Utilization of Data-Driven Methods in Solar Desalination Systems: A Comprehensive Review

Mohammad Alhuyi Nazari1, Mohamed Salem2*, Ibrahim Mahariq3, Khaled Younes3 and Bashar B. Maqableh4

1Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran, 2School of Electrical and Electronic Engineering, Universiti Sains Malaysia (USM), Penang, Malaysia, 3College of Engineering and Technology, American University of the Middle East, Kuwait, 4Engineering and Technology Department, American College of the Middle East, Kuwait City, Kuwait

Renewable energy sources have been used for desalination by employing different technologies and mediums due to the limitations of fossil fuels and the environmental issues related to their consumption. Solar energy is one of the most applicable types of renewable sources for desalination in both direct and indirect ways. The performance of solar desalination is under effects of different factors which makes their performance prediction difficult in some cases. In this regard, data-driven methods such as artificial neural networks (ANNs) would be proper tools for their modeling and output forecasting. In the present article, a comprehensive review is provided on the applications of different data-driven approaches in performance modeling of solar-based desalination units. It can be concluded that by employing these methods with proper inputs and structures, the outputs of the solar desalination units can be reliably and accurately forecasted. In addition, several recommendations are produced for the upcoming work in the relevant areas of the study.

Keywords: solar desalination, artificial neural network, data-driven methods, renewable energies, review

INTRODUCTION

Fresh water is absolutely essential for human societies since they rely on it for development and survival (Zheng, 2017). Around 71% of the earth is covered with water; however, about 96.5% of this water is in the brackish form or saline, which means that it cannot be directly used for irrigation and drinking, and just less than 1% of fresh water resources are within human reach (Tiwari et al., 2003; Chauhan et al., 2021). Regarding the uneven distribution of fresh water in different regions of the world, the increase in demand due to population growth, and the essence of water for human survival and activities, desalination has gained more importance in recent years. Desalination is known as a treatment process of water that includes salt removal from saline water to make it appropriate for drinking (Mito et al., 2019). Desalination of water with the salinity more than normal levels is one of the ways (Tzen and Morris, 2003), and probably the most applicable one, to overcome the mentioned problems related to the unavailability of fresh water. The nature of the desalination process is energy-

Abbreviations: ANFIS, adaptive neuro-fuzzy inference system; ANN, artificial neural network; BP, backpropagation; FF, feed forward; GA, genetic algorithm; ICM, imperialist competition method; MLP, multilayer perception; MLR, multiple linear regression; PSO, particle swarm optimization; PV, photovoltaic; RBF, radial basis function; RF, random forest; RO, reverse osmosis; SVM, support vector machine.
consuming, and it is crucial to properly supply the required energy. Improving the efficiency of the systems and utilizing the renewable energy sources are recommended to solve the problems related to the energy demand of desalination systems. Renewable energy sources can be used for desalination in direct and indirect ways. In direct approaches, thermal energy is mainly used for water evaporation and reducing the salinity of water, while renewable energies can be used for indirect desalination by producing electricity and applying the power in reverse osmosis (RO) technologies (Caldera et al., 2016). Among the renewable energy sources, solar energy is attractive for the desalination purpose since it can be used in different ways such as thermal technologies or photovoltaic/RO systems.

Numerous studies have been performed on the various kinds of solar-based desalination systems to find the influential factors and improve their performance (Mostafa et al., 2020). Depending on the type of solar desalination, the factors affecting the performance can be differed. Solar radiation is one of the most important factors on the output of the systems. For instance, Joseph et al. (2005) found that by increasing the solar radiation from 400 W/m² to 900 W/m², the efficiency of a single-stage solar desalination system increased from 15 to 26%. In addition to solar radiation, the components of the system and their configuration affect the performance of these systems. As an example, Altarawneh et al. (2020) investigated the performance of a solar still composed of two parabolic troughs and two rectangular absorbers under different working conditions. They found that the rim angle of the troughs can influence the productivity of the desalination. Moreover, it was observed that reducing the pressure could remarkably improve the productivity of the desalination system. In another work, Geng et al. (2021) investigated the performance of an RO system powered by a solar dish Stirling engine. They found that by an increment in the temperature of the absorber, productivity of water increased, while there was an optimum temperature at which the exergy efficiency of the system reached its maximum value. In addition to the technical aspects, solar desalinations have been investigated from the economic point of view. For instance, Kettani and Bandelier (2020) carried out techno-economic assessment on large-scale solar powered desalination systems in Morocco by considering photovoltaics (PVs) and concentrated solar power (CSP) for supplying energy. They found that using the PV/RO system without a storage unit is the cheapest configuration today and by 2030. In another work (Zheng and Hatzell, 2020), solar thermal desalination was thermo-economically analyzed, and it was found that construction costs of solar collectors were the largest total investments of the system. Other types of desalination systems have been modeled by using data-driven methods. For instance, Faegh et al. (Faegh et al., 2021) applied different artificial neural network (ANN)-based methods to model the gain output ratio and heat transfer rate of the evaporator and evaporative condenser of a heat pump-assisted desalination system and found that the R-squared of the models were more than 0.91 for all the outputs.

As mentioned in the previous paragraph, the performance of desalination units is affected by several elements such as the applied technology, operating conditions, and the properties of the saline or brackish water. Since the experimental works are costly and time-consuming, it would be useful to propose models for performance prediction and assessment of the desalination systems. Data-driven methods, with outstanding ability in modeling of complex systems, would be attractive options for performance forecasting of desalination systems (Gao et al., 2007; Chauhan et al., 2020; Adda et al., 2021). These methods have shown their outstanding performance in a wide variety of applications such as predicting the properties of materials (Ramezanizadeh et al., 2019a; Ramezanizadeh et al., 2019b), fault diagnosis (Venkatasubramian and Chan, 1989), etc. (Rezaei et al., 2018). Current works focus on providing a comprehensive review on the applications of data-driven methods in modeling the performance of various solar desalination systems, which is performed for the first time. In addition, a table is prepared that summarizes the main findings of the reviewed works, inputs of the proposed models, applied approaches, and algorithms, which will be useful for the scholars working on the similar fields of study. Finally, according to the knowledge of the authors and the investigation of the previous studies, some suggestions are recommended for future works in the relevant subjects. The findings and information represented in this study will facilitate upcoming works to concentrate on the modeling of desalinations systems, especially the ones using solar energy.

MOSTLY USED DATA-DRIVEN METHODS

There are different data-driven methods used in modeling of energy systems. The mostly used approaches in energy system modeling are multilayer perceptron (MLP) ANN, adaptive neuro-fuzzy inference system (ANFIS), radial basis function (RBF), and support vector machines (SVMs). In this regard, these approaches are briefly described in the following subsections.

Multilayer Perceptron Artificial Neural Network

The structure of MLP is shown in Figure 1. As shown in this figure, there are three main layers in the simplest form of this network including input, hidden, and output. However, the hidden layer may be composed of more layers. In each node of this network, a weight vector is used to make connection between the current node and the ones in the upcoming layer. In the primary layer of the network, the summation of the values is sent to the next layer, which plays a role as inputs of that layer. Assuming that the vector of X is the model input and \( n_j \) is applied as the \( j \)th node, the input in the upcoming layer is written as Eq. 1

\[
  n_j = \sum_{i=1}^{n} \omega_{ji}x_i + \theta_j \quad j = 1, 2, \ldots, K
\]

where \( \theta_j \), \( \omega_{ji} \), and \( K \) are the threshold of the \( j \)th node, the weight value of the node, and the number of nodes, respectively.
Subsequently, $f$ as a transfer function is applied to provide the overall inputs in the upcoming layer as represented in Eq. 2

$$y_j = f(n_j) = f\left(\sum_{i=1}^{n} \omega_{ji} x_i + \theta_j \right) \quad j = 1, 2, \ldots, K \quad (2)$$

Different functions can be used in this step with its own features and characteristics. By multiplying the linking weight and the output of the hidden layer, the output of the nodes will be determined. It should be noted that the architecture of the network including the number of hidden layers and neurons is dependent on the problem complexity, the noise of data, and the shares of data used for the test and validation of the model (Du and Swamy, 2006). By applying an iterative process, neurons are added in the procedure of training till it reaches the optimum state. The training process plays a key role in modeling by using this approach. Predicting the process by using this method is conducted by adjusting weight and bias values. Backpropagation (BP) is one of the mostly used training algorithm for adjusting these values (Goh, 1995). The main advantages of ANNs are their ability in synthesizing algorithms through the process of learning, providing solution for nonlinear problems, and robustness of the models; however, the main disadvantages are the necessity of training for each problem, requirement for multiple tests to find the most appropriate architecture, and large data requirement for training the network (Navarro, 2013).

**Adaptive Neuro-Fuzzy Inference System**

The schematic of ANFIS in a simple form with two inputs and one output is illustrated in Figure 2. In this architecture, five layers are considered. The first layer of this model is applied in order to change the inputs to fuzzy sets and projects the variables on fuzzy membership in the range between 0 and 1. In the second layer, the signals of the input are generated; furthermore, values of membership function weight will be checked. In the next layer of this network, normalized firing strength of each node is obtained. Subsequently, the outputs are changed to crisp sets in the fourth layer. Finally, the outputs are determined in the last layer of the proposed network. This layer of the network contains one node which is used to sum up the input signals provided by the prior layer.

There are some advantages in the ANFIS method such as its ability in capturing the nonlinear structure of a procedure and fast learning capacity. In addition, this approach has both linguistic and numerical knowledge. In comparison with MLP ANN, ANFIS is more transparent for the users and results in less memorization error (Şahin and Erol, 2017); however, ANNs can
have superior performance in accuracy of model outputs for test
data compared with ANFIS (Atmaca et al., 2001).

**Radial Basis Function**

The RBF network has some advantages such as fast performance,
a simple structure, and high estimation. The structure of this
network is shown in Figure 3. Similar to MLP, there are three
main layers in this network. The nodes are connected to the
previous one in each layer of the network. In the first layer, input
variables are assigned to the nodes. Subsequently, they are
transferred to the next layer. At the final stage, the weighted
links are used to transfer the data to the third layer. In the hidden
layer of these networks, RBF plays the role of activation functions
to produce the vector distance multiplied by the

In Eq. 3, $\Delta_j$ is the weight factor, $X$ is the input vector, $Z$ is the
RBF, and $\xi_j$ refers to standard deviation. To calculate the standard
deviation, the following equation is used:

$$\xi = \frac{\theta_m}{\sqrt{\Lambda}}$$

(4)

In Eq. 4, $\theta_m$ is the maximum distance between the centers and
$\Lambda$ refers to number of centers. In the last layer of the network,
weights of the signals are obtained by using the previous layer data

$$y = \sum_{j=1}^{N} \omega_j Z_j$$

(5)

In Eq. 5, $\omega_j$ refers to the value of the weight vector determined in
the training process. Despite some advantages of RBF networks
compared with MLP ANN such as a faster training process, their
accuracy in modeling the test data may be lower compared with
MLP ANN (Markopoulos et al., 2016).

**Support Vector Machine**

SVM can be applied for regression and prediction in different
systems (Sreedhara et al., 2019). By considering that $N_s$ is the
number of data set samples and the inputs of $x_k \in R^n$ and $K =
1, 2, ..., N$ and the outputs are $y_k \in R$, the SVM formulation is as
follows (Ramezanizadeh et al., 2019a; Essa et al., 2020):

$$y = w^T \phi(x) + b$$

(6)

In Eq. 6, $b$ and $w$ are the bias and weight, respectively
(Ahmadi and Mahmoudi, 2016; Ramezanizadeh et al., 2019c). $\phi(x)$
denotes a nonlinear function which is applied to transfer $x_k$ to a high-dimension space. Generally, this changes to an optimization problem which can be expressed as follows:

$$F(w) = \frac{1}{2} w^T w + \gamma \sum_{k=1}^{N_s} e_k^2$$

subject to

$$y_k = w^T \phi(x_k) + b + e_k \quad k = 1, 2, ..., N_s$$

(8)

In Eq. 7, $\gamma$ and $e_k$ are the regularization parameter and error
value, respectively (Ahmadi and Mahmoudi, 2016; Ramezanizadeh et al., 2019c). Eqs. 7 and 8 can be rewritten as follows:

$$y = \sum_{k=1}^{N_s} \alpha_k K(x, x_k) + b$$

where $\alpha_k$ is the Lagrange multiplier and $K(x, x_k)$ is the kernel
function. In some studies (Essa et al., 2020), the RBF kernel
function is used, which is defined as follows:
\[ K(x, x_k) = \exp \left( -\frac{\| x_k - x \|^2}{\sigma^2} \right) \] (10)

In this equation, two parameters including \( \sigma \) and Lagrange multipliers must be determined. One of the main advantages of SVM methods for modeling is their ability in providing nonlinear solutions, while the main problem associated with this approach is the requirement for knowledge about the kernel that must be used.

Generally, mean square error (MSE) and R-squared are used in evaluation of regression and predictive models, which are as follows:

\[ MSE = \frac{1}{ns} \sum_{i=1}^{n_s} (\text{predicted value} - \text{actual value})^2 \] (11)

\[ R^2 = 1 - \frac{\sum_{i=1}^{n_s} (y_i^{\text{actual value}} - y_i^{\text{predicted value}})^2}{\sum_{i=1}^{n_s} (y_i^{\text{actual value}} - \bar{y}^{\text{actual value}})^2} \] (12)

where \( ns \) is the number of samples used in regression.

APPLICATIONS OF DATA-DRIVEN METHODS IN SOLAR DESALINATIONS

There are three main principle approaches used for desalination, which are thermal, pressure, and electrical. Thermal distillation can be considered as the oldest approach in which water with high salinity is boiled and the generated steam is collected. The condensed form of the collected steam can now be used as fresh water. In the electrical approach, electrical current is used to separate the salt and water. In these types of desalination units, a permeable membrane is used, in which ions move across it by use of electric current as a driving force. In the RO type of desalination, pressure acts as a driver for moving water through a selectively permeable membrane, leaving the salt behind (Parise, 2011). The majority of the desalination market belongs to thermal and RO types. Although the majority of the installed capacity of desalination systems is in the RO type, there are some benefits in thermal desalinations. For instance, the waste heat of plants can be used for the thermal desalination units, which leads to a high overall efficiency of the system. The majority of the studies performed on the applications of data-driven methods in solar desalination systems have focused on thermal types (Elsheikh et al., 2021). For instance, Zarei and Behyad. (2019) employed ANN to model the output of a humidification–dehumidification-type solar desalination used for humidifying the interior space of the greenhouse and supplying fresh water. The inputs of the model were width and length of the seawater greenhouse, front evaporator height, and the roof transparency, and the output was water yield of the system. Their different structures with one and two hidden layers were examined. They observed that applying one hidden layer with nine neurons led to the highest exactness with \( R^2 \) of 0.997. In addition to the architecture of the model, the applied functions and optimization methods could affect the outputs of the models proposed for solar desalinations. For instance, Nazari et al. (2020) compared the performance of ANN with and without the imperialist competition algorithm (ICA) optimization method in forecasting energy and exergy efficiencies and productivity of single-slope solar stills. They noticed that using the optimization method led to significant reduction in mean absolute errors of the model in predicting the mentioned outputs by up to 54.3% for water productivity. In another work (Mashaly and Alazba, 2017a), the output of an inclined passive solar still fed by agricultural drainage water was modeled by applying ANN with different architectures and multiple linear regression (MLR). The inputs for the modeling of the instantaneous thermal efficiency were relative humidity, ambient temperature, solar radiation, wind speed, feed temperature and its total dissolved solids, and feed mass flow rate. They found that ANN outperforms MLR and the best structure was in the case of using six neurons in the hidden layer. In addition to differing numbers of neurons in the hidden layer, it would be useful in terms of exactness enhancement by changing the number of hidden layers (Ramezanizadeh et al., 2019b); however, it must be considered that an increase in the number of hidden layers may lead to overfitting.

The applied method and algorithm are among the most important factors that influence the exactness of the data-driven methods in forecasting the outputs of solar stills (Mashaly and Alazba, 2015; Mashaly and Alazba, 2017b; Mashaly and Alazba, 2018a; Mashaly and Alazba, 2018b; Mashaly and Alazba, 2019a). For instance, Wang et al. (2021) used random forest (RF), ANN, and multilinear regression to forecast the productivity of the system based on time, solar radiation intensity, wind speed, temperatures of feed water, basin plates, salt water, cover, and ambient temperature. They found that using RF led to the prediction with the least error compared with others. In order to reach further exactness, the Bayesian optimization algorithm was applied to search the most appropriate hyperparameters which led to significant enhancement in the accuracy of the ANN-based model by increasing the determination coefficient from 0.7098 to 0.9614. In another study (Essa et al., 2020), the performance of ANN with the Harris Hawk optimizer was compared with the traditional ANN and SVM in predicting the productivity of an active solar still. In their models, ambient temperature, time, speed of wind, solar irradiance, and velocity of vapor were considered as inputs. They found that ANN outperformed SVM and could be further enhanced by using the optimizer. In their work, the R-squared values of the model for ANN and SVM were 0.9703 and 0.9701, respectively, while this value for the ANN-based model coupled with the optimizer reached 0.9834. Improved accuracy of the models through the coupling optimizer can be attributed to better adjustment of the parameters affecting the performance of the modeling approach. In another work, performance of ANFIS, ANN, and Multiple Regression (MR) in forecasting the performance of an inclined passive solar still was compared. In all the proposed models, solar radiation, relative humidity, feed flow rate, and total dissolved solids of brine and feed were used as inputs. The utilized function in the structure of data-driven
methods is another influential factor. As an example, Mashaly and Alazba (2017c) tested different membership functions including triangle, trapezoid, Pi curve, and difference between two sigmoidal functions in ANFIS-based models to propose a model with the highest exactness. In their models, inputs were dissolved solids of the feed and brine, feed flow rate, relative humidity, and solar radiation. They found that the Pi curve and triangle membership functions can provide outputs with higher accuracy compared with the others. In cases of using these methods, the correlation coefficient of the regression for training data sets was around 0.999. The most proper function in the structure of networks for modeling can be dependent on the physics of the problem which can be obtained by testing different types of functions.

In modeling the system with data-driven methods, it is essential to consider all the effective elements as inputs. In this regard, some models have included more inputs to reach better accuracy or improved the comprehensiveness. As an example, Abujazar et al. (2018) used wider variables such as cloud cover, day and month numbers, number of hours per day, difference between the temperatures of inner and outer surfaces of glass in addition to the factors used in the majority of the studies such as ambient temperature, solar radiation, humidity, wind speed, and temperatures of water, basins, and vapor to forecast productivity of an inclined stepper solar still. In their work, cascaded forward ANN with different numbers of neurons and a linear model and regression were used. They found that the ANN model was more reliable in predicting the productivity of the system. The values of root-mean-squared error (RMSE) for regression, the linear model, and the ANN-based model were 50.21, 80.36, and 41.01, respectively. Despite more comprehensiveness of this model compared with previously mentioned ones, it can be further improved by considering other factors such as the specifications of the system such as the dimensions of different parts and properties of the materials affecting the performance of the systems.

Solar desalination can be integrated with other components to reach higher productivity. Data-driven methods are applicable for performance forecasting of these systems (Bagheri et al., 2020). As an example, Bagheri et al. (2021) used ANN to model a solar desalination system composed of PVs, a heater, a battery, a cylindrical parabolic collector, etc. The panel was applied to supply the power of the heater used in the tank that was employed for preheating the saline water prior to its entrance to the collector. In the collector, saline water was further heated before entering the still. The schematic of the system is shown in Figure 4. By testing different architectures of the network and by varying the number of neurons in the hidden layer, they found that the highest accuracy of the model was obtained in the case of using 24 neurons with an $R^2$ of 0.993. These methods can be developed for other hybrid systems such as solar/wind RO desalination technologies in the near future; however, more
inputs such as wind speed and other factors affecting the systems must be considered. Since the inputs of the systems are increased for hybrid technologies, the modeling process would be more complicated.

Data-driven methods are employable for modeling the dynamic performance of solar desalination systems. In a study carried out by Sohani et al. (2021), different ANNs including backpropagation (BP), feedforward (FF), and RBF were used to estimate water temperature and hourly water production of a solar still with enhanced design. The inputs of their models were wind speed, ambient temperature, received radiation from the Sun, and water depth in the basin. Comparison of the estimated data and the corresponding actual values revealed that RBF and FF were the most powerful approaches in predicting water temperature and hourly water production, respectively. Despite its novel idea in dynamic modeling of a solar desalination, the comprehensiveness of their model was limited and could be further enhanced by considering other inputs such as wind speed and feed temperature.

Utilizing nanofluids in solar stills can improve their performance. Intelligent methods can be applied for accurate evaluation of these solar stills. Kandel et al. (2021) tested various data-driven methods including ANN, Support Vector Regression (SVR), linear SVR, and RF to model the performance of a double-slope solar still utilizing the carbon black nanofluid in 1.5% wt concentration. The inputs of the proposed model were air ambient temperature, solar radiation, wind speed, vapor temperature, basin temperature, and temperatures at the glass inlet and outlet. The models were coupled with the Bayesian optimization algorithm to tune the approaches and obtain the outputs with the highest accuracy. They found that all the proposed models were able to predict the performance of the system with relatively high exactness; however, utilizing RF led to the highest accuracy. The performance of the nanofluidic solar desalination system integrated with other modules can be modeled by data-driven methods. For instance, Bahiraei et al. (2020) used ANN coupled with the genetic algorithm (GA) and Imperialist Competition Method (ICM) to model the performance of a nanofluidic solar still integrated with a thermoelectric module. The inputs of their model were time, solar radiation, ambient temperature, power of the applied fan, concentration of the nanofluid, and temperatures of water, glass, and basins, while the output of the proposed models was water productivity. They observed that the exactness of the model through coupling the mentioned optimization approaches significantly improved, while using ICM was more influential in terms of accuracy enhancement. In addition to the optimization method, the algorithm used for modeling affects the exactness of the predicted values of nanofluidic solar desalinations. For instance, Bahiraei et al. (2021) used Particle Swarm Optimization (PSO)-ANFIS and PSO-ANN for modeling the performance of a solar still with Cu$_2$O nanoparticles. The inputs of the models were similar to those of the previous work, while the output of the designed model was efficiency of the system. They found that in both types of models, coupling the optimization methods led to exactness enhancement; however, the maximum accuracy in modeling was observed in the case of using PSO-ANFIS with an $R^2$ of 0.9884. In another work (Mashaly and Alazba, 2016), the performance of MLR and MLP ANN was compared in predicting the instantaneous thermal efficiency of a solar still. They found that using MLP ANN provided a model with higher exactness compared with MLR. Higher exactness of MLP ANN can be attributed to its more complex structure, which enables it to model the complicated systems with better performance.

The outputs of the ANN-based model can be used for designing an optimal condition for the performance of the desalination systems (Azad et al., 2021). As an example, in a study carried out by Porrazzo et al. (2013), an ANN-based optimizing control system was utilized for a solar-powered membrane desalination module. ANN was used for performance prediction of the system under different operating conditions by considering radiation and the rate of feed flow inlet temperature of cold channel as the inputs. Afterward, a control system was implemented to optimize the distillate production of the system. The proposed system allowed to set the feed flow rate at the optimal values in order to reach continuous maximum production of the distillate. As another example, Maleki et al. (2016) applied ANN to forecast the weather condition and optimize a hybrid system, solar-wind-powered RO desalination. By using the outputs of the network and performing optimization, the optimum design of the system was obtained.

To sum up the findings of the study, it can be declared that the accuracy of the models is under the influence of the applied method, optimization algorithm, etc. Generally, intelligent methods such as ANNs are preferred in terms of accuracy due to their more complex structures which enable them to model complicated systems with higher accuracy. In addition, it is found that applying optimization algorithms and coupling them with the intelligent methods improve the accuracy since the parameters affecting the exactness are used in their optimum values. In addition to the abovementioned factors, the considered inputs influence the exactness. Considering more influential factors as the inputs will provide more accurate models (Ahmadi et al., 2018). The other factors that may cause the differences in the model can be attributed to the noise of data, which is inevitable in experimental data used for modeling. In Table 1, the important outcomes of the studies in the topic of this article are provided.

**SUGGESTIONS FOR UPCOMING STUDIES**

Despite the fact that there are several works on utilization of data-driven methods in performance prediction of solar desalination systems, there are some limitations in modeling the outputs of solar desalination systems. For instance, it is difficult to propose comprehensive models with applicability for different types of solar-assisted desalination systems. For this purpose, the type of desalination must be defined as a meaningful variable. In addition, different working conditions may affect the performance of the systems, which must be distinguished and considered in inputs of the models. Furthermore, since the
TABLE 1 | Important findings of the studies on applications of data-driven methods in solar desalination systems.

| Reference                  | System                                | Method                        | Inputs                                                                                           | Findings                                                                 |
|----------------------------|---------------------------------------|-------------------------------|--------------------------------------------------------------------------------------------------|------------------------------------------------------------------------|
| Elsheikh et al. (2021)     | Stepped solar still                   | Long-short term memory (LSTM) ANN | --                                                                                               | $R^2$ of the provided model was 0.9752.                                  |
| Zarei and Behyad (2019)    | Solar greenhouse desalination (humidification--dehumidification) | ANN                            | Width and length of the seawater greenhouse, front evaporator height, and the roof transparency | Using one hidden layer with nine neurons led to the model with the maximum exactness with $R^2$ of 0.997. |
| Nazari et al. (2020)       | Single-slope solar still              | ANN                            | Time, solar radiation, temperatures of amblence, glass, water, and basins                       | Using the ICA optimization approach led to significant reduction in the error of the predicted values (up to a 54.30% reduction in mean absolute error of the model used for water productivity estimation). The maximum $R^2$ value of the model for water productivity was 0.9924. |
| Mashaly and Alazba (2017a) | Inclined passive solar still          | ANN and MLR                    | Relative humidity, ambient temperature, solar radiation, wind speed, feed temperature and its total dissolved solids, and feed mass flow rate | ANN outperformed the MLR in performance prediction of the system. The maximum absolute errors of the ANN and MLR were 8% and 35%, respectively. The $R^2$ value of the model with ANN was 0.949, while it was 0.739 for MLR. |
| Mashaly and Alazba (2019a) | Solar still                           | ANFIS and multiple nonlinear regression | Relative humidity, solar radiation, flow rate of the feed, total dissolved solids of the feed and brine. | The applied function in ANFIS affects the outputs of the model. The maximum value of $R^2$ for training data sets was 0.999 for the ANFIS-based model. The model based on ANN had higher accuracy compared with stepwise regression. The mean absolute relative error of the stepwise regression was around 2.5 times higher than that of the ANN-based model. $R^2$ for ANN was 0.960, while it was 0.902 under optimal conditions of the SWR model. |
| Mashaly and Alazba (2017b) | Solar still                           | ANN and stepwise regression     | Relative humidity, ambient temperature, total dissolved solids in feed water, rate of feed flow, solar radiation, and wind speed | Using the ICA optimization approach led to the model with the maximum exactness with $R^2$ of 0.997. |
| Mashaly and Alazba (2015)  | Solar still                           | ANN                            | Julian day, relative humidity, ambient temperature, total dissolved solids in the feed and brine water, rate of feed flow, solar radiation, wind speed, and temperature of brine water | Using Levenberg–Marquardt as a learning function led to higher accuracy ($R^2 = 0.99437$) compared with resilient backpropagation ($R^2 = 0.9853$) and conjugate gradient backpropagation with Fletcher–Reeves restarts ($R^2 = 0.98941$). |
| Mashaly and Alazba (2018a) | Solar still                           | ANFIS                          | Total dissolved solids in the feed and brine water, solar radiation, relative humidity, and feed flow rate | The membership function affects the accuracy of the forecasted data. The coefficient of correlation was 0.99 under the most accurate conditions. |
| Mashaly and Alazba (2018b) | Solar still                           | ANFIS                          | Total dissolved solids in the feed and brine water, solar radiation, relative humidity, and feed flow rate | The membership function of the model influences the exactness of the predicted values. The $R^2$ value in the most accurate case was 0.999. |
| Wang et al. (2021)         | Tubular solar still                   | ANN, RF, and multilinear regression | Time, solar radiation intensity, wind speed, temperatures of feed water, basin plates, salt water, cover, and ambient temperature | The proposed model based on RF had the highest exactness with a mean absolute error of 5.21%. The $R^2$ value in the most accurate case was 0.9758, while it was 0.9745 for RF and 0.9614 for ANN. Using the optimizer in the ANN increased $R^2$ of the model from about 0.970 to around 0.983. |
| Essa et al. (2020)         | Active solar still                    | ANN, ANN/Harris Hawks, and SVM | Ambient temperature, time, speed of wind, solar irradiance, and velocity of vapor            | The membership function affects the exactness of the proposed models based on ANFIS. The value of $R^2$ under optimal conditions was 0.9999. |
| Mashaly and Alazba (2017c) | Solar still                           | ANFIS with different membership functions | Dissolved solids of the feed and brine, feed flow rate, relative humidity, and solar radiation | Using ANN provided a model with the highest accuracy (RMSE = 22.48). $R^2$ for ANN and the regression were approximately 0.98 and 0.905, respectively. |
| Abujazar et al. (2018)     | Inclined stepped solar still         | ANN, linear model, and regression | Cloud cover, day and month numbers, number of hours per day, difference between the temperatures of inner and outer surfaces of glass, ambient temperature, solar radiation, humidity, wind speed, and temperatures of water, basins, and vapor | (Continued on following page)                                             |
future works is utilizing more recent optimization approaches to applicability. Another point that must be considered in the case of nano extend the application of the models. Furthermore, in the processing of water and mass, $R^2$ of ANN and first principle models were 0.9998 and 0.9978, respectively. $R^2$ of the provided model was 0.993. RBF and FF showed the highest accuracy in forecasting water temperature and hourly water production, respectively. The $R^2$ values for FF, BP, and RBF networks were 0.9631, 0.9425, and 0.9567, respectively.

Despite the mentioned problems and limitations, there are some recommendations that can improve the upcoming studies. First of all, the majority of the works are on thermal desalination modules, while these methods can be developed for the solar-powered RO systems and other desalination systems powered by solar-based hybrid systems such as solar/thermal or solar/wind. In addition, most of the proposed models are applicable for just one type of solar desalination, while their comprehensiveness can be improved by considering more inputs. For instance, using the dimensions of desalination systems is one of the ways that can be used to extend the application of the models. Furthermore, in the case of nanofluidic solar desalination, using the properties of nanofluids such as their concentration and properties of particles can lead to proposing a model with a higher level of applicability. Another point that must be considered in the future works is utilizing more recent optimization approaches to reach higher exactness. In this regard, hybrid optimization algorithms would be attractive options. Furthermore, it would be useful to use data-driven methods for other purposes such as modeling systems from economic and environmental points of view. In addition, the majority of the studies have focused on water productivity as the output of the model, while it would be useful and beneficial to model other technical criteria such as energy and exergy efficiency of the systems. Finally, it is suggested to compare different approaches in terms of the required time for the training process.

**CONCLUSION**

In the provided article, applications of data-driven methods in solar desalination system modeling are provided. Different variables have been used as the inputs in the models proposed for solar desalination systems including solar radiation, ambient...
conditions, etc. The main findings of this review article are as follows:

- Compared with the correlation, intelligent methods can model the solar desalination systems more accurately.
- Different parameters such as productivity, energy, and exergy efficiency can be modeled by using the intelligent methods.
- The accuracy of the suggested models is influenced by different elements such as the applied method and algorithm and the considered inputs.
- Coupling optimization methods with the models will improve the accuracy due to adjusting the hyperparameters to their optimum values.
- In addition to the applied method for modeling, the type of the optimization algorithm influences the exactness of the models.
- Operating conditions such as solar radiation and relative humidity in addition to the properties of the feed and saline water are among the most important factors that must be used as inputs.
- The outputs of the models, obtained by intelligent methods, can be used to optimize the systems.
- Most of the studies have considered water productivity as the output of the model, while it would be beneficial to consider other technical criteria such as energy and exergy efficiency of the system.
- In addition to technical criteria, considering other factors such as environmental and economical parameters as outputs of the models would be useful.
- It is suggested to compare the intelligent models in terms of required time and calculations for the training process with different algorithms and approaches.
- Applying hybrid optimization algorithms, with more proper ability in finding optimal solutions, can lead to more precise models.

**AUTHOR CONTRIBUTIONS**

MA and MS have designed the work and contributed in writing and implementation of the work. IM and KY have contributed in implementation of the work and edition. BM edited the manuscript and contributed in writing.

**FUNDING**

This work was partially supported by Universiti Sains Malaysia under Short-term grant No. 304/PELECT/6315330.

**REFERENCES**

Abujazar, M. S. S., Fatihah, S., Ibrahim, I. A., Kabed, A. E., and Sharif, S. (2018). Productivity Modelling of a Developed Inclined Stepped Solar Still Based on Actual Performance and Using a Cascaded Forward Neural Network Model. *J. Clean. Prod.* 170, 147–159. doi:10.1016/J.JCLEPRO.2017.09.092

Adda, A., Hanini, S., Bezari, S., Laidi, M., and Abbas, M. (2021). Modeling and Optimization of Small-Scale NF/RO Seawater Desalination Using the Artificial Neural Network (ANN). *Environ. Eng. Res.* 27, 200383. doi:10.4491/ eeer.2020.383

Ahmadi, M. A., and Mahmoudi, B. (2016). Development of Robust Model to Estimate Gas-Oil Interfacial Tension Using Least Square Support Vector Machine: Experimental and Modeling Study. *J. Supercri. Fluids* 107, 122–128. doi:10.1016/J.SUPERFLU.2015.08.012

Ahmadi, M., Hajizadeh, F., Rahimzadeh, M., Shafii, M., Chamkha, A., Lorenzini, G., et al. (2018). Application GMDH Artificial Neural Network for Modeling of Al2O3/water and Al2O3/Ethylene Glycol thermal Conductivity. *Int. J. Heat Technol.* 36, 773–782. doi:10.18280/ijht.360301

Altarawneh, I., Batlha, M., Rawadieh, S., Amlaief, M., and Tarawneh, M. (2020). Solar Desalination under Concentrated Solar Flux and Reduced Pressure Conditions. *Solar Energy* 206, 983–996. doi:10.1016/J.SOLENER.2020.06.058

Atmaca, H., Cetisli, B., and Yavuz, H. S. (2001). “The Comparison of Fuzzy Inference Systems and Neural Network Approaches with ANFIS Method for Fuel Consumption Data,” In Second International Conference on Electrical and Electronics Engineering Papers ELECO, 6, 1–4.

Azad, A., Aghaei, E., Jalali, A., and Ahmadi, P. (2021). Multi-objective Optimization of a Solar Chimney for Power Generation and Water Desalination Using Neural Network. *Energ. Convers. Manage.* 238, 114152. doi:10.1016/J.ENCONMAN.2021.114152

Bagheri, A., Esfandiari, N., Honarvar, B., and Azdarpour, A. (2020). First Principles versus Artificial Neural Network Modelling of a Solar Desalination System with Experimental Validation. *Math. Comput. Model. Dyn. Syst.* 26, 453–480. doi:10.1080/13873954.2020.1788609

Bagheri, A., Esfandiari, N., Honarvar, B., and Azdarpour, A. (2021). ANN Modeling and Experimental Study of the Effect of Various Factors on Solar Desalination. *J. Water Supply Res. Technol. Aqua* 70, 41–57. doi:10.2166/AQUA.2020.085

Bahai, M., Nazari, S., Moayedi, H., and Safarzadeh, H. (2020). Using Neural Network Optimized by Imperialist Competition Method and Genetic Algorithm to Predict Water Productivity of a Nanofluid-Based Solar Still Equipped with Thermoelectric Modules. *Powder Techn.* 366, 571–586. doi:10.1016/J.POWTEC.2020.02.055

Baharne, M., Nazari, S., and Safarzadeh, H. (2021). Modeling of Energy Efficiency for a Solar Still Fitted with Thermoelectric Modules by ANFIS and PSO-Enhanced Neural Network: A Nanofluid Application. *Powder Techn.* 385, 185–198. doi:10.1016/J.POWTEC.2021.03.001

Caldera, U., Bogdanov, D., and Breyer, C. (2016). Local Cost of Seawater RO Desalination Based on Solar PV and Wind Energy: A Global Estimate. *Desalination* 385, 207–216. doi:10.1016/J.JDESAL.2016.02.004

Chauhan, R., Dumka, P., and Mishra, D. R. (2020). Modelling Conventional and Solar Earth Still by Using the LM Algorithm-Based Artificial Neural Network. *Int. J. Ambient Energ.* 1–8. doi:10.1080/01430750.2019.1707113

Chauhan, V. K., Shukla, S. K., Turkey, I. V., and Singh Rathore, P. K. (2021). A Comprehensive Review of Direct Solar Desalination Techniques and its Advancements. *J. Clean. Prod.* 284, 124719. doi:10.1016/J.JCLEPRO.2020.124719

D, K., and S, M., M. (2006). *Neural Networks in a Softcomputing Framework*. Springer Science and Business Media.

Elshiekh, A. H., Katekar, V. P., Muskens, O. L., Deshmukh, S. S., Elaziz, M. A., and Dabour, S. M. (2021). Utilization of LSTM Neural Network for Water Production Forecasting of a Stepped Solar Still with a Corrugated Absorber Plate. *Process Saf. Environ. Prot.* 148, 273–282. doi:10.1016/J.PSEP.2020.09.068

Esa, F. A., Abu Elnazir, M., and Elsheikh, A. H. (2020). An Enhanced Productivity Prediction Model of Active Solar Still Using Artificial Neural Network and Harris Hawks Optimizer. *Appl. Therm. Eng.* 170, 115020. doi:10.1016/J.APPLTHERMENG.2020.115020

Faeih, M., Behnam, P., Shafii, M. B., and Khialedani, M. (2021). Development of Artificial Neural Networks for Performance Prediction of a Heat Pump Assisted Humidification-Dehumidification Desalination System. *Desalination* 508, 115052. doi:10.1016/J.JDESAL.2021.115052

Gao, P., Zhang, L., Cheng, K., and Zhang, H. (2007). A New Approach to Performance Analysis of a Seawater Desalination System by an Artificial Neural Network. *Desalination* 205, 147–155. doi:10.1016/J.JDESAL.2006.03.549

Geng, D., Cui, J., and Fan, L. (2021). Performance Investigation of a Reverse Osmosis Desalination System Powered by Solar Dish-Stirling Engine. *Energ. Rep.* 7, 3844–3856. doi:10.1016/J.ENERGY.2021.06.072
