IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON BIG DATA ANALYTICS IN INTERNET OF THINGS AND CYBER-PHYSICAL SYSTEMS

Received May 17, 2017, accepted June 4, 2017, date of publication June 19, 2017, date of current version November 14, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2716978

Water Desalination Fault Detection Using Machine Learning Approaches: A Comparative Study

MORCHED DERBALI[1,2], SEYED M. BUHARI[2], GEORGIOS TSARAMIRSIS[2], MILOS STOJMENOVIC[3], H. JERBI[4], M. N. ABDELKRIM[1], AND MOHAMMAD H. AL-BEIRUTTY[5]

[1]Unite de recherché Modelisation, Analyse et Commande Des Systemes, ENIG, 6029 Gabes, Tunisie
[2]Faculty of Computing and IT, King Abdulaziz University, Jeddah 23218, Saudi Arabia
[3]Faculty of Informatics and Computing, Singidunum University, 160622 Belgrade, Serbia
[4]Deanship of the Scientific Research, University of Hail, Baqaa 53962, Saudi Arabia
[5]Centre of Excellence in Desalination Technology, King Abdulaziz University, Jeddah 23218, Saudi Arabia

Corresponding author: Milos Stojmenovic (mstojmenovic@singidunum.ac.rs)

ABSTRACT The presence of faulty valves has been studied in the literature with various machine learning approaches. The impact of using fault data only to train the system could solve the class imbalance problem in the machine learning approach. The data sets used for fault detection contain many independent variables, where the salient ones were selected using stepwise regression and applied to various machine learning techniques. A significant test for the given regression technique was used to validate the outcome. Machine learning techniques, such as decision trees and deep learning, are applied to the given data and the results reveal that the decision tree was able to obtain more than 95% accuracy and performed better than other algorithms when considering the tradeoff between the processing time and accuracy.

INDEX TERMS Machine learning, stepwise regression, fault detection, water desalination.

I. INTRODUCTION

Automatic error diagnosis in water desalination can decrease the damage and increase the efficiency of the factory and ensure the high quality of the produced water. Various water desalination approaches exist but in this paper we will focus on Direct Contact Membrane Distillation (DCMD). Membrane distillation (MD) refers to a thermally driven process, where only vapour molecules are transported through porous hydrophobic membranes. MD, a popular water distillation approach, is used in real distillation factories. This research is based on data collected from an actual industrial system implemented as DCMD and is described in [1] and [2]. The data includes ten parameters, Tin hot, Tout hot, Tin cold, Tout cold, Pin hot, Pout hot, Pin cold, Pout cold, Qin and Qout. This set of data was recorded while faults were manually introduced in the heater, cooler, hot valve, cold valve, hot pump, and cold pump. This allowed us to accurately differentiate between faulty and normal data. It is worth noting that only one fault was introduced at a time during the data collection phase.

Recent work in [3] and [4], utilizes the same dataset but with only eight out of ten parameters as that work aimed to produce a proof of concept of the benefits of using machine learning to automate the fault detection process. That work used separate normal and abnormal data for each area such as the heart, cooler, etc. The data was processed with different machine

learning approaches such as linear/quadratic discriminant analysis, K nearest neighbor (KNN), neural networks, decision trees, naïve based classifier and support vector machine (SVM). The research used a confusion matrix to summarize that decision trees produce the best result with an F-measure value equal to 99.6875%.

In this paper, we utilize the dataset from [3] and [4], but include all available features. The salient features were extracted using stepwise regression. Principal Component Analysis (PCA) was used to handle correlation among various variables. Section 2 contains a comprehensive literature review. Methodology used in this research is summarized in Section 3. The processing of the dataset along with its discussion is provided in Sections 4 and 5. Section 6 concludes the paper.

II. LITERATURE REVIEW

Historically manufacturing processes have varied due to advancements in science and technology. Work such as [5]

can be found in literature that deals with various types of faults, such as blockage, sensor faults, operator faults, utility faults, process and wear, etc. The main concern is that different studies have been conducted for different faults but the data available from any single study is no more than two years old. Fusion of faults from different manufacturing environments is the salient aim of [6]. They propose a solution to the issues of Identification and definition of potential faults and their features in a chemical process, and fault diagnosis without data samples. The proposed framework contains an expert system that analyzes different historical data for feature extraction. The actual historical data, collected from the cloud of a chemical corporation, along with the detected faults are introduced into the Fault Detection and Diagnosis (FDD) learning system. Thus, multiple data sources can be fed into the FDD for learning purposes. Further, the online data from any mechanical system can be supplied to the learned FDD system to detect various numbers of faults and also combination of different faults. Results show that they have a 20% missed detection rate and the accuracy of diagnosis is 91.7%.

One of the earlier works considering the Multi-Stage Flash (MSF) water desalination process was the development of its steady-state mathematical model [7], [8]. The relationships between various parameters like thermal performance ratio, specific heat transfer surface area, etc were determined. Assuming constant heat transfer surface area per stage, the impact of the variation of water's physical properties like temperate, salinity, etc are analyzed. The model was found to conform to data obtained from six MSF desalination plants.

Said *et al.* [9] enhanced MSF operations by optimizing the design and operation parameters. The correlations between water temperature, fouling and freshwater demand were studied using a polynomial optimization model. Based on the factors studied, the total daily operating cost of MSF was reduced due to the optimized number of stages and enhanced flowrate.

A Real-time expert system for fault diagnosis in MSF was studied in [10] and [11]. An SDG based qualitative model is used to obtain the 'if-then' rules of the knowledge base of the expert system. These rules are evaluated using fuzzy logic. Another research presented in [12] studied fault diagnosis in MSF by combining qualitative-based SDG and quantitative-based Dynamic Partial Least Square (DPLS) methods. This combination works even when the faulty data is unavailable, which is a common concern in this area of research. Partial Least Square (PLS) uses the Principal Components (PCs) obtained from PCA, which reduces the dimensions of the input variables. PCs are generated using linear combination of various original variables. Given X as the expected output and $\bar{X}$ as the estimated output value:

$$X = TP^T + E \ldots \tag{1}$$

Where T is the score vector, P is the loading matrix and E is the residual matrix between X and $\bar{X}$.

The obtained PCs, represented as p_i, are used to obtain the score vector.

$$t_i = Xp_i^T \tag{2}$$

The PLS approach is used to maximize the covariance between input variable X and output variable Y. Here, X is the PCs obtained using PCA. Similar to Eq. 1, Y is decomposed into

$$Y = UP^T + F \tag{3}$$

Where F is the residual matrix of Y.

PLS combined with the ARMAX (AutoRegressive Moving Average with eXogenous inputs) time series model generates Dynamic PLS (DPLS). DPLS uses the past values of the PCs along with the relationships between the source variables and the target variables. Thus, the required data set for DPLS could be obtained even when there are external disturbances or set-point changes. The difference between the estimated value from the DPLS system and the measured one is called residual.

$$r_i = y_i - \hat{y}_i \tag{4}$$

The accurate detection of faults is vital for fault reduction in the system. This detection is performed by monitoring the residuals using the CUSUM [12] approach. The CUSUM method provides an out-of-control signal if the sums of the larger values of the sample readings cross a predefined threshold. Minimum jump size and threshold size are considered as parameters for the CUSUM method. 60 datasets with 30 variables each were given as input to the system. Among these 30 variables, twelve are selected as target variables by DPLS. This reduction of variables occurs since certain ones have no impact on faults. The relation of many faults to certain variables cause fault detection by CUSUM to be unsuccessful. Thus, certain classifiers are used to identify the fault. According to the results, DPLS with the CUSUM approach performs equal well when compared to neural network based systems in detecting faults.

III. METHODOLOGY

In this research, the flow of operation is performed as described in Figure 1. Raw data, pre-recorded from the desalination system, is obtained and used as input to the feature selection system. The important features are selected from the raw data. With these salient features, the raw data is labelled as normal or faulty data using an expert system. This combined normal and faulty data is passed to the learning system, which learns using decision trees, regression or other approaches described below. Then, the test data is fed into the learned system and the data is classified as normal or faulty data.

The impact of collecting normal samples along with faulty samples cause lots of issues. Within a given dataset, the number of different faulty samples and normal samples used might not reflect the actual composition of data in a specific period of time.

TABLE 1. Multiple linear regression (without pre-processing).

	Adjusted R-Square value	Variables with higher p-value	Mean Square Error
Dataset 1	0.9255	tic, toc, pih, pic, poc and qc	0.05387335
Dataset 2	0.7863	Pih	0.4273828

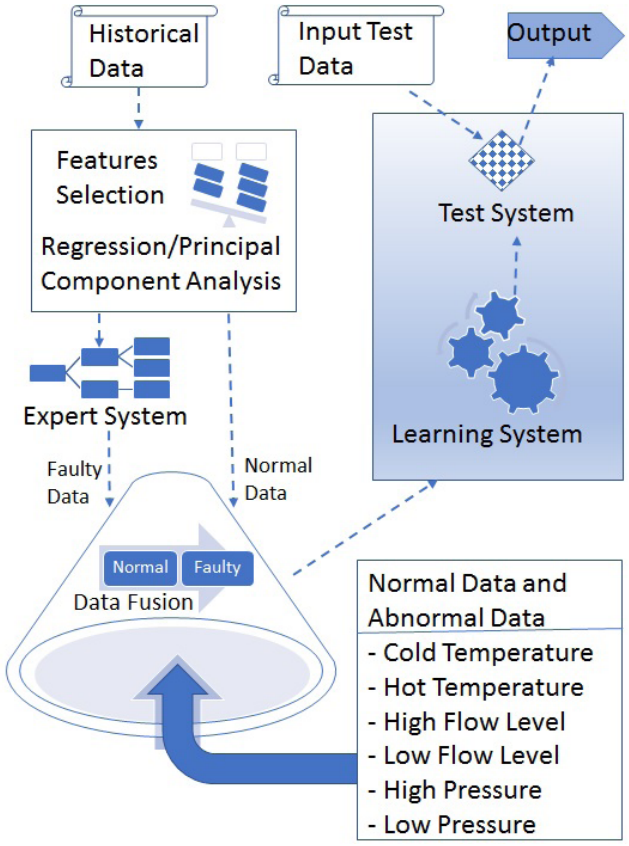

FIGURE 1. Methodology.

A. PRELIMINARIES

The dataset was gathered and compiled using two methods: (1) data was collected from the system while specifically adding a fault to a certain valve; (2) data was collected from a simulated system using Simulink. The first dataset contains 552 samples and the second 55004. The second dataset was created for the purpose of testing the behaviour of the selected methods with large datasets as it was not possible to collect more data from the real system. Additionally, two extra features were removed as we could not accurately replicate them in the generated data. The presence of different independent variables with varied range could cause an impact on the various approaches studied. Thus, normalization of the dataset is performed along the column so as to have a common level being maintained for all the rows of the dataset.

B. PRE-PROCESSING

Normalization needs to be performed when joining the actual faulty data with the simulated, non-faulty data of the amalgamated data set, as seen in Table 1.

Different normalization approaches like normalization, standardization and transformation are applied separately on the dataset in order to study their impacts. Standardization of the given variable is performed as:

$$x_{ij} = \frac{(X_{ij} - \bar{x_j})}{s_j} \tag{5}$$

where X_{ij} *is the* i^{th} *observation of the* j^{th} *variable* $\bar{x_j}$ *is the mean and* s_j *is the standard deviation*

We identify the correlation matrix for all the given independent variables as:

$$coeff_{AB} = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}} \tag{6}$$

Similar to [13], a threshold range of 0.1 to 0.9 is chosen to decide on the inclusion of an independent variable.

C. MULTIPLE LINEAR REGRESSION

Given more than one independent variable (X) and one dependent variable (Y), a linear relation between them could be generated using Multiple Linear Regression. This relationship is given by

$$Y = XB + \varepsilon \tag{7}$$

The B matrix represents the coefficients of the regression. This is given by

$$B = \left(X^T X\right)^{-1} X^T Y \tag{8}$$

The problem with the above equation is that $X^T X$ is singular. This could be because the number of independent variables and the number of observations vary from one another. There is also a possibility of collinearity among variables.

A check for normality and influential elements is done using a Q-Q plot and Cook's distance respectively. Cook's distance is used to measure the influence due to leverage and outliers. Here, leverage indicates the influence on the prediction if the observed value is increased by a unit. Cook's distance is calculated as:

$$C_i = \left|\frac{r_i}{p+1}\right|\left(\frac{p_{ii}}{1-p_{ii}}\right)$$
$$Where\ r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \tag{9}$$

Here, h_{ii} is the hat matrix, which is the measure of leverage.

D. FEATURE EXTRACTION

The presence of different valves in the desalination plant, along with other heterogeneous instrumentation could enlarge the number of features to be studied. The process of learning with all the features could be performed using various algorithms like multiple linear regression discussed above, but it would be efficient to extract only the salient features and impart them as inputs to various algorithms. The process of feature extraction could be performed using all possible regression approaches or automatic methods. All possible regression approach used adjusted R^2, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values to select features. Automatic methods in regression include Stepwise regression with either forward or backward approaches.

E. STEPWISE REGRESSION

Selection of salient features using stepwise regression is affected if the independent variables are highly correlated. In such situations with highly correlated data, the variance is increased and it is indicated as multi-collinearity in regression literature [14]. The presence of multi-collinearity in our dataset is studied using PCA and PLS.

F. PCA BASED APPROACH FOR CLASSIFICATION

Principal Component Analysis is mainly used to reduce the dimensionality of the given problem. The basic concept behind the PCA is the use of orthogonality of vectors and to obtain the vectors considering the variance of projections. This causes,

$$X = TP \quad (10)$$
$$\textit{where } T \textit{ is the orthogonal score}$$
$$\textit{and } P \textit{ is load}$$

Principal components are obtained from the given independent variables and they are used for classification. PCA can do the classification considering the multicollinearity among independent variables which is one of the shortcomings in the linear regression discussed above. Thus, the correlated independent variables are converted to non-correlated principal components. At the same time, we could also reduce the number of variables involved.

Principal component analysis is performed using the following steps:

1) Standardize the independent variables, as specified in eqn. 1.
2) Transform the data by multiplying the it (nx p matrix; n observations with p independent variables) with the eigenvectors (p x p matrix).

$$Z = XA \quad (11)$$

After identifying the salient principal components involved, it was determined that there is no one-to-one relationship between a principal component and the original independent variables. Thus, reducing the principal component does not reduce the considered independent variables and thus, all the independent variables are considered in the PCA.

G. PCR BASED APPROACH

The impact of a selected independent variable over the dependent variable is not considered in PCA. Principal Component Regression (PCR) is used to relate the selected principal components with the dependent variable (Y). The coefficients can be represented as

$$B = P(T^T T)^{-1} T^T Y \quad (12)$$

The regression equation is represented as:

$$Y = XB$$
$$\textit{where } X = TP \quad (13)$$

H. PLSR BASED APPROACH

Partial Least Squares Regression (PLSR) extracts features in such a way that the selected X and Y have higher covariance. This helps in achieving the same prediction error as the PCR approach but with a fewer number of components. The impact of normalization on the PLSR approach is to be studied. As described in [15], PLSR is highly sensitive to noise in the dataset, possibly caused by irrelevant variables.

I. DEEP LEARNING BASED APPROACH

Non-linear representations of data could be trained using a deep learning approach. Deep learning solves the problems of neural networks, which are mainly composed of vanishing gradients and may suffer from overfitting related problems. This approach overcomes such problems due to various activation functions, using huge amounts of data and handling dropouts. Deep learning combines multiple neural networks with non-linear activation functions.

IV. OUR APPROACH

This paper addresses the method to achieve higher accuracy when for water desalination data, using various techniques.

A. DATA COLLECTION

As described earlier, two datasets are used: one with 552 observations and the other with 55004. The first dataset comprises of *state* as dependent variable and {*tih*, *toh*, *tic*, *toc*, *pih*, *poh*, *pic*, *poc*, *qh* and *qc*} as the independent variables. The second dataset is made of the following variables: *state* as the dependent variable and {*tih*, *toh*, *tic*, *toc*, *pih*, *poh*, *pic* and *poc*} as the independent variables. Given the above dataset, 75% of the data available is considered for training and the remaining for testing. To obtain the final result, the data selection from the list is randomized and the process is repeated 10 times.

Data is processed using the following procedures and the outcome is compared.

1) Decision Trees.
2) Multiple linear regression without any pre-processing of data.

TABLE 2. Multiple linear regression (with pre-processing).

	Adjusted R-Square value	Variables with higher p-value	Mean Square Error
Dataset 1 (with normalization)	0.9255	tic, toc, pih, pic, poc and qc	0.07296927
Dataset 2 (with normalization)	0.5365	NULL	0.4634422

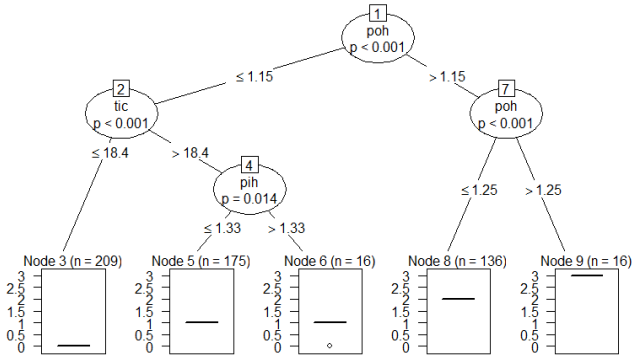

FIGURE 2. Outcome of the Decision Tree.

3) Multiple linear regression with standardization of data.
4) Principal Component Analysis and Principal Component Regression.
5) Deep Learning.

B. DECISION TREE

Decision tree algorithm when applied on dataset 1, with or without normalization, was able to classify the data with an accuracy of 97.1%. Dataset 2, with or without normalization, was able to classify the data with an accuracy of 95.12%.

C. MULTIPLE LINEAR REGRESSION (WITHOUT PRE-PROCESSING)

Considering the significance level of 0.05, certain independent variables have higher p-values. The observation from these two datasets, shown in Table 1, indicate that dataset 2 has only 78.63% variance according to the regression equation. The presence of a fewer number of independent variables and more observations have caused the p-value of only one element to increase. Upon observing the datasets, it is clear that the values of dataset 1 have substantial differences in their mean values. The values of *qh* and *qc* in dataset 1 are higher and thus their coefficients are 9.475e-05 and -2.960e-04 respectively, causing very little impact.

D. MULTIPLE LINEAR REGRESSION (WITH PRE-PROCESSING)

The given datasets 1 and 2 are standardized using equation 1. The mean squared error is high for dataset 2 but the p-values are all acceptable given the level of significance. In dataset 1, 92.55% of the variations of observations are explained by the equation of regression, with or without normalization.

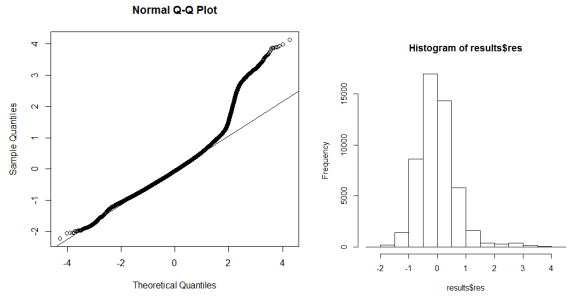

FIGURE 3. Q-Q Plot and Histogram for residuals with non-normalized dataset 2.

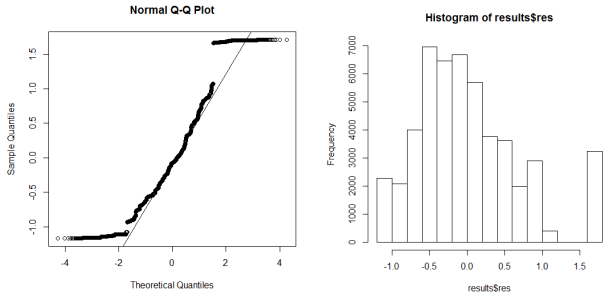

FIGURE 4. Q-Q Plot and Histogram for residuals with normalized dataset 2.

Meanwhile, this value drops to 78.63% without normalization and still down to 53.65% for dataset 2 with normalization. The reason behind this substantial variation of observations was explored further. Residuals are to be checked whether they are normally distributed and have equal variance. The normal distribution is verified using a Q-Q plot as shown in Figure 3 and 4.

It is observed from figures 3 and 4 that normalization of the dataset causes the normality to be affected. Thus, causing the drop in the variations of equations explained by the regression line. Thus, we could conclude that normalization could not be expected to improve the adjusted R-square value always, as shown in Table 2.

Normality checks of residuals obtained from both datasets is tested using:

(1) Shapiro-Wilk normality test: Due to restriction on the sample size (3 to 5000), this test is applied only on dataset 1. This confirms that the sample is obtained from a population that follows a normal distribution.

(2) Kolmogorov-Smirnov test: The resultant p-value indicates that the distribution is normal.

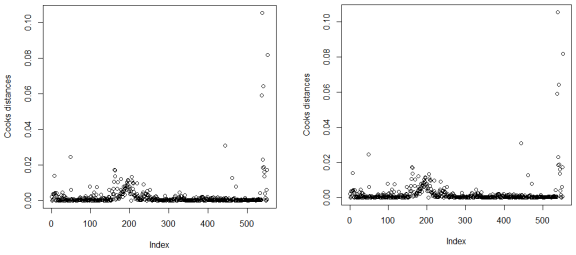

FIGURE 5. Q-Q Cook's distance for non-normalized and normalized dataset 1.

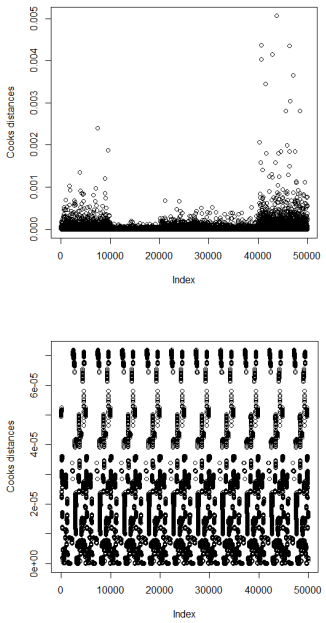

FIGURE 6. Cook's distance for non-normalized and normalized dataset 2.

Cook's distance was measured for both the normalized and non-normalized datasets. There was no specific indicator of influential elements identified by Cook's distance, as shown in figures 5 and 6.

E. FEATURE SELECTION USING ALL POSSIBLE REGRESSION APPROACHES

The effects of different features was tested using *regsubsets* in the *leaps* library. The outcomes of Adjusted R^2 and that of BIC differ significantly, which prevented us from reaching a decisive conclusion. Figure 7 shows the inclusion of a variable using black and exclusion of a variable using white. Observing the top-most row which provides the minimum Adjusted R^2 or BIC, it is obvious that the variables selected using different criteria are different. To explore further, stepwise regression was used.

F. STEPWISE REGRESSION

In order to reduce the number of independent variables involved in regression, stepwise regression was applied on both the datasets. Forward as well as backward stepwise

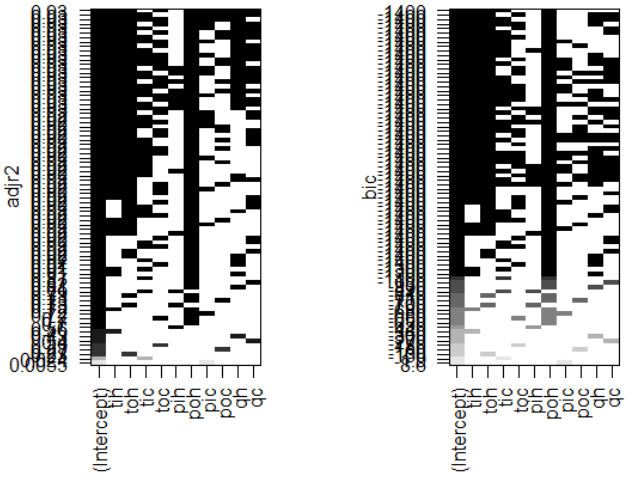

FIGURE 7. Feature selection in dataset 1 using Adj R^2 and BIC.

regression are used to reduce the variables involved in regression. Based on the selected variables, adjusted r-square and mean square error were calculated, as shown in Table 3. Based on the observations provided in Table 3, the selection of five variables (tih, toh, toc, poh, qh and qc) for dataset 1 are sufficient to match the required level of significance. But, dataset 2 has still issues in achieving the required level of significance, so Principal Component Regression is considered to solve this issue.

G. PRINCIPAL COMPONENT ANALYSIS AND PRINCIPAL COMPONENT REGRESSION

Principal component analysis is applied on the datasets, with and without normalization. PCA considered 10 components with different combinations of the given 10 independent variables in dataset 1, shown in Table 4. Considering only 5 principal components, we could achieve 99% percent of the variations.

The correlation between various principal components and the original dependent variable is showed in the Table 5. It could easily be observed from Table 5 that PC1, PC2, PC3 and PC5 are correlated better with the dependent variable.

Similarly, PCA was performed on the non-normalized dataset 2, shown in Table 6.

Considering only 5 principal components, we could achieve 97.77% percent of the variations. It is worth mentioning that using only the first principal component achieves only 37.3% variance. The correlation between various principal components and the original dependent variable, both for normalized and non-normalized data, is showed in Tables 7 and 8.

For the non-normalized data described in Table 7, it could easily be observed that PC1, PC2, PC3 and PC5 are correlated better with the dependent variable.

For the normalized data described in Table 8, it could easily be observed that PC1, PC2, PC3, PC4 and PC8 are correlated better with the dependent variable.

TABLE 3. Stepwise regression (features selection).

	Process	Selected variables	Adjusted R-Square value	Variables with higher p-value	Mean Square Error
Dataset 1 (without normalization)	Forward Stepwise Regression	poh, tic, qc, toh, tih, qh	0.9256	NULL	0.05422843
Dataset 1 (with normalization)	Forward Stepwise Regression	poh, tic, qc, toh, tih, qh	0.9256	NULL	0.0734502
Dataset 1 (without normalization)	Backward Stepwise Regression	tih, toh, toc, poh, qh, qc	0.9256	NULL	0.05421195
Dataset 1 (with normalization)	Backward Stepwise Regression	tih, toh, toc, poh, qh, qc	0.9256	NULL	0.07342789
Dataset 2 (without normalization)	Forward Stepwise Regression	tic, pic, tih, poh, poc, toh, toc	0.7863	NULL	0.4273944
Dataset 2 (with normalization)	Forward Stepwise Regression	tih, toh, poc, tic, toc, pic, pih, poh	0.481	poh (0.186)	0.5189058
Dataset 2 (without normalization)	Backward Stepwise Regression	tih, toh, tic, toc, pih, poh, pic, poc	0.7863	pih (0.243)	0.4273828
Dataset 2 (with normalization)	Backward Stepwise Regression	tih, toh, tic, toc, pih, poh, pic, poc	0.5365	NULL	0.4634422

TABLE 4. Principal component analysis.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	2.5227	1.3886	0.8553	0.6772	0.6563	0.2053	0.1842	0.0947	0.0385	0.00431
Proportion of Variance	0.6364	0.1928	0.0731	0.0458	0.0430	0.0042	0.0033	0.0009	0.0001	0
Cumulative Proportion	0.6364	0.8292	0.9024	0.9482	0.9913	0.9955	0.9989	0.9998	1	1

TABLE 5. Correlation of principal components.

state	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
1	0.703791	-0.22535	-0.10956	0.078408	-0.59519	0.024693	0.071948	0.012308	0.04893	-0.00759

TABLE 6. PCA on non-normalized dataset 2.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.727	1.4075	1.2074	1.0221	0.59556	0.32026	0.24415	0.12503
Proportion of Variance	0.373	0.2476	0.1822	0.1306	0.04434	0.01282	0.00745	0.00195
Cumulative Proportion	0.373	0.6206	0.8029	0.9334	0.97777	0.99059	0.99805	1

H. DATASET 1

An adjusted R-Squared result of 0.9255 was obtained when using all of the principal components as input to PCR regression. Thus, the Adjusted R-Square obtained is the same as the one achieved in Multiple Linear Regression. At the same time, the principal components PC8 and PC10 have higher p-value compared to the level of significance. Considering only PC1, PC2, PC3 and PC5 as independent variables in linear

TABLE 7. Non-normalized data correlation.

state	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
1	-0.53645	-0.24493	-0.12582	0.087715	-0.63721	-0.010211	-0.067015	-0.066207

TABLE 8. Normalized data correlation.

state	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
1	0.605662	-0.15250	-0.17943	0.290933	0.058265	0.043966	-0.02822	-0.153298

TABLE 9. Partial least squares regression (non-normalized dataset 1).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
SS Loadings	1	1	1.259	1.184	1.004	1.021	1.008	1.101	1.006	1
Proportion of Variance	0.1	0.1	0.126	0.118	0.1	0.102	0.101	0.11	0.101	0.1
Cumulative Variance	0.1	0.2	0.326	0.444	0.545	0.647	0.748	0.858	0.958	1.058

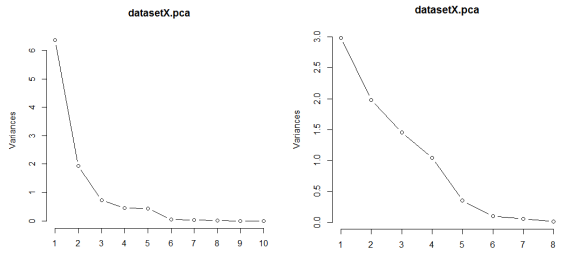

FIGURE 8. Variances with Principal Components for dataset 1 and 2.

regression, the achieved Adjusted R-Square value is 0.9117, where all of the obtained p-values were negligible.

The comparison between all PCs based regression schemes and the selected PCs based regression was studied using ANOVA. The selected PCs based regression gave an F-value of 17.931 and a very small p-value with 547 degrees of freedom.

I. DATASET 2

Including all the principal components into the regression as PCR, resulted with an Adjusted R-Squared value of 0.5365. Thus, the Adjusted R-Square obtained is the same as the one achieved in Multiple Linear Regression. All the Principal components have achieved the required p-value. Using PC1, PC2, PC3 and PC5 as principal components, The adjusted R-Square value drops to 0.4256. The comparison between all PCs based regression and the selected PCs based regression was studied using ANOVA. The results for the selected PCs based regression gave an F-value of 2990.1 and a very small p-value with 49999 degrees of freedom.

J. PARTIAL LEAST SQUARES REGRESSION (PLSR)

Similar to PCR, PLSR is also said to handle multicollinearity efficiently. PLSR was applied both on normalized and non-normalized data of both datasets. The proportion of variance shown in Table 9, was obtained for the non-normalized dataset 1. The proportion of variance shown in Table 10, was obtained for normalized dataset 1.

A similar pattern of having cumulative variance of more than 1 was observed for both the normalized and non-normalized versions of dataset 2, but the number of principal components required to achieve 95% explained variance (R^2) increased from 4 to 6, when the dataset was normalized. The impact of normalization on reducing the number of principal components is not consistent.

K. DEEP LEARNING

Deep learning was used to test the learning curve of the given datasets, in both normalized and non-normalized form. Deep Belief Network (DBN) provided by deepnet package of R language was used for purpose. The hidden nodes were given as 1 or 25 or 100 nodes each in one to five hidden layers. The number of hidden nodes in each hidden layer is maintained as constant. The outcome of Deep Learning on different datasets are given in Table 11.

With multiple hidden layers, we cannot maintain the number of nodes per hidden layer to be 1 because of connectivity problems. Increasing the number of hidden layer nodes to 10 reduces the error rate. But, it is clear that the normalization of data helps in improving the error rate. Adding many hidden layers just increases the time taken to complete the task. Thus, it could easily be concluded not to go beyond three hidden layers. The poor performance of the deep learning in this training might be due to lack of substantial data, as it is known that big data is required for deep learning to train the system better. Dataset 1 has very minimal data and thus causes Deep Learning to perform relatively poorly. The performance on dataset 2 with a more substantial amount of data is comparable with a few other techniques considered in this research.

TABLE 10. Partial least squares regression (normalized dataset 1).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
SS Loadings	1.059	1.591	1.012	1.125	1.668	1.043	3.418	1.303	1	1
Proportion of Variance	0.106	0.159	0.101	0.113	0.167	0.104	0.342	0.13	0.1	0.1
Cumulative Variance	0.106	0.265	0.366	0.479	0.645	0.75	1.091	1.222	1.322	1.422

TABLE 11. Deep learning.

	Number of hidden nodes per layer	Dataset 1 (Non-normalized)	Dataset 1 (Normalized)	Dataset 1 - Average Time in Seconds	Dataset 2 (Non-normalized)	Dataset 2 (Normalized)	Dataset 2 - Average Time in Seconds
Error Rate with One hidden Layer	1	0.3423913	0.4090669	0.01	0.7	0.4318072	0.78
	10	0.3423913	0.2858785	0.03	0.7	0.4230542	1.24
	20	0.3423913	0.2858785	0.04	0.7	0.4318072	1.81
Error Rate with two hidden layers	1	Weight size not equal to network size			Weight size not equal to network size		
	10	0.3423913	0.2858785	0.04	0.7	0.4318072	2.36
	20	0.3423913	0.2858785	0.08	0.7	0.4230542	4.00
Error Rate with three hidden layers	1	Weight size not equal to network size			Weight size not equal to network size		
	10	0.3423913	0.2858785	0.06	0.7	0.4230542	3.58
	20	0.3423913	0.2858785	0.09	0.7	0.4230542	6.04
Error Rate with four hidden layers	1	Weight size not equal to network size			Weight size not equal to network size		
	10	0.3423913	0.4090669	0.08	0.7	0.4230542	4.78
	20	0.3423913	0.2858785	0.12	0.7	0.4230542	8.39
Error Rate with five hidden layers	1	Weight size not equal to network size			Weight size not equal to network size		
	10	0.3423913	0.4090669	0.09	0.7	0.4230542	6.15
	20	0.3423913	0.2858785	0.15	0.7	0.4318072	10.95

TABLE 12. Discussion summary

	Time (Sec) - Dataset 1	**Accuracy OR Adj R^2 - Dataset 1**	**Time (Sec) - Dataset 2**	**Accuracy OR Adj R^2 - Dataset 2**
Decision Tree	0.06	0.971	2.25	0.9512
Multiple Linear Regression	0.04	0.9255	0.12	0.5365
Stepwise Regression	0.14	0.9256	1.47	0.7863
PCR	0	0.9117	0.13	0.4256
Deep Learning	0.06	0.7141215	3.58	0.5769458

V. RESULTS AND DISCUSSION

The summary of all of the above processes is given in Table 12. In order to compare the performance of various approaches, the given datasets are tested with and without normalization of the data. Furthermore, the error rate, accuracy, adjusted R^2 and processing time are analysed. Better accuracy is expected from any process given an acceptable execution time. This duration is vital when the processing of the test data is to be in real time. Adjusted R^2 provides information regarding the variation in the system.

The accuracy of the decision tree algorithm remained similar on both normalized and non-normalized data. In multiple linear regression, the variance explained by the regression dropped for dataset 2. Normalizing dataset 2 and applying multiple linear regression still caused the R^2 value to drop further. This was reflected in the high mean squared error.

To enhance the regression analysis, variable selection was performed using all possible regression approaches. Criteria like Adjusted R^2 and BIC gave different results based on variable selection, making us to try stepwise regression for variable selection. Stepwise regression identified

five variables that are significant in dataset 1 but variable selection by forward and backward stepwise regression was not conclusive for dataset 2. This concern was further studied using PCA. PCA was able to achieve 99% and 97.77% accuracy using only five principal components. The outcome of PCA was applied to PCR regression along with an F-test to study its validity. In order to handle multicollinearity, PLSR was applied both on normalized and non-normalized datasets. It was observed that the number of principal components required to achieve 95% accuracy induced the variance increase from 4 to 6 for the normalized dataset. Furthermore, deep learning was applied on the same datasets. Deep learning with 1 to 5 hidden layers, with 1 or 25 or 100 nodes each, was trained and tested. The error rate was reduced when the number of hidden layers was increased to 10 but further increases in hidden layer numbers did not help with accuracy while increasing the computation time. Similarly, adding more than three hidden layers did not enhance the accuracy.

Overall, it can be observed that decision trees provided better accuracy than other algorithms in this application, given our data sets. Even though the time required by regression is shorter for both datasets, it yields poorer accuracy, especially for dataset 2. Considering the trade-off between processing time and accuracy, decision trees outperform all others.

VI. CONCLUSION

Water desalination is a vital industry especially, in environments that are facing lack of pure water. There is very limited research in the area of automatic fault detection in water purification plants, mainly due to the lack of datasets that include data from various faulty states. This research studies the process of identifying faults in water desalination. Two different datasets were used with very similar features. One with the real values collected from the plant and one with generated values. The second dataset was created for the purpose of testing the behavior of the selected method with larger data. These two datasets were analyzed using different approaches such as decision trees, multiple linear regression, stepwise regression, principal component analysis, partial least squares regression and deep learning, both with and without normalization. Multiple linear regression was found to be inappropriate for the second dataset. Further, salient features from the given dataset were selected with PCA. Based on the analysis, five principal components are fo0und to be significant features to achieve 97.77% accuracy for both datasets with or without normalization. Deep learning in the form of deep belief network was tried with varied number of hidden layers and hidden nodes per layer. The outcome of deep learning was not in accordance with our expectations in dataset 1, mainly due to lack of sufficient data. Meanwhile, the performance of deep learning for dataset 2 is comparable with the other considered techniques. Overall, the decision trees algorithm outperforms others considering the trade-off between processing time and accuracy.

REFERENCES

[1] E. E. Tarifa and N. J. Scenna, "A methodology for fault diagnosis in large chemical processes and an application to a multistage flash desalination process: Part II," *Rel. Eng. Syst. Safety*, vol. 60, no. 1, pp. 41–51, 1998.

[2] S. Bouguechaa, S. Alya, M. Al-Beiruttya, M. Hamdia, and A. Boubakrib, "Solar driven DCMD: Performance evaluation and thermal energy efficiency," *Chem. Eng. Res. Des.*, vol. 100, pp. 331–340, Aug. 2015.

[3] M. Derbali, A. Fattouh, S. Buhari, G. Tsaramirsis, H. Jerbi, and M. N. Abdelkrim, "Discriminant analysis based fault detection in desalination systems," presented at the 3rd Int. Conf. Math. Sci. Comput. Eng., 2016.

[4] M. Derbali, A. Fattouh, H. Jerbi, and M. NaceurAbdelkrim, "Improved fault detection in water desalination systems using machine learning techniques," *J. Theor. Appl. Inf. Technol.*, vol. 9, p. 2, Oct. 2016.

[5] B. C. S. Cauvin, B. Heim, S. Gentil, and L. Travé-Massuyès, "Model based diagnostic module for a FCC pilot plant," *Oil Gas Sci. Technol.-Rev.*, vol. 60, no. 4, pp. 661–679, 2005.

[6] Y. Shu, L. Ming, F. Cheng, Z. Zhang, and J. Zhao, "Abnormal situation management: Challenges and opportunities in the big data era," *Comput. Chem. Eng.*, vol. 91, pp. 104–113, Aug. 2016.

[7] H. El-Dessouky, H. I. Shaban, and H. Al-Ramadan, "Steady-state analysis of multi-stage flash desalination process," *Desalination*, vol. 103, pp. 271–287, Dec. 1995.

[8] E. E. Tarifa and N. J. Scenna, "Fault diagnosis for a MSF using a SDG and fuzzy logic," *Desalination*, vol. 152, pp. 207–214, Feb. 2003.

[9] S. A. Said, M. Emtir, and I. M. Mujtaba, "Flexible design and operation of multi-stage flash (MSF) desalination process subject to variable fouling and variable freshwater demand," *Processes*, vol. 1, no. 3, pp. 279–295, 2013, doi: 10.3390/pr1030279.

[10] E. E. Tarifa, D. Humana, S. Franco, S. L. Martínez, A. F. Núñez, and N. J. Scenna, "Fault diagnosis for MSF dynamic states using neural networks," *Desalination*, vol. 166, pp. 103–111, Aug. 2004.

[11] E. E. Tarifa, D. Humana, S. Franco, S. L. Martínez, A. F. Núñez, and N. J. Scenna, "Fault diagnosis for a MSF using neural networks," *Desalination*, vol. 152, pp. 215–222, Feb. 2003.

[12] S. J. Ahn, C. J. Lee, Y. Jung, C. Han, E. S. Yoon, and G. Lee, "Fault diagnosis of the multi-stage flash desalination process based on signed digraph and dynamic partial least square," *Desalination*, vol. 228, pp. 68–83, Aug. 2008.

[13] W. Zhou, C. Wu, D. Chen, Y. Yi, and W. Du, "Automatic microaneurysm detection using the sparse principal component analysis-based unsupervised classification method," *IEEE Access*, vol. 5, pp. 2563–2572, 2017.

[14] S. Maitra and J. Yan. (2008). *Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression, Casualty Actuarial Society, 2008 Discussion Paper Program*. [Online]. Available: https://www.casact.org/pubs/dpp/dpp08/08dpp76.pdf

[15] *Partial Least Squares Regression*, accessed on Jan. 7, 2017. [Online]. Available: http://www.vcclab.org/lab/pls

MORCHED DERBALI, photograph and biography not available at the time of publication.

SEYED M. BUHARI, photograph and biography not available at the time of publication.

GEORGIOS TSARAMIRSIS, photograph and biography not available at the time of publication.

MILOS STOJMENOVIC, photograph and biography not available at the time of publication.

H. JERBI, photograph and biography not available at the time of publication.

M. N. ABDELKRIM, photograph and biography not available at the time of publication.

MOHAMMAD H. AL-BEIRUTTY, photograph and biography not available at the time of publication.