Soft Computing (2021) 25:7459–7478
https://doi.org/10.1007/s00500-021-05706-4

METHODOLOGIES AND APPLICATION

A hybrid of six soft models based on ANFIS for pipe failure rate forecasting and uncertainty analysis: a case study of Gorgan city water distribution network

Seyed Mehran Jafari[1] · Abdol Reza Zahiri[1] · Omid Bozorg Hadad[2] · Mahmoud Mohammad Rezapour Tabari[3,4]

Accepted: 17 February 2021 / Published online: 8 March 2021

Abstract

The pipes as one of the main and important components of a water distribution network break during operation due to various factors. Developing models for pipes failure rate prediction can be one of the most important tools for managers and stakeholders during optimal operation of the water distribution network. In this study, the statistical and soft models such as Linear Regression, Generalized Linear Regression, Support Vector Machine, Feed Forward Neural Network (FFNN), Radial-Based Function Neural Network (RBFNN), and Adaptive Neuro-Fuzzy Inference System (ANFIS) were studied in order to predict the pipes failure rate based on the characteristics of Gorgan city water distribution network including diameter, length, age, installation depth, and number of failures of each pipe. In order to determine the optimal values of the parameters of each model, appropriate error indices including correlation coefficient (R), Mean Square Error (MSE), and Correlation Mean Square Error Ratio (CMSER) for training and test data were calculated, and the values of the parameters related to the model with the highest value of the CMSER index were considered as the model optimal values. Furthermore, in the validation stage, the values of R and MSE error indices for each of the above models were considered as a criterion for selecting the most appropriate model for predicting pipe failure rate. The findings show that among the soft and statistical models investigated, ANFIS with MSE of 0.071 and R of 0.92 can predict the failure rate of the studied network pipes more efficiently and more accurately than other models. Yet, despite the superiority of this model over other models, this model cannot accurately predict the failure rate of the studied network pipes due to its relatively high MSE value. Therefore, a new approach was developed based on the hybridization of trained models to provide a more efficient model for a more accurate prediction of the pipe failure rates of water distribution network. In this approach, the values of the network pipe failure rate predicted by each of the soft and statistical models are considered as independent input variables, and the observational failure rate values are considered as the dependent output variable of the ANFIS model. A comparison between the values of non-hybrid model validation data indices and the results of the proposed hybrid prediction model reveals that the use of the developed hybrid model increased the R error value from 8.1% (compared to the ANFIS model) to 260% (compared to the RBFNN model). It also decreased the MSE error value from 37% (compared to the FFNN model) to 58% (compared to the RBFNN model). Moreover, the hybrid model, compared to the superior non-hybrid ANFIS model, decreased MSE error rates by 45%. The findings show that the proposed model can significantly raise the accuracy of predicting the failure rate of pipes, compared to other existing models.

Keywords Failure rate · Prediction · Gorgan · Water distribution network · Hybrid model · Adaptive neuro-fuzzy inference system

1 Introduction

Nowadays, the majority of cities in different countries provide the people with the facilities and services of water distribution networks. These networks, which are among the most vital infrastructures in residential areas, fulfill the need for the drinking water in industrial and health sectors. During the operation phase, water distribution network pipes undergo physical failures for various reasons. The breakage probability of older pipes in dispersive soil is very high under unstable and harsh weather conditions. There are many factors affecting the failure of pipes. They

Extended author information available on the last page of the article can be divided into three categories: (1) Pipe characteristics including diameter, material, age, coating, and defects related to pipe construction; (2) Environmental conditions including soil type, soil characteristics, corrosion, frost, rainfall, climate change, and temperature, and (3) implementation conditions including pressure, loading conditions, installation depth, etc. (Barton et al. 2019; Rajeev et al. 2014). Economic, social, and environmental consequences of the physical failure of urban water distribution network pipes have always been a challenge for the management of water distribution networks. Therefore, the correct prediction of the failure rate of network pipes during the operation phase can solve many of the challenges managers and operators are facing with.

Over the past decades, a number of studies have used various models to predict the failure rate of water distribution network pipes. The comparison between physical and statistical models shows that physical models are used for very important parts of the network due to the high cost of collecting the required input information such as internal corrosion of the pipe and loads applied to the pipe. However, statistical models are mostly used for small networks due to the availability of the required information (Rajani and Kleiner 2001; Kleiner and Rajani 2002).

One of the first attempts in order to use statistical models in pipe failure modeling of water distribution network using linear regression was performed by Shamir and Howard (1979). Then, based on the parameters affecting the failure of pipes such as material, age, diameter, length, temperature, and soil conditions, various researchers were able to predict the failure rate of urban water distribution network pipes with a relatively good accuracy. They used a wide range of statistical models such as Multivariate Linear Regression Models (MLRM), Generalized Linear Regression (GLR), Nonlinear Regression (NLR), Logistic Regression (LR), Time Exponential Model (TEM), and Proportional Hazard Model (PHM) (Soltanjalili et al. 2011; Gasemnezhad et al. 2014; Shin et al. 2015; Faris Hamdala and Sagar 2016; Kakoudakis et al. 2017; Robles-Velasco et al. 2020).

In addition to statistical models, soft models such as Artificial Neural Network (ANN), Genetic Algorithm (GA), and Fuzzy Interference System (FIS) models are also widely used in modeling and predicting pipe failure rates. The use of these models has increased the accuracy of predicting the failure rate of pipes. ANN model for modeling the failure of water distribution network pipes was initially used in 1999 by Sacluti. Afterward, other researchers used this model in order to predict the pipe failure rates, optimal water leakage management, decision-making for investment, and modernizing the urban water distribution network. They have shown that the ANN model can be used to predict the failure rate of network pipes with an appropriate precision (Sacluti 1999; Mounce et al. 2002; Tabesh et al. 2009; Ho et al. 2010; Jafar et al. 2010; Asnaashari et al. 2013; Harvey et al. 2013; Sattar and Gharabaghi 2015; Sattar et al. 2016, 2019; Kerwin et al. 2019).

Besides the ANN models, FIS models are widely used in water science as well. The risk of water quality failure in the water distribution network, the risk of failure of the water distribution network pipes, and the potential for leakage in the water distribution network have been investigated by various researchers using FIS models (Sadiq et al. 2007; Fares and Zayed 2010; Islam et al. 2011; Valis 2013; Zangenehmadar and Moselhi 2016; Pandey et al. 2020).

In recent decades, various studies have been carried out on the application of statistical and soft models to finding leak position, calibration of pipe roughness coefficient, water quality prediction, and network pipe failure rate prediction. The results of this research show that hybrid models can be used with great precision in predicting various phenomena (Kapelan et al. 2003; Tu et al. 2005; Berardi et al. 2008; Xu et al. 2011, 2013; Soltani and Tabari 2012; Farmani et al. 2017; Tabari and Malekpour Shahraki 2018; Tavakoli et al. 2019; Malekpour and Tabari 2020; Tabari et al. 2020).

A review of studies on the application of hybrid models in various sciences shows that these models have a very high ability to predict phenomena with acceptable accuracy. The use of combination models to predict the failure rate of water distribution network pipes has been studied by a few researchers in the last decade. A review of their studies shows that, firstly, the number of models used in the hybrid model is limited to only two models, and secondly, only a combination of neural network and genetic models is used, and thirdly, none of the capable statistical models have not been used in previous hybrid models.

Using more models, according to the specific strengths of each model, can increase the accuracy of the hybrid model to predict the pipe failure rate. In addition, the use of statistical models along with intelligent models, due to the high ability of statistical models in predicting phenomena based on statistical data, can also improve the prediction accuracy of the hybrid model.

Therefore, the main purpose of the present study is to develop a new hybrid model in order to more accurately and acceptably predict of the pipes failure rate. In the hybrid model developed in this study, in order to overcome the limitations of combination models of previous studies and thus increase the accuracy of pipe failure rate prediction, the number of models used in the hybrid model is increased to six models. Also, statistical models have also been used. The statistical and intelligent models which are used in the study include Linear Regression (LR), Generalized Linear Regression (GLR), Support Vector Machine (SVR), Feed Forward Neural Network (FFNN), Radial-Based Function Neural Network (RBFNN), and Adaptive Neuro-Fuzzy Inference System (ANFIS). Therefore, the innovation of the proposed hybrid model compared to other developed models, which will improve the prediction accuracy, can be expressed as follows:

- Development of an approach based on the hybrid use of the capabilities of soft simulation models
- Simultaneous application of statistical models in the simulation model developed to cover the weaknesses of soft models
- Determination of the degree of uncertainty of the proposed model in predicting the failure rate of pipes
- Extraction of parameters affecting the uncertainty of prediction of failure rate

In this study, the single prediction models along with the proposed hybrid model were used to predict the failure rate of pipes in the real water distribution network of Gorgan City. Among the parameters affecting the failure of water distribution network pipes, in this study, due to the lack of access to the values of all parameters affecting the pipe failure rate, only information on diameter, length, installation depth, age, and number of historical failures has been used as effective variables. In order to compare the results of different models and the proposed hybrid model, the appropriate error indices were performed. The results obtained from the implementation of the innovated model show an increase in the accuracy of the hybrid model in predicting the failure rate of water distribution network pipes. Based on this correct prediction, the failure rate of pipes can significantly reduce the cost of operation and maintenance of urban water distribution networks.

2 Case study

The Gorgan City with a longitude of 54° 26′ 38″ E and latitude of 36° 50′ 33″ N is located in the north of Iran. This city has a population of about 48,054 people and is 3000 years old. Statistics pipe failure events in Gorgan water distribution network show that the number of pipe failures in this city's network is relatively high due to network wear-out. This has caused many problems in the operation of the water distribution network. This study focused on part of Gorgan water distribution network due to its high number of failures and selected as case study. Thus, the prediction of failure rate in this area can lead to improved management of water distribution network in the future (Fig. 1). According to the statistics available in Golestan Water and Sewage Company, this study used the network pipe information in a period of four years (2015–2018), which includes diameter (D), length (L), age (Ag), depth of installation (DI), and the number of failures of each pipe. The characteristics of the water distribution network are presented in Table 1.

3 The structure of the proposed methodology

In order to development of model for predicting the failure rate of water distribution network pipes in the study area, the hybrid of statistical and soft models was applied. The models used in the combination of the hybrid model proposed in this research include two statistical models (linear regression, generalized linear regression and support vector machine) and four intelligent models including Support Vector Machine, Feedforward Neural Network, Radial-based Neural Network, and System Neuro-Fuzzy Inference (adaptive). The structure of the proposed approach is shown in Fig. 2.

According to this figure, the proposed methodology consists of three parts:

- Extract the failure rates of the water distribution network pipes: In this stage, the specifications of the network pipes including length, diameter, age, depth of installation, and number of previous failures are collected, and then, the failure rate of network pipes is calculated based on Eq. (1).
- Development soft and statistical model and their simulation: Using the failure rate data obtained from the previous step, the optimal parameters of each statistical and intelligent model are calculated, and then, the best mode of each model (for each of the studied six models) is extracted. According to the selected superior models, the failure rate of the pipes is predicted based on the input data set (specified from the first stage). It should be noted that in this step, the most efficient structure of the prediction model is introduced for use in the hybrid model.
- Implementation of the proposed hybrid model: In this stage, the innovated hybrid model is implemented using the predicted failure rate values with each of the intelligent soft and statistical models (as input parameters) and the observed failure rate values (as output parameter).
- Compare models and uncertainty analysis: Finally, in order to evaluate the performance of the hybrid model to predict the failure rate of pipes and compare it with other models, the values of error indices of intelligent, statistical, and proposed hybrid models are calculated and the superior model is selected. Also, using the Monte Carlo simulation method (MCS), the uncertainty

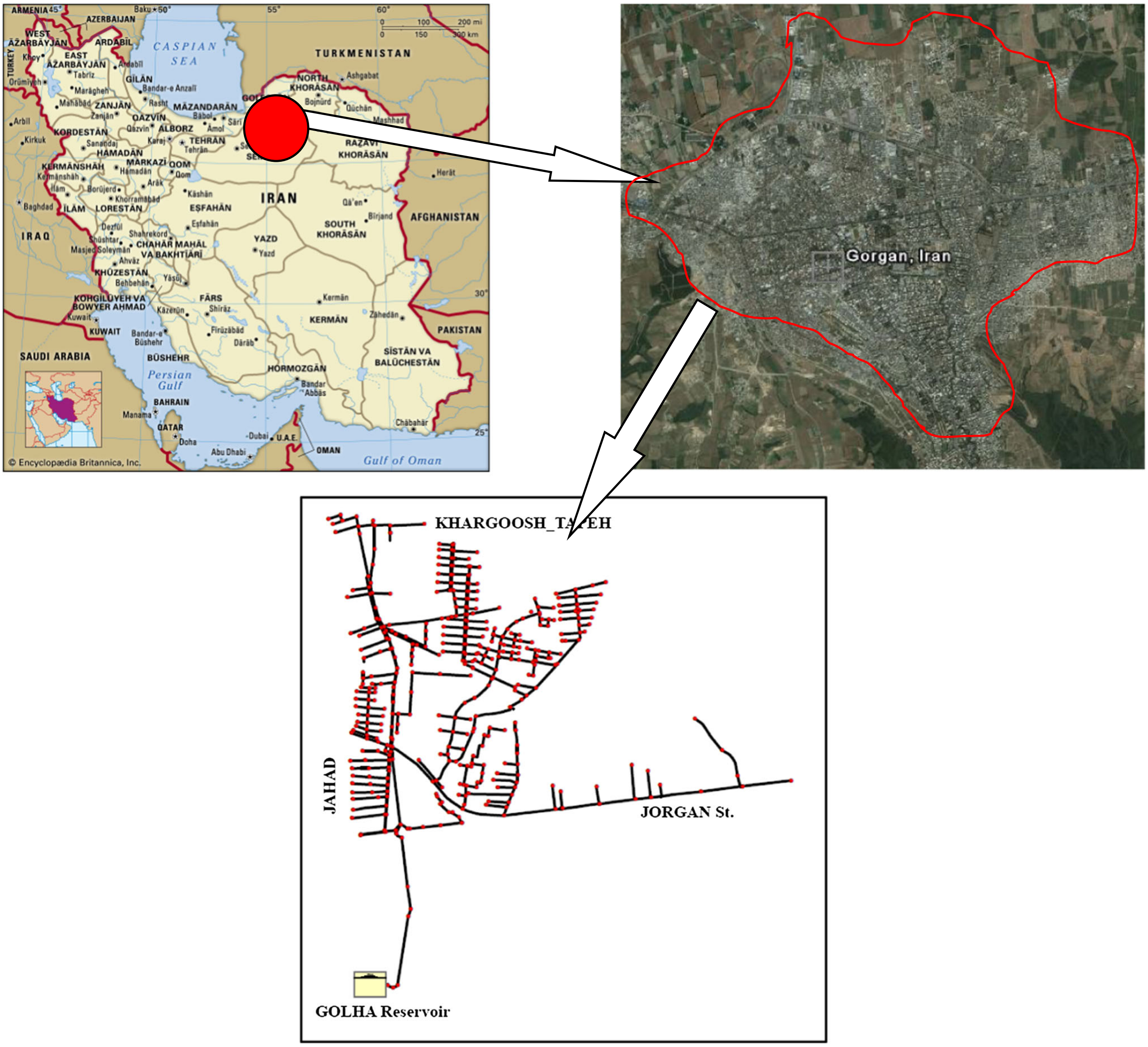

Fig. 1 Location of a part of Gorgan city water distribution network

Table 1 Characteristics of the studied water distribution network

Parameters	Value
Number of reservoirs	1
The time period studied (year)	2014–2018
Diameter range (mm)	63–500
Total length of pipes (m)	80,072
Number of pipes	1547
Number of failures	169
Installation depth range (m)	0.4–2.8

of the proposed model is determined and the effect of each of the input parameters to the forecast model in creating uncertainty of model is determined.

3.1 Parameters affecting pipe failure

The first step in developing simulation model for predicting a phenomenon is to identify the parameters or variables that affect the phenomenon and to convert those variables into quantitative measures to be used in the model. The phenomenon in this study is to predict the failure rate of urban water distribution network pipes. The main problem in developing simulation models for predicting the failure rate of water distribution network pipes is the lack of data or lack of access to the required and accurate data.

Sattar et al. (2016) focused on the prediction of pipe failure rate. They considered the pipe failure time to be just a function of the five parameters of diameter, length,

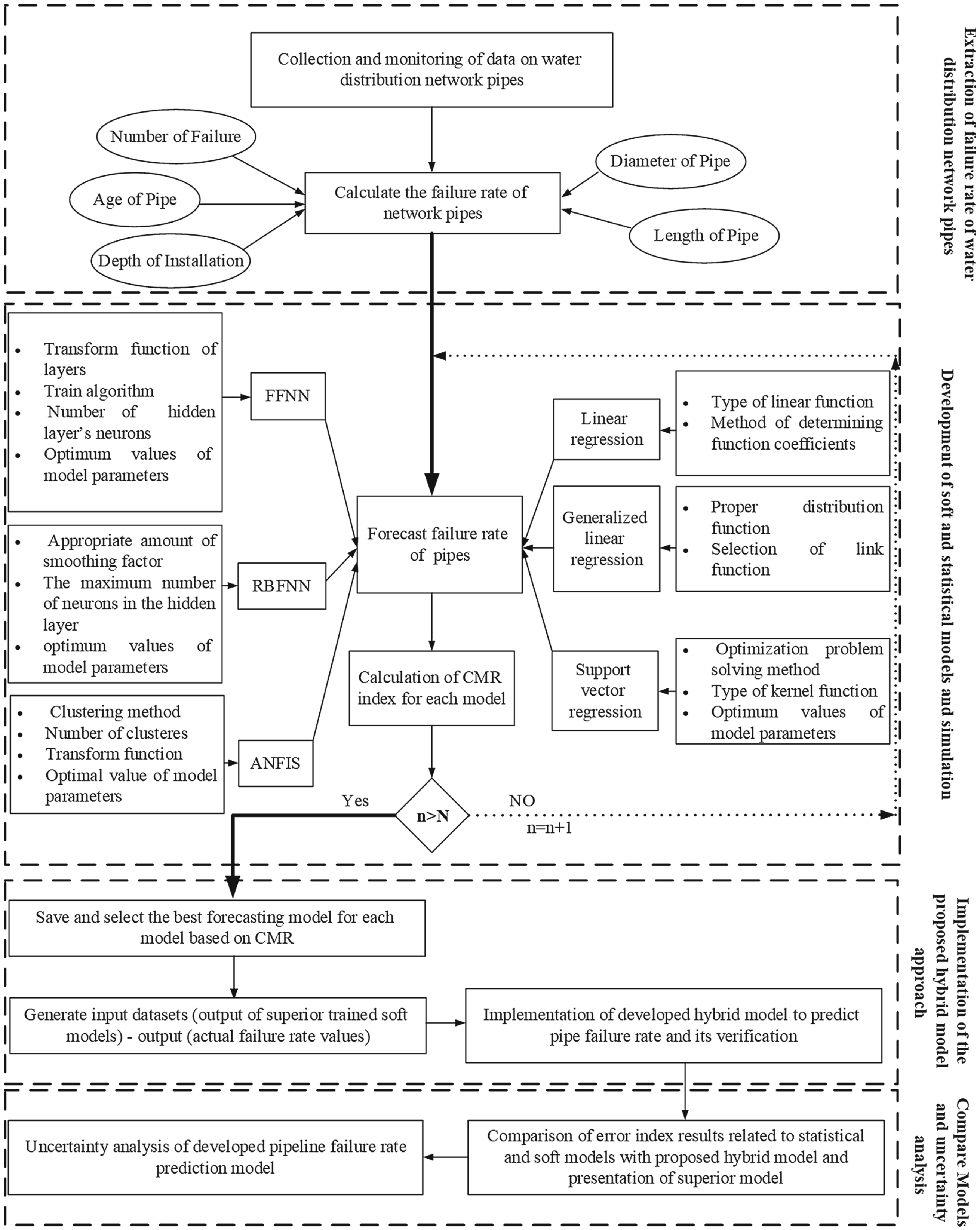
Extraction of failure rate of water distribution network pipes
Collection and monitoring of data on water distribution network pipes
Number of Failure
Age of Pipe
Depth of Installation
Calculate the failure rate of network pipes
Diameter of Pipe
Length of Pipe
Development of soft and statistical models and simulation
Transform function of layers
Train algorithm
Number of hidden layer's neurons
Optimum values of model parameters
FFNN
Appropriate amount of smoothing factor
The maximum number of neurons in the hidden layer
optimum values of model parameters
RBFNN
Clustering method
Number of clusteres
Transform function
Optimal value of model parameters
ANFIS
Forecast failure rate of pipes
Calculation of CMR index for each model
Linear regression
Type of linear function
Method of determining function coefficients
Generalized linear regression
Proper distribution function
Selection of link function
Support vector regression
Optimization problem solving method
Type of kernel function
Optimum values of model parameters
Yes
n>N
NO
n=n+1
Implementation of the proposed hybrid model approach
Save and select the best forecasting model for each model based on CMR
Generate input datasets (output of superior trained soft models) - output (actual failure rate values)
Implementation of developed hybrid model to predict pipe failure rate and its verification
Compare Models and uncertainty analysis
Uncertainty analysis of developed pipeline failure rate prediction model
Comparison of error index results related to statistical and soft models with proposed hybrid model and presentation of superior model

Fig. 2 Structure of the proposed approach to predict the pipe failure rates

number of failures, cathodic protection, and lining of each pipe. Aydogdu and Firat (2015) used three parameters length, diameter, and age of the pipe as the parameters affecting the failure rate of the network pipes in order to predict the failure rate of water distribution network pipes. The parameters of diameter, material, and age of the pipe were considered by Rogers (2011) and Wang et al. (2009) to estimate the risk of pipe failure and to predict pipe failure rate in the water distribution networks, respectively. Therefore, according to the available information related to the characteristics of the studied network and the effective parameters that proposed in previous research, this study considered the parameters of diameter (D), length (L), depth of installation (DI), and age (Ag) as independent variables affecting pipe failure. Then, the failure rate of each pipe (FR) is calculated (in terms of the number of failures per length unit in each year) with Eq. (1) and based on the data obtained from the number of failures of each pipe in the study area during the four-year period:

$$\mathrm{FR} = \frac{\text{Total number of failures during study period}}{\text{Length of pipe } \times \text{study period}} \tag{1}$$

3.2 Introducing the soft models used in hybrid model

In this study, various regression and soft models including Linear Regression Model (LR), Generalized Linear Regression Model (GLR), Support Vector Regression (SVR), Feedforward Neural Network (FFNN), Radial-Based Function Neural Network (RBFNN), and Adaptive Neuro-Fuzzy Inference System (ANFIS) were used to predict the failure rate of water distribution network pipes.

3.2.1 Linear regression model (LR)

In linear regression, the parameters of a linear model are estimated via an objective function and the value of the variables. The coefficients of independent variables are determined by the least square method. Due to the presence of extreme values in some observational data, in order to reduce the effects of extreme data and to search for the appropriate function simultaneously, the use of the balanced weight (BW) squares method to calculate robust least square (RLS) provides a better solution compared to other methods.

In this method, it is important to select the appropriate weight function from different weight functions such as Andrews, Bisquare, Cauchy, etc., to calculate the weight of each independent variable. The relationships related to this method are as follows:

$$w_i = \begin{cases} 1, & |u_i| < 1 \\ 0, & |u_i| \geq 1 \end{cases} \tag{2}$$

$$u_i = \frac{r_{\text{adj}}}{K \cdot S} \tag{3}$$

$$r_{\text{adj}} = r_i / \sqrt{1 - h_i} \tag{4}$$

where r_i: the calculated error using the least squares method, h_i: the leverage parameter. This parameter adjusts the error calculated by the square method by reducing the weight of data containing high h_i. K: compatibility constant. This parameter is determined by the type of weight function. S: smart variance. This parameter is equal to $\frac{\text{MAD}}{0.6745}$ (the MAD is the mean absolute deviation), u_i: the standardized error, w_i: the smart weight of each independent variable.

3.2.2 Generalized linear regression (GLR)

Linear regression models describe a linear relationship between a response and one or more predictive terms. Many times, however, a nonlinear relationship exists. Nonlinear regression describes general nonlinear models. A special class of nonlinear models, called generalized linear models, uses linear methods.

In generalized linear models, these characteristics are generalized as follows:

- At each set of values for the predictors, the response has a distribution that can be normal, binomial, Poisson, gamma, or inverse Gaussian, with parameters including a mean μ.
- A coefficient vector b defines a linear combination Xb of the predictors X.
- A link function f defines the model as $f(\mu) = Xb$

In this study, the probability distribution curve of the pipe failure rate was drawn to determine the type of distribution function governing the response or dependent variable (the failure rate of the water distribution network pipes). This shows that this variable is relatively well compatible with the Poisson distribution function. Its link function is considered also a Log function ($f(\mu) = \log(\mu)$).

3.2.3 Support vector regression model (SVR)

The SVR model is a version of the support vector machine (SVM) model that performs regression instead of data classification. In modeling a phenomenon using the SVR method for an observational set with N observations where x_n is the number of input parameters and y_n is the number of output parameter, the goal is to find the best linear function as follows:

$$f(x) = y_n = WX_n^T + b \tag{5}$$

To find the best function $f(x)$, the following optimization problem must be solved:

$$\begin{aligned}&\text{Minimize}\left\{\frac{1}{2}WW^T + C\sum_{n=1}^{N}(\xi_n + \xi_n^*)\right\}\\&\text{Constrain:}\\&y_n - (WX_n^T + b) \le \varepsilon + \xi_n\\&(WX_n^T + b) - y_n \le \varepsilon + \xi_n^*\\&\xi_n^*, \xi_n \ge 0\end{aligned} \tag{6}$$

where ε is the range of soft margin; ξ_n^* and ξ_n are slack variables; C is a positive number that controls the amount of penalty. It is obtained by using from Eq. (7) if the Gaussian kernel function is used. Otherwise, the best value is obtained through trial and error method and by using the values of error indices. The W is the optimal weight vector, and b is the optimum oblique vectors.

$$C = \frac{iqr(Y)}{1/349} \tag{7}$$

In Eq. (8), *iqr* is equal to the difference between the 75th and 25th quarter of the target vector. In order to solve the problem of linear and nonlinear optimization in SVR problems, in addition to the quadratic programming method, in cases where the number of observations is high, one of the decomposition methods, sequential minimal algorithm, and interactive single data algorithm can be used. In this study, to select the optimal parameter values of the SVR model, this model was developed in MATLAB2018b program and it was executed 100 times. Then, the optimal values of the model parameters were determined based on the CMSER error index.

3.2.4 Feedforward neural network (FFNN)

ANN is a relatively new method for soft computing. It is inspired by the structure of the human brain, and it models the brain synaptic connections and neural structure. With the advancement of artificial neural networks, various types of neural networks have been developed such as Adaline Network, Multilayer Perceptron Network, Feedforward Network, Radial-Based Function Neural Network, Regression-Based Network, etc. Each of them is suitable for specific applications. The FFNN network often has an input layer containing input variables, i.e., independent variables, and one or more hidden layers of sigmoid neurons. Each layer is receptive the output of the previous layer, and the last layer is the network output containing output variables or dependent variables with linear functions.

In order to create an FFNN network, it is necessary to determine the network characteristics such as network dimensions, the number of hidden layers, the number of neurons in each layer, the type of transfer functions, and network training methods. The transfer functions are used in the feedforward network such as Logsig, Tansig, and Purelin, which are selected depending on the data type and problem specifications. To train the FFNN network, various algorithms such as gradient descent, gradient descent with momentum, variable learning rate, conjugated gradient, and quasi-newton algorithms are used. Although determining the network specifications has a great impact on the efficiency of the network, there is no specific method to determine these specifications. Therefore, the most appropriate network dimensions can be obtained by comparing between the simulation results of different networks with real values.

The feedforward network of this research consists of a hidden layer with Tansig transfer function and an output layer with linear Purelin transfer function. To determine the best values of the FFNN model parameters, the developed model was executed 100 times, and then, the best values of the model parameters were determined based on the CMSER error index.

3.2.5 Radial-based function neural network

These types of networks are a type of monitored neural network that has a feedforward structure and consists of an input layer, a hidden layer, and an output layer. In this network, the number of neurons in the hidden layer is determined based on the input parameters. The main advantage of RBF networks is their hidden layer, which has nodes called RBF. In each RBF unit, there are factors that determine the position, deviation, or width of the center of functions. Equation (8) shows the transfer function for a radial basis neuron:

$$a = \text{radbase}(n) = e^{-n^2} \tag{8}$$

where n is the vector distance between its weight vector (w) and the input vector (P), multiplied by bias (b), and a is the output value of the RBF function, which provides a value between zero and one.

The output layer in the RBF network has a linear transfer function (purelin). The output of the RBF network is estimated through the following equations:

$$\text{output} = \sum_{j=1}^{M} \acute{W}_j F(D_j) + \acute{b} \tag{9}$$

$$F(D_j) = \exp\left[-(D_j b_j)^2\right] \tag{10}$$

$$D_j = \sqrt{\sum_{i=1}^{N} \left(w_{ij} - x_i\right)^2} \tag{11}$$

$$b_j = \frac{0.8326}{\sigma} \tag{12}$$

where $\acute{W_j}$ is the weight vector that connects the output layer to the hidden layer; $F\left(D_j\right)$ is the output of neuron j from hidden layer, which is considered as the transfer function of the hidden layer. σ is the smoothing factor; $\acute{b}$ is the bias vector related to the output layer; and M and N are the number of input and the number of neurons in the hidden layer, respectively.

To develop a Radial-Based Function Neural Network, it is necessary to determine the value of smoothing factor and the desired error index. The smoothing factor determines the width of an area in the input space to which each neuron responds. This factor should be large enough that neurons respond strongly to overlapping regions of the input space. However, smoothing factor should not be so large that each neuron is effectively responding in the same large area of the input space.

3.2.6 Adaptive neuro-fuzzy inference system (ANFIS)

Adaptive Neuro-Fuzzy Inference System is a type of artificial neural network based on Sugeno fuzzy inference system. This method was developed in the early 1990s. It combines the capabilities of an artificial neural network and a fuzzy inference system. An ANFIS model has five layers. The first layer captures the inputs and determines the membership functions of the inputs according to the defined rules. In the second layer, in each rule, the degrees of membership of the inputs are multiplied to determine the weight of rule. The normalization of the obtained weights from the previous stage is performed in the third layer using Eq. (13). In the fourth layer, depending on the weight of each rule and the corresponding output value, the weighted output is calculated using Eq. (15), and finally, the output of ANFIS model in the fifth layer is calculated based on Eq. (15).

$$\bar{w}_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{13}$$

$$\bar{w}_i \times f_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \times \left(a_i \cdot x_1 + b_i \cdot x_2 + \cdots + c_i\right) \tag{14}$$

$$y = \sum_{i=1}^{n} \bar{w}_i \times f_i \tag{15}$$

where i is counter of rules; n is the number of rules; w_i is the weight of each rule; $\bar{w}_i$ is the normalized weight of each rule; f_i is the output linear function of the Sugeno system; a_i, b_i and c_i are the constant coefficients of the output linear function; x and y vector are the input and output parameters, respectively.

To create an ANFIS model, the correct choice of parameters of membership functions and introducing appropriate fuzzy rules to the model is very important. For this purpose, the use of classification methods can be effective. There are various algorithms such as grid partitioning (GP), subtractive clustering (SC), and fuzzy clustering method (FCM) for data clustering. The most appropriate method is chosen depending on the type of problem. In addition, according to the structure of the ANFIS model, the appropriate training algorithm is chosen from two methods. The first method is back propagation that is based on the gradient descent of the total error squares, and the second method is the hybrid method which is a combination of the gradient descent method and the minimum error squares.

In this research, the network was trained by using a combination of gradient descent method and minimum error squares. The data were clustered by using FCM algorithm. This method works based on minimizing the following objective function:

$$J_m = \sum_{i=1}^{D} \sum_{j=1}^{N} \mu_{ij}^{m} \left\| x_i - c_j \right\|^2 \tag{16}$$

where D is the number of data point; N refers to the number of clusters; m is fuzzy partition matrix exponent for controlling the degree of fuzzy overlap; x_i is the ith data point; c_j is the center of the jth cluster; and μ_{ij} is the degree of membership of x_i in the jth cluster.

3.3 The proposed innovative hybrid model

Due to the complexity of some phenomena as well as the impact of various factors on that phenomenon, the performance of soft and statistical models in predicting a phenomenon always has limitations. These limitations lead to inconsistencies in modeling results using soft and statistical models with observational values. As a result, the accuracy of individual prediction models decreases. In addition, the hybrid models proposed by the researchers also have limitations such as the number of models used in the hybrid model and the non-use of statistical models.

In the present study, six different models including two statistical models and four smart soft models, which were introduced in the previous section, have been used to develop a innovative hybrid model. The use of different models in this research has eliminated the limitations and weaknesses of previous hybrid models. Therefore, the innovation of the proposed hybrid model of the present study is the use of more models (including six models), the use of statistical models, and the use of a combination of statistical and intelligent models in the hybrid model. The executive steps of hybrid developed model are as follows (Fig. 3):

- Gathering the urban water distribution network data including pipe diameter, pipe length, number of failures per pipe, pipe age, and pipe installation depth, as well as their division into three data sets entitled training, testing, and validation data.
- Training single models and predicting failure rate: using observational values related to independent and dependent variables, statistical and soft models were performed 100 times. The best prediction models were selected based on CMSER index values related to training and test data sets.
- Generation of input–output data sets for the innovative hybrid model: the failure rates values predicted based on the best selected single models from the previous step are considered as input variables to the hybrid model.
- Extract the parameters of the hybrid model: by running the hybrid model 100 times and comparing the MSE and R error indices of the training and testing stage, the model with the highest R and the lowest MSE was presented as the best hybrid model.

3.4 The model evaluation indices

There are various error indices in order to measure the accuracy and appropriateness of the fitted models. Due to the importance of R (for investigating the goodness of model data fit) and MSE (for examining the difference between observational and predicted values), among the available error indices, these indices were used in this study to evaluate the suitability and accuracy of the model in predicting the failure rate of water distribution network pipes. Equations (17) and (18) are used to measure R and MSE error indices, respectively.

$$R = \frac{\sum_{i=1}^{n}(x_{mi} - \bar{x}_m)(x_{ci} - \bar{x}_c)}{\sqrt{\sum_{i=1}^{n}(x_{ci} - \bar{x}_c)^2 \sum_{i=1}^{n}(x_{mi} - \bar{x}_m)^2}} \quad (17)$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(x_{mi} - x_{ci})^2}{n} \quad (18)$$

where x_{mi} is the observed value; x_{ci} is the predicted value; n is the number of observations data; $\bar{x}_m$ and $\bar{x}_c$ are the arithmetic mean of the observed and predicted values, respectively.

According to the definitions of the above indices, the lower value of MSE index and the higher value of R index lead to the higher accuracy of the prediction model. The value of R-indicator only shows that the behavioral status of the predicted values matches with the actual values. As a result, based on this error index, the magnitude of the error value, i.e., the difference between the prediction and observation values, cannot be examined. Therefore, it is necessary to be considered the MSE error index properly in measuring the accuracy of the prediction models.

In this research, the CMSER error index (Eq. (19)), which is defined by the ratio R to MSE, was proposed in

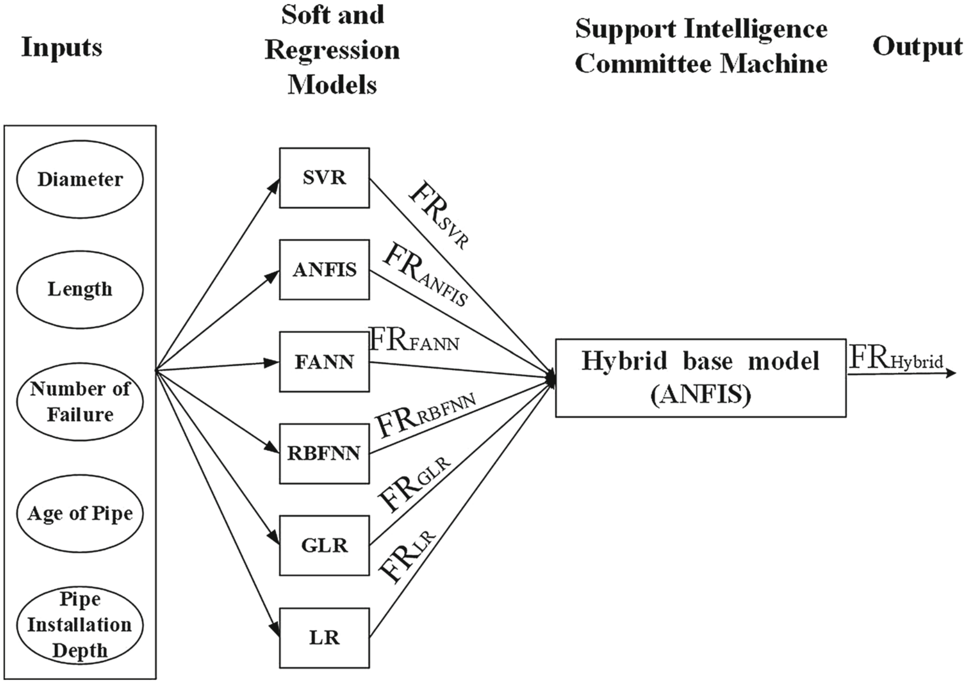

Fig. 3 Structure of the proposed hybrid model to predict pipe failure rates

order to simultaneous use of the benefits of the R and MSE error indicators.

$$\mathrm{CMSER} = \frac{R}{\mathrm{MSE}} \tag{19}$$

In this study, in order to measure the accuracy of different models and also to prevent overfitting, the observational data are divided into three sub-sets: training dataset (including 70 percent of data), testing dataset (including 15 percent of data), and validation dataset (including 15 percent of data). Therefore, according to the proposed approach, each model was re-iterated 100 times, and R, MSE, and CMSER error index values were estimated in each iteration in order to select the best values for the parameters of each model. Then, based on the sum of the CMSER index related to training and testing dataset of each iteration, the model that has the highest CMSER_i^m value is selected as the most appropriate prediction model.

$$\mathrm{CMSER}_i^m = \frac{R_i^m}{\mathrm{MSE}_i^m}, \quad i = 1, 2, \ldots, 100, \quad m = 1, 2, \ldots, M \tag{20}$$

where R_i^m and MSE_i^m are the correlation coefficient and root square error related to the ith iteration of the mth model, respectively.

After selecting the best models, the following three indices were used to evaluate the predictive power of the models based on the validation data (Tropsha et al. 2003):

a) The K or K' index, which indicates the slope of the regression line passing through the origin of the coordinates between the observed and predicted values or between the predicted and observed values. From the above two indices, at least one of the values of K or K' should be close to one.

$$K = \sum_{i=1}^{n} (x_{ci} \times x_{mi}) / x_{ci}^2 \tag{21}$$

$$K' = \sum_{i=1}^{n} (x_{ci} \times x_{mi}) / x_{mi}^2 \tag{22}$$

b) The m or n index, which indicates the coefficient of determination of the regression line between observed and predicted values, or vice versa. m and n values should be less than 0.1.

$$m = \left(R^2 - R_o^2\right)/R^2 \tag{23}$$

$$n = \left(R^2 - R_o^{\prime 2}\right)/R^2 \tag{24}$$

where

$$R_o^2 = 1 - \left(\sum_{i=1}^{n} x_{ci}^2 (1-K)^2 \Big/ \sum_{i=1}^{n} (x_{ci} - \bar{x}_c)^2\right) \tag{25}$$

$$R_o^{\prime 2} = 1 - \left(\sum_{i=1}^{n} x_{mi}^2 (1-K')^2 \Big/ \sum_{i=1}^{n} (x_{mi} - \bar{x}_m)^2\right) \tag{26}$$

c) The R_{m1} or R_{m2} indices, which indicate the cross-validation condition, should be greater than 0.5 and are defined as follows:

$$R_{m1} = R^2\left(1 - \sqrt{\left|R^2 - R_o^2\right|}\right) \tag{27}$$

$$R_{m1} = R^2\left(1 - \sqrt{\left|R^2 - R_o^{\prime 2}\right|}\right) \tag{28}$$

Given that each of the above indices shows a part of the model's ability to prediction, therefore, it is necessary to examine the above three indicators and simultaneously satisfy their conditions to demonstrate the ability of the prediction models (Golbraikh and Tropsha 2002).

3.5 Model uncertainty analysis

As previously mentioned, the failure rate of water distribution network pipes depends on various independent parameters such as length, diameter, material, installation depth, and number of previous failures. The pipe failure in different water distribution networks does not occur similarly and at a constant rate, and according to different effective parameter values, the occurrence of this phenomenon is different in miscellaneous networks.

The uncertainty of the independent input parameters of the model, such as length, installation depth, environmental conditions, and especially the number of failures of the network pipes, causes the uncertainty of the model output, i.e., the failure rate of network pipes. Therefore, it is necessary and important to measure the uncertainty of the failure rate predicted by the model for correct decision-making in the operation of the network.

There are several ways to determine the uncertainty of the prediction model output. In this study, the Monte Carlo Simulation method (MCS) is used. The MCS method is a numerical method for calculating output uncertainty, and it was developed by Ulma and Neuman in 1946 for military uses (Frey and Patil 2002). In this method, in order to measure the model output uncertainty, it is necessary to know the uncertainty of the input variables or, in other words, the type of probability density function of the input parameters.

The model must be run several times before preforming the statistical analysis using MCS method. Each time the model is run, a definite output is obtained. Then, by using the definite outputs, the probability density function of the model output and its uncertainty are estimated. The model output uncertainty is measured via using Eq. (29) (Walker 1931):

$$\text{Uncertainty}\% = \frac{100 \times \text{MAD}}{\text{median}(P)} \tag{29}$$

where P is the mean values of the model output and MAD is the mean absolute deviation, which is estimated from Eq. (30):

$$\text{MAD} = \frac{1}{J}\sum_{i=1}^{J}|P_i - \text{median}(P)| \tag{30}$$

where J is the number of times the model was run.

Equation (31) estimates the model output uncertainty, which is a result of the uncertainties of the independent input variables. The contribution of each input variable to the model output uncertainty is different and can be calculated through least square linearization method (Verbeeck et al. 2006). Equation (31) is the general form of the equation used in this method. This equation is a multiple regression between the deviation of each input parameter from the mean value of that parameter and the output of the model.

$$y = w_1\Delta v_1 + w_2\Delta v_2 + \cdots + w_i\Delta v_i + b \tag{31}$$

where v_i is the model input parameter; y is the model output; Δv_i is the difference of each random sample of input variable i (i.e., v_i) with the mean value of all random samples of parameter i (i.e., m_{v_i}); w_i is the regression coefficient between the calculated model output and the input parameters; and b is the regression constant coefficient.

Using multiple regression analysis, the regression coefficients are determined between the calculated model output (pipes failure rate) and the input variables. Thus, the sensitivity coefficients of each input variable i can be obtained by Eq. (32):

$$S_{v_i} = 100 \times \frac{w_i^2 \cdot \sigma_{\Delta v_i}^2}{\sum_{i=1}^{n} w_i^2 \cdot \sigma_{\Delta v_i}^2} \tag{32}$$

where $\sigma_{\Delta v_i}$ is the variance of Δv_i and n is the number of the model input variables.

4 Results and discussion

In this section, the optimal values of each prediction model parameters in order to make the appropriate model, and predict the failure rate of water distribution network pipes are presented. Then, using the proposed approach, the results of the hybrid model were investigated in comparison with individual soft models. Finally, the uncertainty of the superior model was discussed in detail.

4.1 Linear regression model (LR)

According to the explanations provided in the linear regression section, multiple linear regression method with RLS method and dual weighted squares were used in this study in order to calculate the minimum prediction error value. In the first step of linear regression modeling, the appropriate fitness function was selected from the following functions: linear, quadratic, pure quadratic, and interaction functions. This selection was performed by running the multiple linear regression model for different functions and by calculating the value of the CMSER index for each of the functions. As shown in Fig. 4, the best function was the pure quadratic function with the highest value of the CMSER index.

In order to select the type of weight function in the dual weighted squares method, the linear regression model of this study was performed for different weight functions, and then, the values of CMSER error index were calculated. Finally, as shown in Fig. 5, the Cauchy weight function was selected as the best weight function of the linear regression model due to having the highest value of CMSER index.

Based on the results obtained from the LR model, the multiple linear regression model was developed and performed to predict the failure rate of network pipes with the pure quadratic fitness function and Cauchy weight function. The MSE and R values for the validation data were obtained equal to 0.0863 and 0.51, respectively.

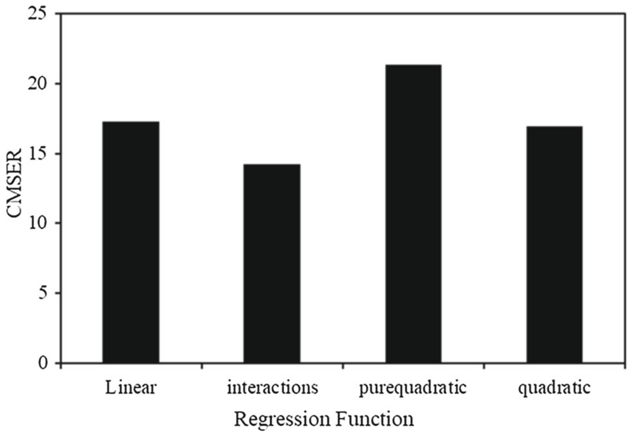

Fig. 4 CMSER error index values for different fitness functions in multiple linear regression model

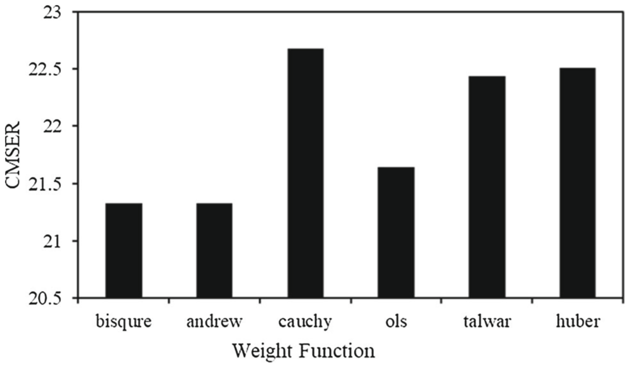

Fig. 5 CMSER index values of multiple linear regression model with different weight functions

4.2 Generalized linear regression model (GLR)

There are three sections in the structure of a generalized linear regression model. The first section involves selecting the type of distribution function governing the independent variable. According to the failure rate date recorded for the water distribution network in this study, the Poisson probability distribution function is considered. Therefore, Log function is the suitable as link function for the Poisson distribution function.

According to the Poisson probability distribution function and the Log link function, the generalized linear model for predicting the water distribution network pipe failure rate was developed. The MSE and R error index of the validation data were obtained equal to 0.0897 and 0.35, respectively.

4.3 Support vector regression model (SVR)

In order to establish the best SVR model, it is necessary to determine various parameters such as C coefficient, kernel function type, kernel function scale, ε value. Due to the multiplicity of model parameters, choosing the best values for the mentioned parameters is only possible using the optimization method. For this purpose, the SVR model was prepared and implemented for different optimization methods and kernel function. Then, CMSER index values were calculated for each situation. Based on CMSER error index, the best optimization method in the SVR model is the LIQP method with the Gaussian kernel function (Fig. 6). The optimal values of the other parameters of the SVR model are shown in Table 2.

4.4 Feedforward neural network (FFNN)

To select the appropriate training algorithm, the feedforward neural network with a one hidden layer, the number of different neurons, and ten training algorithms were created in the MATLAB-R2018b. The value of the MSE and R indicators related to the training and testing dataset was calculated for 100 times the execution of each of the different structures of the FFNN, and the best model of each structure was extracted based on the CMSER index.

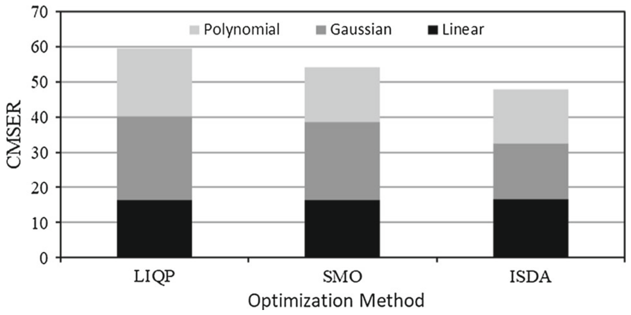

Fig. 6 CMSER index value for different optimization methods and Kernel functions

The results show that among the studied training algorithms, the algorithm called Trainlm leads to the creation of a FFNN model with the highest value of the CMSER index. Based on this algorithm, the number of neurons suitable for the hidden layer of the selected structure is selected to be 20 (Fig. 7).

Therefore, the specifications of the trained FFNN are: an input layer with 4 input variables (diameter, length, depth of installation, and age), a hidden layer with 20 neurons with tansig transfer function, an output layer with a dependent variable (failure rate of each pipe) and purelin linear transfer function, and the type of algorithm used to train the FFNN model being the Trainglm algorithm.

It should be noted that the initial weight and oblique vectors are considered the same for different structures of the FFNN model to compare the results of the models and have the same initial conditions. The results of MSE and R error index of the validation data for the best structure were obtained equal to 0.060 and 0.69, respectively.

4.5 Radial-based function neural network (RBFNN)

Two parameters of smoothing factor and desired error index should be determined in the construction of RBF neural network model in order to stop the training process. In this algorithm, two conditions are considered for stopping the training process, which include: minimum error, which is considered zero, and the number of hidden layer neurons. The maximum allowable number of neurons in the hidden layer is equal to the number of observation dataset.

To determine the most appropriate smoothing factor, the RBF neural network model was developed and run in the MATLABR2018b for different values of smoothing factor

Table 2 Optimal parameters of LIQP method in SVR model

Optimization method	Kernel function	Kernel Scale	Standard	Epsilon(σ)	Bias(b)	BoxCo. (C)
LIQP	Gaussian	0.0023	Yes	0.005	0.014	309.98

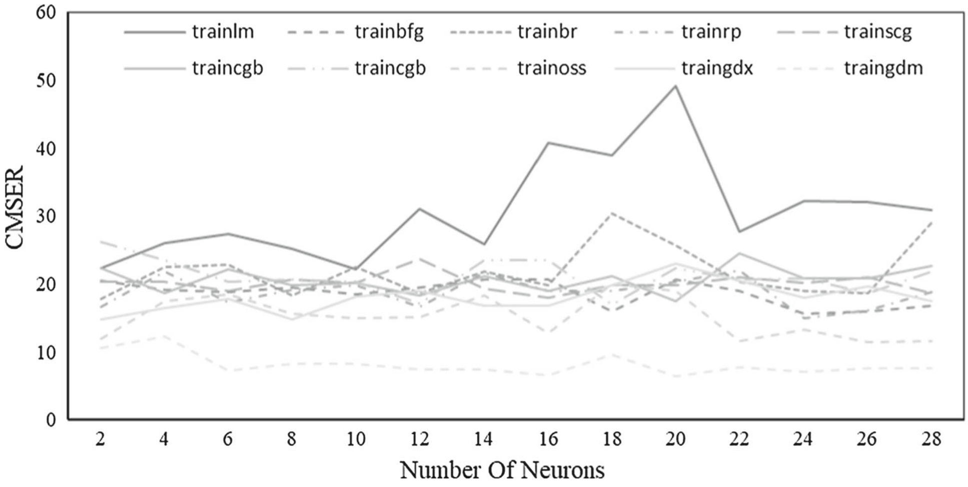

Fig. 7 Comparison of the results of different training algorithms per different number of hidden layer neurons based on the CMSER index

(1-100) and number of neurons. Studies in this research show that if the smoothing factor is chosen to be greater than 49, overfitting occurs. Therefore, to prevent overfitting, the range of smoothing factor was considered to be between 1 and 49.

The value of the CMSER index obtained for different values of the smoothing factor and the number of hidden layer neurons shows that the maximum values of the CMSER index occur in the smoothing factor values between 6 and 18 (Fig. 8).

Investigation of CMSER error index versus the number of different hidden layer neurons for smoothing factor between 6 and 18 shows that with the increasing number of neurons, the CMSER index values increase and MSE related to training dataset decreases. This analysis is not true for testing dataset and for the number of neurons over 9, and the CMSER error index for the testing dataset is reduced. This means that there is an overfitting in the process of model training. Therefore, the range of the number of hidden layer neurons for model training in this study is considered to be between 2 and 9 (Fig. 9).

Then, the structure of RBF neural network model was prepared and implemented with a smoothing factor between 6 and 18 for different amounts of hidden layer neurons between 1 and 9. The result of trained RBFNN model indicates that the value of R and MSE error indices of the validation was obtained equal to 0.276 and 0.0931, respectively. Therefore, the specifications of the RBFNN model for predicting the failure rate of water distribution network pipes in this research have five input parameters, a hidden layer with a maximum of 9 neurons and a Radbase transfer function and a smoothing factor equal to 10 and an output layer with one output and purelin linear transfer function.

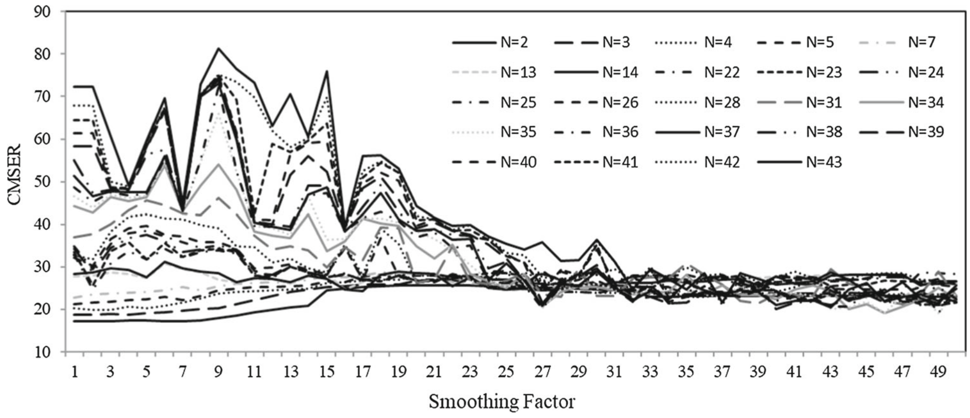

Fig. 8 Variation in CMSER index in comparison with different values of the smoothing factor and the number of neurons in the hidden layer

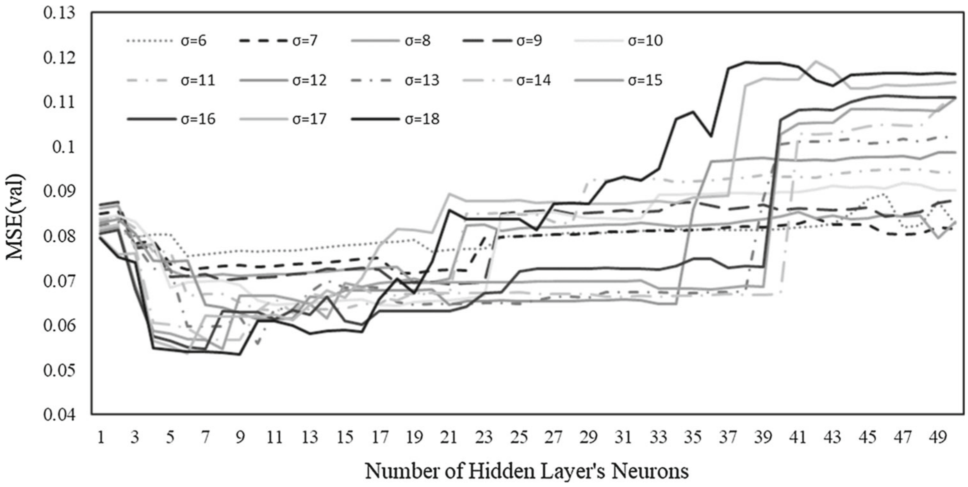

Fig. 9 Variation in MSE error index versus the number of hidden layer neurons for smoothing factor values between 6 and 18

4.6 Adaptive neuro-fuzzy inference system (ANFIS)

By running 100 times of ANFIS model, it can be found that training by using a combination of gradient descent and error squares provides the highest value of the CMSER index. Therefore, the hybrid training method was used to train the ANFIS model. In addition, to select the appropriate clustering method, the selected ANFIS model was iterated 100 times for each of the network grid partitioning, subtractive clustering, and fuzzy clustering. The CMSER index values for each iteration were estimated. The results showed that the fuzzy clustering method had the highest value of CMSER index; thus, it was selected as the appropriate method for data clustering.

In the fuzzy clustering method, it is very important to determine the number of clusters and the overlap of the clusters to create a matrix for cluster separation. After 100 iterations of ANFIS model for different the number of clusters and the power of the matrix, and the estimation of CMSER index, seven clusters were chosen as the best number of clusters, and power equal to 10 was considered as the best power for ANFIS model. After selecting the optimal values of the ANFIS model parameters, the failure rate of the studied network pipes for the validation data was predicted, and the values of R and MSE error indices were estimated as 0.92 and 0.071, respectively.

4.7 Proposed hybrid prediction model

Based on the first and second steps described in the proposed hybrid model, the gathering of water distribution network datasets, as well as training and development of the best model related to each of the soft and statistical models, was carried out.

Using the data collected in the first stage, including the values of the input parameters (diameter, length, installation depth, age, and number of pipe failures), and the best developed models in the second stage, including FFNN, RBFN, SVR, GLR, LR, and ANFIS, in the third step, the values of the failure rate predicted by mentioned six soft models were considered as input data to hybrid model.

The basic model of hybrid prediction model was chosen based on the comparison of CMSER index of the predicted values of pipe failure rate and based on the best soft and regression models. In this study, ANFIS soft model had the highest CMSER index value (12.95) of the validation data and it has been selected as the base model of the hybrid prediction model.

In the fourth step, based on the results of the previous step, the input datasets of the hybrid model were divided into three sub-sets: training, testing, and validation. Then, the optimal parameters of the hybrid model were obtained using 100 times hybrid model execution and comparison of the CMSER error index of the training and testing dataset in different executions. Finally, the MSE and R error indices related to validation data of the hybrid model are calculated and compared with the result of singular soft models. The values of CMSER error index, related to the validation data of the hybrid prediction model, were estimated to be 5.61. Comparing the results of the proposed model with the results of the best selected soft model (i.e., the ANFIS model) shows that the use of the proposed approach can reduce the prediction error of the failure rate of network pipes up to 56.7%.

4.8 Uncertainty analysis of pipeline failure rate prediction model

To evaluate the uncertainty of the developed prediction model, it is first necessary to determine the probability density function (PDF) of each of the input parameters to the model. In order to choose the best PDF, various PDFs were fitted for the observation data of the studied water distribution network. The best PDFs were obtained for the length, diameter, age, and depth of installation of the pipes parameters as follows: Gaussian, log-logistic, normal and log-logistic functions.

After selecting the PDFs for each of the input parameters, for each iteration of MCS method, a random sample of each input parameter value including length, diameter, age, and depth of installation was selected and given to the best soft model for predicting pipe failure rate. Finally, the failure rate was determined based on the values predicted by the six soft models (which are considered as inputs to the hybrid model). In order to determine the appropriate number of simulation iterations in the MCS method, the output variance (i.e., the rate of pipe failure) was estimated in each simulation (Verbeeck et al. 2006). As shown in Fig. 10, the output variance of the model is converged for over 100,000 times simulations. In other words, for the number of simulations over 100,000 times, the variance is constant. Therefore, increasing the number of simulations more than this number does not have any effect on the resulting uncertainty of the model. So, in this study the number of simulations in MCS procedures is considered to be 100,000.

According to the values of random samples in each of the input variables, the MAD value for the hybrid prediction model is equal to 2.72. This shows an uncertainty of 43.59% of the hybrid model for predicting the failure rate of distribution network pipes.

Estimating the contribution of each input variable to the uncertainty of the predicted failure rate shows that pipe length had the largest effect (with a contribution of 71%) on the pipe failure rate uncertainty. After that, the age of pipe, diameter, and installation depth had the next largest impact with a contribution of 24%, 4.5%, and 0.5%, respectively.

4.9 Evaluation of the accuracy and performance of the developed prediction models

This section discusses the accuracy and performance of various developed models to predict the pipe failure rate. The models were compared and analyzed in terms of the error indices mentioned in the previous sections. The values of MSE, R, Nash Sutcliffe (NS), mean absolute error (MAE) for each developed prediction model and for three datasets (training, testing and validation) are calculated and presented in Tables 3, 4 and 5.

The validation data are a part of the observation data, which did not participate in the training process. Therefore, in order to select the best prediction model, the error indices of the validation data in different models are compared with each other. The comparison between the error indices of the hybrid model and six statistical and soft models shows that the use of the hybrid model increases the R index rate by 8.1% (compared to the ANFIS model) to 260% (Compared to the RBFNN model). It also decreases the MSE index rate by 45% (compared to the ANFIS model) to 58% (compared to the RBFNN model). As a result, the use of hybrid model improved the performance and accuracy of the pipe failure rate prediction. The CMSER error index was used to consider the simultaneous effect of R and MSE indices in order to select the superior model. As shown in Fig. 11, the developed hybrid model has the highest CMSER index (25.51) compared to studied soft and regression (statistical) models.

NS error index is used as a criterion for evaluating the predictive power of models which can range from $-\infty$ to one. An efficiency of 1 (NS=1) corresponds to a perfect match of predicted to the observed data. As shown in Table 5, the hybrid model had the maximum value of the NS index of the validation data (equal to 0.985). In addition, the MAE error index is a measure of average absolute deviation between the predicted and observed values. The

Fig. 10 Variation in output variance of the hybrid model for different simulations

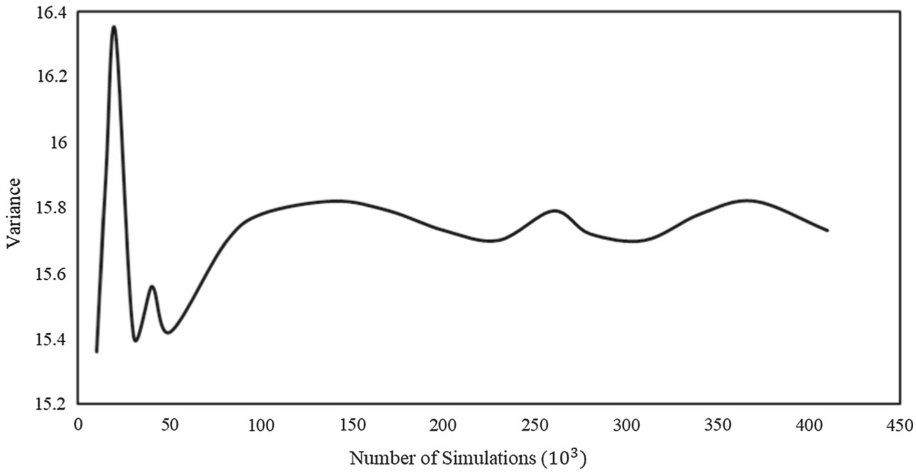

Table 3 Comparison of error indices of different soft models in the training stage

Model/error index	MSE	*R*	NS	MAE
LR	0.0907	0.49	− 1.917	0.0045
GLR	0.0868	0.59	− 1.32	0.0044
SVR	0.0921	0.55	0.13	0.0036
FFNN	0.0826	0.68	0.46	0.202
RBFNN	0.0624	0.91	0.788	0.0024
ANFIS	0.037	0.98	0.976	0.00088
Hybrid	0.035	0.974	0.987	0.035

Table 4 Comparison of error indices of different soft models in the testing stage

Model/error index	MSE	*R*	NS	MAE
LR	0.0597	0.71	0.493	0.0027
GLR	0.56	0.75	0.407	0.0024
SVR	0.067	0.93	0.58	0.003
FFNN	0.055	0.84	0.417	0.031
RBFNN	0.077	0.49	0.23	0.0045
ANFIS	0.078	0.62	0.098	0.0052
Hybrid	0.033	0.993	0.97	0.001

Table 5 Comparison of error indices of different soft models in the validation stage

Model/error index	MSE	*R*	NS	MAE
LR	0.0863	0.51	− 1.77	0.005
GLR	0.0897	0.35	− 3.53	0.0057
SVR	0.066	0.62	0.43	0.0026
FFNN	0.062	0.69	0.44	0.0453
RBFNN	0.0931	0.276	− 12.67	0.0051
ANFIS	0.071	0.92	0.732	0.0044
Hybrid	0.039	0.995	0.985	0.0012

lower value of this index means higher precision and accuracy of the model. According to Table 5, the hybrid model has the lowest value of this index for the validation data (equal to 0.0012).

In addition to the indices presented in Tables 3, 4 and 5 and Fig. 11, in order to compare the prediction power of statistical and soft models to estimate the failure rate of water distribution network pipes, the values of the validation criteria that recommended by Tropsha et al. (2003) for validation data were estimated (Table 6).

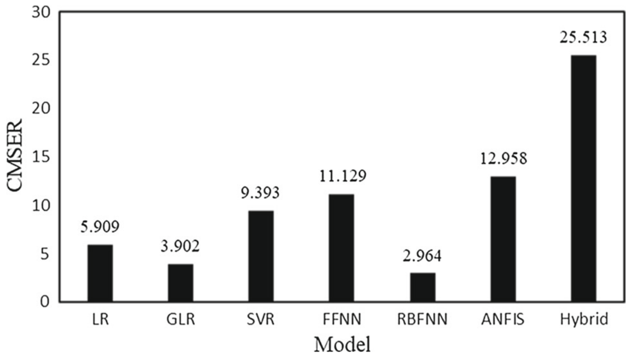

Fig. 11 Comparison of validation data CMSER error index for different developed soft models

Based on the description provided in Sect. 3.4, developed soft models are considered valid for prediction the failure rate of pipes if they satisfy some or all of the required conditions. As observed in Table 5, only the hybrid model satisfies all of the related validation criteria; thus, proposed hybrid model has strong prediction power and was not random correlations.

The above error indices only show the mean value of the errors of the predicted models and cannot reveal the accurate assessment of the distribution of errors. Therefore, for a more efficient evaluation of the models, cumulative distribution functions (CDF) of the simulated and observed values of pipe failure rates of various soft models in validation stage were used (Tabari and Zarif Sanayei (2019); Yoon et al. (2011)). As shown in Figs. 12, 13, 14, 15, 16, 17 and 18, the lowest deviation of the predicted values from the observed values was related to the hybrid model. This shows the accuracy, performance, and more reliability of this model compared to the other soft models. The CDF deviation of ANFIS model was almost equal and slightly higher than the hybrid model, which shows the great similarity of CDFs of these two models. The highest deviation of CDFs is related to the SVR model.

5 Conclusion

In order to eliminate the limitations, deficiencies and thus improve the predicting accuracy of pipe failure rate in the water distribution network of this research, the new hybrid model with combination of six different models (four soft models and two statistical models) including Adaptive Neural Fuzzy Inference System (ANFIS), Fit Forward Neural Network (FFNN), Radial-Based Function Neural Network (RBFNN), Support Vector Regression (SVR), Generalized Linear Regression (GLR) and Polynomial Linear Regression (LR) models was developed. The development process of the proposed hybrid model consists

Table 6 External validation statistical measures for developed soft models based on the validation dataset

Model/error index	R_{m2}	R_{m1}	n	m	K'	K
LR	0.264	0.267	− 0.197	− 0.184	1.27	0.62
GLR	0.219	0.44	0.614	− 0.0132	0.95	0.84
SVR	0.43	0.59	0.23	0.057	1.53	0.59
FFNN	0.58	0.66	− 0.0025	− 0.035	1.046	0.88
RBFNN	0.0187	0.0288	0.99	− 6.23	1.43	0.37
ANFIS	0.52	0.59	0.10	0.29	0.982	0.838
Hybrid	0.918	0.929	0.0038	0.0025	0.96	1.032

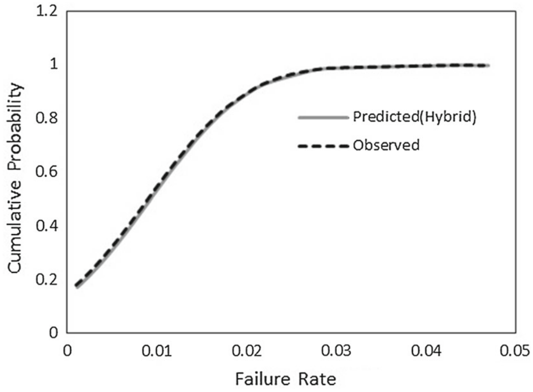

Fig. 12 Cumulative distribution function of the observed and predicted pipe failure rates using hybrid model

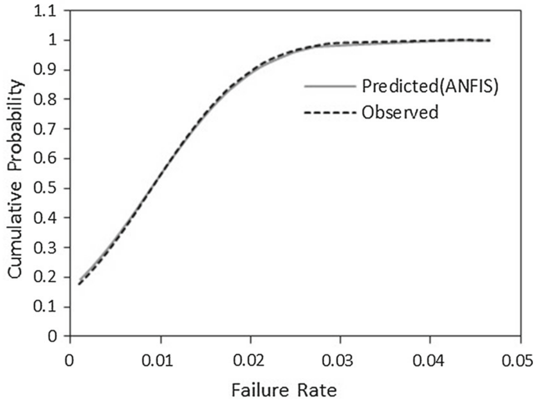

Fig. 13 Cumulative distribution function of the observed and predicted pipe failure rates using ANFIS model

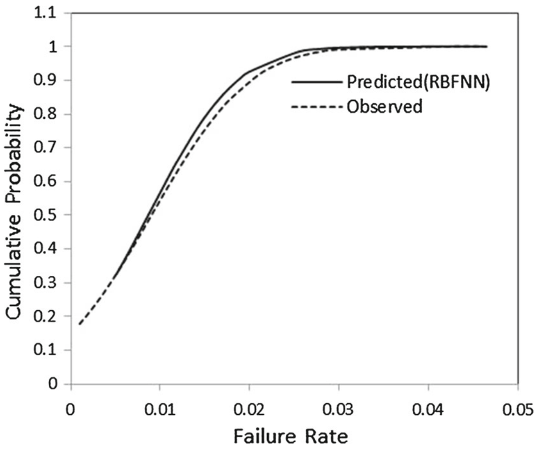

Fig. 14 Cumulative distribution function of the observed and predicted pipe failure rates using RBFNN model

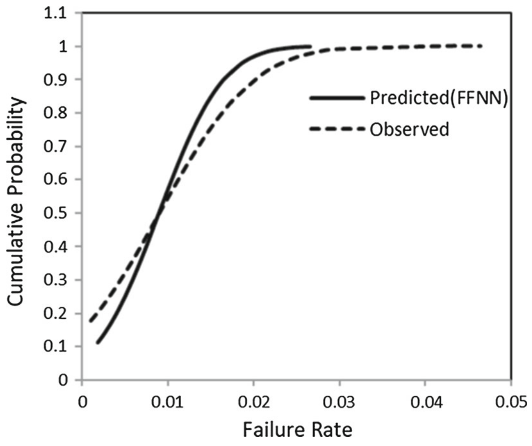

Fig. 15 Cumulative distribution function of the observed and predicted pipe failure rates using FFNN model

of three parts. First, the failure rate values are extracted based on the characteristics of network pipes such as diameter, length, age, installation depth, and number of previous failures. Then, the best model is determined based on the optimal parameters values of each soft model and the values of the predicted failure rate. Finally, the hybrid model is developed and implemented by combining the results of superior soft models as input.

In order to study the accuracy, performance, and prediction power of proposed hybrid model and compare with different developed models, in this research, different error indices and validity criteria were used. These error indices examine the accuracy and ability of the soft models from two perspectives, the average value of the errors and distribution of errors. In terms of mean error, the new hybrid model increased the R index from 8.1% (compared to the ANFIS model) to 260% (compared to the RBFNN model), and it reduced the MSE error by 37%-45% compared to other soft models. Therefore, the hybrid model had the highest value of the CMSER index, which simultaneously

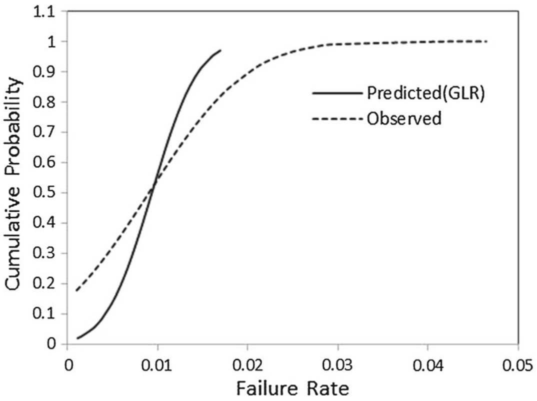

Fig. 16 Cumulative distribution function of the observed and predicted pipe failure rates using GLR model

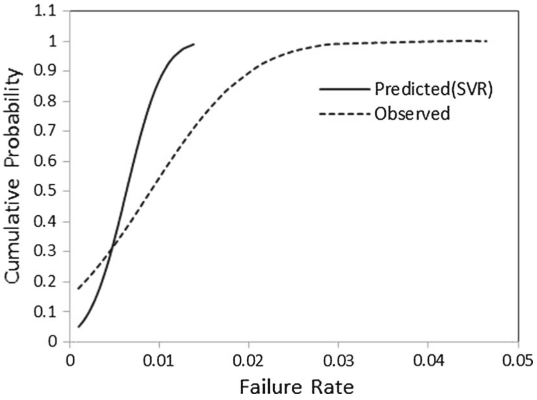

Fig. 17 Cumulative distribution function of the observed and predicted pipe failure rates using SVR model

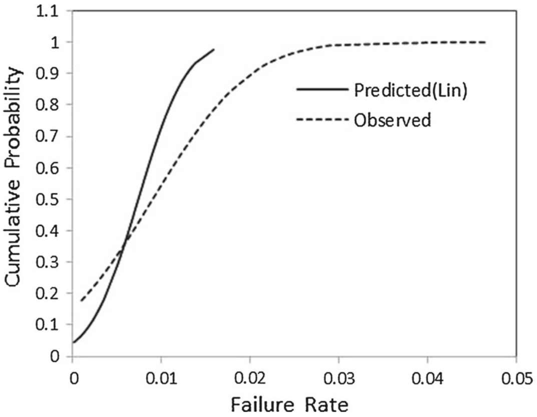

Fig. 18 Cumulative distribution function of the observed and predicted pipe failure rates using LR model

shows the effect of two indicators R and MSE. In addition, based on validation criteria, the hybrid model simultaneous satisfies all of the related validation criteria. Therefore, in terms of mean error, the proposed hybrid model has more accuracy and ability to predict the pipe failure rate compared to other soft models.

From the point of view of distribution error, the CDF of the values predicted by the new developed hybrid model was more consistent with the CDFs of observed values. Therefore, in terms of distribution error, the hybrid model had a better performance and accuracy compared to other soft models. Therefore, due to the appropriate performance and capability of the hybrid model in this research, this model has been selected as the most appropriate model and can detect the pipe failure rate of the urban water distribution network with a very high accuracy.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Asnaashari A, McBean EA, Gharabaghi B, Tutt D (2013) Forecasting watermain failure using artificial neural network modelling. Can Water Resour J 38(1):24–33

Aydogdu M, Firat M (2015) Estimation of failure rate in water distribution network using fuzzy clustering and LS-SVM methods. Water Resour Manag 29(5):1575–1590

Barton NA, Farewell TS, Hallett SH, Acland TF (2019) Improving pipe failure predictions: factors effecting pipe failure in drinking water networks. Water Res 164:114926. https://doi.org/10.1016/j.watres.2019.114926

Berardi L, Giustolisi O, Kapelan Z, Savic DA (2008) Development of pipe deterioration models for water distribution systems using EPR. J Hydroinf 10(2):113–126

Fares H, Zayed T (2010) Hierarchical fuzzy expert system for risk of failure of water mains. J Pipeline Syst Eng Pract 1(1):53–62

Faris Hamdala K, Sagar GY (2016) Statistical analysis of pipe breaks in water distribution systems in Ethiopia, the case of Hawassa. IOSR J Math 12(3):127–136

Farmani R, Kakoudakis K, Behzadian Moghadam K, Butler D (2017) Pipe failure prediction in water distribution systems considering static and dynamic factors. Proc Eng 186:117–126

Frey HC, Patil SR (2002) Identification and review of sensitivity analysis methods. Risk Anal 22:553–578

Gasemnezhad S, Motiee H, Moosavi Nodoushan MS (2014) Prediction of damage rate of urban drinking water network pipes by using and developing statistical models. In: Iranian water and sewerage science and engineering congress (**in Persian**)

Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20(4):269–276

Harvey R, McBean EA, Gharabaghi B (2013) Predicting the timing of water main failure using artificial neural networks. J Water Resour Plan Manag 140(4):425–434

Ho CI, Lin MD, Lo SL (2010) Use of a GIS-based hybrid artificial neural network to prioritize the order of pipe replacement in a water distribution network. Environ Monit Assess 166(1–4):177–189

Islam MS, Sadiq R, Rodriguez MJ, Francisque A, Najjaran H, Hoorfar M (2011) Leakage detection and location in water distribution systems using a fuzzy-based methodology. Urban water J 8(6):351–365

Jafar R, Shahrour I, Juran I (2010) Application of Artificial Neural Networks (ANN) to model the failure of urban water mains. Math Comput Model 51(9–10):1170–1180

Kakoudakis K, Behzadian K, Farmani R, Butler D (2017) Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering. Urban Water J 14(7):737–742

Kapelan ZS, Savic DA, Walters GA (2003) A hybrid inverse transient model for leakage detection and roughness calibration in pipe networks. J Hydraul Res 41(5):481–492

Kerwin S, de Soto BG, Adey BT (2019) January. Performance comparison for pipe failure prediction using artificial neural networks. In 6th international symposium on life-cycle civil engineering, IALCCE 2018. CRC Press/Balkema, pp 1337–1342

Kleiner Y, Rajani B (2002) Forecasting variations and trends in water-main breaks. J Infrastruct Syst 8(4):122–131

Malekpour MM, Tabari MMR (2020) Implementation of supervised intelligence committee machine method for monthly water level prediction. Arab J Geosci 13(19):1–14. https://doi.org/10.1007/876s12517-020-06034-x

Mounce SR, Day AJ, Wood AS, Khan A, Widdop PD, Machell J (2002) A neural network approach to burst detection. Water Sci Technol 45(4–5):237–246

Pandey P, Dongre S, Gupta R (2020) Probabilistic and fuzzy approaches for uncertainty consideration in water distribution networks: a review. Water Supply 20(1):13–27

Rajani B, Kleiner Y (2001) Comprehensive review of structural deterioration of water mains: physically based models. Urban Water 3(3):151–164

Rajeev P, Kodikara J, Robert D, Zeman P, Rajani B (2014) Factors contributing to large diameter water pipe failure. Water Asset Manag Int 10(3):9–14

Robles-Velasco A, Cortés P, Muñuzuri J, Onieva L (2020) Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliab Eng Syst Saf 196:106754. https://doi.org/10.1016/j.ress.2019.106754

Rogers PD (2011) Prioritizing water main renewals: case study of the Denver water system. J Pipeline Syst Eng Pract 2(3):73–81

Sacluti FR (1999) Modelling water distribution pipe failures using artificial neural networks. MSc. thesis, Department of Civil and Envir. Eng. University of Alberta, Canada

Sadiq R, Kleiner Y, Rajani B (2007) Water quality failures in distribution networks risk analysis using fuzzy logic and evidential reasoning. Risk Anal Int J 27(5):1381–1394

Sattar AM, Gharabaghi B (2015) Gene expression models for prediction of longitudinal dispersion coefficient in streams. J Hydrol 524:587–596

Sattar AM, Gharabaghi B, McBean EA (2016) Prediction of timing of watermain failure using gene expression models. Water Resour Manag 30(5):1635–1651

Sattar AM, Ertuğrul ÖF, Gharabaghi B, McBean EA, Cao J (2019) Extreme learning machine model for water network management. Neural Comput Appl 31(1):157–169

Shamir U, Howard CD (1979) An analytic approach to scheduling pipe replacement. J Am Water Works Assoc 71(5):248–258

Shin H, Kobayashi K, Koo J, Do M (2015) Estimating burst probability of water pipelines with a competing hazard model. J Hydroinf 18(1):126–135

Soltani J, Tabari MMR (2012) Determination of effective parameters in pipe failure rate in water distribution system using the combination of artificial neural networks and genetic algorithm. J Water Wastewater 23(83):2–15 **(In Persian)**

Soltanjalili M, Bozorg-Haddad O, Mariño MA (2011) Effect of breakage level one in design of water distribution networks. Water Resour Manag 25(1):311–337

Tabari MMR, Malekpour Shahraki MM (2018) Reservoir water level prediction using supervised intelligent committee machine method, case study: Karaj Amirkabir Dam. Iran Water Resour Res 14(5):15–30 **(in Persian)**

Tabari MMR, Zarif Sanayei HR (2019) Prediction of the intermediate block displacement of the dam crest using artificial neural network and support vector regression models. Soft Comput 23(19):9629–9645

Tabari MMR, Azari T, Dehghan V (2020) A supervised committee neural network for the determination of aquifer parameters: a case study of Katasbes aquifer in Shiraz plain, Iran. Soft Comput. https://doi.org/10.1007/s00500-020-05487-2

Tabesh M, Soltani J, Farmani R, Savic D (2009) Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. J Hydroinf 11(1):1–17

Tavakoli R, Najafi M, Sharifara A (2019) Artificial neural networks and adaptive neuro-fuzzy models for prediction of remaining useful life. arXiv:1909.02115

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22(1):69–77

Tu MY, Tsai FTC, Yeh WWG (2005) Optimization of water distribution and water quality by hybrid genetic algorithm. J Water Resour Plan Manag 131(6):431–440

Valis KPU (2013) Application of fuzzy logic for failure risk assessment in water supply system management. CEST

Verbeeck H, Samson R, Verdonck F, Raoul L (2006) Parameter sensitivity and uncertainty of the forest carbon flux model FOUG: a Monte Carlo analysis. Tree Physiol 26:807–817

Walker H (1931) Studies in the History of the Statistical Method. Williams and Wilkins Co., Baltimore, pp 24–25

Wang Y, Zayed T, Moselhi O (2009) Prediction models for annual break rates of water mains. J Perform Constr Facil 23(1):47–54

Xu Q, Chen Q, Li W, Ma J (2011) Pipe break prediction based on evolutionary data-driven methods with brief recorded data. Reliab Eng Syst Saf 96(8):942–948

Xu Q, Chen Q, Ma J, Blanckaert K (2013) Optimal pipe replacement strategy based on break rate prediction through genetic programming for water distribution network. J Hydro Environ Res 7(2):134–140

Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK (2011) A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J Hydrol 396:128–138

Zangenehmadar Z, Moselhi O (2016) Application of neural networks in predicting the remaining useful life of water pipelines. In: Pipelines 2016, pp 292–308

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Seyed Mehran Jafari[1] · **Abdol Reza Zahiri**[1] · **Omid Bozorg Hadad**[2] · **Mahmoud Mohammad Rezapour Tabari**[3,4]

✉ Mahmoud Mohammad Rezapour Tabari
mrtabari@umz.ac.ir

Abdol Reza Zahiri
zahiri.areza@gmail.com

[1] Department of Water Engineering, University of Gorgan Agriculture Sciences and Natural Resources, Gorgan, Iran

[2] Department of Water Engineering, University of Tehran, Tehran, Iran

[3] Faculty of Technology and Engineering, University of Mazandaran, Mazandaran, Iran

[4] Center of Excellence in Risk Management and Natural Hazards, Isfahan University and Technology, Isfahan, Iran