Predictive Regression Models for Solar Energy Harvesting and Sustainable, Low Energy, Highly Efficient Solar-Desalination Systems

Hisham A. Maddah*
Department of Chemical Engineering, Faculty of Engineering—Rabigh Branch
King Abdulaziz University, Jeddah 21589, Saudi Arabia
*Corresponding author. hmaddah@kau.edu.sa

***Abstract*—There has been a growing trend for developing predictive solar-desalination models. However, forecasting productivities of solar stills of different designs remains a challenge. Herein, we developed predictive machine learning (ML) models for predictions of a double-slope still productivity based on experimental results. Trained datasets were taken from earlier designed passive and/or active solar stills used to treat brackish/wastewater with 45% TDS. FGSVM, EBoT, and SEGPR regression models showed the least possible MSE's (<138) indicating their reliability to accurately predict distillate amounts in double-slope still designs. The highest accuracy of SEGPR trained model with (R^2=1) and very low RMSE<8 shows its promise in predicting the performance of such similar solar-desalination systems. The novelty of this work is associated with paving the way towards creating a unified theoretical model that would provide the key to maximize still efficiencies and distillate-water outputs from supervised ML models allowing tuning of the correct parameters.**

Keywords—supervised regression; machine learning; solar still; desalination; distillation; renewable energy

I. INTRODUCTION

The utilization of solar radiation for solar-desalination systems and/or power conversion, in general, is still not widely industrialized due to the relatively high installation costs and low conversion rates. There should be innovative research on the application of supervised machine learning (ML) and cross-validation (CV) techniques. This would help in materials selection [1], solar harvesting [2], and solar-desalination *via* developing novel technology methodologies capable of comprehensively analyzing the available patents and literature datasets. Such created algorithms facilitate the advancement of current solar-desalination technologies for the commercialization of large-scale solar stills, benefiting the community as well as companies' technical R&D centers and business interests [3]. In the late 19th century, various studies [4], [5] discovered the use of "solar stills" as a promising and emerging water distillation technology. The advantage of solar stills is that they only utilize solar radiation as an abundant, free, environmentally friendly, and easily employed thermal energy source for seawater desalination and/or industrial water purification applications [4], [5]. The produced distilled water is potable high-quality water due to the complete removal of total dissolved solids (TDS), inorganic, and organic contaminants [6].

Solar intensity, wind velocity, air temperature, water-glass temperatures, water surface area, basin/absorber area, feed temperature, glass cover angle and transparency, and water level in the tank (feed flowrate) are some of the critical parameters controlling the performance and productivity of a solar still [7]. Wang et al. [8] observed that saltwater temperature, basin temperature, and solar radiation are the most important predictors (40.87, 32.43, and 18.2%, respectively), for productivity prediction of tubular solar still.

There has been a growing trend for using ML and AI models for modeling and simulation in environmental and energy engineering. Mashaly et al. [9] utilized an artificial neural network (ANN) approach for the construction of a mathematical model as a useful and valuable tool for the prediction of solar still productivity. Passive solar still fed with agricultural drainage water was studied for prediction of its instantaneous thermal efficiency. The model predicted experimental results accurately, with minimum errors confirmed from the coefficient of determination (R^2=0.96) [9]. However, forecasting the potential productivities of different designed solar stills remains a challenge to be modeled *via* built-in and pre-existing ML toolboxes. This is because productivity depends on many parameters that need to be considered both implicitly and explicitly, to ensure model adequacy [9]. ML is an alternative way of dealing with complex nonlinear problems [10] such as prediction of the solar still productivity [11], rather than using conventional numerical with complex modeled systems or inaccurate regression models [12]. Based on experimental data, the conventional methods utilized for prediction of the still performance include: (i) numerical models based on solutions of differential equations of heat and mass transfer [13], [14], (ii) regression models capable of predicting the relationship between multi-dependent variables (inputs) and the independent output [15], [16], and (3) trained models constructed from ML and artificial intelligence (AI) built-in toolboxes used for energy and solar-desalination systems [17], [18].

Srivastava et al. [19] found an evident relationship between water temperatures and distilled output as a function of solar insolation, which impacted water levels and temperature. Random forest (RF) and ANN non-linear ML techniques were previously applied to tubular solar still [8] to generate prediction models estimating water productivity. ANN model achieved optimal predictions with very high

accuracy confirmed from determination coefficients (R^2 > 0.997), which were found to be much higher than MLR models. Such models hold the promise in forecasting productivity and effectively design solar stills for the highest distillate outputs [20].

This study aims to develop an accurate predictive model based on supervised ML tools (MATLAB) and previous experimental results for predictions of a double-slope still productivity. Collected data are taken from previously conducted experiments in double-slop solar still 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) [21]. Input variables include basin (T_B), glass (T_g), and water (T_w) temperatures, as well as average water-glass temperature difference (T_w–T_g), were correlated to the water distillates. A thorough comparison between the different trained/tested model results has been established to assess the performance of the developed models. The most reliable and promising models with high efficiency ranging from 79% to 95% are then selected for further analysis to choose the optimum model for performance forecasting.

II. Methods and Equations

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 [9].

The ordinary least squares (OLS) method [22] identifies the optimal values of β_n from finding the parameters that minimize 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 [23].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

$$\min \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} (y_i - Y)^2 \quad (2)$$

Decision Trees and Ensemble: The decision tree builds 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 breaks down datasets into smaller and smaller subsets while simultaneously an associated decision tree is incrementally developed. Regression trees are designed to approximate real-valued functions built through a process known as binary recursive partitioning [24].

Solar still datasets were gathered from previous experimental work [21] and that the collected datasets included measurements of the still basin, glass cover, and water temperatures against water distillates.

Collected datasets covered six different solar still experiments (with or without reflectors and/or PCM) conducted at Solapur (Maharashtra), India; using a double-slope design and a basin area of 0.62 m^2 with a highest recorded efficiency ~42.1% [21]. The original datasets containing 48 numbers have been expanded to 144 numbers from correlating the same distillate outputs to ±0.1 of the original T_w, T_g, and T_B values (from taking advantage of non-considerable, but inevitable experimental errors). The curated datasets were then divided randomly into two groups: 80% for training and 20% for testing

Various supervised ML regression learners from the toolbox in MATLAB [25] have been selected in training/testing labeled datasets. Defined depended 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 data is water distillate, which has been correlated to the inputs.

The training datasets (80% from curated data) consist of seven matrices of [123×1], each matrix represents an input parameter or the distillate output. Testing datasets had the same inputs and were taken as 20% from the curated data to predict already known distillate outputs for checking trained model accuracy.

To assess the accuracy of the prediction models, the coefficient of determination (R^2), the mean square error (MSE), the root mean square error (RMSE), and the mean absolute error (MAE), were calculated using Eqs. (3)–(6), respectively [26]–[28]. RMSE and MAE values (ranging from 0 to ∞) allow us to demonstrate more accurate prediction results, where the higher R^2 values show the greater similarities between observed and predicted values [9].

The selection of the best models was carried out by checking whether the R^2>0.7, and thereby by predicting distillates of testing datasets. Only those models which met the R^2>0.7 condition were kept for further analysis. 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 compared to stepwise linear regressors. Figure 1 shows a flowchart illustrating the development and selection of optimal supervised ML models for accurate prediction of distillates correlated with T_w–T_g.

$$R^2 = \frac{\left[\sum_{i=1}^{n}(x_{o,i} - \bar{x}_o)(x_{p,i} - \bar{x}_p)\right]^2}{\sum_{i=1}^{n}(x_{o,i} - \bar{x}_o)^2 \times \sum_{i=1}^{n}(x_{p,i} - \bar{x}_p)^2} \quad (3)$$

$$MSE = \frac{\sum_{i=1}^{n}(x_{o,i} - x_{p,i})^2}{n} \quad (4)$$

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{o,i} - x_{p,i}\right)^2}{n}} \quad (5)$$

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n}\left|x_{o,i} - x_{p,i}\right|}{n} \quad (6)$$

where $x_{o,i}$ and/or x_o is the observed value; $x_{p,i}$ and/or x_p is the predicted value; $\bar{x}_o$ is the averaged observed values; $\bar{x}_p$ is the averaged predicted values; and n is the number of observations.

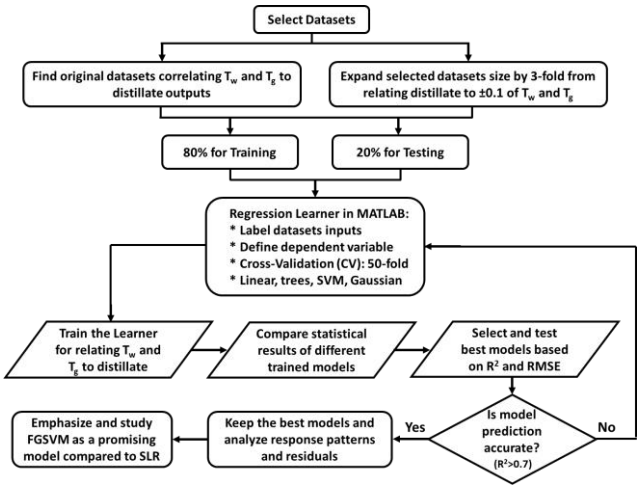

Figure 1. Flowchart for development of supervised machine learning (ML) models for prediction of distillates correlated with water-glass temperature.

III. Results and Discussion

Both FGSVM and SLR regression models have been utilized to estimate 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 Figure 2(A) and Figure 2(B). Testing datasets have also confirmed the model validity in predicting water distillates Figure 2(C). Testing results showed that the FGSVM model correctly predicted most of the distillate outputs with only two outliers, Figure 2(D). Conversely, the SLR model had many outliers and was not considered as a good model built from training 80% of the datasets.

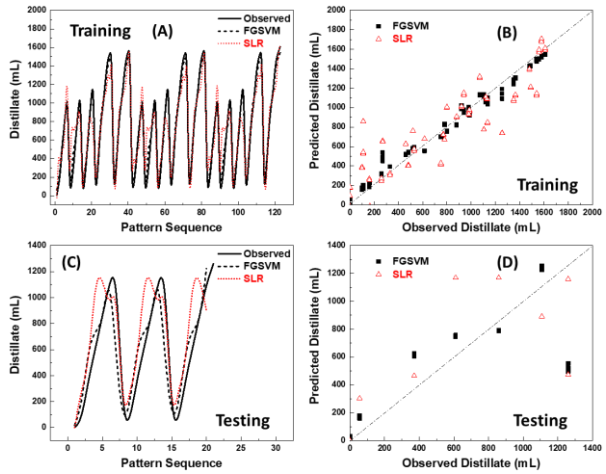

Figure 2. Comparison between [observed vs. predicted] distillate values using FGSVM and SLR models: (A) and (B) from training datasets; (C) and (D) from testing datasets.

Moreover, the other supervised ML trained models (e.g. FT, MT, EBoT, EBaT, SEGPR) suggest that the different tried models have not perfectly predicted distillate outputs (as compared to the observed distillates, yellow line in Figure 3(A)), except for the SEGPR trained model which was able to correctly predict each distillate value owing to it is ideal (R^2=1) and very low RMSE<8. The other models including the FT, MT, and EBaT trained models had the most outliers due to their more scattered values of [predicted vs. observed], indicated by the predicted values found to be away from the linear ideality or the plotted linear line [observed:predicted]=[1:1], Figure 3(B).

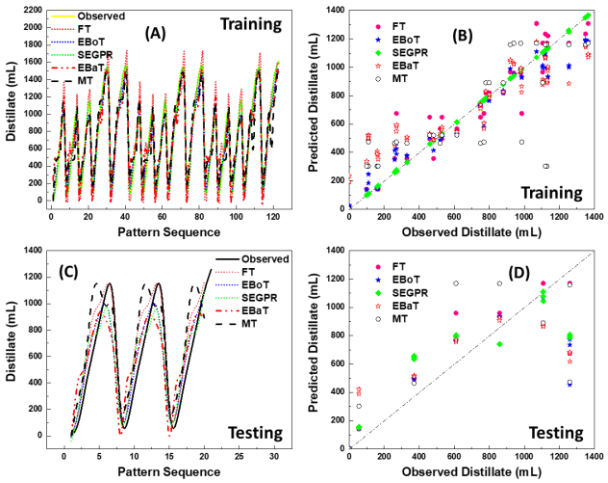

Figure 3. Comparison between [observed vs. predicted] distillate values using FT, MT, EBoT, EBaT, and SEGPR models with their obtained residuals: (A) and (B) from training datasets; (C) and (D) from testing datasets.

According to the testing datasets analysis, SEGPR and EBaT models showed the least number of outliers in their distillate prediction for the testing datasets, Figure 3(D). These results were also in agreement with the observed model trends over the tested pattern sequence as shown in Figure 3(C). None of the trained models were able to predict the last few points of the testing datasets because of the already observed deviation in the training models. Despite that the perfect predictions of SEGPR training model, the model was still unable to accurately predict the observed distillates in the testing datasets due to differences in the recognized patterns.

The correlation between T_w–T_g and the water distillates is initiated from the plotted training trends of dependent (distillate) and independent (T_w–T_g) parameters, as shown in Figure 4(A). Typically, distillate outputs should be proportionally related to T_w–T_g. This has been observed from the trained datasets in Figure 4(A) with only three outliers at

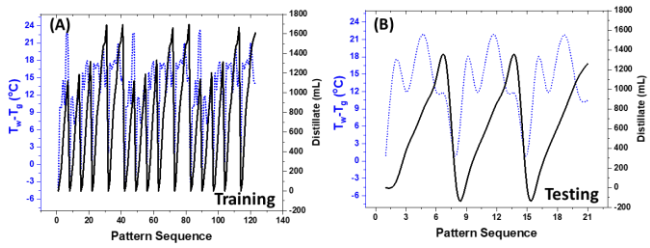

Figure 4. The observed relationship between the independent variable (T_w–T_g) and the dependent variable (distillate) found in (A) training and (B) testing datasets.

the overestimated T_w–T_g=23°C corresponding to the low distillate outputs of ~1100 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 similar a pattern as shown in Figure 4(B). It was noticed that distillate outputs increase after T_w–T_g took place with the distillate curve being super-positioned by 2 points (or 2 hrs 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, it will take some time to reach the dew point for vapor condensation. Ideally, it is desired to have a very high T_w and very low T_g to promote h_{ewg} from the high-temperature gradient, influencing evaporation and condensation rates, and water productivity.

Calculated statistical errors of various applied regression models used in the prediction of water distillates are shown in Table I. 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.

TABLE I. STATISTICAL ERRORS OF VARIOUS APPLIED REGRESSION MODELS FOR THE PREDICTION OF T_w–T_G AND WATER DISTILLATES*

Error / Model	Trained Machine Learning Models						
	SLR	FT	MT	FGSVM	EBoT	EBaT	SEGPR
RMSE	298.59	174.25	296.4	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	89158	30362	87902	13781	19093	58186	59.29
MAE	230.64	102.98	205.7	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; Reliable models are considered with R^2 > 0.90 that show the minimum MSE or RMSE.

IV. CONCLUSION

We developed novel energy-based predictive models *via* applying supervised ML on previous experimental datasets. Training datasets were collected from previously carried experiments in passive and/or active double-slope solar still (with a basin area of 0.62 m^2 and maximum η~42.1%) for the treatment of brackish/wastewater with 45% TDS. Water-glass temperature difference (T_w–T_g) was correlated to water distillates using input variables which were simply taken as a basin (T_B), glass (T_g), and water (T_w) temperatures, which were trained/tested corresponding to their experimentally observed water distillates (output). The FT, MT, and EBaT trained models had the most outliers showing low accuracy of regression trees models in finding the relationship between T_w–T_g and productivities. However, FGSVM, EBoT, and SEGPR showed the least possible MSE indicating the reliability of these ML models for accurate predictions. The trained FGSVM model showed high accuracy of prediction with only two outliers as compared to the SLR model with many outliers. We estimated the highest accuracy for SEGPR trained model owing to it is ideal (R^2=1) and very low RMSE<8. The novelty of this work is associated with paving the way towards creating energy-based theoretical models for the prediction of solar-desalination still outputs and their corresponding convective, evaporative, and radiative heat transfer coefficients.

ACKNOWLEDGMENT

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] H. A. Maddah, V. Berry, and S. K. Behura, "Cuboctahedral stability in Titanium halide perovskites via machine learning," *Comput. Mater. Sci.*, 2020.

[2] H. A. Maddah, V. Berry, and S. K. Behura, "Biomolecular photosensitizers for dye-sensitized solar cells: Recent developments and critical insights," *Renewable and Sustainable Energy Reviews.* 2020.

[3] A. J. C. Trappey, P. P. J. Chen, C. V. Trappey, and L. Ma, "A machine learning approach for solar power technology review and patent evolution analysis," *Appl. Sci.*, 2019.

[4] T. Arunkumar, K. Vinothkumar, A. Ahsan, R. Jayaprakash, and S. Kumar, "Experimental Study on Various Solar Still Designs," *ISRN Renew. Energy*, 2012.

[5] D. Kumar, P. Himanshu, and Z. Ahmad, "Performance Analysis of Single Slope Solar Still," *Int. J. Mech. Robot. Res.*, 2013.

[6] A. Saxena and N. Deval, "A high rated solar water distillation unit for solar homes," *J. Eng. (United Kingdom)*, 2016.

[7] P. Kalita, A. Dewan, and S. Borah, "A review on recent developments in solar distillation units," *Sadhana - Acad. Proc. Eng. Sci.*, 2016.

[8] N. Wang, Y., Kandeal, A. W., Swidan, A., Sharshir, S. W., Abdelaziz, G. B., Halim, M. A., ... & Yang, "Prediction of tubular solar still performance by machine learning integrated with Bayesian optimization algorithm," *arXiv Prepr.*, 2020.

[9] A. F. Mashaly and A. A. Alazba, "Neural network approach for

predicting solar still production using agricultural drainage as a feedwater source," *Desalin. Water Treat.*, 2016.

[10] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*. 1996.

[11] P. Gao, L. Zhang, K. Cheng, and H. Zhang, "A new approach to performance analysis of a seawater desalination system by an artificial neural network," *Desalination*, 2007.

[12] M. S. S. Abujazar, S. Fatihah, I. A. Ibrahim, A. E. Kabeel, and 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. Clean. Prod.*, 2018.

[13] Y. Gong, X. L. Wang, and L. X. Yu, "Process simulation of desalination by electrodialysis of an aqueous solution containing a neutral solute," *Desalination*, 2005.

[14] H. Ben Bacha, T. Damak, M. Bouzguenda, A. Y. Maalej, and 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*, 2003.

[15] X. Wang and K. C. Ng, "Experimental investigation of an adsorption desalination plant using low-temperature waste heat," *Appl. Therm. Eng.*, 2005.

[16] G. Yuan, L. Zhang, and H. Zhang, "Experimental research of an integrative unit for air-conditioning and desalination," *Desalination*, 2005.

[17] A. F. Mashaly and A. A. Alazba, "MLP and MLR models for instantaneous thermal efficiency prediction of solar still under hyper-arid environment," *Comput. Electron. Agric.*, 2016.

[18] A. F. Mashaly, A. A. Alazba, A. M. Al-Awaadh, and M. A. Mattar, "Predictive model for assessing and optimizing solar still performance using artificial neural network under hyper arid environment," *Sol. Energy*, 2015.

[19] N. S. L. Srivastava, M. Din, and G. N. Tiwari, "Performance evaluation of distillation-cum-greenhouse for a warm and humid climate," *Desalination*, 2000.

[20] O. O. Badran and M. M. Abu-Khader, "Evaluating thermal performance of a single slope solar still," *Heat Mass Transf. und Stoffuebertragung*, 2007.

[21] 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 Br.*, 2019.

[22] Guru99, "R Simple, Multiple Linear and Stepwise Regression," 2020. [Online]. Available: https://www.guru99.com/r-simple-multiple-linear-regression.html.

[23] P. Paisitkriangkrai, "Linear Regression and Support Vector Regression," *The University of Adelaide*, 2012. [Online]. Available: https://cs.adelaide.edu.au/~chhshen/teaching/ML_SVR.pdf.

[24] S. Sayad, "Decision Tree - Regression," *Data Science: Predicting the Future, Modeling & Regression*. [Online]. Available: https://www.saedsayad.com/decision_tree_reg.htm.

[25] Mathworks, "Statistics and Machine Learning Toolbox ™ User's Guide R2017a," *MatLab*, 2017.

[26] M. M. Rahman and B. K. Bala, "Modelling of jute production using artificial neural networks," *Biosyst. Eng.*, 2010.

[27] M. Zangeneh, M. Omid, and A. Akram, "A comparative study between parametric and artificial neural networks approaches for economical assessment of potato production in Iran," *Spanish J. Agric. Res.*, 2011.

[28] A. A. Alazba, M. A. Mattar, M. N. ElNesr, and M. T. Amin, "Field assessment of friction head loss and friction correction factor equations," *J. Irrig. Drain. Eng.*, 2011.

Full Name	Email address	Position	Research Interests	Personal website (if any)
Hisham A. Maddah	hmaddah@kau.edu.sa	Assistant Professor	Photovoltaic, renewable energy,	hmaddah.com