omega^T \phi\left(x\right) + b \tag{3}$$

Knowing the following definitions, ω is a weight vector, $\phi(x)$ is a selected function for data mapping of x from a low dimension to a high dimension space, and b is an up or down numeric value. SVM regression adopts ε-insensitive function, where training data were assumed to follow a linear trendline with an accuracy associated with the ε value. Thus, function minimization might be optimized by converting the problem to an objective function as shown in the following [48]:

$$\min \frac{1}{2}\left\|\omega\right\|^2 + \frac{C}{2}\sum_{i=1}^{m}\left(\xi_i^2 + \xi_i^{*2}\right) \tag{4}$$

Under constraints:

$$\begin{aligned} &\omega^T \varphi\left(x_i\right) + b - y_i \le \varepsilon + \xi_i, \quad i = 1,2,\ldots,m \\ &y_i - \omega^T \varphi\left(x_i\right) - b \le \varepsilon + \xi_i^*, \quad i = 1,2,\ldots,m \\ &\xi_i, \xi_i^* \ge 0, \qquad\qquad i = 1,2,\ldots,m \end{aligned} \tag{5}$$

where ξ_i, ξ_i^* is the relaxation factor, which should be equal to 0 when there is no error in the fitting. The first term (left term) of the function shown in Eq. (4), for optimization purposes, allowed generalizing the model from the improved fitting smoothness. The second term (right term) of the function shown in Eq. (4) reduced the error and that when $C > 0$, there would be errors in the estimated regression with penalty indicated by the error ε [46]. The structure of the SVM regression is shown in Fig. 1. There should be an appropriate selection of the model that determines the most suitable kernel function for the data characteristics [48]. This would ensure accurate training based on the constructed kernel function type and relevant parameters [46].

2.3. Decision trees and ensemble

The decision tree built regression models from observations of datasets attributes or predictors (represented in the branches as a decision or terminal nodes) to reach conclusions about the numerical target variable continuous values (represented in the leaf nodes). It broke down the datasets into smaller and smaller subsets while simultaneously developing incremental associated decision trees. Regression trees approximated real-valued functions which were built through a process known as binary recursive partitioning [49]. This is an iterative process that split the data into partitions with the continuous splitting of each partition into smaller groups as the regression moves up each branch. The goal was to select the split that minimize the sum of the squared deviations from the mean in the two separate partitions [49–51]. Further, ensemble methods *via* bagging (bootstrap aggregating) offered strong models and better selection for very large datasets. This is usually carried out *via* successively training models to concentrate on records receiving inaccurate predictions, where predictors were combined by a weighted majority vote. Trained datasets were generated from random sampling and taken as inputs for the regression trees to calculate the average from used models and determine the predictions of the new data [50–52].

3. Methods and equations

The solar still datasets were gathered from previous experimental results found in the literature [32]. The collected datasets included measurements of the still basin, glass cover, and water temperatures against the water distillates from 10:00 to 16:00. The collected datasets covered six different solar still experiments (with or without reflectors and/or PCM) which were conducted at Solapur (Maharashtra) in India using a double-slope design and a basin area of 0.62 m^2 with a highest recorded efficiency of ~42.1% [32]. The solar still productivities were determined by calculating the cumulative amount of collected water over the selected period. The temperatures of water, basin, and glass were measured using thermocouples which recorded the temperatures with the corresponding distillate outputs at 1 h intervals [32].

The collected original datasets containing 48 numbers were expanded to 144 numbers from correlating the same distillate outputs to ±0.1 of the original T_w, T_g, and T_B (taking advantage of the non-considerable, but inevitable experimental errors). This approach of datasets expansion

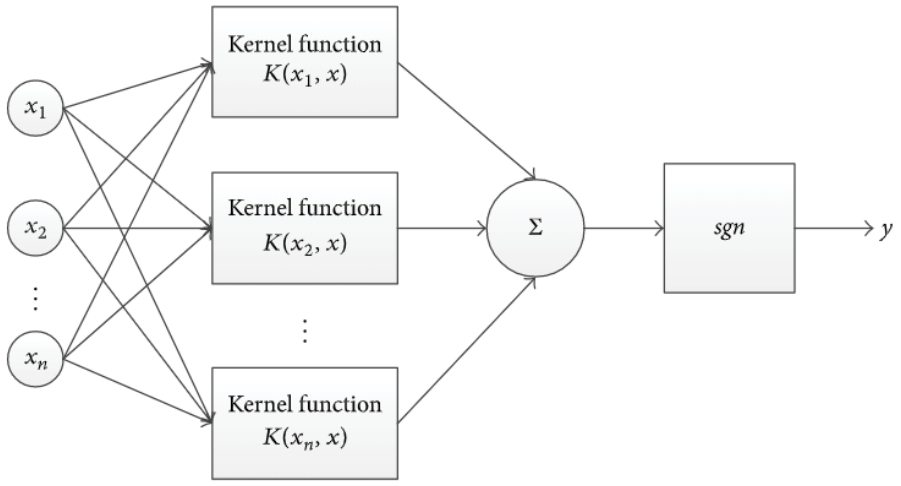

Fig. 1. Schematic diagram of SVM regression, adapted from [46].

by a 3-fold allowed us to have a better ML training/testing analysis. The curated datasets were then divided randomly into two groups: 80% for training and 20% for testing to check the validity and reliability of the built ML models in predicting the water distillates in double-slope designs (from being related to the pre-recorded T_w, T_g, and T_B from the experimental results found in the literature [32]).

Various supervised ML regression learners from the toolbox in MATLAB [53] were selected in training/testing labeled datasets. Linear, tree regression, SVM, and Gaussian process regression models (GPRM) were utilized to establish the trained models with a CV of 50-fold [53–56]. It is worth mentioning that the defined dependent variables (inputs) include: (i) basin temperature (T_B); (ii) average water temperature (T_w); (iii) inner-side glass temperature (T_{g_in}); (iv) outer-side glass temperature (T_{g_out}); (v) average glass temperature (T_g); and (vi) water-glass temperature difference ($T_w - T_g$). The only investigated output was the water distillate which was correlated to the studied inputs including the $T_w - T_g$ as the focus of our ML analysis. Identifying the correlation between $T_w - T_g$ and the water distillates is important to understand the impact of the water-glass temperature differences on the evaporation/condensation rates in solar stills. The daily radiation and the other atmospheric parameters such as wind velocity and humidity were not involved in our ML analysis since the collected datasets were taken from experiments carried out at the same location/time [32]. The water quality parameters of the feed and the distillate including the initial concentration of TDS, pH, alkalinity, hardness, chlorides concentration, and turbidity were listed in the earlier work [32] since these parameters were crucially important to evaluate the water safety standards for safe human consumption [57].

The trained models used in predicting the still distillate (in systems with or without reflectors and/or PCM) were utilized in the testing analysis of different supervised ML models. The training datasets (80% from the curated data) consist of seven matrices of [123 × 1], each matrix representing an input parameter or the distillate (output). This is also equivalent to saying that a one [123 × 7] matrix was curated considering the mentioned inputs and the distillate output to relate $T_w - T_g$ to the water productivity. The Testing datasets had the same inputs and were taken as 20% from the curated data to predict the known distillate for checking trained model accuracy.

For the assessment of the models' prediction ability, it is quite common to measure the models' validity *via* various statistical metrics including coefficient of determination (R^2), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and residual; which were calculated using their mathematical definitions in Eqs. (6)–(10) [58–60]. Knowing that the observed value is symbolized as $x_{o,i}$ and/or x_o; $x_{p,i}$ and/or x_p refer to the values predicted by the ML model; predicted value $\overline{x}_o$ is the experimentally obtained or observed value from averaging; $\overline{x}_p$ is the theoretically estimated or predicted value from averaging; and n refer to the datasets size or number of experimental observations. RMSE and MAE metric numbers (ranging from 0 to ∞) allowed us to demonstrate more accurate prediction results where more similarities arise between the trendlines of both experimental and predicted samples with high R^2 (reaching an identical pattern when R^2 = 1) [31]. Once the regression learners were trained, the statistical errors (e.g., RMSE, R^2, MSE, MAE) were then obtained from the different trained models and compared with one another.

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(x_{o,i} - \overline{x}_o\right)\left(x_{p,i} - \overline{x}_p\right)\right]^2}{\sum_{i=1}^{n}\left(x_{o,i} - \overline{x}_o\right)^2 \times \sum_{i=1}^{n}\left(x_{p,i} - \overline{x}_p\right)^2} \tag{6}$$

$$\text{MSE} = \frac{\sum_{i=1}^{n}\left(x_{o,i} - x_{p,i}\right)^2}{n} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{o,i} - x_{p,i}\right)^2}{n}} \tag{8}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left| x_{o,i} - x_{p,i} \right|}{n} \tag{9}$$

$$\text{Residual} = x_o - x_p \tag{10}$$

The selection of the best models (e.g., SLR = Stepwise linear regression, FT = Fine-Trees, MT = Medium-Trees, FGSVM = Fine-Gaussian-SVM, EBoT = Ensemble-Boosted-Trees, EBaT = Ensemble-Bagged-Trees, SEGPR = Squared-Exponential-Gaussian-Process-Regression) were carried out by checking whether the $R^2 > 0.7$ or not and by predicting the distillates from the testing datasets. Only those models which met the previous conditions were kept for further analysis to compare their predicted response patterns and residuals (from the testing datasets) to the observed experimental results (from the training datasets). Lastly, the FGSVM ($R^2 > 0.95$) trained model was chosen for in-depth analysis against the SLR ($R^2 > 0.68$) to show the promise behind selecting support vector machines as regressors when compared with stepwise linear regressors. Fig. 2 shows a flowchart illustrating the development and selection of the optimal supervised ML models for accurate prediction of distillates correlated with $T_w - T_g$.

4. Results and discussion

The FGSVM and the SLR regression models have been utilized to estimate the solar still distillates. The trained FGSVM model showed high accuracy of prediction as compared to the SLR model due to its higher R^2 and lower statistical errors (RMSE and MAE), as shown in Fig. 3A and B. Testing datasets had also confirmed the model's reliability in predicting the water distillates (Fig. 3C). Testing results showed that the FGSVM model correctly predicted most of the distillate outputs with only two outliers (Fig. 3D). Conversely, the SLR model had many outliers and was not considered as a good model built from training 80% of the datasets.

Moreover, the other supervised ML trained models (e.g., FT, MT, EBoT, EBaT, SEGPR) suggested that the different tried models had not perfectly predicted distillate outputs when compared with the observed distillates yellow line in Fig. 4A. However, there was an exception in which the SEGPR trained model was able to correctly predict each distillate value owing to its ideal ($R^2 = 1$) and very low RMSE < 8. The other models including the FT, MT, and EBaT trained models had shown the most outliers due to their more scattered values of [predicted vs. observed], indicated by the deviation of predicted values from the linear ideality or the plotted linear line [observed:predicted] = [1:1], Fig. 4B. The residuals of the trained models from Fig. 4C were also in good agreement with the obtained training prediction results. This explained the found low accuracy of the utilized regression trees training models in comparison to the ideal FGSVM and SEGPR gaussian training models for predicting water distillates.

According to the testing datasets analysis, the SEGPR and EBaT models showed the least number of outliers in their distillate prediction for the testing datasets, Fig. 4E. These results were also in agreement with the observed

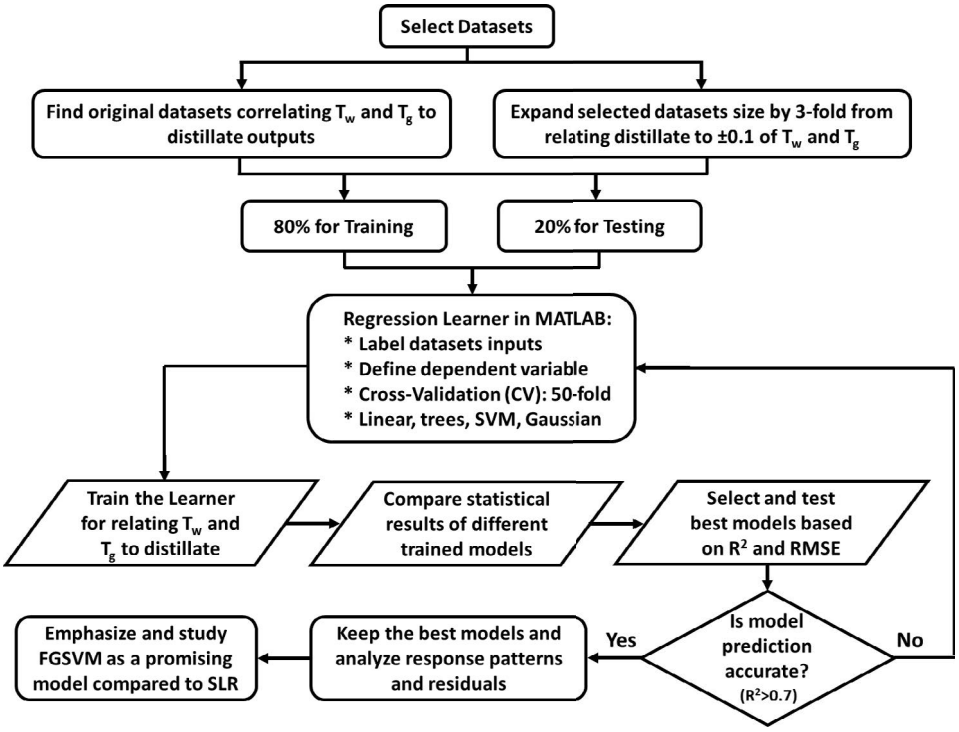

Fig. 2. Flowchart for developing and selecting optimal supervised machine learning (ML) models for accurate prediction of distillates correlated with water-glass temperatures.

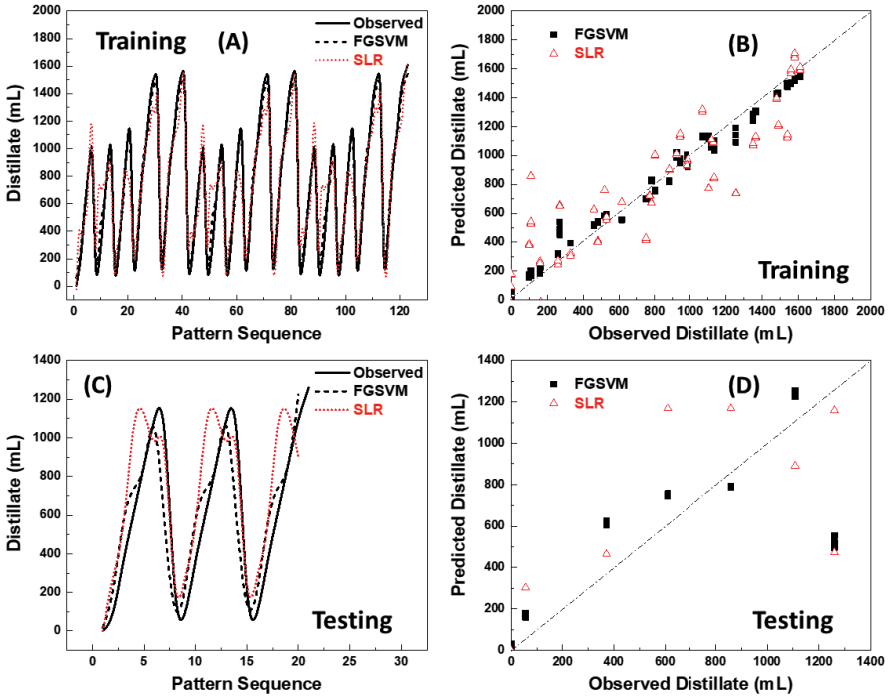

Fig. 3. Comparison between (observed vs. predicted) distillate values using FGSVM and SLR models: (A) and (B) from training datasets; (C) and (D) from testing datasets.

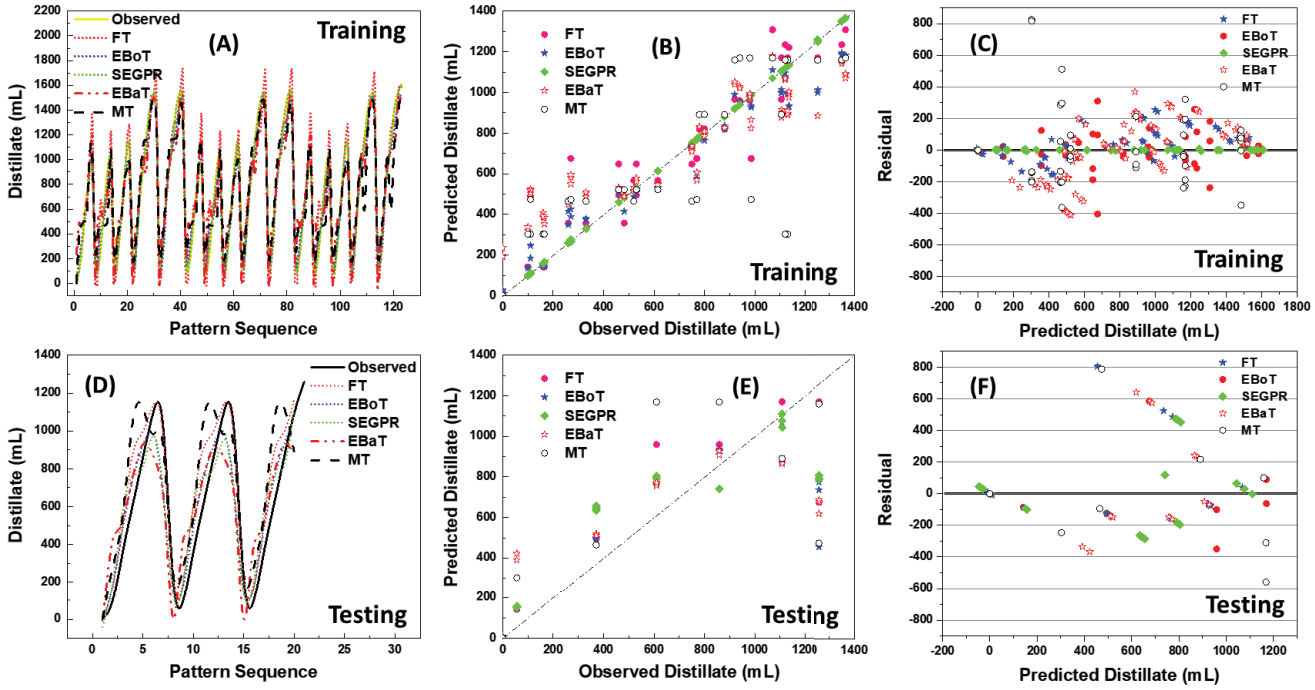

Fig. 4. Comparison between (observed vs. predicted) distillate values using FT, MT, EBoT, EBaT, and SEGPR models with their obtained residuals: (A)–(C) from training datasets; (D)–(F) from testing datasets.

model trends over the tested pattern sequence as shown in Fig. 4D. None of the trained models were able to predict the last few points of the testing datasets because of the previously observed deviation in the training models. Despite that the SEGPR training model had perfect predictions, the model was still unable to accurately predict the observed distillates for the testing datasets due to the recognized differences in patterns. Nevertheless, the SEGPR

had shown the expected least residuals (i.e., prediction errors) among the other tested models as illustrated in Fig. 4F. Similarly, residuals of the predicted distillates for the FGSVM and the SLR trained models and their testing were calculated and plotted in Fig. 5A and B, respectively. The FGSVM training and testing residual results were very close to the zero value indicating fewer prediction errors than the errors observed from the SLR model. The less scattered results of the FGSVM in Fig. 5B were found to be closer to the zero lines and with very few outliers confirming the model reliability. The author was able to predict the relationship between $T_w - T_g$ and water distillate from the testing datasets and based on the experimentally observed and ML prediction models, as shown in Fig. 5C and F. It is worth mentioning that these results were taken from 10:00 to 14:00 from the experiment datasets reported in the literature [32]. Data fitting in Fig. 5C was done using the "ORIGIN Software" with a third-order polynomial tool function which gave us $R^2 > 0.74$. There was a semi-proportional relationship between $T_w - T_g$ and water distillate whereas that a noticeable increase in the water-glass temperature difference resulted in the maximum water distillate at time 14:00. The exponential fitting of the data in Fig. 5F with [$R^2 > 0.94$ for 'observed and FGSVM' and $R^2 \approx 0.75$ for SLR], showed an increase in outputs with T_B.

The correlation between $T_w - T_g$ and the water distillates was initiated from the plotted training trends of dependent (distillate) and independent ($T_w - T_g$) parameters, as shown in Fig. 5D. The distillate outputs should be proportionally related to $T_w - T_g$. This relationship was observed from the trained datasets in Fig. 5D with only three outliers at the overestimated $T_w - T_g$ = 23°C which correspond to the low distillate outputs of ~1,100 mL. However, the rest of the training pattern was accurate showing the expected proportionality which was validated using the trained models for the testing datasets generating a similar pattern as shown in Fig. 5E. Note that the testing patterns sequence from 1 to 7 (or 8 to 14, 15 to 21) in Fig. 5E were attributed to the time from 10:00 to 16:00 with a 1 h increment. It was noticed that the distillate outputs increased after $T_w - T_g$ took place with the distillate curve being super-positioned by 2 points (or 2 h from 14:00 to 16:00). This delay for the highest outputs, based on the testing analysis, might.

be attributed to the fact that the water evaporation/condensation process takes some time to be accelerated at higher temperatures. Once the $T_w - T_g$ is at its peak, water vapors will take some time to reach the dew point or droplet formation. Water droplets form with vapor condensation (the dew point) at which the still inner or basin temperature is much higher than the surrounding temperature facilitating the water to glass evaporative heat transfer coefficient (h_{ewg}). To sum it up, there should always be high incident solar radiation to raise T_w higher than the relative T_g or the glass (or cover) temperature. Such high differences in temperatures should be favored to boost water evaporation and reach the dew point (condensation) promoted by the high h_{ewg} for increased water productivity [57].

As mentioned, the predicted relationships between $T_w - T_g$ and/or T_B against water distillate based on the FGSVM and the SLR trained models are shown in Fig. 5C

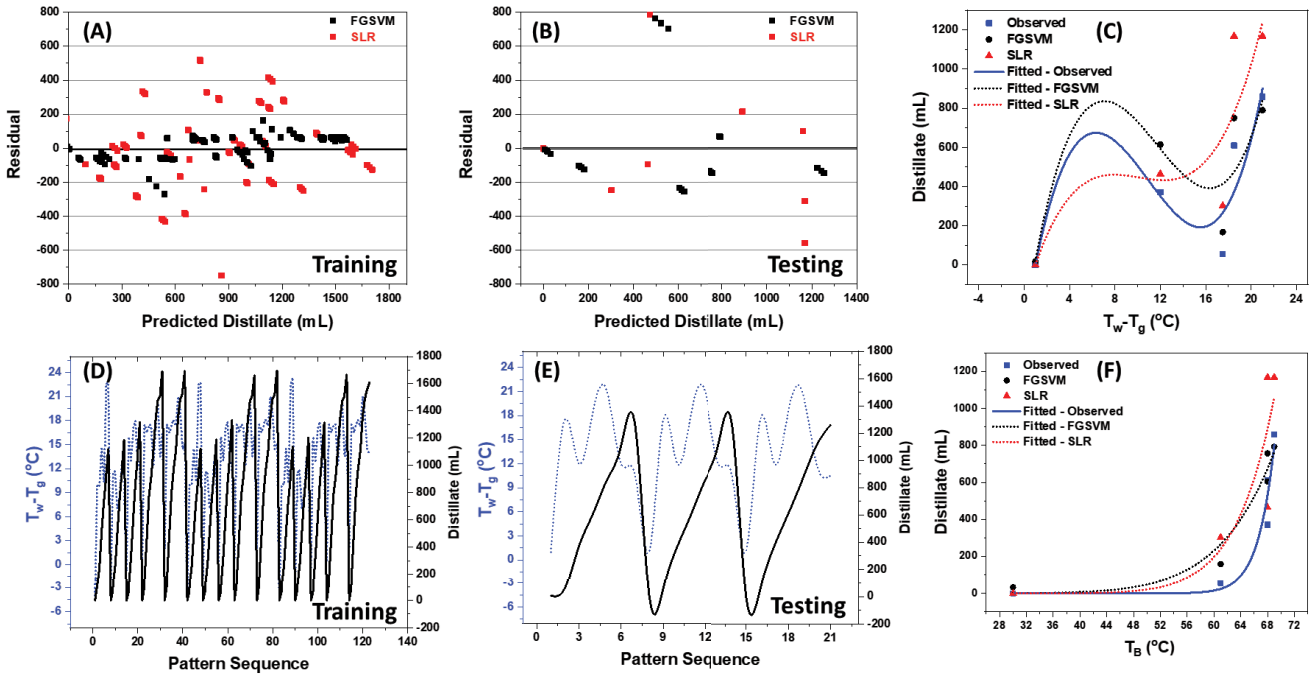

Fig. 5. Residuals of FGSVM and SLR regression models for selected (A) training and (B) testing datasets; (C) Predicted relationship between $T_w - T_g$ and water distillate based on FGSVM and SLR trained models (using third-order polynomial fitting); the observed relationship between the independent variable ($T_w - T_g$) and the dependent variable (distillate) found in (D) training and (E) testing datasets; (F) Predicted relationship between T_B and water distillate based on FGSVM and SLR trained models (using exponential fitting; $y = ax^b$).

Table 1
Statistical errors of various applied regression models for the prediction of $T_w - T_g$ and water distillates

Model	SLR	FT	MT	FGSVM	EBoT	EBaT	SEGPR
RMSE	298.59	174.25	296.48	117.39	138.18	241.22	7.70
R^2	0.68	0.89	0.69	0.95	0.93	0.79	1.00
MSE	89,158	30,362	87,902	13,781	19,093	58,186	59.29
MAE	230.64	102.98	205.75	84.75	96.09	187.18	4.03

SLR = Stepwise Linear Regression, FT = Fine-Trees, MT = Medium-Trees, FGSVM = Fine-Gaussian-SVM, EBoT = Ensemble-Boosted-Trees, EBaT = Ensemble-Bagged-Trees, SEGPR = Squared-Exponential-Gaussian-Process-Regression; Cross-Validation (CV): 50-fold; The author considered that reliable models should have minimum MSE or RMSE with $R^2 > 0.90$, which were shown in bolded numbers for the best three found regressions models.

and F, which were plotted using a third-order polynomial and an exponential fitting in ORIGIN, respectively. The FGSVM trained models were found to be more reliable for $T_w - T_g$ against water distillate, whereas the SLR models predicted almost a similar pattern (superposition) for T_B against water distillate; suggesting that both models were valid to consider when it comes to the prediction of water outputs based on water, glass, or basin temperatures in such double-slop solar still systems. The calculated statistical errors of the various applied regression models which were used in the prediction of water distillates are shown in Table 1. The FGSVM, EBoT, and SEGPR showed the least possible mean square errors indicating the reliability of these ML models for accurate predictions of future datasets from double-slope passive or active solar stills.

5. Conclusion

This work laid out novel statistical techniques for the development of highly accurate predictive supervised ML models for the water productivity in double-slope solar stills based on previous experimental datasets. The model development was initiated *via* training the collected datasets from earlier conducted experiments in double-slope solar stills (η~42.1%) for the treatment of brackish/wastewater containing 45% TDS. Input variables were taken as the basin temperature (T_B), the glass (T_g), and the water (T_w) temperatures which were trained/tested against their experimentally observed water distillates (output) for prediction of the water productivity. A semi-proportional relationship between the water-glass temperature difference ($T_w - T_g$) and the water distillate was established. A noticeable increase in $T_w - T_g$ and T_B parameters resulted in the maximum water distillate at time 14:00. The FT, MT, and EBaT trained models had the most outliers showing low accuracy of regression trees. The FGSVM trained model was found to be more reliable than the SLR models for $T_w - T_g$ against the water distillate. This observation was linked to the FGSVM [trained vs. tested] calculated residuals which were very close to the zero value indicating fewer prediction errors. A highly accurate predictive model with ($R^2 = 1$) and low RMSE < 8 was created using the SEGPR training for the prediction of the distillate. However, the FGSVM and the SLR models were found to be more reliable in predicting $T_w - T_g$ against the water distillate and T_B against the water distillate, respectively. The built trained models can be further optimized by using inputs as the water flowrate, the insulator properties, and the surroundings (environmental) conditions for more accurate predictions. Such theoretical models would guide us towards tuning the correct parameters correlated with the convective, the evaporative, and the radiative coefficients for maximizing the distillate-water outputs in double-slope solar stills.

Nomenclature (i.e., abbreviations, notation and units)

Abbreviation/unit

ML	—	Machine learning
CV	—	Cross-validation
AI	—	Artificial Intelligence
MLR	—	Multilinear regression
ANN	—	Artificial neural network
RO	—	Reverse osmosis
PCM	—	Phase change materials
T_B	—	Basin temperature
T_g	—	Glass temperature
T_w	—	Water temperature
$T_w - T_g$	—	Average water-glass temperature difference
RF	—	Random forest
BOA	—	Bayesian optimization algorithm
SLR	—	Stepwise linear regression
γ	—	Predicted (dependent) parameter
X_i	—	Predictor (independent variable)
β_0	—	Intercept
β_i	—	Coefficient on the *i*th predictor
OLS	—	Ordinary least squares
MSE	—	Mean square error
y_i	—	Actual value
$\hat{y}_i$	—	Predicted value
SVM	—	Support vector machines
$f(x)$	—	Regression function f: $R^D \rightarrow R$
x	—	Training point
b	—	Numeric value (up or down)
ω	—	Weight vector
$\phi(x)$	—	Function to map data x from low-to-high dimensional space
ε	—	Training data fitting accuracy
ζ_i, ζ_i^*	—	Relaxation factors equal to zero with no fitting errors
$C > 0$	—	Extent of the penalty for a sample out of error ε

GPRM	—	Gaussian process regression models
T_{g_in}	—	Inner-side glass temperature
T_{g_out}	—	Outer-side glass temperature
T_g	—	Average glass temperature
R^2	—	Coefficient of determination
RMSE	—	Root mean square error
MAE	—	Mean absolute error
FT	—	Fine-Trees
MT	—	Medium-Trees
FGSVM	—	Fine-Gaussian-SVM
EBoT	—	Ensemble-Boosted-Trees
EBaT	—	Ensemble-Bagged-Trees
SEGPR	—	Squared-Exponential-Gaussian-Process-Regression
x_o	—	Observed value
x_p	—	Predicted value
$\overline{x}_o$	—	Averaged observed values
$\overline{x}_p$	—	Averaged predicted values
n	—	Number of observations
h_{ewg}	—	Water to glass evaporative heat transfer coefficient

Acknowledgement

The author would like to acknowledge the Deanship of Scientific Research (DSR) at King Abdulaziz University (KAU) for their support and motivation to complete this work.

References

[1] B. Gupta, T. Kumar Mandraha, P. Edla, M. Pandya, Thermal modeling and efficiency of solar water distillation: a review, Am. J. Eng. Res., 2 (2013) 203–213.
[2] G.N. Tiwari, H.N. Singh, R. Tripathi, Present status of solar distillation, Sol. Energy, 75 (2003) 367–373.
[3] H. Maddah, A. Chogle, Biofouling in reverse osmosis: phenomena, monitoring, controlling and remediation, Appl. Water Sci., 7 (2016) 2637–2651.
[4] H.A. Maddah, A.M. Chogle, Applicability of low pressure membranes for wastewater treatment with cost study analyses, Membr. Water Treat., 6 (2015) 477–488.
[5] K. Sampathkumar, T.V. Arjunan, P. Pitchandi, P. Senthilkumar, Active solar distillation—a detailed review, Renewable Sustainable Energy Rev., 14 (2010) 1503–1526.
[6] E. Cuce, P.M. Cuce, A. Saxena, T. Guclu, A.B. Besir, Performance analysis of a novel solar desalination system – Part 1: The unit with sensible energy storage and booster reflector without thermal insulation and cooling system, Sustainable Energy Technol. Assess., 37 (2020) 100566, doi: 10.1016/j.seta.2019.100566.
[7] T. Arunkumar, K. Vinothkumar, A. Ahsan, R. Jayaprakash, S. Kumar, Experimental study on various solar still designs, ISRN Renewable Energy, 2012 (2012) 1–10.
[8] D. Kumar, P. Himanshu, Z. Ahmad, Performance analysis of single slope solar still, Int. J. Mech. Robot. Res., 3 (2013) 66–72.
[9] P. Kalita, A. Dewan, S. Borah, A review on recent developments in solar distillation units, Sadhana – Acad. Proc. Eng. Sci., 41 (2016) 203–223.
[10] A. Saxena, N. Deval, A high rated solar water distillation unit for solar homes, J. Eng. (United Kingdom), 2016 (2016) 1–8.
[11] O.O. Badran, M.M. Abu-Khader, Evaluating thermal performance of a single slope solar still, J. Heat Mass Transfer, 43 (2007) 985–995.
[12] Y. Wang, A.W. Kandeal, A. Swidan, S.W. Sharshir, G.B. Abdelaziz, M.A. Halim, A.E. Kabeel, N. Yang, Prediction of tubular solar still performance by machine learning integrated with Bayesian optimization algorithm, Appl. Therm. Eng., 184 (2021) 116233, doi: 10.1016/j.applthermaleng.2020.116233.
[13] H.E.S. Fath, M. El-Samanoudy, K. Fahmy, A. Hassabou, Thermal-economic analysis and comparison between pyramid-shaped and single-slope solar still configurations, Desalination, 159 (2003) 69–79.
[14] H.A. Maddah, V. Berry, S.K. Behura, Cuboctahedral stability in Titanium halide perovskites via machine learning, Comput. Mater. Sci., 173 (2020) 109415, doi: 10.1016/j.commatsci.2019.109415.
[15] H.A. Maddah, V. Berry, S.K. Behura, Biomolecular photosensitizers for dye-sensitized solar cells: recent developments and critical insights, Renewable Sustainable Energy Rev., 121 (2020) 109678, doi: 10.1016/j.rser.2019.109678.
[16] A.J.C. Trappey, P.P.J. Chen, C.V. Trappey, L. Ma, A machine learning approach for solar power technology review and patent evolution analysis, Appl. Sci., 9 (2019) 1478, doi: 10.3390/app9071478.
[17] A.K. Jain, J. Mao, K.M. Mohiuddin, Artificial neural networks: a tutorial, Computer, 29 (1996) 31–44.
[18] P. Gao, L. Zhang, K. Cheng, H. Zhang, A new approach to performance analysis of a seawater desalination system by an artificial neural network, Desalination, 205 (2007) 147–155.
[19] M.S.S. Abujazar, S. Fatihah, I.A. Ibrahim, A.E. Kabeel, S. Sharil, Productivity modelling of a developed inclined stepped solar still system based on actual performance and using a cascaded forward neural network model, J. Cleaner Prod., 170 (2018) 147–159.
[20] Y. Gong, X.L. Wang, L.X. Yu, Process simulation of desalination by electrodialysis of an aqueous solution containing a neutral solute, Desalination, 172 (2005) 157–172.
[21] H. Ben Bacha, T. Damak, M. Bouzguenda, A.Y. Maalej, H.B. Ben Dhia, A methodology to design and predict operation of a solar collector for a solar-powered desalination unit using the SMCEC principle, Desalination, 156 (2003) 305–313.
[22] X. Wang, K.C. Ng, Experimental investigation of an adsorption desalination plant using low-temperature waste heat, Appl. Therm. Eng., 25 (2005) 2780–2789.
[23] G. Yuan, L. Zhang, H. Zhang, Experimental research of an integrative unit for air-conditioning and desalination, Desalination, 182 (2005) 511–516.
[24] A.F. Mashaly, A.A. Alazba, MLP and MLR models for instantaneous thermal efficiency prediction of solar still under hyper-arid environment, Comput. Electron. Agric., 122 (2016) 146–155.
[25] A.F. Mashaly, A.A. Alazba, A.M. Al-Awaadh, M.A. Mattar, Predictive model for assessing and optimizing solar still performance using artificial neural network under hyper arid environment, Sol. Energy, 118 (2015) 41–58.
[26] G.M. Ayoub, L. Malaeb, Developments in solar still desalination systems: a critical review, Crit. Rev. Environ. Sci. Technol., 42 (2012) 2078–2112.
[27] R.S. Adhikari, A. Kumar, G.D. Sootha, Simulation studies on a multi-stage stacked tray solar still, Sol. Energy, 54 (1995) 317–325.
[28] R.S. Adhikari, A. Kumar, M.S. Sodha, Thermal performance of a multi-effect diffusion solar still, Int. J. Energy Res., 15 (1991) 769–779.
[29] H.A. Maddah, Modeling and designing of a novel lab-scale passive solar still, J. Eng. Technol. Sci., 51 (2019) 303–322.
[30] H.A. Maddah, Highly efficient solar still based on polystyrene, Int. J. Innov. Technol. Explor. Eng., 8 (2019) 3423–3425.
[31] A.F. Mashaly, A.A. Alazba, Neural network approach for predicting solar still production using agricultural drainage as a feedwater source, Desal. Water Treat., 59 (2016) 28646–28660.
[32] S.V. Kumbhar, Double slope solar still distillate output data set for conventional still and still with or without reflectors and PCM using high TDS water samples, Data Brief, 24 (2019) 103852, doi: 10.1016/j.dib.2019.103852.
[33] H. Li, Z. Liu, K. Liu, Z. Zhang, Predictive power of machine learning for optimizing solar water heater performance: the potential application of high-throughput screening, Int. J. Photoenergy, 2017 (2017) 1–10.

[34] H.S. Aybar, A Review of Desalination by Solar Still, In: Solar Desalination for the 21st Century, NATO Security through Science Series C: Environmental Security, 2007, pp. 207–214, doi: 10.1007/978-1-4020-5508-9_15.
[35] E. Isaksson, Solar Power Forecasting with Machine Learning Techniques, KTH Royal Institute of Technology School of Engineering Sciences, Degree Project in Mathematics Second Cycle, 30 Credits, Stockholm, Sweden, 2018, pp. 1–46.
[36] J. Li, B. Pradhan, S. Gaur, J. Thomas, Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells, Adv. Energy Mater., 9 (2019) 1901891, doi: 10.1002/aenm.201901891.
[37] E. Mathioulakis, K. Voropoulos, V. Belessiotis, Modeling and prediction of long-term performance of solar stills, Desalination, 122 (1999) 85–93.
[38] K. Voropoulos, E. Mathioulakis, V. Belessiotis, Analytical simulation of energy behavior of solar stills and experimental validation, Desalination, 153 (2003) 87–94.
[39] N.S.L. Srivastava, M. Din, G.N. Tiwari, Performance evaluation of distillation-cum-greenhouse for a warm and humid climate, Desalination, 128 (2000) 67–80.
[40] A. Sohani, S. Hoseinzadeh, S. Samiezadeh, I. Verhaert, Machine learning prediction approach for dynamic performance modeling of an enhanced solar still desalination system, J. Therm. Anal. Calorim., (2021) 1–12, doi: 10.1007/s10973-021-10744-z.
[41] University of Leeds, Stepwise Linear Regression: School Of Geography. Available at: http://www.geog.leeds.ac.uk/courses/other/statistics/spss/stepwise/
[42] N.R. Draper, H. Smith, Applied Regression Analysis, 3rd ed., John Wiley & Sons, United Kingdom, 2014.
[43] Guru99, R Simple, Multiple Linear and Stepwise Regression, 2020. Available at: https://www.guru99.com/r-simple-multiple-linear-regression.html
[44] P. Paisitkriangkrai, Linear Regression and Support Vector Regression, The University of Adelaide, 2012. Available at: https://cs.adelaide.edu.au/~chhshen/teaching/ML_SVR.pdf
[45] V.N. Vapnik, The Nature of Statistical Learning Theory, 1995. Available at: https://www.dais.unive.it/~pelillo/Didattica/Artificial%20Intelligence/Old%20Stuff/2015–2016/Slides/SLT.pdf
[46] H. Wang, D. Xu, Parameter selection method for support vector regression based on adaptive fusion of the mixed Kernel function, J. Control Sci. Eng., 2017 (2017) 1–12.
[47] MathWorks, Understanding Support Vector Machine Regression, 2020. Available at: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html
[48] S. Ghassem Pour, F. Girosi, Joint Prediction of Chronic Conditions Onset: Comparing Multivariate Probits with Multiclass Support Vector Machines, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, pp. 185–195.
[49] S. Sayad, Decision Tree – Regression, Data Science: Predicting the Future, Modeling & Regression. Available at: https://www.saedsayad.com/decision_tree_reg.htm
[50] Frontline Solvers – Frontline Systems, Regression Trees, 2020. Available at: https://www.solver.com/regression-trees
[51] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and regression trees, Taylor & Francis Group, Boca Raton, 2017.
[52] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Animal Ecol., 77 (2008) 802–813.
[53] Mathworks, Statistics and Machine Learning Toolbox™ User's Guide R2017a, MatLab, 2017.
[54] S.B. Kotsiantis, Supervised machine learning: a review of classification techniques, Informatica (Ljubljana), 31 (2007) 249–268.
[55] Y. Baştanlar, M. Ozuysal, Introduction to Machine Learning: miRNomics: MicroRNA Biology and Computational Analysis, Springer Nature, Switzerland, 2014.
[56] O. Simeone, A brief introduction to machine learning for engineers, IEEE Trans. Cognit. Commun. Networking, 4 (2018) 648–664.
[57] R. Pillai, A.T. Libin, M. Mani, Study into solar-still performance under sealed and unsealed conditions, Int. J. Low-Carbon Technol., 10 (2015) 354–364.
[58] M.M. Rahman, B.K. Bala, Modelling of jute production using artificial neural networks, Biosyst. Eng., 105 (2010) 350–356.
[59] M. Zangeneh, M. Omid, A. Akram, A comparative study between parametric and artificial neural networks approaches for economical assessment of potato production in Iran, Spanish J. Agric. Res., 3 (2011) 661–671.
[60] A.A. Alazba, M.A. Mattar, M.N. ElNesr, M.T. Amin, Field assessment of friction head loss and friction correction factor equations, J. Irrig. Drain. Eng., 138 (2012) 166–176.