Prediction model for the leakage rate in a water distribution system

Burak Kizilöz
Kocaeli Metropolitan Municipality Water and Sewerage Administration, İzmit, Kocaeli, Turkey
*Corresponding Author. E-mail: bkiziloz@isu.gov.tr

ABSTRACT

Leakages cause real losses in water distribution systems (WDSs) from transmission lines, storage tanks, networks, and service connections. In particular, the amount of leakage increases in aging networks due to pressure effects, resulting in severe water losses. In this study, various artificial neural network (ANN) models are considered for determining monthly leakage rates and the variables that affect leakage. The monthly data, which are standardized by Z-score for the years 2016–2019, are used in these models by selecting four independent variables that affect the leakage rate regarding district metered areas and pressure metered areas in WDSs. The pressure effects are taken into consideration directly as input. The model accuracy is determined by comparing the predicted and measured data. Furthermore, the leakage rates are estimated by directly modelling the actual data with ANNs. Consequently, it is found that the model results after data standardization are somewhat better than the original nonstandardized data model results when 30 neurons are used in a single hidden layer. The reason for the higher accuracy in the standardized case compared with previous modelling studies is that the pressure effect is taken into consideration. The suggested models improve the model accuracy, and hence, the methodology of this paper supports an improved pressure management system and leakage reduction.

Key words: artificial neural network, leakage rate prediction, pressure management, water distribution system, water losses

HIGHLIGHTS

- The data standardized by Z-score.
- Pressure effect has been directly taken into consideration for the first time.
- The model accuracy has been developed by augmenting neurons numbers in single hidden layer.
- The highest accuracy has been achieved with the least input in models.
- Pressure and age effect use directly in modelling.

LIST OF ABBREVIATIONS

The following symbols are used in this paper:

| Acronym | Definition |
|---------|------------|
| ANN     | Artificial neural network |
| ANP     | Average network pressure |
| AWWA    | American Water Works Association |
| DMA     | District metered area |
| FFBP    | Feed forward back propagation |
| IWA     | International Water Association |
| LR      | Leakage rate |
| MAN     | Mean age of network |
| MDN     | Mean diameter of network |
| FMS     | Pressure management system |
| PRV     | Pressure reducing valve |
| R²      | Coefficient of determination |
| SCADA   | Supervisory control and data acquisition system |
| SI      | Scatter index |
| SIV     | System input volume |
| TNL     | Total network length |
| TSIV    | Total system input volume |
| WDS     | Water distribution system |

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (http://creativecommons.org/licenses/by/4.0/).
INTRODUCTION

Water losses in water distribution systems cause increase in operational cost of water utilities and, in turn, increase the water price. It is predicted that the amount of water leakage in water distribution systems (WDSs) is 48 billion m³ per year around the world (Kingdom et al. 2006). Water utilities try to apply certain techniques besides modernization programs in networks to control and reduce high levels of water losses. Each water utility should prioritize reliable water loss studies by modernizing its water distribution system. Reliable water loss method developments that use modern techniques will help to reduce losses on a planned basis, save energy, reduce water production costs, improve water quality and increase investments.

According to the American Water Works Association (AWWA) and the International Water Association (IWA) Water Balance and Terminology, water loss consists of apparent losses (non-physical losses and management losses) and real losses (physical losses) (AWWA 2003). Real losses consist of leakage on the transmission and/or distribution mains, real losses from raw water mains and treatment works, leakage and overflows at transmission and/or distribution storage tanks and leakage on service connections up to the point of customer metering (Alegre et al. 2016). Leakage is a key parameter for water loss.

Leakages in WDSs can be categorized as reported leakages, unreported leakages and background leakages (Lambert 2003). Reported leakages can be defined as emerging and visible leakages; unreported leakages as non-surface leakages that are detectable by acoustic devices; and background leakages as non-surface leakages that are acoustically undetectable. Leakage removal by timely detection is also significant for water loss levels. A literature review concerning this subject shows that various methods have been preferred in several studies (Xue et al. 2020; Hu et al. 2021).

To reduce water leakage, a pressure management system – a well-known system with low cost – is implemented in WDSs (Kanakoudis & Gonelas 2016; Samir et al. 2017). High leakage rates are observable at high pressure levels because the leakage rate is a function of pressure (Kanakoudis & Muhammetoglu 2014). Pressure management is achieved by dividing the WDS into smaller and more manageable areas (DMAs) (Kanakoudis & Muhammetoglu 2014). The pressure is reduced and controlled by installing pressure reducing valves (PRVs) at the critical points in the DMA. Using a pressure management system (PMS) and DMAs, it is possible to monitor a system 24 hours per day via the supervisory control and data acquisition system (SCADA), which can prevent losses by reducing leakages and breaks. Leakage reduction helps to protect limited water sources, minimize the quantity of refined water, pump less water and minimize power consumption.

The leakage rate (LR) is the ratio between the total system input water volume and the water loss. The leakage rate varies depending on the pipe age, material quality, hole geometry on the pipe surface, operating pressure and similar factors. Marchis & Milici (2019) examined leakages in laboratory environments by using rectangular and circular cracks in polyethylene pipes of different sizes at various pressure levels and then evaluated the experimental study results with Toricelli, International Water Association (IWA) standards and their modifications as well as Cassa formulations. Niu et al. (2018) modelled the leakage rate in Tianjin water supply networks through the principal component regression method. The researchers took the network factors into account in their studies, such as maintenance cost, annual average water pressure, pipe material, valve replacement cost, pipe age, and pipe diameter. They obtained the adjusted $R^2$ value as 0.72 through the developed leakage rate–leakage factors model. AL-Washali et al. (2018) analysed the leakage rate by using the minimum night flow analysis in Zarqa intermittent supply system. The researchers indicated that the one-day minimum night flow analysis should not be used to predict the leakage rate due to customer tanks will fill overnight. Leu & Bui (2016) developed, through the Bayesian method, a leakage prediction model in the Taipei WDS. According to the model results, the pipe age, construction activity, ground movement and pressure fluctuation have significant roles in leakage. Jang et al. (2018) predicted the leakage rates in WDSs by using certain statistical analysis methods, such as ANNs, Z-scores and principal components. The pipe length/junction, demand energy rate, number of water leaks, mean diameter, pipe rate deterioration and water supply quantity and junction parameters were used as input variables in the model. The best determination coefficient (0.55) was estimated by an ANN model with multiple hidden layers and 24 neurons (Jang et al. 2018). However, the pressure effect, the most important parameter regarding the leakage, was not taken into consideration in the modelling.

The study aims to predict the leakage rate through the artificial neural networks that are applied today in many science fields with the ability to solve complex problems successfully. For this purpose; (i) in this study, the pressure parameter and network age that are directly related to the leakage have been taken into consideration.
for the first time as model input for the LR prediction, (ii) the ideal ANN architecture has been developed by analysing the effective of each parameter one by one, (iii) the combination that provides the highest model accuracy by using the least input parameters has been researched, (iv) and finally, the original data has been standardized through the Z-score technique, as in similar studies, to increase the prediction model accuracy and the calculations have been repeated for the determined ANN model combinations.

**METHODOLOGY**

In this study, the artificial neural networks, one of the artificial intelligence methods, has been used to predict the leakage rate according to the following steps:

1. The parameters of İzmit’s water distribution system that monthly measured and recorded between 2016 and 2019 have been collected for the model study.
2. The original data has been standardized through the Z-score technique to increase the prediction performance of ANN models.
3. The ANN models with single input have been fictionalized to determine the model input parameters. In these models, an ideal ANN structure has been designed for the LR prediction by increasing the neuron numbers in the single hidden layer from 5 to 30.
4. The effective parameters such as system input volume (TSIV), total network length (TNL), mean age of networks (MAN), mean diameter of networks (MDN), and average network pressure (ANP) have been selected using single input models for the LR prediction.
5. Various model combinations have been developed to predict the LR with minimal input by increasing the model input numbers. The best prediction model has been obtained as TSIV/TNL-ANP-MAN-MDN combination.
6. The performance criteria such as R², SI, and G-value have been used to analyse the model accuracy.
7. The LR has been predicted using original data in TSIV/TNL-ANP-MAN-MDN combination.
8. The prediction accuracy of models obtained through the original data has been evaluated with the same performance criteria.
9. The ideal LR prediction model has been identified by comparing all model performance results.
10. The accuracy of all prediction models developed through the applied methodology and selected parameters is higher than the prediction models with six inputs suggested by Jang et al. (2018). Also, the methodology suggested in the study has been summarized in Figure 1.

**Data standardization via the Z-score**

The Z-score can be used when performing analysis to distinguish the differences and the distributions of variables (Jang & Choi 2017). The variable values, x, can be standardized by taking the mean difference of each variable value and dividing it into the standard deviation.

\[ z = \frac{x - \mu}{\sigma} \]  

In this equation, z is the standardized data value, μ is the mean data and σ is the standard deviation. The Z-score technique allows the variables in all data sets to be accumulated into a common variable range. In addition, this technique indicates by how many standard deviations the variables deviate from the mean. By means of this technique, the raw data are converted to a standardized value score with a standardized deviation of 1 and a mean of 0. Hence, comparing the standardized values and variables becomes easier.

**Artificial neural networks**

In recent years, the development of ANNs has accelerated to help cognitive science by imitating the working principle of nervous systems. ANNs can be classified in accordance with their topologies (e.g., single-layer and multilayer feed-forward networks). Single- and multilayer feed-forward networks have been widely used in studies to better understand hydraulic engineering problems (Kızılöz et al. 2015) and to determine the complex structures of WDS components (Jang et al. 2018).

The architectural structure of an ANN is composed of artificial neurons, which allow data transfer between layers in the forward and backward directions. Each neuron in the network is connected to the others by weights.
The weights are the parameters used to establish the effects of inputs on outputs. The key of the network is to calculate the required optimum weight values by propagating the error in accordance with the training algorithm of the given weights.

The transfer function used in this study calculates the effect of all inputs and weights. This function calculates the net neuron input. The total net input collected in a neuron (net) is obtained by the following expression:

$$\text{net} = \sum_{i=1}^{n} w_{ji} x_i + b$$  \hspace{1cm} (2)

where $x_i$ is the neuron input value, $w_{ji}$ is the weight coefficient, $n$ is the total number of input neurons, and $b$ is the threshold value.

In addition, the activation function helps to determine the neuron output by processing the net input obtained from the transfer function. Selecting the correct activation function significantly affects the network performance and the success rate. The sigmoid function is generally selected as the activation function for multilayer perceptron models. The neuron output calculated by means of this function is given as follows:

$$y_i = f(\text{net}) = \frac{1}{1 + e^{-\text{net}}}$$  \hspace{1cm} (3)

In this study, a three-layer feed-forward back-propagation network (FFBP) model is considered, with a sigmoid function in the hidden layer and a linear function in the output layer. While the forward transfer of data through the model to obtain a result in the output layer is called feed-forward, the backward transfer of the network into the input layer – if there is an error between the actual and target outputs – is called back-propagation. In the back-propagation stage, all weights are readjusted according to the error correction rule (Haykin 1999). When the error reaches the actual value, the iteration ends, and the target value is calculated as the output. Each model needs to be trained before estimation. In this study, the Levenberg-Marquart back-propagation (trainlm) algorithm is considered a fast and precise algorithm for the training process (Kızılöz et al. 2015). The Levenberg-Marquart algorithm is based mainly on the least-squares method that uses the maximum neighbourhood. This method includes the best features of the Gauss-Newton and gradient descent algorithms for adjusting the weights, and it is expressed by the following equation in accordance with the LM algorithm after various iterations.
approximations and optimizations.

\[ \Delta w = (J + \mu I)^{-1} J^T e \]  

(4)

where \( w \) is the weighting factor, \( J \) is the Jacobian matrix, \( I \) is the unit matrix, \( \mu \) is a constant coefficient larger than zero and \( e \) is the error. If \( \mu \) is too large, the method behaves like the gradient-descent method; if it is too small, it behaves like the Gauss-Newton method.

In this study, the MATLAB software was used to calculate the ANN prediction model design and result. The model inputs included TSIV/NL, ANP, MAN, and MDN variables, and the model output was the LR. For each model application, the data was randomly divided into training (55%), validation (35%), and test data (10%) through the algorithm defined in the MATLAB program (Kızılöz et al. 2015; Şişman & Kızılöz 2020). The most important issue in the ANN application is to decide the hidden layer and neuron numbers. Many studies in the literature have preferred a single hidden layer due to the higher number of hidden layers do not improve the model performance (Kızılöz et al. 2015). All ANN models in this study were installed as a single hidden layer (Şişman & Kızılöz 2020). There is no mathematical test to determine the neuron numbers in the hidden layer for ANN design. Generally, the numbers are determined through trial-and-error methods. The ANN models suggested in this study are chosen on the basis of various numbers of neurons, such as 5, 10, 20 and 30, in one hidden layer. A typical FFBP network consists of an input layer, one or more hidden layers and an output layer, as shown in Figure 2.

### Evaluation of the ANN model performance

Accuracy evaluation is possible through a comparison of the model estimation value using ANNs and the measured value in estimating the LR. In this study, certain performance functions, such as the coefficient of determination (\( R^2 \)), scatter index (SI), and G-value, have been used to determine the model accuracy. The calculation methods of all the performance functions for all the validation data sets are given by the following expressions:

\[
R^2 = \left[ \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \right]^2
\]  

(5)

\[
SI = \frac{\text{RMSE}}{\bar{x}}
\]  

(6)

\[
G = \left( 1 - \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \times 100
\]  

(7)

where \( x_i \) is a data value, \( y_i \) is the estimated data value and \( n \) is the number of validation data values. Finally, \( \bar{x} \) and \( \bar{y} \) are the means of the measurement and estimation data.

### Figure 2

A FFBP-ANN structure with a single hidden layer used in this study.
Data collection and description of the study area

Izmit is the second largest district of Kocaeli and is selected as the study area. The district has had 363,416 people, 160,135 water consumers and 30,840,477 m³ of water supply since 2018. Here, 67 sections of DMAs and 84 sections of PMAs, as shown in Figure 3, were installed in 2014 to reduce the water loss rate of 45.40%. While the total network length of the district is 1,114 km (Kızılöz & Şisman 2021), the network length in the DMAs is 56,639 km. In addition, all water meters in the DMAs have been replaced entirely by smart water metres to remove the apparent loss effect. As a result of WDS hydraulic model studies of the district in question at the end of 2018, the water loss rate was reduced up to 29.70% by dividing the WDS into DMAs and PMAs. In particular, the pressure management system has been very useful in the WDS, where the losses were minimized, reducing the leakages of mains and service connections that could not be detected.

To analyse the leakage rate in the modelling study, 1,357 data measurements were taken on a monthly basis between 2016 and 2019 in the DMAs and PMAs. The effective factors affecting leakage in the WDS divided into the DMAs are as follows: the average pipe diameter, water supply quantity, district characteristics, pipe length, frequency of leaks, water pressure in the pipes and network configuration (Jo et al. 2016). In this study, certain variables are used for modelling that directly express the real losses in DMAs and PMAs, such as the total system input volume (TSIV), total network length (TNL), mean age of networks (MAN), mean diameter of networks (MDN), average network pressure (ANP), and leakage rate (LR). The TSIV and TNL represent the total monthly measured values, and the MAN, MDN and ANP represent the average monthly measurements.

The descriptive summary statistics used in the prediction models for the variables are given in Table 1.

The DMA comparisons were made using the average monthly variable measurements in each DMA. The leakage rates (LRs) were calculated by dividing the water losses by the TSIV. The largest rate was 0.64 in DMA No. 35, while the smallest rate was 0.05 in DMA No. 63 (Figure 4). An analysis of the LR rates for DMAs has shown that the rate is above 0.50 in eleven of the DMAs, between 0.3 and 0.5 in twenty-eight and between 0.2 and 0.3 in fourteen. It is necessary to identify the detection failures in DMAs with very high LR rates by means of active leakage control activities through acoustic devices, to replace aging networks that break down frequently and to revise the ideal operating pressure after these studies. The LR in fourteen DMAs was successfully maintained under 0.2. The LR may be minimized by reducing the pressure at regular intervals in accordance with the minimum night flow due to the 24-hour monitoring by the SCADA system.

Figure 3 | General view of the DMAs and PMAs in the study area.
Table 1 | WDS variables used in the study area

| Variables | Unit | Min   | Mean  | Max   | Median | S.D.  |
|-----------|------|-------|-------|-------|--------|-------|
| TSIV      | m³   | 1.094 | 23,109| 72,016| 22,140 | 12,462|
| TNL       | km   | 0.43  | 7.74  | 38.18 | 6.97   | 5.20  |
| MAN       | year | 4.88  | 11.54 | 25.71 | 11.35  | 3.71  |
| MDN       | mm   | 81.25 | 135.31| 237.14| 132.50 | 32.11 |
| ANP       | m    | 4.03  | 41.05 | 79.57 | 38.90  | 13.25 |
| LR        | –    | 0.002 | 0.32  | 0.75  | 0.31   | 0.15  |

Figure 4 indicates that the minimum network pressure is 5.6 m in DMA No. 44 and the maximum network pressure is 75.39 m in DMA No. 2. In the study area, the network pressure of fourteen DMAs is above 50 m. The mean pipe age of all the DMAs is 11.54; the greatest pipe age is 25.71 in DMA No. 27, and the least is 4.88 in DMA No. 23. While DMA No. 6 has the greatest network length, 38.18 km, DMA No. 35 has the smallest length, 0.43 km. By generating smaller DMA areas, the LR can be controlled and reduced. The maximum mean system input volume, 60,882 m³, is that of DMA No. 2. The maximum mean pipe diameter is that of DMA No. 16, 237.14 mm, and the minimum diameter is that of DMA No. 54, 81.25 mm. The average data regarding the dependent and independent variables that affect the LR are shown in Figure 4.

RESULTS AND DISCUSSION

Z-score analysis

The standardized data were obtained by means of the Z-score method in the estimation of the LR by the ANNs method. A standardized analysis method was implemented using a total of 1,357 data points on a monthly basis for various variables that affected the leakage in 67 DMAs. The analysis results indicated that the Z-scores of 66 data points were outside the range of ±3; that is, these data were outliers from the average and were removed before the analysis. When analysing the distribution of the removed data, it was found that there were 27 data points from the MDN, 20 from the MAN, 3 from the ANP, 15 from the TSIV/TNL (km) and 1 from the LR. On the other hand, 1,291 pieces of monthly data were used in this study for LR estimation by standardizing 1,357 pieces of raw data in the DMAs. The Z-score results regarding all variables in the DMAs and PMAs are shown in Figure 5.

Artificial neural networks

To identify the effective variables in the LR estimates, single-input single-output ANN models were established by using standardized data. The monthly data collected from the DMAs and PMAs were randomly divided into 55% for training (710 data points), 35% for validation (452 data points) and 10% for testing (129 data points). Similar training, validation and testing data sets were used for all models. The Levenberg-Marquart method of back-propagation was selected for the training algorithm by using the Neural Net Fitting toolbox in MATLAB. Before each training process, the models were initialized with irregular initial weights and biases (Kizilöz et al. 2015). In this study, different numbers of neurons (such as 5, 10, 20 and 30) were used in the hidden layer for the models. The best model with four inputs was developed by means of the best model variables with a single input. The performance of the prediction models is given in Tables 1 and 2. Subsequently, the same model with four inputs was established by using the same ANN methodology as for the original data, and finally, the best prediction model was determined as a result of the performance evaluation for the models obtained from the original and standardized data.

ANN model performance and optimal model selection

To separately analyse the effects of physical parameters such as the MDN, MAN, TNL, ANP and TSIV, which are related to the LR, ANN models with a single input and single output were established by using data with removed outliers. The model accuracy was evaluated by comparing the predicted LR values with the measured LR values. The performance functions given in Table 2 were used for the model accuracy evaluations. The best criterion for how well the model results fit in a linear curve is the coefficient of determination, R², in the regression analysis process. A higher R² value means that the prediction models are more accurate. If the SI is small, the model
Figure 4 | Monthly average data of the DMAs.
results are closer to a 1:1 (45°) straight line. When the G-value approaches 100, the model prediction accuracy is considered excellent (Kim et al. 2010).

The single input ANN models on LR prediction are available in Table 2. When the first models are analyzed, it is seen that the performances of MDN, MAN, and TNL are better than ANP and TSIV. The performances of the models with single input suggested in this study are higher than the ones given in the study conducted by Jang et al. (2018). In addition, the neuron numbers were increased up to 30 starting from 5 neurons in the hidden layer, and the model performances were evaluated accordingly. The highest accuracy was obtained by using 30 neurons in the hidden layer described in Table 2 for prediction models with a single input using different neuron numbers (such as 5, 10, 20, and 30). If more than 30 neurons are available in the hidden layer, the model performances have decreased, so they are not included in this study. The model performances based on the neuron numbers in the hidden layer can be seen in Table 3. The model results indicate that pressure, diameter, and age are the effective parameters of the leakage rates.

The TSIV/TNL-ANP-MAN-MDN prediction models with four inputs and a single output were obtained with a higher accuracy by using the independent variables, which are effective on leakage rates, as shown in Table 2. The prediction model has a higher accuracy than the other applied neuron numbers, according to the performance evaluations of $R^2$, the SI and the G-value, if 30 neurons are available in a single hidden layer using data eliminated and standardized with the Z-score (Table 3).

The ANN [30] prediction model has a lower scattering value, which is 15.223 in all models. The LR prediction models that use discretization of the outlier data through the Z-score technique are shown in Figure 6. The LR

### Table 2 | Performance of ANN models with a single input variable

| Model          | Input combinations | $R^2$   | SI (%)  | G-value |
|----------------|--------------------|---------|---------|---------|
| ANN [30]       | MDN                | 0.6946  | 99.603  | 69.443  |
| ANN [30]       | MAN                | 0.6931  | 99.922  | 69.247  |
| ANN [30]       | TNL                | 0.6589  | 105.273 | 65.865  |
| ANN [30]       | ANP                | 0.2661  | 154.706 | 26.283  |
| ANN [30]       | TSIV               | 0.2307  | 158.386 | 22.734  |

### Table 3 | Performance of ANN models with four input variables

| Data          | Model       | Input                  | $R^2$   | SI (%)  | G       |
|---------------|-------------|------------------------|---------|---------|---------|
| Standardized  | ANN [5]     | TSIV/TNL               | 0.5603  | 27.511  | 55.931  |
|               | ANN [10]    | ANP                    | 0.78525 | 20.657  | 75.154  |
|               | ANN [20]    | MAN                    | 0.8350  | 16.851  | 83.466  |
|               | ANN [30]    | MDN                    | 0.8658  | 15.223  | 86.506  |
| Original      | ANN [5]     |                        | 0.5495  | 32.438  | 54.719  |
|               | ANN [10]    |                        | 0.7383  | 24.759  | 73.626  |
|               | ANN [20]    |                        | 0.8231  | 20.378  | 82.129  |
|               | ANN [30]    |                        | 0.8586  | 18.160  | 85.870  |
prediction model with 30 neurons in the hidden layer in Figure 6 has the highest coefficient of determination of 0.8658 and the highest G-value of 86.506.

The most accurate results were obtained by means of the ANN in [30] in comparison with other neuron numbers when the same number of original data values were used as the input in TSIV/TNL-ANP-MAN-MDN, the best model for LR prediction (see Table 3). This LR prediction model has a higher accuracy than the other applied neuron numbers provided that 30 neurons are available in the hidden layer in accordance with the $R^2$, SI and G-value performance evaluations. Different LR prediction models based on the original data with various numbers of neurons in the hidden layer are shown in Figure 7. The prediction model of the ANN in [30] has the highest $R^2$ value of 0.8586 and the lowest scattering index of 18.160.

The most accurate model results corresponding to the measured LR were achieved when there were 30 neurons in the hidden layer of the suggested models for both the original and standardized data sets. In the case of using 30 neurons in the hidden layer with outlier data removal, the G-value, scattering index (SI) and coefficient of determination, $R^2$, are slightly better than for the original data.

When comparing the model results with the study of Jang et al. (2018), it was found that the prediction accuracy was higher. They obtained the best model result by using 24 neurons in multiple hidden layers with 6 principal component analysis data inputs ($R^2 = 0.5516$ and G-value $= 52.4$). In this study, the monthly leakage rates were predicted with higher accuracy through the ANN models with comparatively less input and fewer neurons.

The prediction models with a pressure variable have higher model accuracy, which derives from the effect of pressure on the leakage rate being higher than that of the other variables. Various examples from the literature are as follows: the leakage in water distribution systems changes directly with pressure (Bonthuys et al. 2020); while a small amount of leakage occurs at low pressure, excessive leakage occurs at high pressure (Marchis & Milici 2019); reducing the leakage in WDSs can be achieved by controlling the pressure through a pressure management system (Jafari-Asl et al. 2020); on the other hand, pressure management can reduce the system input

Figure 6 | Scatter graph for the TSIV/TNL-ANP-MAN-MDN model with standardized data: (a) ANN in [5]; (b) ANN in [10]; (c) ANN in [20]; (d) ANN in [30].
volume (SIV) amount due to water loss and a decrease in demand (Kravvari et al. 2018); and pressure regulation and replacing old water supply networks in a planned way prevents leakages (Leu & Bui 2016).

**CONCLUSIONS**

In this study, the monthly leakage rate in the water distribution system (WDS) of İzmit district (Kocaeli/Turkey) was predicted through artificial neural network (ANN) models. The model input variables were determined to be the ANP, TSIV/TNL, MAN and MDN, and the goal was to achieve the highest prediction accuracy with the least input in this way. The pressure effect was considered as an input for the first time for model performance improvement up to 57.41%, according to the previous studies in the literature. In this study, the model performance improvement was achieved with data standardization by suitable methods and with an increase in the preferred neuron numbers. Also, higher prediction accuracies can be obtained through the model structure with one hidden layer designed in this study.

The developed models clearly revealed the relationship between leakage and pressure. It is understood from the study that pressure is a significant factor for modelling and that pressure management should be taken into account by water utilities to reduce water losses by preventing leakages in water distribution systems (WDSs). According to the models, the other factor influencing leakages is the network age. An increase in the leakage rates has been observed in old networks under high pressure effects due to the reduced resistance to pressure. It is necessary to control the operating pressure to certain levels by taking the network age into consideration to reduce leakage rates.

In conclusion, the leakage rates can be predicted through the suggested models by taking into consideration the network pressure and network age as a reference, and these models provide important information for water utilities. In the suggested models, the pressure evaluation and age appear to be the variables with the greatest effect.
on leakage, and this shows that water utilities should give priority to the replacement of old networks and the application of a pressure management system for operable and sustainable management policies.

**DATA AVAILABILITY STATEMENT**

All relevant data are included in the paper or its Supplementary Information.

**REFERENCES**

Alegre, H., Baptista, J. M., Cabrera, E., Cubillo, F., Duarte, P., Hirner, W., Merkel, W. & Parena, R. 2016 *Performance Indicators for Water Supply Services*, 3rd edn. IWA Publishing, London, UK.

AL-Washali, T., Sharma, S., AL-Nozaily, F., Haidera, M. & Kennedy, M. 2018 Modelling the leakage rate and reduction using minimum night flow analysis in an intermittent supply system. *Water* 11 (1), 48.

AWWA 2005 *Best Practice in Water Loss Control: Improved Concepts for 21st Century Water Management*. American Water Works Association, Denver, CO, USA. Available from: https://www.awwa.org/Portals/0/AWWA/ETS/Resources/WLCFlyerFinal.pdf?ver=2015-02-10-083650-287.

Bonthuys, G. J., Van Dijk, M. & Cavazzini, G. 2020 Energy recovery and leakage-reduction optimization of water distribution systems using hydro turbines. *Journal of Water Resources Planning and Management* 146 (5), 04020026.

De Marchis, M. & Milici, B. 2019 Leakage estimation in water distribution network: effect of the shape and size cracks. *Water Resources Management* 33 (3), 1167–1183.

Haykin, S. 1999 Neural networks: a comprehensive foundation by Simon Haykin. *The Knowledge Engineering Review*.

Hu, X., Han, Y., Yu, B., Geng, Z. & Fan, J. 2021 Novel leakage detection and water loss management of urban water supply network using multiscale neural networks. *Journal of Cleaner Production* 278.

ISU (Kocaeli Water and Sewage Administrative General Directorate) 2018 *Kocaeli Water and Sewage Administrative: Annual Activity Report for Water Management*. Kocaeli, Turkey.

Jafari-Asl, J., Sami Kashkooli, B. & Bahrami, M. 2020 Using particle swarm optimization algorithm to optimally locating and controlling of pressure reducing valves for leakage minimization in water distribution systems. *Sustainable Water Resources Management* 6 (4).

Jang, D. & Choi, G. 2017 Estimation of non-revenue water ratio for sustainable management using artificial neural network and Z-score in Incheon, Republic of Korea. *Sustainability* 9 (11), 1933.

Jang, D., Park, H. & Choi, G. 2018 Estimation of leakage ratio using principal component analysis and artificial neural network in water distribution systems. *Sustainability* 10 (3), 1–13.

Jo, H. G., Choi, G. W. & Jang, D. W. 2016 Development of the Non-revenue Water analysis equation through the statistical analysis of main parameter in waterworks system in Incheon City. *Crisis and Emergency Management* 12 (11), 63–75.

Kanakoudis, V. & Gonelas, K. 2016 Assessing the results of a virtual pressure management project applied in Kos Town water distribution network. *Desalination and Water Treatment* 57 (25), 11472–11483.

Kanakoudis, V. & Muhammetoglu, H. 2014 Urban water pipe networks management towards non-revenue water reduction: two case studies from Greece and Turkey. *Clean – Soil, Air, Water* 42 (7), 880–892.

Kim, D. H., Ryu, D. W., Choi, Y. M. & Lee, W. J. 2010 Application of Kriging and inverse distance weighting method for the estimation of Geo-Layer of Songdo Area in Incheon. *Journal of the Korean Geotechnical Society* 26, 5–19.

Kingdom, B., Liemberger, R. & Marin, P. 2006 *The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries – How the Private Sector Can Help: A Look at Performance-Based Service Contracting*. Water Supply and Sanitation Sector Performance Indicators Board Discussion Paper Series. IWA Publishing, London, UK.

Kizilöz, B. & Şişman, E. 2021 Exceedance probabilities of non-revenue water and performance analysis. *International Journal of Environmental Science and Technology*.

Kizilöz, B., Çevik, E. & Aydoğan, B. 2015 Estimation of scour around submarine pipelines with Artificial Neural Network. *Applied Ocean Research* 51, 241–251.

Kravvari, A., Kanakoudis, V. & Patelis, M. 2018 The Impact of pressure management techniques on the water age in an urban pipe network – The Case of Kos City Network. *Proceedings* 2 (11), 699.

Lambert, A. 2003 Assessing non-revenue water and its components: a practical approach. *Water* 21 (2), 50–51.

Leu, S. S. & Bui, Q. N. 2016 Leak prediction model for water distribution networks created using a bayesian network learning approach. *Water Resource Management* 30 (8), 2719–2733.

Samir, N., Kansoh, R., Elbarki, W. & Flei, A. 2017 Pressure control for minimizing leakage in water distribution systems. *Alexandria Engineering Journal* 56 (4), 601–612.

Şişman, E. & Kizilöz, B. 2020 Artificial neural network system analysis and Kriging methodology for estimation of non-revenue water ratio. *Water Supply* 20, 1871–1883.

Wang, C., Zhang, Y., Wei, X. & Gao, X. 2018 Leakage rate model of urban water supply networks using principal component regression analysis. *Transactions of Tianjin University* 24 (2), 172–181.

Xue, Z., Tao, L., Fuchun, J., Riehle, E., Xiang, H., Bowen, N. & Singh, R. P. 2020 Application of acoustic intelligent leak detection in an urban water supply pipe network. *Journal of Water Supply: Research and Technology – AQUA* 69 (5), 512–520.

First received 13 February 2021; accepted in revised form 10 June 2021. Available online 23 June 2021