International Journal of Pressure Vessels and Piping 189 (2021) 104274

Contents lists available at ScienceDirect

International Journal of Pressure Vessels and Piping

journal homepage: http://www.elsevier.com/locate/ijpvp

Predicting burst pressure of defected pipeline with Principal Component Analysis and adaptive Neuro Fuzzy Inference System

Hieu Chi Phan [a], Huan Thanh Duong [b,*]

[a] Le Quy Don Technical University, 236 Hoang Quoc Viet, Hanoi, 100000, Viet Nam
[b] Faculty of Engineering, Vietnam National University of Agriculture, Trau Quy, Gia Lam, Hanoi, 100000, Viet Nam

ARTICLE INFO

Keywords:
Burst pressure
Corrosion
Soft computing
Principal component analysis
Adaptive neuro fuzzy inference system

ABSTRACT

Pipeline is an important and valuable infrastructure for transporting oil and gas which expanding a long distance and working in a corrosive environment. Consequently, corrosion becomes one of the most critical threads for metal material pipeline. The high internal pressure in an oil and gas pipeline is the additional factor leading to the high risk of bursting. Various models predicting the burst pressure of defected pipeline have been developed in literature. However, evaluating burst pressure of defected pipe is a nonlinear mechanical problem with the appearance of the stress concentration, accuracy of the existing models is not high and the issue still open. The application of data-driven approach with soft computing and machine learning has been a potential and promising approach. This paper investigates the application of Adaptive Neuro Fuzzy Inference System (ANFIS) and a data transforming technique for dimension reduction and noise elimination, the Principal Component Analysis (PCA). The PCA has demonstrated its ability in noise removal for the database and ANFIS provides an improvement in the accuracy of the prediction. The developed model is the combination of ANFIS and PCA, the ANFIS-PCA model, has overwhelmed other existing models by archiving the correlation of determination at 0.9919 and the Root Mean Square Error decreases to 0.9883 MPa. Observations on the difference network configurations and number of epochs also provided.

1. Introduction

Transporting oil and gas with high strength steel pipeline is a commonly choice because of the continuous and stable characteristics. These structures expand to a long distance with various severe environments which can directly affect to the mechanical capacity of them. Some of the hazards for the pipeline can be named as toxic components in the transported materials, corrosive soil, ice gouging, land sliding or earth quake etc. Among such factors, corrosion is the most significant cause leading to the reduction of pipe reliability and various studies have been conduct to understand the mechanism of failure of defected pipe. These defects significantly reduce the capacity of pipe-shaped structures especially when they suffer from internal pressure.

Various models have been developed to predict the burst pressure of pipes under the internal pressure which can be mainly categorized as analytical or empirical approaches. The analytical approach which based on mechanical theory faces a difficult to deal with the appearance of defects causing the stress concentration and the locally change of pipe shape. A few researches have been attempted to provide the analytical equation to predict burst pressure compared to the empirical approach [1–3]. The common approach is to develop empirical model. The empirical models partly based on the mechanical relationship between input variables and the experiment or simulation data from to optimized the factors in an explicit equation. The factors may be limited to less than 10 and the accuracy of the model depend on both the structure of the equations and the database used to developed them. Most of the existing models in literature belong to this category [4–6].

However, recent researches revealed that both analytical and empirical models only provided a medium evaluation metric on predicting burst pressure compared to Finite Element Analysis (FEA) or experiments [6–8]. Phan et al. [6] has recognized the limitations of the existing models and revisited 3 of them based on an optimization algorithm with the optimized variables are the empirical factors in the equation and the objective function is the error on the database obtained from FEA. The factors adjustment based on a database has significantly improved the models. Keshtegar et al. [7] has a wonderful validation for more than 30 existing models which depicts the high Mean of Absolute Error, (i.e. MAE) of such models, ranging from 3.183 MPa up to more

* Corresponding author.
E-mail address: dthuan@vnua.edu.vn (H.T. Duong).

https://doi.org/10.1016/j.ijpvp.2020.104274
Received 11 June 2020; Received in revised form 21 November 2020; Accepted 27 November 2020
Available online 30 November 2020

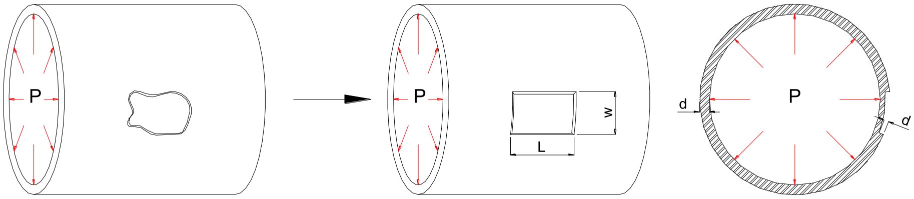


Fig. 1. Dimensions (depth × length × width, d × L × w) of the idealized defects in the pipe suffered from internal pressure P.

than 30.249 MPa and about a quarter of models have MAE larger than 10 MPa on the test dataset. This is unacceptable with a database has the mean of burst pressure at 24.448 MPa. Amaya-Gómez et al. [8] provide a summary of ratios of predict-to-test burst pressures of 22 models and there are models produce low means of this ratio at around 0.7–0.8 along with the coefficient of variances are significantly high at 0.16 to 0.31. These studies have demonstrated an urge of improving the accuracy of burst pressure models.

The recent surge of data-driven models, which are heavily depend on the relationship of input the database, provides another approach for the prediction task in engineering and science. There are various techniques in machine learning and soft computing have been developed for predicting the output. Some of them are Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) and Bayesian Network, Adaptive Neuro Fuzzy Inference System (ANFIS). Despite of the success of these models in other fields, machine learning and soft computing have not gain much concerned for predicting pipe burst pressure. There are rarely seen researches conducted this approach, some of few studies are [9–11]. The ANN has been applied for pipe with both single and multiple defects in Ref. [9,10]. Silva et al. [9] uses neural network to predict the burst pressure of multiple corrosion defects using the database from FEA. Chin et al. [10] conducted ANN for the pipe with single defect based on a database collected from literature. This approach has also been implemented with another machine learning model, the SVM. None of such data-driven models for have tried to preprocessing the database to obtain better results, the boundaries of the data-driven model, which naturally set by the range of the inputs is limitedly discussed.

To the best of our knowledge, the application of ANFIS for the prediction of burst pressure under internal pressure have not been developed. ANFIS is firstly introduced by Jiang [12] with various applications for a variety of engineering problems such as but not limit to geotechnique [13,14] or finding material properties [8,15,16] etc. This soft computing method is a hybrid system of a neural network and Fuzzy Inference System (FIS) through the training process to optimized the membership functions, MF, of inputs and the rule weight.

A drawback of the ANFIS is the exponential decrease of number of rules due to the increase of number of inputs and this leads to the explosive in computational cost. Consequently, the use of ANFIS may need a dimension reduction technique and the Principal Component Analysis (PCA) which is a data transformation aimed to the ranked the transformed data due to their variances is applied in this paper. Inputs with least variables thus can be removed or the feature selection is implemented. The PCA also well-known for the capacity of noise reduction in database and consequently improve the accuracy of the prediction [16–19].

In this study, the database is collected globally from various studies available in literature. The application of ANFIS to predict the burst pressure of the defected pipeline is implemented with and without PCA to observe the effectiveness of this technique. Along with this investigation, various network configurations for ANFIS will be conducted in a grid search. Comparison study of ANFIS and other available models also provided to illustrate the effectiveness of the developed models.

Table 1
The reference models.

	Model	Equation
1	Netto et al. (2005) [4]	$P = P_0 \times \left(1 - 0.9435\left(\frac{d}{t}\right)^{1.6}\left(\frac{L}{D}\right)^{0.4}\right)$
2	ASME B31G (2012) [5]	$P = P_0 \times \left(\frac{1-\frac{d}{t}}{1-\frac{d}{tM^*}}\right)$
3	Gajdoš and Šperl (2012) [3]	$P = P_0 \times \left(\frac{1-\frac{\pi d}{4t}}{1-\frac{d}{t}}\right)$
4	Modified PCORRC (2004) [21]	$P = P_0 \times \left(1 - \frac{d}{t}\left(1 - \exp\left(-0.157\frac{L}{\sqrt{Dt}}\right)\right)\right)$
5	Phan et al. (2017) Model 1 [6]	$P = P_0 \times \left(1 - 0.88555\left(\frac{d}{t}\right)^{0.98077}\left(\frac{L}{D}\right)^{0.31053}\right)$
6	Phan et al. (2017) Model 2 [6]	$P = P_0 \times$ $\left(\frac{1-0.92126\frac{d}{t}}{1-0.92126\frac{d}{t}\left(1+0.06361\frac{L^2}{Dt}\right)^{-2.75485}}\right)$
7	Phan et al. (2017) Model 3 [6]	$P = P_0 \times \left(1 - \frac{1.24678\frac{d}{t}}{1+12.6739\frac{t}{L}}\right)$

*M: Folias factor $M = \sqrt{1 + 0.6275\frac{L^2}{Dt} - 0.003375\frac{L^4}{D^2t^2}}\, for\ \frac{L^2}{Dt} \le 50$; $M = 0.032\frac{L^2}{Dt} + 3.3\, for\ \frac{L^2}{Dt} > 50$

2. Material and methods

2.1. Existing burst pressure models

In general, both analytical and empirical models agree with the formation of burst pressure P model for defected pipe as the multiple of the burst pressure of the intact pipe, P_0, and the reduction factor, f, as in Eq. (1).

$$P = f(D, t, d, L, w) \times P_0(D, t, \sigma) \tag{1}$$

$$P_0 = \frac{2t\sigma}{D} \tag{2}$$

The intact burst pressure is a function of material strength or allowable stress, σ, and pipe dimension including diameter, D, and pipe wall thickness, t. The widely accepted format of P_0 is the Barlow's formula in Eq. (2) with minor variations. Allowable stress of material may be the yield stress σ_y (e.g. Ref. [1,2]) and a large number of models use the ultimate tensile strength σ_u (e.g. Refs. [4–6]).

The defect, which has the complex and random shape, is commonly idealized to be rectangle-shaped as in Fig. 1. The defect dimensions are: defect depth, L; and sometimes there is an appearance defect width, w (e.g. Ref. [20]). The reduction factor is a function of normalized defect dimensions based on pipe dimensions such as d/t, L/D or $\frac{L}{\sqrt{Dt}}$ [4–6,21]. A variety formats of the burst pressure equations mostly deprived from these factors. There are 7 existing equations given in Table 1 are chosen as the examples of the burst pressure equation format in Eq. (1) and they

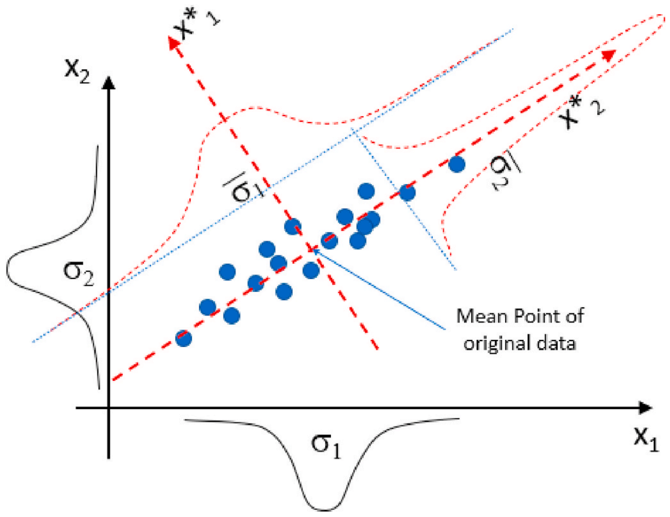


Fig. 2. Illustration of PCA with a 2 inputs database.

are latter used as the reference models for the proposed model.

2.2. Principal Component Analysis (PCA)

The main idea of the PCA technique is to orthogonally project data with n dimensional inputs and m samples, X_{nxm}, to the principal subspace with n_1 dimensions in which the variance of the transformed data is maximized or the sum of square of projection errors is minimized [22]. Fig. 2 illustrates the basic idea of PCA with a database of 2 inputs x_1 and x_2 with the variances are σ_1 and σ_2, respectively. By a projection to the new space, inputs of database now are the principle components x^*_1 and x^*_2 with the variances are $\overline{\sigma_1}$ and $\overline{\sigma_2}$. The variance of x^*_1 is maximized with a large $\overline{\sigma_1}$ and the sum of square of data points to x^*_2 is minimized with the minimized $\overline{\sigma_2}$. The small value of sum of square of projection errors leads to the ignorable of the principle component x^*_2 and the large variance of x^*_1 implied the capacity of this principal component to explain the variance of the overall database.

Mathematically, if the covariance matrix, CV, is expressed as:

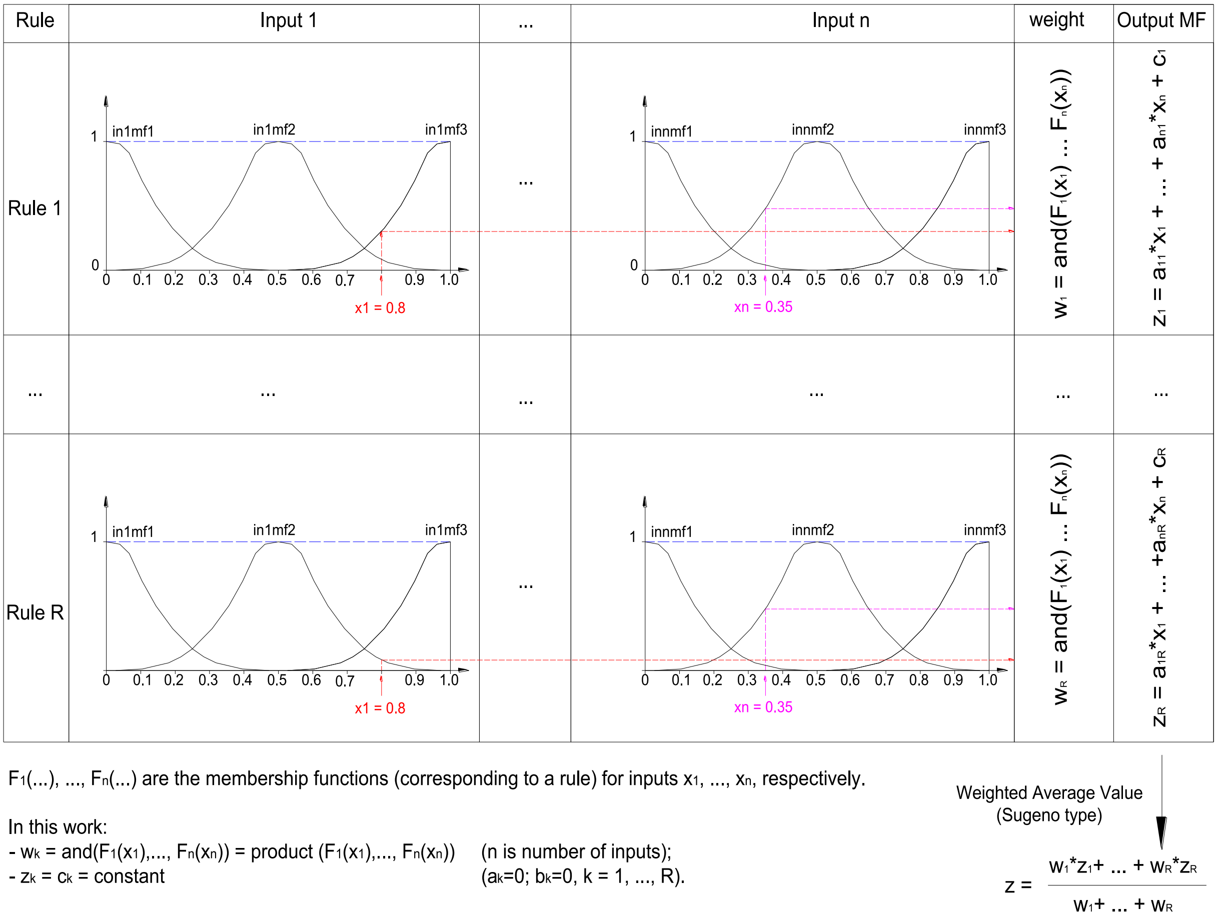
$$CV_{nxn} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x})(x_i - \overline{x})^T = \frac{1}{m}\overline{\overline{X}} \times \overline{\overline{X}}^T \quad (3)$$

where: x_i is column vector corresponding to data of each datapoint and $\overline{x}$ is the mean vector of the inputs $\overline{x} = \frac{1}{m}\sum_{j=1}^{m}x_j$. Consequently, $\overline{\overline{X}}$ is the normalized matrix of database X.

Designating λ_j and e_j where i = [1, …,n] as the eigenvalues and eigenvectors of the CV matrix where:

$$CV \times e_j = \lambda_j \times e_j \quad (4)$$

Denoting U_{nxn1} is a matrix where X_{nxm} can be projected to the principal n_1 dimension subspace by $U^T_{n\times n1}\overline{\overline{X}}_{n\times m}$. It is proofed that if the first n_1 eigenvectors of CV matrix, e_j where j = [1, …,n_1], are used to establish the U_{nxn1} = [e_j] matrix, the maximum of variance of the transformed data or the sum of square of projection errors as in the definition of the PCA can be satisfied. The eigenvalue of principal component ith, λ_i, is the explanation of this principal component to the variance of all database. Thus, we have $\sum_{i=1}^{n}\lambda_i = 1$ and $\sum_{i=1}^{n1}\lambda_i$ is the cumulative explanation of the selected principal components. The


Rule
Input 1
...
Input n
weight
Output MF
Rule 1
in1mf1
in1mf2
in1mf3
1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.9 1.0
x1 = 0.8
innmf1
innmf2
innmf3
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
xn = 0.35
w1 = and(F1(x1) ... Fn(xn))
z1 = a11*x1 + ... + an1*xn + c1
...
Rule R
wR = and(F1(x1) ... Fn(xn))
zR = a1R*x1 + ... +anR*xn + cR
Weighted Average Value (Sugeno type)
z = (w1*z1+ ... + wR*zR) / (w1+ ... + wR)

F_1(...), ..., F_n(...) are the membership functions (corresponding to a rule) for inputs x_1, ..., x_n, respectively.

In this work:
- w_k = and($F_1(x_1)$,..., $F_n(x_n)$) = product ($F_1(x_1)$,..., $F_n(x_n)$) (n is number of inputs);
- z_k = c_k = constant (a_k=0; b_k=0, k = 1, ..., R).

Fig. 3. Inputs with Gaussian MFs with R rules and Sugeno output weighted average method in FIS.

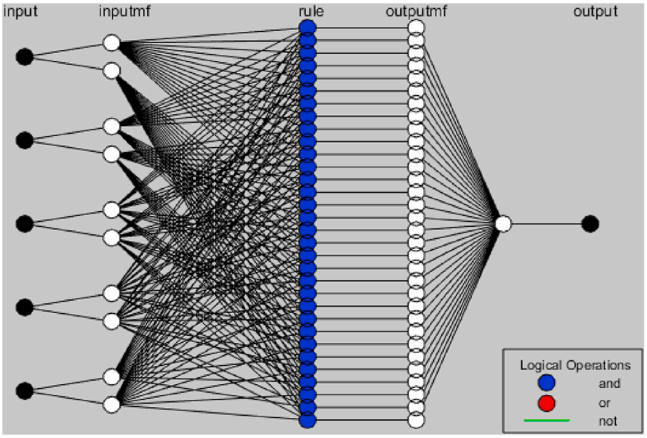


Fig. 4. Example of an ANFIS network structure.

database X_{nxm} is transformed to the new subspace as matrix $X^*_{n1\times m}$ by:

$$X^*_{n1\times m} = U^T_{n\times n1} \times \overline{\overline{X}}_{n\times m} \tag{5}$$

2.3. Adaptive Neuro Fuzzy Inference System (ANFIS)

Introduced since 1993 by Jang [12], the ANFIS is the hybrid soft computing method which combines the Fuzzy Inference System FIS which has the advantage of explaining the patterns in the database and the neural network which has the capacity of accurately predict the output [23]. The FIS based on the linguistic-interpretable fuzzy rules and input MFs for decision making. An illustration of the Sugeno FIS can be observed in Fig. 3 with a system with n input 1, …, n; each of them has 3 Gaussian MFs. Assuming that there is a sample with n inputs [x_1; …; x_n], in the example from Fig. 3, it is equivalent to [0.8; …; 0.35]. This sample can be fuzzified into 3 MFs of input 1 = 0.8 (i.e. in1mf1, in1mf2 and in1mf3) and input n = 0.35 (i.e. innmf1, innmf2 and innmf3). Linguistically, the MFs [in1mf1, in1mf2, in1mf3] can be interpreted as [low, medium, high] based on its values, respectively. Assuming that there is an existing set of R rules used to define the output MFs (i.e. [out1mf1, out1mf2, …, out1mfR]) of the output such as:

Rule 1: If (Input 1 is in1mf3) and … and (Input n is innmf2) then (Output = out1mf1 is z1)

…

Rule *k:* … … … …

….

Rule R: If (Input 1 is in1mf2) and … and (Input n is innmf2) then (Output = out1mfR is z_R)

The output thus has R MFs of [*out1mf1*, …, *out1mfR*] = [z_1, …, z_R] (constants in this study) and corresponding weights w_1, …, w_R. The crisp output of the FIS can be found by weighted average as in Fig. 3 and Eq. (6). The weights, w_k with k = 1, …, n, can be found by a given manner (i.e. *and* method in Fig. 3 or *and* logical gate in Fig. 4) which is the product of membership functions corresponding to a given rule (i.e. F_i(.) with i = 1, …,n) of the n variables in this study. The full version of z_k for Suneo type of FIS is given in Eq. (7) but simplified to be $z_k = c_k$ (i.e. $a_{k1} = a_{k2} = … a_{kn} = 0$) in this study.

$$z = \frac{\sum_{k=1}^{R} w_k z_k}{\sum_{k=1}^{R} w_k} \tag{6}$$

$$z_k = \sum_{i=1}^{n} a_{ki} + c_k \tag{7}$$

Without the training process, the use of FIS is hard to provide proper MFs of inputs and the weights w_k corresponding to each rule. The neutral network thus applied to tune the Gaussian MFs (i.e. by adjusting the mean and standard deviation) and the weights of rules to minimize the error on the training set. In this study, the Root Mean Square Error, RMSE, is used as the objective function of this optimization. An example of the structure of the ANFIS is given in Fig. 4 with the first layer is the crisp input, which is fuzzified with the inputmf layer which contain a selection of nodes, each of these nodes is presented for the MF of each input. This layer is connected to the next layer with a set of nodes represented for the set of rules. The inputmf of inputs are combined with the rule and then the output MFs are obtained at the outputmf layer. This layer then converged to a node as the weighted average z as in Fig. 3. The training and validating processes in this study are implemented on Matlab®.

3. Results and discussion

3.1. Data collecting

Database used for training ANFIS contained 217 samples is gathered from published studies including Ma et al. [24] (79 experiment samples), Shuai et al. [25] (39 FEA samples and 14 experiment samples), Phan et al. [6] (28 FEA samples), Freire et al. [26] (17 experiment samples) and Cronin [27] (40 experiment samples). Descriptive statistics of the database are given in Table 2. The database covered a wide range of steel grades including X42, X46, X52, X56, X60, X65, X80, X100 and some anonyms due to the missing data. The Ultimate Tensile Strength ranges from 309 MPa to 886 MPa (there is none of N/A for σ_u). Different pipe sizes are collected from both experiment and simulation with diameters ranges from 76.2 mm to 1320 mm and wall thicknesses from 2 mm to 25.4 mm. The data contains both intact and defected pipes with depth to thickness ratio varies within [0, 0.8] and length of the defects are up to 1432.560 mm. Width of the defects are not included in the input data because this input is observed to have minor effect to the pipe burst pressure and absence in many available models in literature as can be seen in Table 1 and reviewed in Ref. [7,8,25,27]. The predicted variable or the burst pressure ranges from 3.570 MPa to 35.968 MPa. This set of data ranges (bolded in Table 2) is used as the boundary of the developed model. The database is split into train set and test set with the ratio of 8:2. The train set uses 173 samples for the training the model and the test set with 44 samples uses for validating the trained models.

Table 2
Descriptive statistics of the database.

#	Metric	D	t	d	L	σ_u	P_{actual}
1	count	217	217	217	217	217	217
2	mean	482.906	9.397	4.424	314.023	554.132	15.279
3	std	242.888	4.506	2.998	294.417	84.741	6.752
4	**min**	**76.200**	**2.000**	**0.000**	**0.000**	**309.000**	3.570
5	25%	323.600	6.400	2.620	99.060	481.130	10.210
6	50%	324.104	8.585	3.750	243.000	542.000	13.580
7	75%	508.000	9.800	6.818	466.700	598.900	21.100
8	**max**	**1320.000**	**25.400**	**15.410**	**1432.560**	**886.000**	35.968

Table 3
Explain variances of Principle Component.

Principle Component	Explain Variance λ_i	Cumulative Explanation $\sum_{i=1}^{n1}\lambda_i$
X^*_1	0.6865	0.6865
X^*_2	0.2910	0.9776
X^*_3	0.0221	0.9997
X^*_4	0.0003	0.9999
X^*_5	0.0001	1.0000

3.2. Grid searching for the best ANFIS

As can be seen in Table 3, the most significant principle x^*_1 account for more than 68% of the overall variance of the database followed by x^*_2 with 29.1%. The combination of x^*_1 and x^*_2 thus can explain most of the total variance (97.76%). x^*_3, x^*_4 and x^*_5 are the least significant level with the explain variance at 2.21%, 0.03% and 0.01%, respectively.

Grid searching for both ANFIS with PCA (designated as ANFIS-PCA) and without PCA are conducted and provided in Table 4. Along with the development of ANFISs with original database without the PCA (i.e. Attempt 1), the transformed databases with 4, 3 and 2 principal components used in attempts 2, 3 and 4, respectively.

In each attempt, several trials are implemented with the adjustment of Gaussian MFs. The number of epochs of the training process is regularly chosen at 10000 with some exceptions on the Attempt 2 when this value intentionally changed to observe the effect of epoch to the result. The relative difference of RMSEs of train and test set provides an intuition of the overfitting, the scenario that the model well predicted on train set but badly on the test set, if it is occurred.

The best models in this attempt is Model 1.1, which contain 2 MFs for each input, has the RMSE on the train and test set are 1.4723 MPa and

Table 4
Grid search for selecting best configuration for the network.

Attempt	Trial/Model	Number of Inputs, n	Epochs	Number of input MFs	$RMSE_{train}$ (MPa)	$RMSE_{test}$ (MPa)	Relative difference of RMSE[a]
1	1.1	Original (5)	10000	2 2 2 2 2	1.4723	2.3513	0.5970
1	1.2	Original (5)	10000	3 3 3 3 3	0.92875	10.0764	9.8494
2	**2.1**	**4**	**100000**	**2 2 2 2**	**0.6805**	**0.9883**	**0.4523**
2	2.2	4	10000	2 2 2 2	0.7378	1.0485	0.4211
2	2.3	4	1000	2 2 2 2	0.8459	1.1304	0.3363
2	2.4	4	100	2 2 2 2	1.1456	1.4254	0.2442
2	2.5	4	10	2 2 2 2	1.1380	1.1372	−0.0007
2	2.6	4	10000	5 2 2 2	0.4058	1.2902	2.1794
2	2.7	4	10000	3 3 3 3	0.1843	2.2299	11.0993
3	3.1	3	10000	2 2 2	5.7025	6.3969	0.1218
3	3.2	3	10000	3 3 3	4.2414	18.7202	3.4137
3	3.3	3	10000	4 4 4	3.0930	44.4064	13.3571
4	4.1	2	10000	2 2	5.8088	6.5054	0.1199
4	4.2	2	10000	3 3	5.2718	6.2452	0.1846
4	4.3	2	10000	4 4	4.7500	4.7499	0.0000
4	4.4	2	10000	5 5	4.1742	7.1229	0.7064

[a] $(RMSE_{test} - RMSE_{train})/RMSE_{train}$.

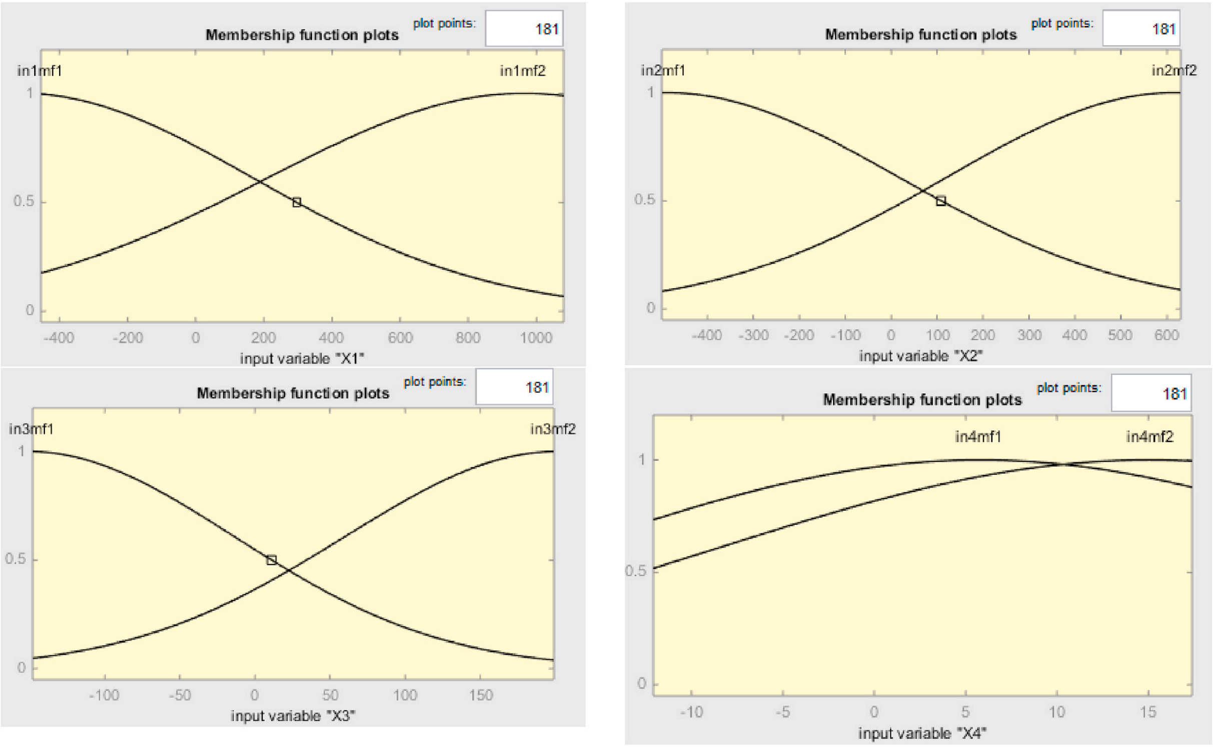


Fig. 5. MFs of input variables in Model 2.1.

Table 5
Rules for the ANFIS-PCA with the corresponding output MF.

#	Linguistic Rules	Output MF
1	1. If (X1 is in1mf1) and (X2 is in2mf1) and (X3 is in3mf1) and (X4 is in4mf1) then (P_(MPa) is out1mf1 = z_1)	z_1 = -145.9760
2	2. If (X1 is in1mf1) and (X2 is in2mf1) and (X3 is in3mf1) and (X4 is in4mf2) then (P_(MPa) is out1mf2 = z_2)	z_2 = 216.7739
3	3. If (X1 is in1mf1) and (X2 is in2mf1) and (X3 is in3mf2) and (X4 is in4mf1) then (P_(MPa) is out1mf3 = z_3)	z_3 = −12.1173
4	4. If (X1 is in1mf1) and (X2 is in2mf1) and (X3 is in3mf2) and (X4 is in4mf2) then (P_(MPa) is out1mf4 = z_4)	z_4 = 32.2241
5	5. If (X1 is in1mf1) and (X2 is in2mf2) and (X3 is in3mf1) and (X4 is in4mf1) then (P_(MPa) is out1mf5 = z_5)	z_5 = -47.7218
6	6. If (X1 is in1mf1) and (X2 is in2mf2) and (X3 is in3mf1) and (X4 is in4mf2) then (P_(MPa) is out1mf6 = z_6)	z_6 = 99.7615
7	7. If (X1 is in1mf1) and (X2 is in2mf2) and (X3 is in3mf2) and (X4 is in4mf1) then (P_(MPa) is out1mf7 = z_7)	z_7 = −184.8438
8	8. If (X1 is in1mf1) and (X2 is in2mf2) and (X3 is in3mf2) and (X4 is in4mf2) then (P_(MPa) is out1mf8 = z_8)	z_8 = 296.7603
9	9. If (X1 is in1mf2) and (X2 is in2mf1) and (X3 is in3mf1) and (X4 is in4mf1) then (P_(MPa) is out1mf9 = z_9)	z_9 = −139.0048
10	10. If (X1 is in1mf2) and (X2 is in2mf1) and (X3 is in3mf1) and (X4 is in4mf2) then (P_(MPa) is out1mf10 = z_{10})	z_{10} = 161.9820
11	11. If (X1 is in1mf2) and (X2 is in2mf1) and (X3 is in3mf2) and (X4 is in4mf1) then (P_(MPa) is out1mf11 = z_{11})	z_{11} = −77.0989
12	12. If (X1 is in1mf2) and (X2 is in2mf1) and (X3 is in3mf2) and (X4 is in4mf2) then (P_(MPa) is out1mf12 = z_{12})	z_{12} = 110.0268
13	13. If (X1 is in1mf2) and (X2 is in2mf2) and (X3 is in3mf1) and (X4 is in4mf1) then (P_(MPa) is out1mf13 = z_{13})	z_{13} = −72.6329
14	14. If (X1 is in1mf2) and (X2 is in2mf2) and (X3 is in3mf1) and (X4 is in4mf2) then (P_(MPa) is out1mf14 = z_{14})	z_{14} = 114.1636
15	15. If (X1 is in1mf2) and (X2 is in2mf2) and (X3 is in3mf2) and (X4 is in4mf1) then (P_(MPa) is out1mf15 = z_{15})	z_{15} = −73.9586
16	16. If (X1 is in1mf2) and (X2 is in2mf2) and (X3 is in3mf2) and (X4 is in4mf2) then (P_(MPa) is out1mf16 = z_{16})	z_{16} = 118.1119

2.3513 MPa, respectively. Compared with the mean value of burst pressure in Table 2 (15.279 MPa), this evaluation metric implied a reasonable model and further observed in the next part of the study. In model 1.2, RMSE on the test set is 10.0764 MPa, an unacceptable result for the output. The relative difference of RMSE of this model exploded to 9.8494 confirms that the overfitting occurred.

Attempt 2 with only 4 eigenvectors used preprocessing seems to be the best options compared to both original ANFIS and other ANFIS-PCAs models (i.e. in attempts 3, 4). Model 2.1 has the lowest RMSE at 0.9883 on the test set, and thus, chosen to be the final model (Table 4). Because the final model appeared in this Attempt, more observations implemented with the change of epochs and number of MFs in each input. RMSEs in models from 2.1 to 2.5 decrease gradually from 1.1372 MPa to 0.9883 MPa, respectively. It is worth noting that a very appropriate RMSE at 1.1372 MPa on the test set is reached after only 10 epochs in Model 2.5. This error value even higher than RMSE of Model 1.1 with RMSE on the test set at 2.3513 MPa. This observation provides a conclusion that with the proper configuration, ANFIS-PCA can be developed fast and provided a proper model with only few epochs.

Table 6
Comparison of developed ANFIS to reference models on the test set.

Group	Model	RMSE	R2	MAE
ANFIS	ANFIS without PCA	2.3513	0.9516	1.7423
	ANFIS-PCA	**0.9883**	**0.9919**	**0.6917**
Reference models	Netto et al. (2005)	2.4902	0.8829	2.0863
	Gajdoš and Šperl (2012)	2.7434	0.8579	2.1563
	ASME (2012)	2.0516	0.9205	1.5797
	Modified PCORRC (2004)	2.0346	0.9218	1.6026
	Phan et al. 1 (2017)	2.0773	0.9185	1.6927
	Phan et al. 2 (2017)	1.8116	0.9380	1.4444
	Phan et al. 3 (2017)	2.0964	0.9170	1.7515

Other trials on Attempt 2 are the change of number of MF for the most important principal component from 2 to 5 (trial 2.6) and increase MFs of all inputs from 2 to 3 (trial 2.7). In trial 2.6, RMSE on the train set decrease significantly to 0.4058 MPa but this value is slightly increasing to 1.2902 MPa on the test set. This implies a slight overfitting with the increaset of MFs on the most critical principal component x^*_1. This overfitting increases with the relative difference RMSE jumping to 11.0993 in Trial 2.7. However, the RMSE on the test set of this model is reasonable at 2.2299 and slightly larger than this of Model 1.1.

Attempt 3 and 4 not provide any proper configurations for ANFIS with the appearance of high values of RMSEs is consistently larger than 3 and up to 44.4064. The overfitting also occurred with relative difference RMSE increases to 13.3571 in Model 3.3. Analogous to Attempt 2, there is a trend of increasing MFs of input leading to the increase of errors in Attempt 3. This tendency is not appeared in Attempt 4 with a fluctuation of this metric appeared.

Details of the adaptive MFs of input for the ANFIS-PCA are in Fig. 5 and the set of rules for the models are in Table 5 with the corresponding MF for the output. The scatter plots of the predicted versus actual burst pressures of Model 1.1 and 2.1 provided in Fig. 6 illustrates the effective of such models. Most of the data points scatter densely around the 1:1 line in both cases. This implies the capacity of the models to predict the burst pressure close to the actual values. The higher density of ANFIS-PCA model around 1:1 line compared with the ANFIS graphically depicts its lower RMSEs (0.9883 compared to 2.3513, respectively).

Comparison of soft computing models with other reference models are given in Table 6 with the outstanding of soft-computing model with the significant improvement in evaluations metrics. The application of PCA effectively improves the coefficient of determination, R^2, the

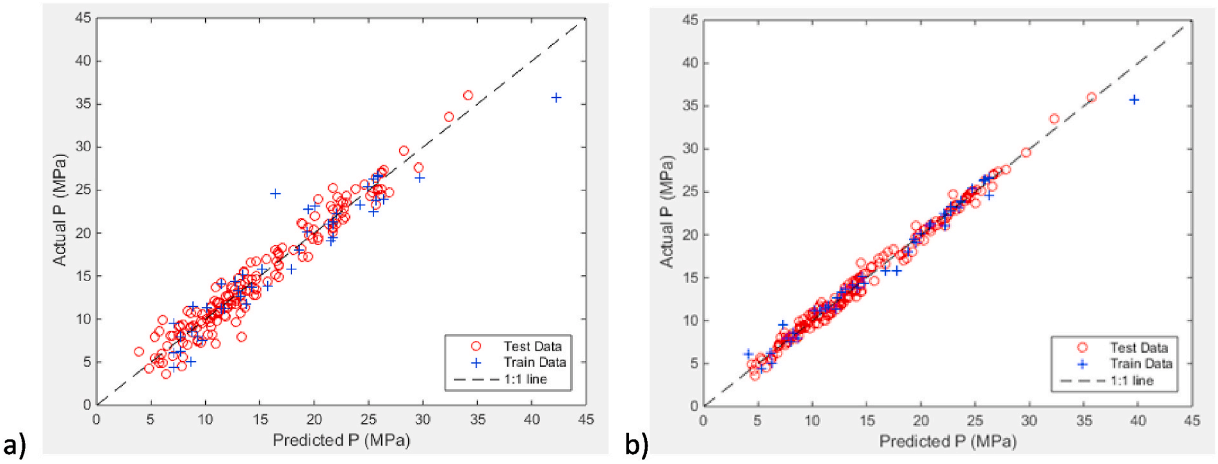


Fig. 6. Scatter plot of predicted with a) Model 1.1 and b) model 2.1 (ANFIS-PCA) versus actual burst pressure.

models from 0.9516 (ANFIS) to 0.9919 (ANFIS-PCA). The ANFIS-PCA model also outdistances other existing models with R^2 ranges from 0.8829 (Netto et al., 2005) to 0.9380 (Phan et al., 2017). The RMSE of ANFIS-PCA at 0.9883 MPa is about a half of those of other models which ranges from 1.8166 MPa to 2.7434 MPa.

4. Conclusions

The study has developed a high accuracy model for burst pressure predicting bases on PCA and ANFIS. The grid search, which tries different configurations for both ANFIS and ANFIS-PCA, is necessary to be applied while both overfitting and low-accuracy models found with inappropriate network configurations and number of MFs used for each input. The increase of number of MFs seems to be useless to improve the quality of the models. In many cases, the increase of this value may lead to high error on test set, overfitting, or both.

The PCA technique has significantly improved the model with the RMSE is cut off a half from 2.3513 MPa to 0.9883 MPa. The chosen model uses a set of 4 most significant principal components which explain up to 0.9999 variance of the database. The ignorance of the last principal component is not only reducing the calculation cost but also removing the unnecessary noise appeared in the database. The model is also tested with different number of epochs. It is interesting that ANFIS-PCA required a limited number of epochs (i.e. 10 epochs) to obtain a model with acceptable evaluation metric in Model 2.5. The evaluation of the ANFIS and ANFIS-PCA models with other existing empirical models shows the advantage of using data-driven models on the evaluation process. Without the PCA, the ANFIS has a R^2 of 0.9516, higher than any of those of references models. With the PCA, this value outdistances other models with R^2 up to 0.9919.

Author statement

Hieu Chi Phan: Conceptualization, Methodology, code writing, model developing, writing manuscript, revising manuscript, Visualization, revising; **Huan Thanh Duong**: Methodology, data collecting, writing manuscript, model validating, Visualization, revising

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant number: 107.02-2020.04.

References

[1] H. Schulze, G. Togler, E. Bodmann, Fracture mechanics analysis on the initiation and propagation of circumferential and longitudinal cracks in straight pipes and pipe bends, Nucl. Eng. Des. 58 (1) (1980) 19–31.
[2] W. Kastner, et al., Critical crack sizes in ductile piping, Int. J. Pres. Ves. Pip. 9 (3) (1981) 197–219.
[3] Ľ. Gajdoš, M. Šperl, Evaluating the Integrity of Pressure Pipelines by Fracture Mechanics, Applied Fracture Mechanics, 2012, p. 283.
[4] T. Netto, U. Ferraz, S. Estefen, The effect of corrosion defects on the burst pressure of pipelines, J. Constr. Steel Res. 61 (8) (2005) 1185–1204.
[5] ASME B31G, A., Manual for Determining the Remaining Strength of Corroded Pipelines, The American Society of Mechanical Engineers, New York, NY, 1991.
[6] H.C. Phan, A.S. Dhar, B.C. Mondal, Revisiting burst pressure models for corroded pipelines, Can. J. Civ. Eng. 44 (7) (2017) 485–494.
[7] B. Keshtegar, M.e.A.B. Seghier, Modified response surface method basis harmony search to predict the burst pressure of corroded pipelines, Eng. Fail. Anal. 89 (2018) 177–199.
[8] R. Amaya-Gómez, et al., Reliability assessments of corroded pipelines based on internal pressure–A review, Eng. Fail. Anal. 98 (2019) 190–214.
[9] R. Silva, J. Guerreiro, A. Loula, A study of pipe interacting corrosion defects using the FEM and neural networks, Adv. Eng. Software 38 (11–12) (2007) 868–875.
[10] K.T. Chin, et al., Failure pressure prediction of pipeline with single corrosion defect using artificial neural network, Pipeline Sci. Technol. 4 (1) (2020) 10–17, 3.
[11] J. Ji, et al., Prediction of stress concentration factor of corrosion pits on buried pipes by least squares support vector machine, Eng. Fail. Anal. 55 (2015) 131–138.
[12] J.-S. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans. Syst. Man Cybernet 23 (3) (1993) 665–685.
[13] J. Hou, M. Zhang, M. Tu, Prediction of surface settlements induced by shield tunneling: an ANFIS model, in: Geotechnical Aspects of Underground Construction in Soft Ground, CRC Press, 2008, pp. 567–570.
[14] A. Mottahedi, F. Sereshki, M. Ataei, Overbreak prediction in underground excavations using hybrid ANFIS-PSO model, Tunn. Undergr. Space Technol. 80 (2018) 1–9.
[15] A. Seitllari, M. Naser, Leveraging artificial intelligence to assess explosive spalling in fire-exposed RC columns, Comput. Concr. 24 (3) (2019) 271–282.
[16] H.-B. Ly, et al., Improvement of ANFIS model for prediction of compressive strength of manufactured sand concrete, Appl. Sci. 9 (18) (2019) 3841.
[17] I.E. Bell, G.V. Baranoski, Reducing the dimensionality of plant spectral databases, IEEE Trans. Geosci. Rem. Sens. 42 (3) (2004) 570–576.
[18] B. Poon, M.A. Amin, H. Yan, PCA based face recognition and testing criteria, in: 2009 International Conference on Machine Learning and Cybernetics, IEEE, 2009.
[19] C. Sun, et al., Noise reduction based on robust principal component analysis, J. Comput. Inf. Syst. 10 (10) (2014) 4403–4410.
[20] Y. Chen, et al., Electrochemiluminescence sensor for hexavalent chromium based on the graphene quantum dots/peroxodisulfate system, Electrochim. Acta 151 (2015) 552–557.
[21] Y.-p. Kim, et al., The evaluation of failure pressure for corrosion defects within girth or seam weld in transmission pipelines, in: 2004 International Pipeline Conference, American Society of Mechanical Engineers Digital Collection, 2004.
[22] C.M. Bishop, Pattern Recognition and Machine Learning, springer, 2006.
[23] S. Sivanandam, S. Deepa, Principles of Soft Computing (With CD), John Wiley & Sons, 2007.
[24] B. Ma, et al., Assessment on failure pressure of high strength pipeline with corrosion defects, Eng. Fail. Anal. 32 (2013) 209–219.
[25] Y. Shuai, J. Shuai, K. Xu, Probabilistic analysis of corroded pipelines based on a new failure pressure model, Eng. Fail. Anal. 81 (2017) 216–233.
[26] J. Freire, et al., Part 3: burst tests of pipeline with extensive longitudinal metal loss, Exp. Tech. 30 (6) (2006) 60–65.
[27] D.S. Cronin, Assessment of Corrosion Defects in Pipelines, 2000.