International Journal of Pressure Vessels and Piping 191 (2021) 104384

Contents lists available at ScienceDirect

International Journal of Pressure Vessels and Piping

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

Predicting pipeline burst pressures with machine learning models

Hieu Chi Phan [a,*], Ashutosh Sutra Dhar [b]

[a] Le Quy Don Technical University, 236 Hoang Quoc Viet, Hanoi, 100000, Viet Nam
[b] Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St John's, NL, A1B 3×5, Canada

ARTICLE INFO

Keywords:
Pipeline integrity
Burst pressure
Machine learning
Random forest
Support vector machine
Artificial neural network

ABSTRACT

Establishing an accurate model to predict burst pressure is desired, which has been developed for decades. Although various models have been developed, errors unavoidably appear in the prediction of burst pressures because of the uncertainty in both input variables and nonlinear relationship of such variables to the burst pressure. Consequently, machine learning models, which is a data-driven approach, are potential alternatives. In this paper, various machine learning models such as Random Forest, Support Vector Machine, and Artificial Neural Network are examined to predict the burst pressure, gathering databases available in the literature. The applications of these models are investigated to identify the advantages and limitations of the models. The machine learning models showed a significant improvement in the prediction of the burst pressures compared to the available reference models. However, some drawbacks of the models should be carefully considered, including an increase of error with the unfamiliar data and the fluctuations within the overall trend in the parametric study.

1. Introduction

Pipelines are essential and valuable assets for transporting various materials such as water, wastewater, oil, and gas. These structures are exposed to various hazards in the service life. Corrosion has been identified as a critical hazard for metal pipelines. The corrosion causes metal loss, causing defects on the pipe wall. Various studies and standards are available to assess the capacity of the pipelines with wall defects under the loading of internal pressure (known as the "burst pressure"). While the equations to predict the burst pressure of the intact pipe (i.e., pipe without corrosion) are explicit, based on the theoretical Barlow's equation and Tresca/von Mises criteria, there are uncertainties in the models for the burst pressure of corroded pipes. Decades of research resulted in the development of a number of different models for the burst pressure of the corroded pipes.

Keshtegar and Seghier [1] provided a thorough literature review of up to 33 burst pressure models developed since 1980. However, none of the models was capable of predicting the burst pressure with sufficient accuracy. Amaya-Gómez et al. [2] illustrated the weakness of different models by conducting predictions using 22 available models and comparing the results with databases from various sources. The means of the prediction-to-test ratio were as low as 0.7. Only three out of the 22 models in this review had prediction-to-test ratios larger than 0.9. Phan et al. [3] provided an observation of overestimation by 2 out of 3 existing models with their developed database. The low predictability by the existing models with respect to the available database implies that the problem with the limitations of the existing models is not fully developed.

Different approaches were applied for the development of burst pressure models using the analytical and semi-analytical methods. The analytical method relies on the theory of solid mechanics based on some assumptions to find the relationships for the burst pressure of intact and corroded pipes (e.g. Refs. [4,5]). The semi-analytical methods partly employ the theory of solid mechanics with the incorporation of parameters based on the experimental or simulation data ([3,6–8]). Semi-analytical methods are widely accepted and used by many standards or codes due to the simplicity of the design equations. However, the major limitation of the design equations is the inaccuracy in predicting the pipeline burst pressures. To address the limitation, a machine learning method can be developed based on available data. Although the development of the machine learning method requires specialized skills (e.g., coding, mathematics, and statistics), the developed model, such as the application software (apps), would be easy-to-use by engineers without specialized skills. In the current digital age, apps are considered a viable tool for engineering design and assessment.

The quality of data-driven models depends heavily on the database to

* Corresponding author.
E-mail address: phanchihieu@lqdtu.edu.vn (H.C. Phan).

https://doi.org/10.1016/j.ijpvp.2021.104384
Received 24 January 2021; Received in revised form 1 March 2021; Accepted 15 March 2021
Available online 24 March 2021

develop a regression model. The model varies from conventional statistical regression models (e.g. linear regression) to up-to-date machine learning models such as Random Forest, Support Vector Machine, and Artificial Neural Network (ANN). The fundamental of model development is to minimize the error between the predicted values and the "actual" values. In this field, the machine learning techniques overwhelm classical regression by providing much more flexible methods to be fitted with hundreds to millions of optimized weights. The machine learning methods are commonly incorporated with Finite Element method or experiment results to obtain the predicting model [9–12]. Meanwhile, this approach is not fully applied for burst pressure problems. Some applications can be named as Silva et al. [13], Ji et al. [14], Zolfaghari and Izadi [15], Oh et al. [16], and Phan and Duong [17]. Silva et al. [13] used ANN and FEM to adjust DNV model for pipes containing a single defect and multiple defects with 43 samples. Ji et al. [14] proposed a burst pressure model based on Support Vector Machine techniques with a dataset of 45 samples from Finite Element Analysis. The above machine learning models were developed using a limited database of fewer than 50 samples, and this can be insufficient to account for the nonlinear relationships between input and output variables for burst pressure models. This may lead to overfitting of the models and the low accuracy of the results when dealing with unfamiliar data.

Recent research on the problem, such as Zolfaghari and Izadi [15], Oh et al. [16], and Phan and Duong [17] focuses on developing a proper ANN model without thoroughly comparing it with other machine learning techniques. Additionally, the boundaries of the data-driven models should be clearly stated because the model is only appropriate within the ranges of the training data. Outside of the boundaries, the prediction is analogous to extrapolation. This intuition has rarely been tested in the available machine learning applications. Even though the study of Phan and Duong in Ref. [17] have implemented the search for optimal Adaptive neuro-fuzzy inference system, ANFIS, and the boundary was claimed, the illustration of out-of-boundaries predicting using the trained models has not been provided.

This paper attempts to use a global database with wide ranges of input variables from various available sources in the literature for high-strength steel pipe to investigate burst pressure models based on Machine Learning techniques. Along with the detailed description on developing machine learning models (Support Vector Regression, Random Forest and ANN), the capacity of the models developed using hundreds-size databases (e.g., 217 samples) is investigated. The grid search is implemented to obtain the optimal models, especially for the ANN, which is well known for the uncertainty in choosing the best configuration. These models were then validated with an unfamiliar database of the burst pressure for cast iron pipes provided in our previous work [18]. This evaluation was conducted to validate the model with out-of-boundary input variables.

2. Existing corroded burst pressure models

The existing models were mainly developed from the NG-18 equations and the Buckingham π theorem. The NG-18 equations, based on the works of Kiefner et al. [19] and Maxey et al. [20], aims to predict strength reduction of pipeline based on the longitudinal area of intact pipe (A_o), the longitudinal area of corroded pipe (A_c), Folias factor (M) and flow stress (σ_{flow}):

$$\sigma_{re} = \sigma_{flow} \times \left(\frac{1 - \frac{A_0}{A_c}}{1 - \frac{A_0}{A_c M}} \right) \tag{1}$$

where: M is the Folias factor.

The models derived from the NG-18 equations (such as ASME B31G [6]; CSA [21]; DNV RPF101 [22]; Phan et al. [3] – Model 2) have the general format as:

$$P = P_0 \times \left(\frac{1 - k_1 \frac{d}{t}}{1 - k_2 \frac{d}{tM}} \right) \tag{2a}$$

where: k_1 and k_2 are the factors, d is the depth of the defect, t is the wall thickness, P_0 is the burst pressure of the intact pipe. The burst pressure of the intact pipe can be obtained using the following equation:

$$P_0 = \frac{2t\sigma_{flow}}{D} \tag{2b}$$

where: D is the outside diameter of the pipe.

On the other hand, the models derived from the Buckingham π theorem are developed from the study of Netto et al. [8] (e.g. Wang et al. [23]; Phan et al. [3]– Model 1) and have the general format as in Eq. (3).

$$P = f\left(P_0, \frac{d}{t}, \frac{l}{t}, \frac{w}{D} \right) \tag{3}$$

where: w is the width of the defect.

Another form of the equation is as in Eq. (4) from Pipe Corrosion Criterion – PCORRC [24]:

$$P = f\left(P_0, \left(1 - \frac{d}{t}\right), \frac{l}{\sqrt{Dt}}, 1 - \frac{w}{D} \right) \tag{4}$$

Based on these formats, various models were developed with adjustments of parameters. Details of different models in the literature are available in Refs. [1,2] and not presented in this study to avoid repetition. In this paper, seven models are chosen for comparison (Table 1). These are Netto et al. [8], ASME B31G [6], Gajdoš and Šperl [5], Modified PCORRC (2004), Phan et al. [3] (Model 1, 2 and 3). All these models were developed for high-strength steel pipes based on various fundamental approaches. The Netto et al. [8] model is based on the Buckingham π theorem, while the model in ASME B31G [6] is NG-18-based. The Gajdoš and Šperl [5] model is an analytical model. The Modified PCORRC (2004) was developed using data generated from FEA simulations. Phan et al. [3] evaluated different models based on a database developed through FE analysis using Abaqus^R software and proposed modifications of model parameters to minimize the summation of squared error. A potential weakness of models in Phan et al. [3] is that a limited-size of a database with only 28 samples was employed to determine the model parameters.

3. Machine learning models

Vapnik [25] summarised the learning process from the example of a learning machine as in Fig. 1, where the generator draws random independent vectors x with fixed and unknown distribution function, $F(x)$. The supervisor returns output value, y, with conditional distribution function of $F(y|x)$. The learning $\bar{y}$machine yields the predicted using the function $f(x,\alpha)$ of x vectors and α parameters. The target of the process is

Table 1
Seven reference models available in the literature.

	Model	Equation	
1	Netto et al. (2005) [8]	$P = P_0 \times \left[1 - 0.9435\left(\frac{d}{t}\right)^{1.6}\left(\frac{l}{D}\right)^{0.4}\right]$	(5)
2	ASME B31G (2012)[a] [6]	$P = P_0 \times \left[\frac{1-\frac{d}{t}}{1-\frac{d}{tM}}\right]$	(6)
3	Gajdoš and Šperl (2012) [5]	$P = P_0 \times \left[\frac{1-\frac{\pi d}{4t}}{1+\frac{d}{t}}\right]$	(7)
4	Modified PCORRC (2004)	$P = P_0 \times \left[1 - \frac{d}{t}\left(1 - \exp\left(-0.157\frac{1}{\sqrt{Dt}}\right)\right)\right]$	(8)
5	Phan et al. (2017) Model 1 [3]	$P = P_0 \times \left(1 - 0.88555\left(\frac{d}{t}\right)^{0.98077}\left(\frac{l}{D}\right)^{0.31053}\right)$	(9)
6	Phan et al. (2017) Model 2 [3]	$P = P_0 \times \left(\frac{1 - 0.92126\frac{d}{t}}{1 - 0.92126\frac{d}{t}\left(1 + 0.06361\frac{l^2}{Dt}\right)^{-2.75485}}\right)$	(10)
7	Phan et al. (2017) Model 3 [3]	$P = P_0 \times \left(1 - \frac{1.24678\frac{d}{t}}{1 + 12.6739\frac{t}{l}}\right)$	(11)

$M = 0.032\frac{l^2}{Dt} + 3.3\frac{l^2}{Dt} > 50$ (for).

[a] $M = \sqrt{1 + 0.6275\frac{l^2}{Dt} - 0.003375\frac{l^4}{D^2t^2Dt}\frac{l^2}{Dt}} \leq 50$ (for).

to minimize the difference between the "actual" $\bar{y}y$ and by minimizing the Empirical Risk Function, R_{emp}:

$$R_{emp}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} Q(x_i, \alpha) \tag{12}$$

where: n is the number of samples, Q is the loss function (e.g., least squared error).

3.1. Support vector regression (SVR)

The support vector machine was introduced in the 1990s (Boser et al. [26]; Cortes and Vapnik [27]; Vapnik [28]) as a versatile model applicable for both regression and classification problems. Assuming that x is the input variable with n samples and y is the dependent or predicted variable. The x and y compose the training set for the training process. To solve the nonlinear problem, x is mapped to high-dimensional space with w is the normal vector of a hyperplane to obtain the linear relationship with the predicted variable:

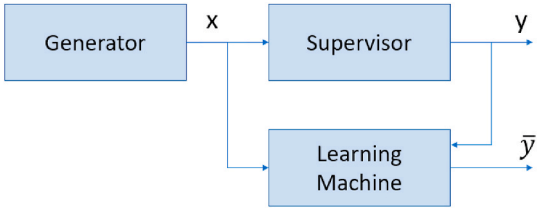

Fig. 1. The model of learning from example (Adapt from Ref. [25]).

$$y(x) = w^T\phi(x) + b \tag{13}$$

The aim of the regression problem here is to fit as many data points as possible to the "street" in Fig. 2, which is defined by the solid line and the margin of ξ_i^* tolerance ε. To obtain a soft margin that is not sensitive to outliers, the ξ_i and are the slack variables used to eliminate the effect of outliers.

The Empirical Risk function in Eq. (12) can be written in the regularized risk function format:

$$\min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \tag{14}$$

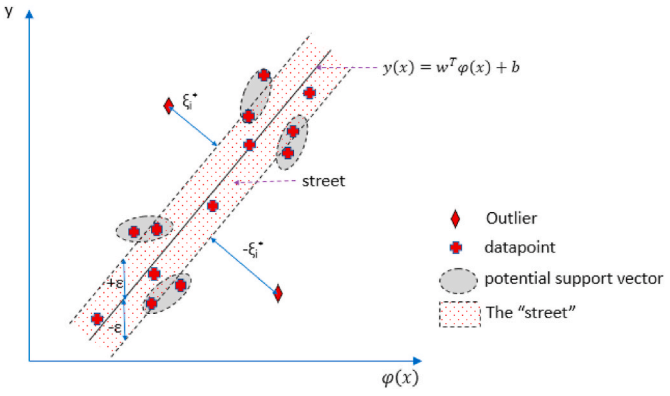

Fig. 2. The Support Vector Machine for regression problem (Adapt from Ref. [29]).

where: C is the hyper-parameter added for balancing the two objectives.

with $y_i - y(x_i) \leq \varepsilon + \xi_i$ and $y_i - y(x_i) \geq -\varepsilon - \xi_i^*$ and. $\xi_i^*, \xi_i \geq 0$

However, it is impossible to solve this solution directly because it is computationally expensive, especially when the training set is large. This leads to the so-call duality problem of SVM with $w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$:

$$\min\left(\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\left(\phi(x_i)^T.\phi(x_j)\right)\right) \tag{15}$$

With: $\sum_{i=1}^{n}\alpha_i y_i = 0$ *and* $0 \leq \alpha_i \leq C(\phi(x_i)^T.\phi(x_j))$To avoid the computational expense of the dot product in Eq. (15), the Kernel function is applied where Kernel functions follow the Mercer's Theorem, that is:

$$K(x_i, x_j) = \left(\phi(x_i)^T.\phi(x_j)\right) \tag{16}$$

Consequently, the predicted variable can be found by:

$$y(x) = \frac{1}{n}\sum_{i,j=1}^{n}\alpha_i K(x_i, x_j) + b \tag{17}$$

where: α_i composes the well-known support vector α.

The commonly used kernel functions are linear, polynomial, Gaussian radial basis, and sigmoid functions. In this study, the polynomial (poly) and Gaussian radial basis (RBF) are investigated in the fine-tuning process, which can be found by the equations shown below:

$$K(x_i, x_j) = \left(x_i^T x_j + 1\right)^m \ (poly) \tag{18}$$

$$K(x_i, x_j) = exp\left(-\gamma\left(x_i - x_j\right)^2\right) \ (RBF) \tag{19}$$

where γ and m is predefined parameters, m is the order of the polynomial, γ decides the shape of the bell in RBF.

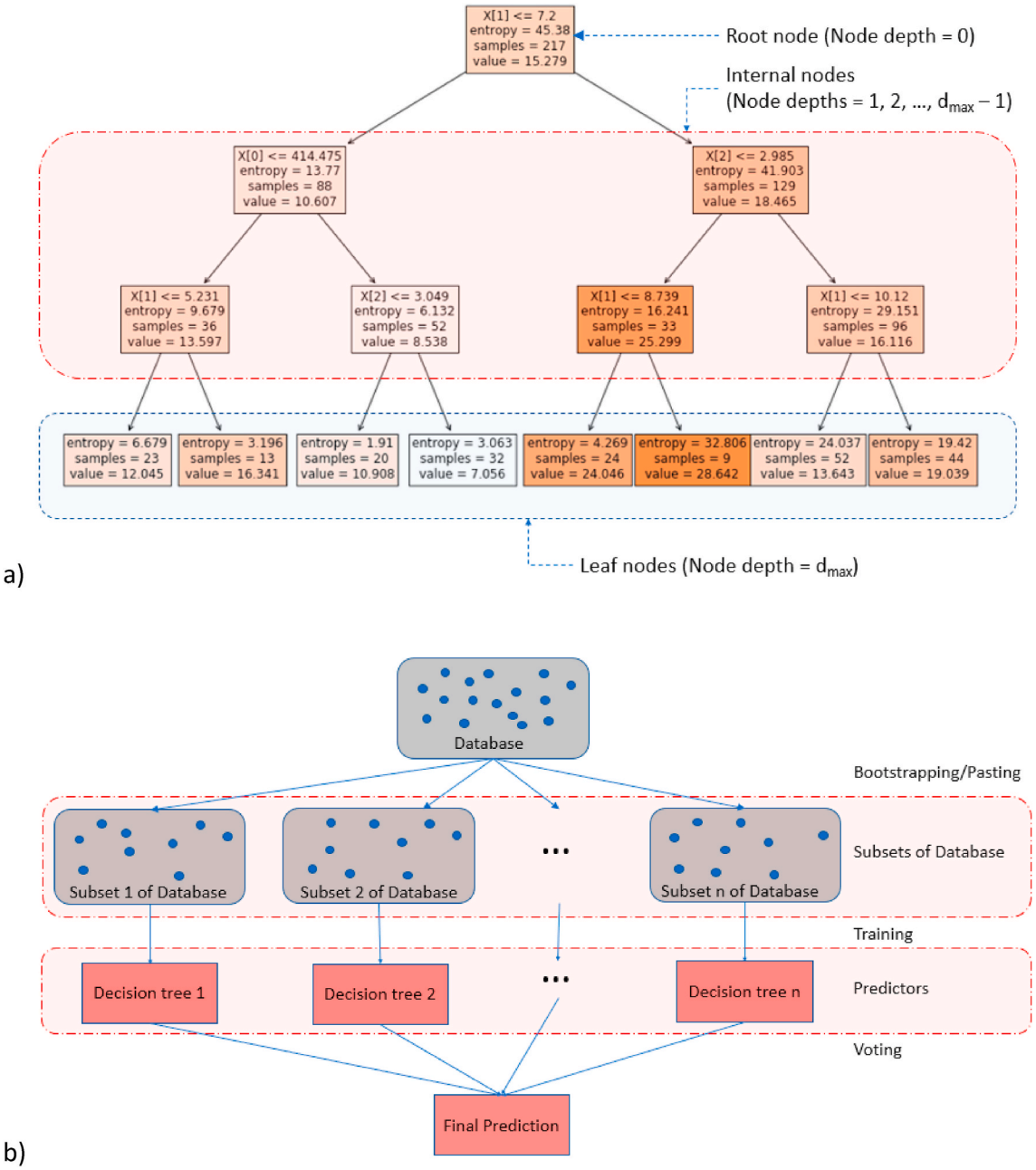

Fig. 3. Example of a) decision tree and b) Scheme of voting procedure in Random Forest (after [32]).

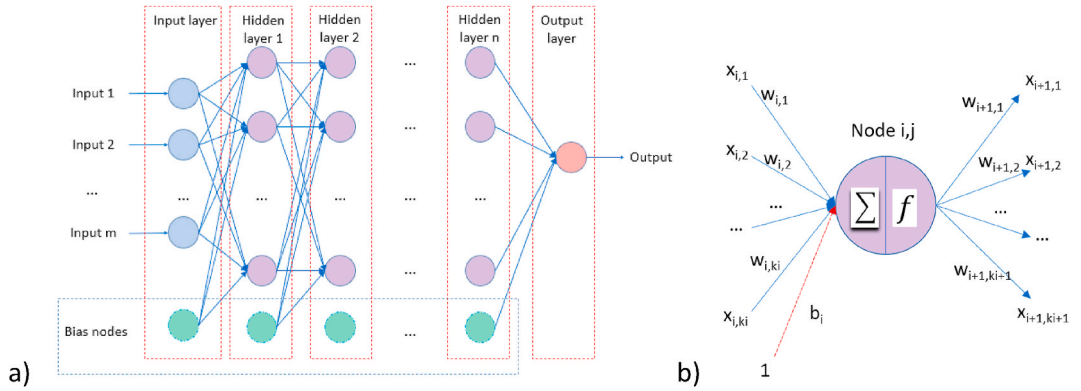

Fig. 4. a) A feedforward ANN and b) a given node in the hidden layer with corresponding weights.

3.2. Random forest regression (RFR)

Random Forest is an ensemble learning method [30], which is one of the most powerful machine learning models (Géron [31]) for the supervised problem. A decision tree in a Random Forest uses the Classification And Regression Tree algorithm, CART, to predict the interested variable with objective or the empirical risk function for regression problem as in Eq. (20) [31].

$$\min\left(J(k,\ t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right}\right) \tag{20}$$

$MSE_{left/right}$where: is the Mean Squared Error of the left/right subset; $m_{left/right}$ is the sample of the left/right subset.

$$MSE_{node} = \sum_{i \in node} \left(\overline{y}_{node} - y_i\right)^2 \tag{21}$$

$$\overline{y}_{node} = \frac{1}{m_{node}} \sum_{i \in node} y_i \tag{22}$$

Fig. 3a provides an illustration of the decision tree for the high-strength steel pipe database with all 217 samples. Starting with the root node, which has no children with a depth of 0, the children nodes have the depth of d+1 with d is the depth of their parent. The nodes without any children are defined as the leaf nodes. The decision tree in Fig. 3a has the maximum depth, $d_{\max} = 3$ with 8 leaf nodes.

The overall database is divided into various sub-databases randomly selected from the overall database. For each sub-database, a decision tree is established with the procedure shown in Fig. 3a. The random selection is the bootstrap aggregating with replacement. Without such replacement, the sampling is a pasting process. A voting process was then conducted to aggregate the predicted values of various decision trees in the random forest (Fig. 3b).

In this study, the number of decision trees, number of leaf nodes in the tree, and the maximum depth are investigated to obtain the best set of hyper-parameters.

3.3. Artificial Neural Networks (ANN)

An ANN with n hidden layers for m features or inputs are provided in Fig. 4. Data is consumed from the input layer with the number of node equals to m+1 with the first m nodes corresponding to the number of features and a bias node.

The jth node at the ith layer receives the signal from other nodes in the preceding layer by sets of signals/functions [$X_{i,1}$, $X_{j,2}$, …, $X_{i,ki}$, 1] and corresponding weights [$w_{i,1}$, $w_{j,2}$, …, $w_{i,ki}$, b_i] where k_i is the number of node in the (i-1)th layer, b_i is the weight of bias node in layer (i-1)th. The summation weighted signals of node (i,j), $x_{i,j}$, can be written as:

$$x_{i,j} = \sum_{q=1}^{ki} X_{i,q} \times w_{i,q} + b_i \tag{23}$$

This summation is then considered as the input of the activation function, f to obtain the signal of the node (i,j):

$$X_{i,j} = f\left(x_{i,j}\right) \tag{24}$$

The chosen activation function in Eq. (24) is the Rectifier function (i.e., Relu), which can be explicitly expressed as:

$$X_{i,j} = f_{\text{Relu}}\left(x_{i,j}\right) = \max\left(0,\ x_{i,j}\right) \tag{25}$$

The signal from the node (i,j) is then transmitted to nodes in the next layer. Generally, in the feedforward ANN, a data sample is taken in the input layer. Each node in the network received signals from the previous layer and transmits the signal to the next layer. In the end, the output layer receives the signal of the last hidden layer that yields the output results.

In the training process, the found output is the predicted value of that sample which contains an error or the difference with the labeled value corresponding to such sample. A set of errors of b samples between predicted and the "true" value is consequently computed as the loss function. The b number is well-known as the batch size. The loss function can be chosen from various options such as Binary Cross Entropy, Mean Squared of Error, Mean Absolute Error, Mean Absolute Percentage Error, Squared Hinge, etc. In this paper, for the regression model, the Mean Squared Error, MSE, is commonly chosen:

$$L_{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - f_i)^2 \tag{26}$$

Once the loss function for each batch is found, the backpropagation, whose function is to adjust the weights in the network to minimize the loss function, is implemented. The process continues until the database is consumed in e number of epochs.

3.4. Hyper-parameter tuning with grid search

There are various hyper-parameters in the machine learning models, which are selected within the training and predicting processes. For instance, the SVR critical hyper-parameters are C in Eq. (15) or type of kernel (e.g. polynomial or RBF). The parameters of interest for RFR model are maximum depth, $d_{\max}$, maximum leaf nodes, or the number of decision trees to satisfy the large number theorem. The number of nodes in a layer and the number of layers in the ANN are the common concerns.

Unfortunately, there is no strict rule to obtain the optimized hyper-parameter of the machine learning models because of its dependence on the training data. The practical approach to find the appropriate parameters of a Machine Learning model is conducting the grid search.

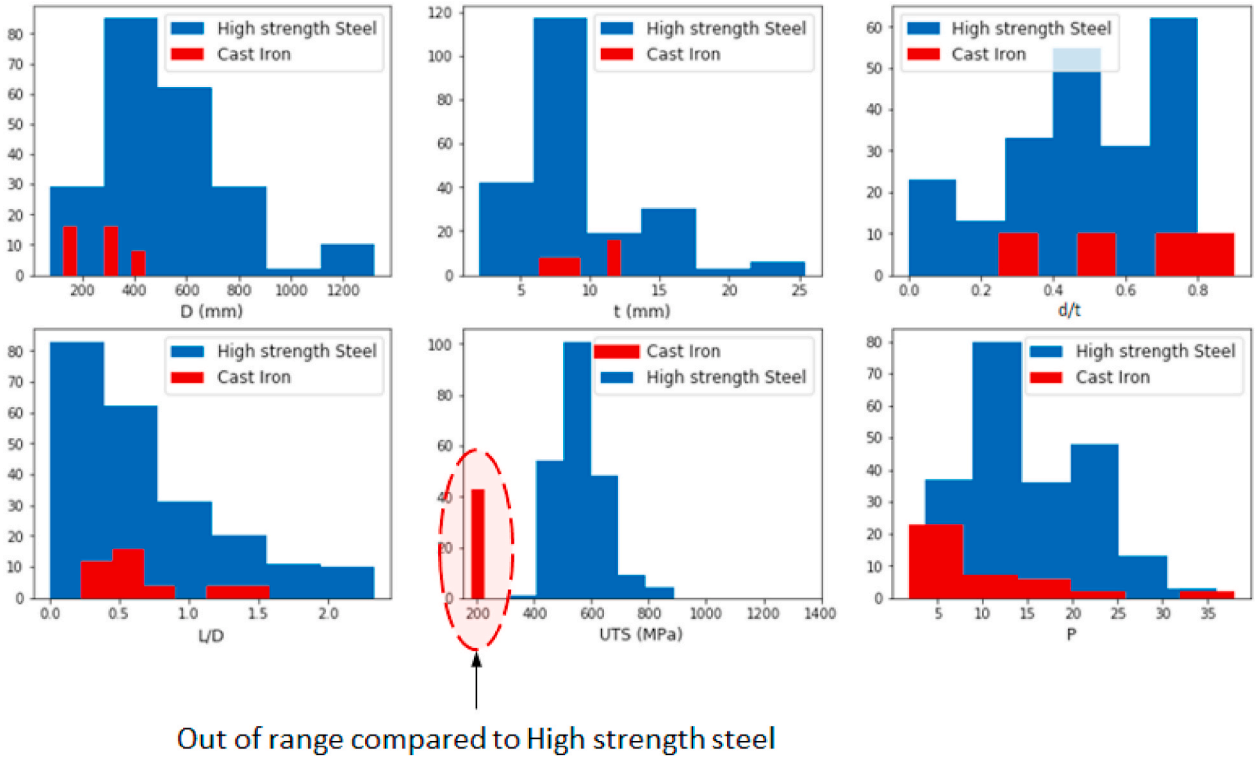

Fig. 5. Histograms of the High strength steel and Cast-Iron pipe database.

Table 2
Ranges of input parameters of the database.

Variable	Unit	Min	Mean	Max
D	mm	76.20	482.90	1320.00
t	mm	2.00	9.40	25.40
d	mm	0.00	4.42	15.41
L	mm	0.00	314.02	1432.56
σ_u	MPa	554.13	309.00	886.00

In the grid search, lists of potential hyper-parameters are predefined to obtain the list of a set of hyper-parameter via combination for searching. For instance, there are 2 hyper-parameter of interest in RFR: maximum depth and maximum leaf nodes, each of them has a list of candidates, such as maximum depths = [2–4] and maximum leaf nodes = [10,20, 30]. Combinations of the element in these lists are used for establishing the model. For example, RFR model 1 has maximum depth = 2 and maximum leaf nodes = 10; RFR model 2 has maximum depth = 1 and maximum leaf nodes = 20. The chosen model (selected from $3 \times 3 = 9$ models) is the one with the minimum Empirical Risk Function in Eq. (12) (MSE for this paper). For ANN, multiple hidden layers are assumed to have the same number of nodes, and the total number of nodes and number of weights for both single and multiple hidden layers are of interest. For the ANN with a single hidden layer, different batch sizes are also investigated.

4. Analysis and results

4.1. Databases and feature space

Two sets of database are used for the development and validation of the machine learning models. The first database is for high strength steel pipes that are collected from Ma et al. [33] (79 samples), Shuai et al. [34] (53 samples), Phan et al. [3] (28 samples), Freire et al. [35] (17 samples), Cronin [36] (40 samples) with a total of 217 samples. It includes a mixture of experimental and simulation data. The second dataset contains 50 samples for cast iron pipes obtained from Phan [18]. The database for the cast iron pipes was generated using FE simulations. The multiple histograms in Fig. 5 show the ranges of diameter and wall thickness as [76.2 mm, 1320 mm] and [2 mm, 25.4 mm], respectively, in the overall database. The range of the ratio of d/t is [0, 0.8] with a mean value of 4.796. The L/D ratio ranges from 0 to 2.34, with a mean value of 0.671. High strength steel includes API grade X42, X46, X52, X56, X60, X65, X80 and X100, with the Ultimate Tensile Strength (UTS) ranging from 309 MPa to 886 MPa. The experimental/simulated burst pressures range from 3.57 MPa to 35.97 MPa. These ranges of input variables provides the boundaries of valid inputs for a data-driven models. A summary of the ranges of parameters is provided in Table 2.

Table 3
Grids searches for machine learning models.

Model	Parameter	Grid	"Best" parameter	"Best" MSE
SVR	C kernel type gamma (for rbf) degree (for poly)	[0.001, 0.1, 1, 10, 100, 1000] poly, rbf [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] [1–6]	100 rbf 0.1 –	1.9656
RFR	Number of decision trees Max leaf nodes Max depth	[100, 1000, 10000] [4, 8, 16, 32, 64, 128, 256] [2–10]	100 64 10	4.3265
Single hidden layer ANN	Batch size Nodes in hidden layer	[10, 15, 20, 25, 30, 35, 40, 45, 50] [8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096]	20 2048	2.1519
Multiple hidden layers ANN	Total layers Nodes in hidden layers	[2–6] [4, 8, 16, 32, 64, 128, 256, 512, 1024]	3 32	2.2611

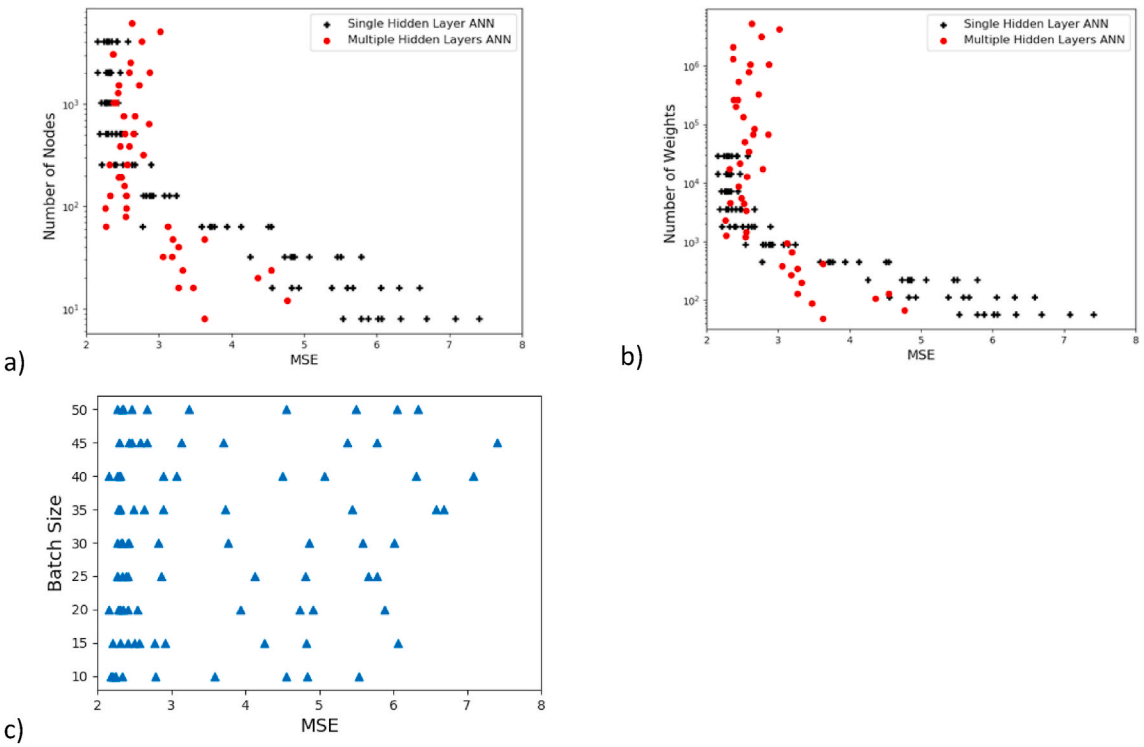

Fig. 6. Scatter plots of MSE versus a) Number of nodes in layer; b) Number of weights for single and multiple hidden layers ANN and c) batch size for single layer ANN.

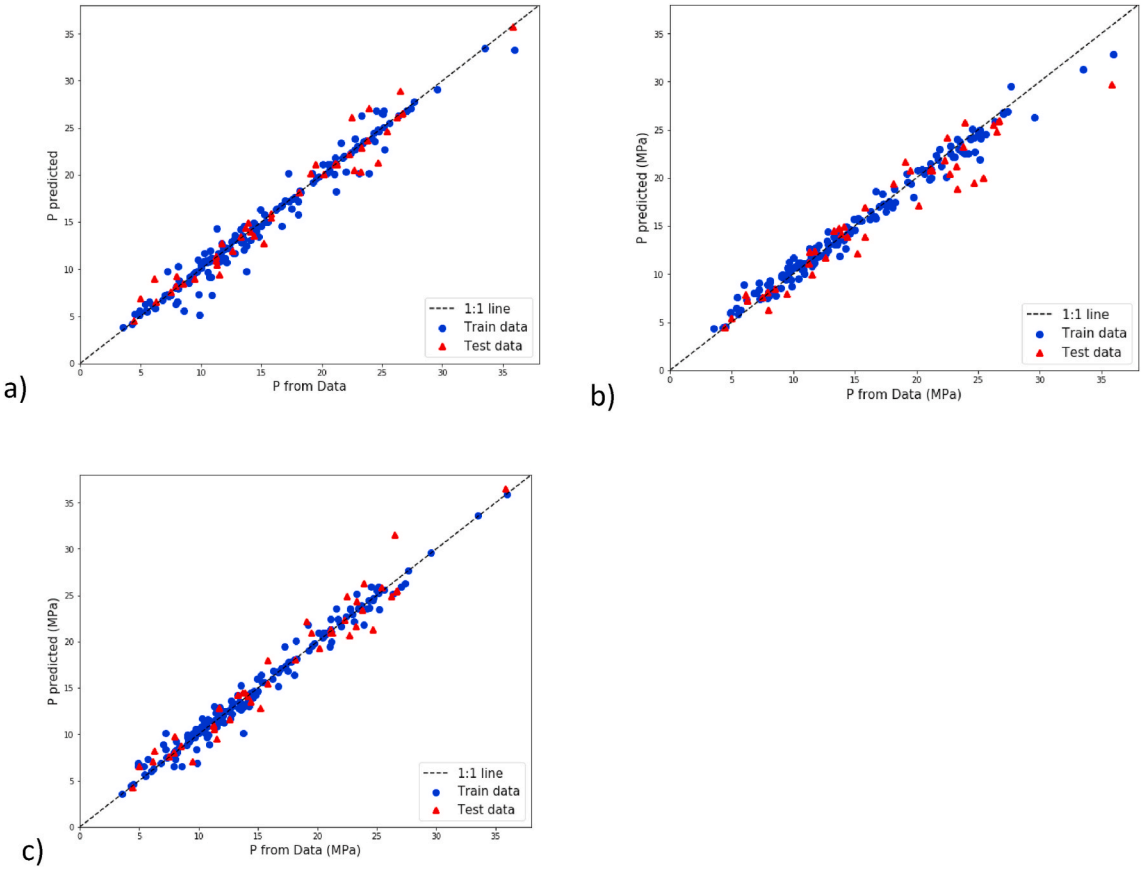

Fig. 7. Experiment/FEA versus predicted burst pressure with, a) SVR model; b) RFR model; c) ANN model.

Fig. 5 also shows that while all other parameters of the cast iron pipe database lie within the ranges of the high-strength steel pipe database, the UTS (or σ_u) of the cast iron is far below the UTS of the high-strength steel. Thus, the cast iron database is outside the boundary of high strength steel database in terms of UTS. The machine learning models were developed using the high-strength steel pipe database and were compared with the cast iron pipe database to examine the performance at the out-of-boundary.

4.2. Model development

Table 3 provides a summary of the grid search employed for the development of the Machine learning models. For the SVR model, the C hyper-parameter, the RBF and the Polynomial Kernel functions and corresponding hyper-parameters (gamma and degree, respectively) are investigated. The "best" or chosen set of parameters are obtained as $C = 100$, kernel type = RBF with gamma = 0.1 with the MSE of 1.9656. For the RFR model, the number of decision trees, maximum leaf nodes and maximum depth are investigated, and the "best" parameters are obtained as: number of decision trees = 100, maximum leaf nodes = 64, and maximum depth of tree = 10. The RFR model has MSE of 4.3265, which is much higher than the SVR model.

For the ANN model development, all trials were conducted with 1000 epochs. For the simple ANN model with a single hidden layer, the best performing model is not necessarily the one with the highest layer dimension. Instead, a model with the batch size and layer dimension of 20 and 2048, respectively, has the lowest MSE. The scatter plot of the layer dimension versus MSE in Fig. 6a shows a reduction of MSE with the increase of the number of nodes in the hidden layer until the layer dimension of 256 nodes. No further improvement is observed with the addition of nodes beyond 256 nodes.

Fig. 6c shows no correlation between the batch size and MSE, indicating that the batch size has no effect on improving the model. However, ANN models with the same configuration but different training processes may yield significantly different errors with the same test set, especially for the small networks with less than 100 nodes. This variation is narrowed with larger networks. The MSE of the best single hidden layer ANN model is 2.1519 on the test set, resulted from 2048 nodes and 20 samples of batch size.

As seen in Table 3, the ANN model with multiple hidden layers is not better than the single hidden layer ANN model. The MSE of the best model with multiple hidden layers is 2.2611 (occurred with 3 hidden layers, each containing 32 nodes), which is comparable to the MSE of 2.1519 for the best model with a single hidden layer. Fig. 6 (a and b) shows that the MSEs are the lowest for the number of nodes higher than about 100 and the number of weights higher than 1000, which are consistent for both multiple hidden layers ANN and single hidden layer ANN. For a higher number of nodes or weights, multiple hidden layers networks yield a wider range of MSE. In general, for a medium-small size database with hundreds of samples, as in this paper, the MSE of both ANN-based models are better than the RFR model but worse than the

Table 4
Validating models on testing data set of High strength steel database.

Group	Model	MSE	R^2	MAE
Machine Learning models	SVR	1.9656	0.9629	0.9263
	RFR	4.2127	0.9205	1.4775
	ANN	2.1519	0.9594	1.2311
Reference models	Netto et al. (2005)	6.2010	0.8829	2.0863
	Gajdoš and Šperl (2012)	7.5261	0.8579	2.1563
	ASME (2012)	4.2090	0.9205	1.5797
	Modified PCORRC (2004)	4.1398	0.9218	1.6026
	Phan et al. 1 (2017)	4.3153	0.9185	1.6927
	Phan et al. 2 (2017)	3.2820	0.9380	1.4444
	Phan et al. 3 (2017)	4.3949	0.9170	1.7515

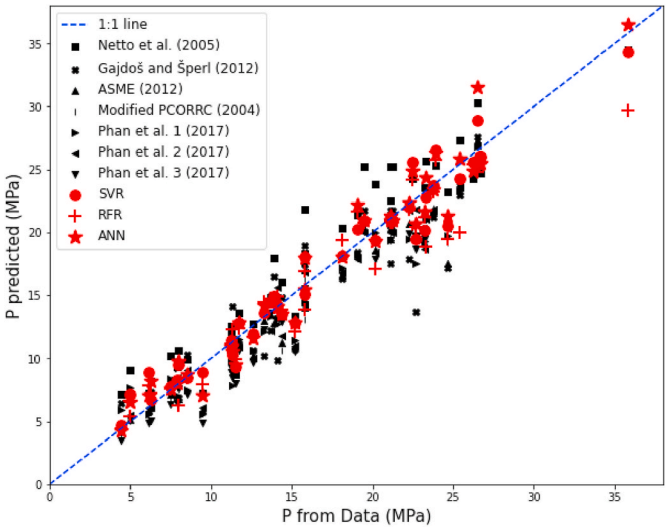

Fig. 8. Comparison of various burst pressure prediction models.

Table 5
Comparison of the models with Cast Iron pipe database.

Group	Model	MSE	R^2	MAE
Machine Learning models	SVR	54.8388	0.2477	4.8410
	RFR	87.6980	−0.2031	8.6329
	ANN	22.1662	0.6959	3.2486
Reference models	Netto et al. (2005)	3.8784	0.9468	1.6518
	Gajdoš and Šperl (2012)	5.7706	0.9208	1.5848
	ASME (2012)	3.5066	0.9519	1.2997
	Modified PCORRC (2004)	3.9479	0.9458	1.3423
	Phan et al. 1 (2017)	7.5026	0.8971	1.8922
	Phan et al. 2 (2017)	3.6147	0.9504	1.1755
	Phan et al. 3 (2017)	12.6398	0.8266	2.1540

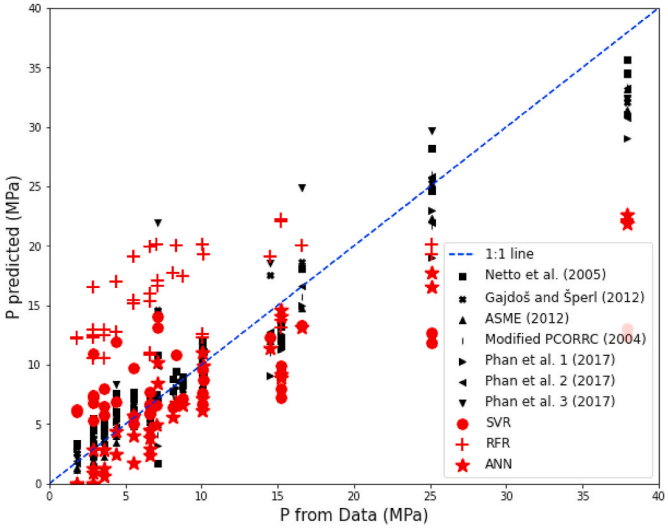

Fig. 9. Comparison of the model predictions with the cast iron pipe database.

SVR model. A single hidden layer ANN with 2048 nodes and 20 batch sizes is selected for the ANN model for further investigation.

4.3. Model comparison and validation

Fig. 7 shows the scatter plot of the predicted values against the data (results from simulations or experiments). The dashed lines present the perfect results when the predicted values are equal to the actual values (i.e., line 1:1). All three selected Machine Learning models successfully predict the burst pressures as all data points are closely concentrated around the 1:1 line. Overfittings of the models are not anticipated since data points from the testing data set (red triangles) narrowly scatter around the 1:1 lines while larger distances of data points from the 1:1 lines are also observed at some burst pressures. The scatter of the data from the 1:1 line is the highest for the RFR model, which is consistent with the higher MSE for the model (Table 3).

The machine learning models, along with other existing models, are validated with the data using three performance measures in Table 4. The performance measures are: the Mean Squared Error (Empirical Risk Function for all developed models), R square (R^2), and Mean of Absolute Error (MAE). Table 4 provides the performance measures of three machine learning models and seven existing models (discussed earlier) calculated using the testing dataset. As seen in the table, all machine learning models have relatively low MSE, R^2 and MAE values compare to the other models (called herein as the "reference model"). R^2 values of the machine learning models are consistently higher than 0.9205, with the maximum mean absolute error for the RFR model (i.e., 1.4775 MPa). The scatter plots for all ten models are presented in Fig. 8. As expected, the machine learning models are less scattered from the 1:1 line in the figure compared to the other models (reference models).

The database for cast iron pipe (Phan [18]) is compared against the predictions using different models in Table 5 and Fig. 9. Table 5 shows that the MSE and MAE are the highest for the machine learning models, identifying the drawbacks of the models in predicting the burst pressures for cast iron pipes. The lowest errors among the machine learning models were for the ANN model that has MSE and MAE of 3.2486 MPa and 22.1662 MPa, respectively, with $R^2 = 0.6959$. The MSE, MAE, and R^2 for the high-strength steel pipe database were 1.2311, 2.4728, and 0.9533, respectively, with the ANN model. The coefficients of determination with the SVR and RFR models were 0.2477 and negative 0.2031,

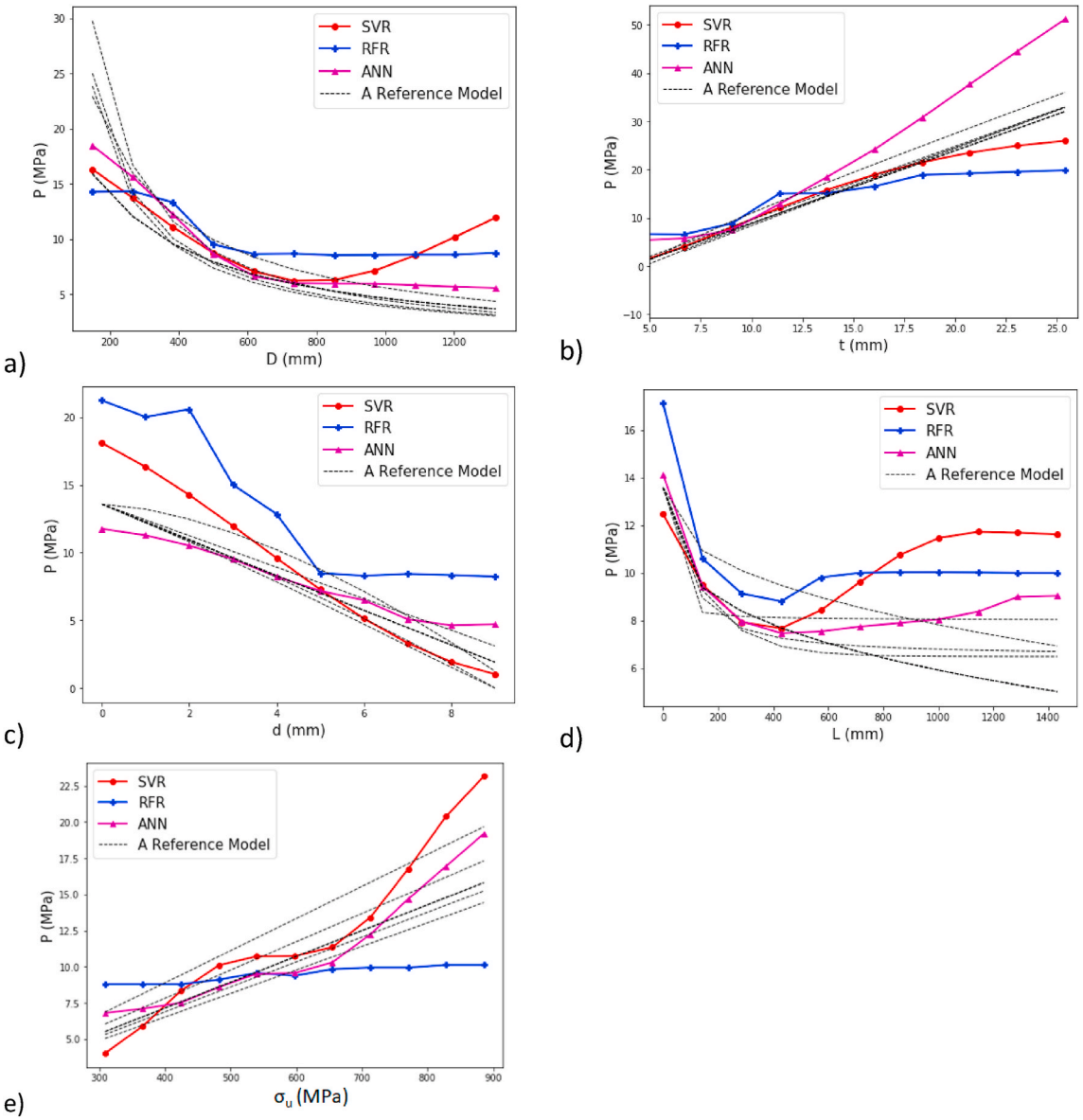

Fig. 10. Parametric study of the burst pressure problem with machine learning models.

respectively, which are very low and unacceptable. Fig. 9 illustrates the low accuracy of machine learning models where the data points are highly scattered around the 1:1 line. It implies that the data-driven models may not work outside the boundaries of data within which the models are developed. The developed models thus should clearly state the boundaries of their applicability to have proper predictions.

The other models (reference models) show stable R^2 values, consistently larger than 0.8266 (model 3 of Phan et al. [3]). Among these models, there are even some improvements in the performance measures. For example, the R^2 of the ASME [6] model increase from 0.9205 for the high-strength steel pipe database to 0.9519 for the cast iron database. Higher accuracy with these models is because the models are based on the theory of solid mechanics that is applicable regardless of materials.

4.4. A parametric study

A parametric study is conducted to examine the burst pressures calculated using various models for a range of parameters. For the parametric study, a control case with D = 550 mm, t = 9 mm, d = 4.7 mm, L = 500 mm, and σ_u = 415 MPa, is randomly chosen. The input variables of models are changed one-by-one, and their effects on the burst pressure are revealed, as shown in Fig. 10. The burst pressure predicted using SVM, RFR, and ANN models for an intact pipe (d = 0) are 18.1 MPa, 21.2 MPa, and 11.8 MPa, respectively. The burst pressure calculated using the analytical model (i.e., Eq. (2a)) is 13.6 MPa.

In Fig. 10, the burst pressures calculated using the reference models (dashed lines in the figure) show smooth changes with the parameters. The smooth input-output relationships are expected because the explicit equations are used in these models. However, the burst pressure calculated using the machine learning models shows fluctuations. In general, negative correlations of the burst pressures with the diameter, defect depth, and defect length are seen. The wall thickness and ultimate stress have a positive relationship with the burst pressure.

Note that the differences between the burst pressures predicted by the machine learning model and those from the reference models are larger near the boundary of the parameters considered. For example, the differences are more significant at the pipe diameter of 150 mm and 1400 mm, while the differences are less for the diameters of 300 mm–600 mm. At D = 150 mm, the differences in the predicted burst pressures are about 16 MPa. For diameters from 300 mm to 600 mm, differences are about 3–4 MPa.

In Fig. 10, the ANN model yields burst pressures that show smooth changes with the least fluctuations. The burst pressured predicted using the ANN model is also the closest to those calculated using the reference models except for larger wall thickness (Fig. 10b). Thus, the ANN model is more suitable for the prediction outside the boundaries of data. The burst pressures predicted using the RFR model show the highest fluctuations and often highest differences from those calculated using the other models. Unexpectedly, the σ_u versus burst pressure line in Fig. 10e calculated using the RFR model is parallel to the x-axis.

5. Conclusion

The paper focuses on investigating the applicability of data-driven models for burst pressure prediction with machine learning techniques. Models have been developed for the high-strength steel pipes with various inputs obtained from 5 databases available in the published literature. Three machine learning models, such as SVR, RFR, and ANN, are successfully developed, showing promising performance indicators. The SVR is found as the "best" model, followed by ANN and RFR models. The RFR model has the lowest R^2 value (0.9205) among the Machine Learning models developed. However, the performance of the machine learning models is found to be limited by the boundaries of input variables with which the models are developed. Consequently, the models developed with high-strength steel pipes data are not applicable for cast iron pipes.

The reference models developed based on the theory of solid mechanics showed lower predictability (higher MSE) than the machine learning models. However, these models performed better with the unfamiliar database (outside the data boundary used for the development of the model). As a result, these models yielded a better prediction of the burst pressure for cast iron pipes. Among the machine learning models, the ANN is found to perform better with the unfamiliar database, while SVR and RFR models performed poorly with the unfamiliar data. Therefore, the input boundaries should be carefully followed when applying the Machine Learning models in predicting the burst pressures.

The study reveals that even though the machine learning models can provide the general trend of the output variables, fluctuation can occur, particularly at the boundaries of the input variables. A database with wide ranges of input variables can be used to develop more comprehensive machine learning models. Future work can also focus on extending the machine learning models through interaction with other advanced techniques such as optimization or reliability analysis.

Funding

This research received no external funding.

Data availability

The data required to reproduce these findings will be made available on request.

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.

References

[1] 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.
[2] R. Amaya-Gómez, et al., Reliability assessments of corroded pipelines based on internal pressure–A review, Eng. Fail. Anal. 98 (2019) 190–214.
[3] 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.
[4] W. Kastner, et al., Critical crack sizes in ductile piping, Int. J. Pres. Ves. Pip. 9 (3) (1981) 197–219.
[5] L. Gajdos, M. Sperl, Evaluating the Integrity of Pressure Pipelines by Fracture Mechanics, Applied Fracture Mechanics, 2012, p. 283.
[6] A. ASME B31G, Manual for Determining the Remaining Strength of Corroded Pipelines, The American Society of Mechanical Engineers, New York, NY, 1991.
[7] F.J. Klever, G. Stewart, C.A. van der Valk, New Developments in Burst Strength Predictions for Locally Corroded Pipelines, American Society of Mechanical Engineers, New York, NY (United States), 1995.
[8] 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.
[9] T.-T. Le, H.C. Phan, Prediction of ultimate load of rectangular CFST columns using interpretable machine learning method, Adv. Civ. Eng. (2020) 2020.
[10] H.T. Duong, et al., Optimization design of rectangular concrete-filled steel tube short columns with Balancing Composite Motion Optimization and data-driven model, in: Structures, Elsevier, 2020.
[11] H.C. Phan, et al., An empirical model for bending capacity of defected pipe combined with axial load, Int. J. Pres. Ves. Pip. (2021) 104368.
[12] T.-T. Le, Practical Machine Learning-Based Prediction Model for Axial Capacity of Square CFST Columns, Mechanics of Advanced Materials and Structures, 2020, pp. 1–16.
[13] 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.
[14] 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.
[15] A. Zolfaghari, M. Izadi, Burst pressure prediction of cylindrical vessels using artificial neural network, J. Pressure Vessel Technol. 142 (3) (2020).
[16] D. Oh, et al., Burst pressure prediction of API 5L X-grade dented pipelines using deep neural network, J. Mar. Sci. Eng. 8 (10) (2020) 766.
[17] H.C. Phan, H.T. Duong, Predicting burst pressure of defected pipeline with principal component analysis and adaptive neuro fuzzy inference system, Int. J. Pres. Ves. Pip. 189 (2021) 104274.

[18] H.C. Phan, Development of novel methods for municipal water main infrastructure integrity management, in: Faculty of Engineering and Applied Science, Memorial University of Newfoundland: Canada, 2019.
[19] J. Kiefner, A. Duffy, Summary of Research to Determine the Strength of Corroded Areas in Line Pipe, Battelle Columbus Laboratories, 1971.
[20] W. Maxey, et al., Ductile fracture initiation, propagation, and arrest in cylindrical vessels, in: Fracture Toughness: Part II, ASTM International, 1972.
[21] Association, C.S., *Oil And Gas Pipeline Systems: CSA Z662-2015*. 2015, Toronto: Canadian Standards Association.
[22] N. Veritas, Corroded Pipelines: DNV Recommended Practice RP-F101, Det Norske Veritas, 1999, 1999.
[23] N. Wang, et al., Transient behaviors of loop heat pipes for alpha magnetic spectrometer cryocoolers, Appl. Therm. Eng. 68 (1–2) (2014) 1–9.
[24] D.R. Stephens, B.N. Leis, Development of an alternative criterion for residual strength of corrosion defects in moderate-to high-toughness pipe, in: 3rd International Pipeline Conference. 2000, American Society of Mechanical Engineers Digital Collection, 2000.
[25] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 2013.
[26] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992.
[27] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[28] V. Vapnik, Statistical Learning Theory, *John Wiley&Sons*. Inc., New York, 1998.
[29] A. Statnikov, A Gentle Introduction to Support Vector Machines in Biomedicine: Theory and Methods, vol. 1, world scientific, 2011.
[30] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, IEEE, 1995.
[31] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, 2019.
[32] T.D. Pham, et al., Predicting the reduction of embankment pressure on the surface of the soft ground reinforced by sand drain with random forest regression, in: IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2020.
[33] B. Ma, et al., Assessment on failure pressure of high strength pipeline with corrosion defects, Eng. Fail. Anal. 32 (2013) 209–219.
[34] 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.
[35] J. Freire, et al., Part 3: burst tests of pipeline with extensive longitudinal metal loss, Exp. Tech. 30 (6) (2006) 60–65.
[36] D.S. Cronin, Assessment of Corrosion Defects in Pipelines, 2000.