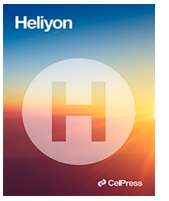
Heliyon 10 (2024) e26399

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Etemadi regression in chemometrics: Reliability-based procedures for modeling and forecasting

Sepideh Etemadi, Mehdi Khashei *

Department of Industrial and Systems Engineering, Isfahan University of Technology (IUT), Isfahan, 84156-83111, Iran

ARTICLE INFO

Keywords:
Forecasting and modeling processes
Chemometrics
Accuracy and reliability-based modeling strategies
Generalization capability
Multiple linear regression

ABSTRACT

The creation of predictive models with a high degree of generalizability in chemical analysis and process optimization is of paramount importance. Nonetheless, formulating a prediction model based on collected data from chemical measurements that maximize quantitative generalizability remains a challenging task for chemometrics experts. To tackle this challenge, a range of forecasting models with varying characteristics, structures, and capabilities has been developed, utilizing either accuracy-based or reliability-based modeling strategies. While the majority of models follow the accuracy-based approach, a recently proposed reliability-based approach, known as the Etemadi approach, has shown impressive performance across various scientific fields. The Etemadi models were constructed through a reliability-based parameter estimation process in such a manner that maximizes the models' reliability. However, the foundation of modeling procedures for chemometrics purposes is built upon the assumption that high generalizability in inaccessible/test data is achieved through the accuracy-based training procedure in which errors in available historical/training data are minimized. After conducting a thorough review of the current literature, we have found that none of the forecasting models for chemometrics purposes incorporate reliability into their modeling procedures. Given the dynamic and highly sensitive nature of chemistry experiments and processes, implementing a reliable model that controls performance criteria variation is a promising strategy for achieving stable and robust forecasts. To address this research gap, this paper introduces several key innovations, which can be highlighted as follows: **(1)** Proposing a general design structure based on a new optimal reliability-based parameter estimation process. **(2)** Introducing a novel risk-based modeling strategy that minimizes the performance variation of models implemented under different conditions in chemical laboratory experiments, to generate a more generalizable model for diverse applications in chemometrics. **(3)** Specifying the degree of influence that each reliability and accuracy factor has in enhancing the generalizability and uncertainty modeling of chemometric models. Empirical evidence confirms the effectiveness and superior performance of reliability-based models compared to accuracy-based models in 78.95% of the cases across various fields, including Pharmacology, Biochemistry, Agrochemical, Geochemical, Biological, Pollutants, Physicochemical Properties, and Gases Experiment. Furthermore, the study's findings demonstrate that the reliability-based modeling approach outperforms the accuracy-based strategy in terms of MAE, MSE, ARV, and RMSE by an average of 4.697%, 5.646%, 5.646%, and 4.342%, respectively. It is also statistically proven that reliability has a more significant impact on improving the generalizability of chemometric models than accuracy. This emphasizes the importance of including reliability as a crucial factor in chemometrics modeling, a consideration

* Corresponding author.
E-mail address: Khashei@cc.iut.ac.ir (M. Khashei).

https://doi.org/10.1016/j.heliyon.2024.e26399
Received 1 September 2023; Received in revised form 18 November 2023; Accepted 12 February 2024
Available online 15 February 2024
2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

that has been overlooked in traditional modeling processes. Consequently, reliability-based modeling approaches can be regarded as a viable alternative to conventional accuracy-based modeling methods for chemical modeling purposes.

1. Introduction

Chemometrics analysis demands the creation of prediction models capable of accurately generalizing and optimizing the understudy process. To achieve this objective, several machine-learning models have been developed within the categories of linear/nonlinear and statistical/intelligent methods. Regression models are recognized as a widely favored statistical approach in chemometrics for determining the correlation between a set of descriptive attributes and a target variable. Commonly utilized regression models encompass Multiple Linear Regression (MLR), Partial Least Square Regression (PLSR), Penalized Regression (PR), and Support Vector Regression (SVR) methods. These chemometrics regression techniques have diverse applications in fields such as Pharmacology [1], Biochemistry [2], Agrochemical [3], Geochemical [4], Biological [5], Toxicity Compound [6], Pollutants [7–10], Physicochemical Properties [11], and Chemical compounds of Food Additives and Ingredients [12] (see Tables 1–3).

From a technical perspective, research and studies in the field of chemometrics can be categorized into three main types: application papers, comparative studies, and techniques based on pre-processing and optimization. Pre-processing-based regression techniques aim to improve the predictive capability of models by either augmenting input data or using feature selection methods, while optimization-based algorithms fine-tune the model's parameters. For example, Courand et al. [13] enhanced the prediction performance of PLSR in forecasting grape berry sugar content by developing a robust version of PLSR capable of handling outliers. Similarly, Yang et al. [14] applied SVR to predict the nitrate-nitrogen isotopic composition and made efforts to improve the model's generalizability. In this regard, Principal Component Analysis (PCA) and Grid Search (GS) techniques were employed to reduce input variables and optimize the model's parameters, respectively. Researchers in the field of chemometrics regression have demonstrated various approaches to achieve highly accurate predictions. Chen et al. [15] applied MLR to estimate the quality of river water and detect pollution sources, employing the absolute principal component score as a feature selection technique. Arthur et al. [16] employed MLR to predict the anti-leukemia activities and toxicities of NCI anticancer compounds, utilizing the Genetic Algorithm (GA) for selecting appropriate features. Similarly, Yang et al. [17] applied MLR and GA to identify relevant input variables and analyze the physical and chemical properties of different materials. In contrast, Kaneko and Funatsu [18] enhanced the predictive accuracy of their chemometric model by implementing Cross-Validation (CV), Grid Search (GS), and theoretical techniques to determine the optimal hyperparameters of the model.

Regression models have proven their effectiveness in various specialized fields, as evidenced by numerous application papers. For example, Lago et al. [19] employed PLSR to estimate cation exchange capacity, while Brown [20] developed a regression model to determine solvent-air partitioning with acceptable precision. In the field of biochemical methane, Raposo et al. [21] used predictive regression models to specify the organic content of biomass samples, and the research concluded that accurate laboratory measurements are essential for reliable predictions. Barra et al. [22] achieved desirable accuracy in estimating the cetane number, one of the most critical quality parameters of diesel fuel, using PLSR. Naguib and Abdallah [23] assessed the predictive capabilities of PLSR in the quality control analysis of pharmaceutical dosage forms in two modes: with and without a UV cutoff area. Ávila et al. [24] applied high-performance regression techniques to estimate various physicochemical attributes associated with meat quality characteristics.

Table 1
Summarized information from the chemistry benchmarks obtained from Kaggle and UCI databases.

Scope	Number of cases	Title of datasets	Sample Size	Number of Attributes	Attribute Type
Pharmacology	2	Drug Consumptions (UCI), Pharmaceutical Drug Spending by countries	1036–1884	4–11	Real-Integer
Biochemistry	2	Chemical element abundances, Basal Metabolic Rate	280–500	4–19	Real-Integer
Agrochemical	2	Chemical Fertilizers, Agriculture & Weather	3084–10151	4–19	Real-Integer
Geochemical	3	Comprehensive database of Minerals, Geochemical Variations in Igneous Rocks–Mining, Multivariate Geochemical classification	1125–3907	9–81	Real-Integer
Biological	2	QSAR Bioconcentration classes, covid19 blood sample and biochemistry tests	779–1133	9–22	Real-Integer
Toxicity Compound	2	QSAR aquatic toxicity, QSAR fish toxicity	546–908	7–8	Real
Pollutants	3	Beijing Multi-Site Air-Quality Data, Shanghai air pollution and weather 2014–2021, Full Scale Waste Water Treatment Plant Data	1382–32907	15–18	Real-Integer
Physicochemical Properties	1	Physicochemical Properties of Protein Tertiary Structure	45730	9	Real
Gases Experiment	1	Gas sensor array temperature modulation	295679	19	Real
Chemical compounds of Food Additives & Ingredients	1	Wine Quality	1599	11	Real

Michanowicz et al. [25] designed a regression model to predict intra-urban NO2, demonstrating the versatility of regression models in diverse fields. Several studies have successfully utilized the MLR model in chemometrics. For instance, Adeniji et al. [26] accurately forecasted the activity of anti-tubercular compounds using MLR, while Xu et al. [27] applied the model to identify suitable areas for cultivating pharmaceutical plants. In a separate study, Nakamura et al. [28] utilized physical and chemical soil properties to predict the concentrations of cadmium, lead, and fluorine with the MLR model. Furthermore, Bertelkamp et al. [29] effectively forecasted the rates of organic micropollutants in the river bank filtration process using MLR, and Ewaid et al. [30] developed a precise MLR for predicting the quality of river water. Additionally, Olaya-Abril et al. [31] successfully estimated the distribution of soil organic carbon, while Palmer et al. [32] employed MLR to estimate the toxicity of samples in the municipal solid waste incineration process. Other successful applications of the MLR model in chemometrics include the identification of compounds responsible for aquatic contamination by Dieguez-Santana et al. [33] and the estimation of phenol removal from wastewater by Mandal et al. [34].

Comparative studies have consistently highlighted the superior performance of regression models in comparison to other machine learning models. In one study, More and Gupta [35] utilized MLR and Non-Linear Regression (NLR) to estimate the chromium removal efficiency in the cathode chamber of a bioelectrochemical system, and the findings revealed the superiority of MLR over NLR. Similarly, Abrougui et al. [36] discovered that MLR outperformed an Artificial Neural Network (ANN) when predicting organic potato crops based on soil chemical properties, a result that could be leveraged to select appropriate tillage techniques and enhance cultivation. Furthermore, Du et al. [37] evaluated the predictive performance of several regression models, which included GRA-MLR, PCA-MLR, and PLSR, combined with feature selection techniques, to estimate the physical and chemical properties of tea stems. The outcomes demonstrated that GRA-MLR outperformed the other two models. In a separate study conducted by Robert et al. [38], the aim was to predict the effectiveness of anticancer drugs. They employed three different models: MLR, modified MLR-Weighted Least Square (MLR-WLS), and enhanced MLR-WLS. Upon analyzing the results, they concluded that the enhanced MLR-WLS exhibited superior performance compared to the other two models. Hosseinzadeh et al. [39] assessed the accuracy of two different models, MLR and ANN, for estimating nutrient recovery from solid waste. Tang et al. [40] predicted the biodegradability of organic chemicals, a crucial parameter for evaluating the environmental persistence of chemicals, using two models, MLR and Support Vector Machines (SVM). Finally, Rendall et al. [41] evaluated the performance of four models, PR, MLR, latent variables regression, and tree-based ensemble methods, to forecast the age of the wine. Their findings indicated that the PR method outperformed the other models.

As previous research indicates, all predictive models in chemometrics are formulated with the aim of maximizing accuracy in historical and training data. The primary objective of these models is to attain a high level of generalizability in a manner that significantly enhances the quality of decision-making. Thus, despite the apparent differences among various modeling approaches, the basic foundation of all these models originates from an accuracy-based identical concept. Although, it seems that the accuracy-based strategy is a practical and logical approach for modeling and forecasting. However, in certain situations, especially when training and testing data do not follow a similar paradigm, it is incorrect to expect the results of the training data to repeat in the test data. Thus, this strategy loses its effectiveness. Additionally, it appears that employing an accuracy-based strategy to model chemistry processes that fall within the category of high-risk and sensitive situations with significant variations is not the most suitable approach. Therefore,

Table 2
The performance criteria in the accuracy- and Etemadi reliable-based models.

Category	Data Set	Etemadi reliability-based MLR				Classic accuracy-based MLR			
		MAE	MSE	RMSE	ARV	MAE	MSE	RMSE	ARV
1) Pharmacology	#1	0.539	0.460	0.678	0.469	0.540	0.462	0.680	0.471
	#2	30015	5679145441	75360	0.861	30097	5703973233	75525	0.865
2) Biochemistry	#3	4.E−11	3.E−21	5.E−11	3. E−15	2.E−10	5.E−20	2.E−10	5. E−14
	#4	0.490	0.346	0.588	0.172	0.488	0.344	0.587	0.171
3) Agrochemical	#5	15175	1008680530	31760	0.648	15294	1047776812	32369	0.673
	#6	0.152	0.048	0.220	0.0086	0.155	0.049	0.222	0.0088
4) Geochemical	#7	1.338	3.454	1.859	0.524	1.358	3.418	1.849	0.518
	#8	7.E−03	9.E−05	0.009	0.006	7.E−03	9.E−05	0.009	0.006
	#9	164.948	35267	187.794	0.642	165.585	35411	188.178	0.645
5) Biological	#10	0.571	0.598	0.773	0.568	0.570	0.595	0.771	0.565
	#11	0.343	0.203	0.451	1.224	0.339	0.206	0.454	1.241
6) Toxicity Compound	#12	0.945	1.625	1.275	0.553	0.950	1.637	1.279	0.557
	#13	0.601	0.633	0.795	0.409	0.597	0.627	0.792	0.406
7) Pollutants	#14	17.422	577.400	24.029	0.067	17.519	584.838	24.183	0.068
	#15	1.346	3.495	1.869	0.966	1.364	3.525	1.878	0.975
	#16	58.464	5084.358	71.305	0.959	60.492	5186.914	72.020	0.978
8) Physicochemical Properties	#17	1.596	5.521	2.350	0.153	1.616	5.583	2.363	0.155
9) Gases Experiment	#18	3.446	22.269	4.719	0.387	3.512	22.804	4.775	0.397
10) Chemical compounds of Food Additives and Ingredients	#19	0.510	0.463	0.680	0.696	0.509	0.459	0.678	0.691

estimating the model's parameters using such a method may not result in a generalizable model.

On the other hand, a novel class of modeling methodology based on the concept of reliability, known as Etemadi[1] models, has recently been introduced for causal forecasting, classification, and time series prediction. In other words, the Etemadi modeling approach is a general methodology that can be implemented on all supervised and unsupervised machine learning methods [42]. These proposed models were constructed using a reliability-based parameter estimation process in a such manner that the models' reliability is maximized instead of its accuracy. The underlying notion of the reliability-based modeling procedure is minimizing the performance variation of models implemented on validation data in order to yield a generalizable model [43,44]. The effectiveness and superiority of generalizability of reliability-based strategy against accuracy-based one, especially in high-risk and highly sensitive domains like as medicine were demonstrated [45]. Additionally, the reliability-based approach averagely outperformed the accuracy-based method in finance, energy, environment, management, transportation, and engineering scopes [46].

Furthermore, some studies directed their attention to assessing the reliability of complex systems. In their research, Li *et al.* [47] integrated the multivariate ensemble model with the hierarchical linkage technique to enhance the precision and efficiency of evaluating system reliability for aeroengine cooling blades. Their study highlighted the superiority of the proposed model over boosting trees, multivariate ensembles, and ANNs. In a different research, Li *et al.* [48] created a physics-informed distributed modeling approach using Extreme Gradient Boosting (XGBoost) for aeroengine rotor systems that have various failure modes and numerous vulnerable locations. Their model demonstrated superior performance compared to SVR, XGBoost, and Multi-Layer Perceptron (MLP). Zhang *et al.* [49] introduced an active extremum Kriging-based multi-level linkage approach that employs active learning techniques to identify the most suitable training samples. They also incorporated a multi-level linkage strategy to consider failure correlations, all aimed at constructing a reliability framework for intricate, dynamic multi-component systems, including aeroengine mechanism systems. Li *et al.* [50] devised a stratified strategy based on deep learning regression to address the challenges associated with correlated relationships and high nonlinearity in the damage assessment of probabilistic combined cycle fatigue. They confirmed the efficacy of their approach through testing on a typical turbine bladed disk, accounting for various uncertainties, including material variations, model uncertainties, and load fluctuations. In another investigation, Roy and Chakraborty [51] conducted a review of the applications of support vector algorithms in structural reliability analysis, considering factors such as computational expenses, the order of failure probability, and dimensionality. In the study by Mazhar *et al.* [52], various machine learning-based charge

Table 3
The achieved improvement by the Etemadi reliable-based model compared to the accuracy-based version.

Category	Data Set	Improvement (%)			
		MAE	MSE	ARV	RMSE
1) Pharmacology	#1	0.230	0.403	0.403	0.202
	#2	0.272	0.435	0.435	0.218
Average		**0.251**	**0.419**	**0.419**	**0.210**
2) Biochemistry	#3	78.712	94.166	94.166	75.852
	#4	−0.285	−0.604	−0.604	−0.302
Average		**39.214**	**46.781**	**46.781**	**37.775**
3) Agrochemical	#5	0.774	3.731	3.731	1.883
	#6	1.937	2.250	2.250	1.131
Average		**1.356**	**2.991**	**2.991**	**1.507**
4) Geochemical	#7	1.433	−1.073	−1.073	−0.535
	#8	−0.680	0.040	0.040	0.058
	#9	0.384	0.407	0.407	0.204
Average		**0.379**	**−0.209**	**−0.209**	**−0.091**
5) Biological	#10	−0.301	−0.516	−0.516	−0.258
	#11	−1.248	1.376	1.376	0.690
Average		**−0.775**	**0.430**	**0.430**	**0.216**
6) Toxicity Compound	#12	0.493	0.733	0.733	0.367
	#13	−0.643	−0.856	−0.856	−0.427
Average		**−0.075**	**−0.062**	**−0.062**	**−0.030**
7) Pollutants	#14	0.551	1.272	1.272	0.638
	#15	1.313	0.866	0.866	0.434
	#16	3.352	1.977	1.977	0.994
Average		**1.739**	**1.372**	**1.372**	**0.689**
8)Physicochemical Properties	#17	1.242	1.108	1.108	0.555
Average		**1.242**	**1.108**	**1.108**	**0.555**
9) Gases Experiment	#18	1.882	2.347	2.347	1.180
Average		**1.882**	**2.347**	**2.347**	**1.180**
10) Chemical compounds of Food Additives and Ingredients	#19	−0.173	−0.792	−0.792	−0.395
Average		**−0.173**	**−0.792**	**−0.792**	**−0.395**
Average		**4.697 (0.02)***	**5.646 (0.02)***	**5.646 (0.02)***	**4.342 (0.03)***

[1] Etemadi in the Persian is equivalent to the reliability.

management systems were examined, and the findings revealed that Long Short-Term Memory (LSTM) outperformed other models in enhancing the reliability and sustainability of the transportation system.

Based on the current literature review, it is evident that no forecasting models in chemometrics integrate the reliability concept into the modeling process. However, considering the dynamic and highly sensitive nature of chemistry experiments and processes, the use of a reliable model capable of controlling performance variation is an effective approach to attain stable and generalizable forecasts. To address this research gap, in this study, the Etemadi reliability-based modeling strategy is implemented for chemometrics purposes. In addition, the performance of the reliability-based modeling strategy is evaluated and compared with its accuracy-based counterpart. The main goal of this study is to assess the effectiveness of the reliability factor on the model's generalizability in chemometrics domains. On the other hand, the level of modeling complexity has a significant impact on the generalizability of models, and this impact varies depending on the model's linearity, non-linearity, statistical or intelligent nature, shallowness or depth, and single or hybrid structure. To minimize the effect of complexity on generalizability, this paper focuses on using a linear regression model to implement the reliability-based modeling strategy. Linear regression is the most popular and commonly used model in chemometrics. Moreover, this paper examines the influence of both reliability and accuracy factors on the quality of decision-making in diverse applications of chemometrics. To do so, we consider 19 benchmark datasets across various fields including Pharmacology, Biochemistry, Agrochemical, Geochemical, Biological, Toxicity Compound, Pollutants, Physicochemical Properties, Gases Experiment, and Chemical compounds of Food Additives and Ingredients. We assess the generalizability of the reliability-based modeling strategy in comparison with the accuracy-based modeling strategy. In summary, the novelty of this paper can be outlined as follows.

1) Proposing a new optimal formulation for the reliability-based parameter estimation process.
2) Applying Etemadi's reliability-based modeling strategy in various chemometrics applications.
3) Evaluating the performance of the reliability-based model, comparing it with the accuracy-based model, and specifying the influence degree of each reliability and accuracy factor on the generalizability of the models.

The rest of the paper is structured as follows: Section 2 poses the problem statement, followed by the presentation of the reliability-based parameter estimation process in Section 3. Section 4 offers a description of the chemistry benchmark datasets considered in the study. Section 5 evaluates and compares the performance of the reliability-based modeling strategy against the accuracy-based one. The discussion, conclusions, and future research potentials are expressed in Sections 6 and 7.

2. Problem statement

In the field of chemometrics, researchers often use a range of statistical and machine-learning models for descriptive and predictive purposes in chemistry, which is inherently an experimental science. In descriptive applications, chemical systems are modeled to learn the basic relationships and structure of the experiments and processes. In predictive applications, chemical systems are modeled to forecast new properties or behaviors of interest. There is currently a significant amount of interest among scholars in the development of novel modeling approaches in the field of chemometrics, which is an application-driven science. Modeling approaches in chemometrics for predictive purposes are developed with the goal of maximizing the generalizability of the model. In other words, the principal objective of statistical and intelligent modeling methods is to achieve high accuracy in unavailable and test data using a generalizable model. To this end, researchers typically focus on two main categories of modeling approaches: those that aim to maximize the accuracy of training data and those that prioritize maximizing the reliability of the model. These two approaches involve distinct strategies for designing parameter estimation processes and cost functions in the learning procedures of statistical and intelligent models. Accuracy-based models are established by maximizing performance metrics on the training data. In contrast, the reliability-based approach is developed by minimizing changes in performance metrics on validation data.

The dominant number of the parameter estimation processes are designed based on maximizing the accuracy in the training data. The Ordinary Least Square (OLS), being the most popular and widely used technique for parameter estimation, is established based on the aforementioned concept and falls into the category of accuracy-based methods. In the OLS method, the parameters of the model are estimated in a way that minimizes the sum of squared differences between the actual and predicted values of the target variable. The ultimate goal of this process is to create a model that can be generalized to test data that is inaccessible during the training phase. In other words, the fundamental assumption of such a procedure is that maximizing accuracy in the training data leads to the maximum generalizability of the model in the test set. While this is a logical and practical strategy, it is not the only one that guarantees high generalizability. On the other hand, previous research has indicated that, in addition to accuracy, reliability is another critical influential factor in the model's generalizability. Reliability is defined as confidence in the model's accuracy, ensuring that the desired results can be reproduced [42]. From another standpoint, in situations where the risks are high and there is a high degree of variability, achieving stable and reliable results is of utmost importance [43,44]. Consequently, to address the changeable nature and high sensitivity of chemistry experiments and processes, this paper applies the concept of reliability to statistical modeling processes. A new formulation of the reliability-based parameter estimation procedure for chemometrics issues is proposed, aiming to provide stable and generalizable forecasts.

3. Reliability-based parameter estimation process

Among chemometrics models, the linear regression model is prevalent and extensively used as a linear statistical method for discovering the relationship between a set of descriptive attributes and a target variable. Specifically, forecasting and analysis are two

principal purposes for utilizing this type of model in chemometrics issues. Typically, the implementation procedure of the regression model comprises three steps as follows: **1)** Feature Selection: Initially selecting the relevant explanatory variables that describe the target variable. **2)** Designing Structure: Specifying the parameter estimation process to determine the coefficients of independent variables in the model. **3)** Forecasting or Analyzing: Ultimately, the regression model can be employed to survey and evaluate the impact rate of each descriptive attribute on the target variable or forecast the desired value of the dependent variable based on specified values of independent variables.

Generally, a k-variable linear regression model involving the dependent variable Y and $k-1$ independent or explanatory variables $X_2, X_3, ..., X_k$ can be written as follows:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + ... + \beta_k X_{kt} + u_t \qquad t = 1, 2, 3, ..., N + n + N^{'} \tag{1}$$

where, β_1 is the intercept, β_2 to β_k are the partial slope coefficients, u is the stochastic disturbance term, and $N + n + N^{'}$ is the total sample size.

The reliability-based modeling approach differs from the accuracy-based strategy by emphasizing the importance of reliability in maximizing a model's generalizability and minimizing uncertainty. As a result, in this type of modeling approach, the emphasis is on maximizing the model's reliability rather than its accuracy. The underlying idea of the reliability-based modeling procedure is estimating the parameters of the model in a way that the performance changes of the implemented models across different data points are minimized. Based on this, in the Etemadi modeling approach, the uncertainty in the parameter estimation process is considered and quantified, aiming to minimize the model's uncertainty. In other words, the concept of reliability plays a crucial role in designing the parameter estimation procedure to minimize the variance of the performance criterion.

The reliability-based modeling process begins by splitting the dataset into three parts: training, validation, and testing, which respectively contain N, n, and N' data points. After that, the validation set is subdivided into n parts. In the first stage, the initial validation data point is added to the training set, and the mean square of differences between the actual and predicted values for this data point is calculated. This process is repeated for the second and subsequent validation data points until all validation data have been included in the training set. Consequently, the final mean square error is calculated based on the training data and all validation data. To achieve the most reliable regression line according to the reliability-based modeling strategy, the mean square errors at each stage based on the validation data points should be roughly equal to each other.

$$MSE_j \cong MSE_{j^{'}} \quad \forall j, j^{'} \quad j \neq j^{'} \tag{2}$$

This can be summarized as follows:

$$nMSE_1 \cong MSE_1 + MSE_2 + MSE_3 + \ldots + MSE_n \tag{3}$$

By substituting the actual and predicted values into Eq. (3), Eq. (4) can be represented in the following manner:

$$\frac{n}{N}\sum_{t=1}^{N}\left(Y_t - \sum_{i=1}^{k}\widehat{\beta}_{1i}X_{it}\right)^2 \cong \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}\left(Y_t - \sum_{i=1}^{k}\widehat{\beta}_{ji}X_{it}\right)^2 \tag{4}$$

The reliability-based regression can be constructed, by minimizing the differences between each pair of mean square errors across all validation data points. The mathematical formulation of this concept is shown in Eq. (5).

$$Min \quad f(\widehat{\beta}_{ji}) = \frac{n}{N}\sum_{t=1}^{N}\left(Y_t - \sum_{i=1}^{k}\widehat{\beta}_{1i}X_{it}\right)^2 - \sum_{j=0}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}\left(Y_t - \sum_{i=1}^{k}\widehat{\beta}_{ji}X_{it}\right)^2 \tag{5}$$

As a consequence, to obtain a reliable regression model with the lowest uncertainty, the unknown parameters of all regression lines must be substituted with reliability-based parameters, as shown in Eq. (6).

$$\widehat{\beta}_{ji} = \widehat{\beta}e_i \quad j = 1, 2, ..., n \quad i = 1, 2, ..., k \tag{6}$$

Where, $\widehat{\beta}e_i$ is the ith parameter of the reliability-based regression. Accordingly, by partially differentiating Eq. (7) with respect to $\widehat{\beta}e_1$, $\widehat{\beta}e_2, \widehat{\beta}e_3, ..., \widehat{\beta}e_k$ according to Eq. (8) and setting the result to zero, k simultaneous equations with k unknown coefficients are yielded, as shown in Eq. (9).

$$Min \quad f(\widehat{\beta}e_i) = \frac{n}{N}\sum_{t=1}^{N}\left(Y_t - \sum_{i=1}^{k}\widehat{\beta}e_i X_{it}\right)^2 - \sum_{j=0}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}\left(Y_t - \sum_{i=1}^{k}\widehat{\beta}e_i X_{it}\right)^2 \tag{7}$$

$$\frac{\partial f}{\partial \beta e_1} = \frac{n}{N}(N\hat{\beta}e_1 + \hat{\beta}e_2\sum_{t=1}^{N}X_{2t} + ... + \hat{\beta}e_k\sum_{t=1}^{N}X_{kt}) - \frac{1}{N}(N\hat{\beta}e_1 + \hat{\beta}e_2\sum_{t=1}^{N}X_{2t} + ... + \hat{\beta}e_k\sum_{t=1}^{N}X_{kt})$$
$$-\frac{1}{N+1}((N+1)\hat{\beta}e_1 + \hat{\beta}e_2\sum_{t=1}^{N+1}X_{2t} + ... + \hat{\beta}e_k\sum_{t=1}^{N+1}X_{kt}) - ... - \frac{1}{N+n}((N+n)\hat{\beta}e_1 + \hat{\beta}e_2\sum_{t=1}^{N+n}X_{2t} + ...$$
$$+\hat{\beta}e_k\sum_{t=1}^{N+n}X_{kt}) = \frac{n}{N}\sum_{t=1}^{N}Y_t - \frac{1}{N}\sum_{t=1}^{N}Y_t - \frac{1}{N+1}\sum_{t=1}^{N+1}Y_t - ... - \frac{1}{N+n}\sum_{t=1}^{N+n}Y_t$$

$$\frac{\partial f}{\partial \beta e_2} = \frac{n}{N}(\hat{\beta}e_1\sum_{t=1}^{N}X_{2t}X_{1t} + \hat{\beta}e_2\sum_{t=1}^{N}X_{2t}^2 + ... + \hat{\beta}e_k\sum_{t=1}^{N}X_{2t}X_{kt}) - \frac{1}{N}(\hat{\beta}e_1\sum_{t=1}^{N}X_{2t}X_{1t} + \hat{\beta}e_2\sum_{t=1}^{N}X_{2t}^2 + ...$$
$$+\hat{\beta}e_k\sum_{t=1}^{N}X_{2t}X_{kt}) - \frac{1}{N+1}(\hat{\beta}e_1\sum_{t=1}^{N+1}X_{2t}X_{1t} + \hat{\beta}e_2\sum_{t=1}^{N+1}X_{2t}^2 + ... + \hat{\beta}e_k\sum_{t=1}^{N+1}X_{2t}X_{kt}) - ... - \frac{1}{N+n}(\hat{\beta}e_1\sum_{t=1}^{N+n}X_{2t}X_{1t}$$
$$+\hat{\beta}e_2\sum_{t=1}^{N+n}X_{2t}^2 + ... + \hat{\beta}e_k\sum_{t=1}^{N+n}X_{2t}X_{kt}) = \frac{n}{N}\sum_{t=1}^{N}X_{2t}Y_t - \frac{1}{N}\sum_{t=1}^{N}X_{2t}Y_t - \frac{1}{N+1}\sum_{t=1}^{N+1}X_{2t}Y_t - ... - \frac{1}{N+n}\sum_{t=1}^{N+n}X_{2t}Y_t$$

.
.
.

$$\frac{\partial f}{\partial \beta e_k} = \frac{n}{N}(\hat{\beta}e_1\sum_{t=1}^{N}X_{kt}X_{1t} + \hat{\beta}e_2\sum_{t=1}^{N}X_{kt}X_{2t} + ... + \hat{\beta}e_k\sum_{t=1}^{N}X_{kt}^2) - \frac{1}{N}(\hat{\beta}e_1\sum_{t=1}^{N}X_{kt}X_{1t} + \hat{\beta}e_2\sum_{t=1}^{N}X_{kt}X_{2t} + ...$$
$$+\hat{\beta}e_k\sum_{t=1}^{N}X_{kt}^2) - \frac{1}{N+1}(\hat{\beta}e_1\sum_{t=1}^{N+1}X_{kt}X_{1t} + \hat{\beta}e_2\sum_{t=1}^{N+1}X_{kt}X_{2t} + ... + \hat{\beta}e_k\sum_{t=1}^{N+1}X_{kt}^2) - ... - \frac{1}{N+n}(\hat{\beta}e_1\sum_{t=1}^{N+n}X_{kt}X_{1t}$$
$$+\hat{\beta}e_2\sum_{t=1}^{N+n}X_{kt}X_{2t} + ... + \hat{\beta}e_k\sum_{t=1}^{N+n}X_{kt}^2) = \frac{n}{N}\sum_{t=1}^{N}X_{kt}Y_t - \frac{1}{N}\sum_{t=1}^{N}X_{kt}Y_t - \frac{1}{N+1}\sum_{t=1}^{N+1}X_{kt}Y_t - ... - \frac{1}{N+n}\sum_{t=1}^{N+n}X_{kt}Y_t$$

Therefore, we have that

$$\widehat{\beta}e_2\left(\frac{n-1}{N}\sum_{t=1}^{N}X_{2t} - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{2t}\right) + ... + \widehat{\beta}e_k\left(\frac{n-1}{N}\sum_{t=1}^{N}X_{kt} - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{kt}\right) = \frac{n-1}{N}\sum_{t=1}^{N}Y_t - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}Y_t$$
$$\widehat{\beta}e_1\left(\frac{n-1}{N}\sum_{t=1}^{N}X_{2t}X_{1t} - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{2t}X_{1t}\right) + ... + \widehat{\beta}e_k\left(\frac{n-1}{N}\sum_{t=1}^{N}X_{2t}X_{kt} - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{2t}X_{kt}\right) = \frac{n-1}{N}\sum_{t=1}^{N}X_{2t}Y_t - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{2t}Y_t$$
. (9)
.
.
$$\widehat{\beta}e_1\left(\frac{n-1}{N}\sum_{t=1}^{N}X_{kt}X_{1t} - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{kt}X_{1t}\right) + ... + \widehat{\beta}e_k\left(\frac{n-1}{N}\sum_{t=1}^{N}X_{kt}^2 - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{kt}^2\right) = \frac{n-1}{N}\sum_{t=1}^{N}X_{kt}Y_t - \sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{kt}Y_t$$

in matrix form, it can be represented as follows:

$$\begin{bmatrix} 0 & A(X_{2t}) - B(X_{2t}) & ... & A(X_{kt}) - B(X_{kt}) \\ A(X_{2t}X_{1t}) - B(X_{2t}X_{1t}) & A(X_{2t}^2) - B(X_{2t}^2) & ... & A(X_{2t}X_{kt}) - B(X_{2t}X_{kt}) \\ ... & ... & ... & ... \\ A(X_{kt}X_{1t}) - B(X_{kt}X_{1t}) & A(X_{kt}X_{2t}) - B(X_{kt}X_{2t}) & ... & A(X_{kt}^2) - B(X_{kt}^2) \end{bmatrix} \begin{bmatrix} \widehat{\beta}e_1 \\ \widehat{\beta}e_2 \\ ... \\ \widehat{\beta}e_k \end{bmatrix} = \begin{bmatrix} A(Y_t) - B(Y_t) \\ A(X_{2t}Y_t) - B(X_{2t}Y_t) \\ ... \\ A(X_{kt}Y_t) - B(X_{kt}Y_t) \end{bmatrix} \quad (10)$$

For $i, i^{'} = 1, 2, ..., k$, we have that:

$$\frac{n-1}{N}\sum_{t=1}^{N}X_{it}=A(X_{it}),\frac{n-1}{N}\sum_{t=1}^{N}X_{it}X_{i't}=A(X_{it}X_{i't}),\frac{n-1}{N}\sum_{t=1}^{N}X_{it}^{2}=A\left(X_{it}^{2}\right)$$

$$\sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{it}=B(X_{it}),\sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{it}X_{i't}=B(X_{it}X_{i't}),\sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{it}^{2}=B\left(X_{it}^{2}\right)$$

$$\frac{n-1}{N}\sum_{t=1}^{N}Y_{t}=A(Y_{t}),\frac{n-1}{N}\sum_{t=1}^{N}X_{it}Y_{t}=A(X_{it}Y_{t}),$$

$$\sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}Y_{t}=B(Y_{t}),\sum_{j=1}^{n}\sum_{t=1}^{N+j}\frac{1}{N+j}X_{it}Y_{t}=B(X_{it}Y_{t}),$$

in this manner, the reliability-based cost function is mathematically formulated for regression as a linear statistical model. By solving Eq. (10), the unknown coefficients of each explanatory variable in the reliability-based regression can be determined. For example, in a 2-variable regression model, the reliability-based parameters can be obtained as follows:

$$\widehat{\beta}e_{1}=\frac{(A(X_{2t}Y_{t})-B(X_{2t}Y_{t}))-\left(\frac{A(Y_{t})-B(Y_{t})}{(A(X_{2t})-B(X_{2t}))}\right)\left(A\left(X_{2t}^{2}\right)-B\left(X_{2t}^{2}\right)\right)}{(A(X_{2t}X_{1t})-B(X_{2t}X_{1t}))} \tag{11}$$

$$\widehat{\beta}e_{2}=\frac{A(Y_{t})-B(Y_{t})}{(A(X_{2t})-B(X_{2t}))} \tag{12}$$

in the final analysis, this paper highlights both the advantages and disadvantages of the Etemadi regression model when contrasted with classical regression models. While the chemometrics literature has introduced various forms of regression for causal modeling in diverse fields, it is notable that despite the apparent distinctions among these methods, they all rely on accuracy-based parameter estimation and learning techniques. Therefore, **(1)** The primary advantage of the Etemadi regression model over classical regression lies in its formulation of the reliability concept in the design of the parameter estimation process and training framework for statistical modeling, resulting in the creation of more accurate and generalizable models. **(2)** The Etemadi regression model adopts a risk-based modeling strategy that, in contrast to classical regression models, seeks to maximize the model's reliability or, equivalently, minimize the model's uncertainty. This approach leads to the development of a more accurate and efficient model that exhibits strong generalizability, particularly for inaccessible/test data in high-risk and sensitive real-world situations. **(3)** The Etemadi regression model considers and quantifies uncertainty in the parameter estimation process, resulting in the minimization of the model's uncertainty. In contrast, the classical regression model disregards uncertainty and operates as a deterministic model. **(4)** The Etemadi regression model is built upon the principle of minimizing the variance of the performance criterion across various data points, leading to a higher reliability level of accuracy in inaccessible or test data. This is distinct from the classical regression model, which focuses on minimizing errors within the available historical or training data. On the other hand, the formulation complexity of the Etemadi regression model, which is built on the reliability-based parameter estimation process, is higher than that of classical regression models. However, the coefficients of the Etemadi regression model are optimally determined.

4. Data description

The purpose of this paper is to conduct a thorough evaluation of the performance of a modeling approach based on reliability as opposed to one based on accuracy. To accomplish this, the study utilizes 19 benchmark datasets acquired from the UCI [53] and Kaggle databases, covering various chemistry fields including Pharmacology, Biochemistry, Agrochemical, Geochemical, Biological, Toxicity Compound, Pollutants, Physicochemical Properties, Gases Experiment, and Chemical compounds of Food Additives and Ingredients. These datasets vary in sample size, ranging from 92 to 295,679, and contain between 4 and 81 explanatory variables of different types, including real and integer, as detailed in Table (1).

5. Empirical result

This research paper employs a random sampling method, where 85% of the data is randomly selected for training and validation, with 10% of that subset held out for validation, and the remaining 15% reserved for testing. To ensure the randomness of the selection does not impact the results, the estimation process is repeated 100 times. Both the proposed Etemadi regression and the classic regression are modeled using Phyton and MATLAB software. Table (2) and Table (3) present the results for the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Average Relative Variance (ARV) of both the reliability-based MLR and classic MLR models. The tables also include the percentage improvement of the proposed model over the classic version for all 19 benchmark datasets. Additionally, the formulas for these evaluation metrics are provided in Eq. (13) to Eq. (16).

$$MAE=\frac{1}{N}\sum_{t=N+n+1}^{N+n+N'}|y_{t}-\widehat{y}_{t}| \tag{13}$$

$$MSE = \frac{1}{N'} \sum_{t=N+n+1}^{N+n+N'} (y_t - \hat{y}_t)^2 \tag{14}$$

$$RMSE = \sqrt{\frac{1}{N'} \sum_{t=N+n+1}^{N+n+N'} (y_t - \hat{y}_t)^2} \tag{15}$$

$$ARV = \sum_{t=N+n+1}^{N+n+N'} (y_t - \hat{y}_t)^2 \Big/ \sum_{t=N+n+1}^{N+n+N'} (y_t - \bar{y}_t)^2 \tag{16}$$

This section presents an analysis and comparison of the outcomes of the Etemadi and conventional models. The comparison is based on the number and percentage of cases in which one model outperforms the other, as well as the improvement in generalization ability. The Etemadi model demonstrates superior performance in 12 cases (63.16%), excelling in all MAE, MSE, RMSE, and ARV metrics. In 15 cases (78.95%), it shows higher performance based on at least one of the evaluation measurements. However, the proposed reliability-based MLR exhibits lower generalizability compared to the conventional accuracy-based model in 4 cases (21.05%). Therefore, these results emphasize the significance and efficacy of incorporating the reliability factor, in addition to accuracy, to enhance the generalizability of linear regression in chemometrics applications.

Upon further analysis, as depicted in Fig. (1), these findings demonstrate that the reliability-based modeling approach exhibited greater generalizability on average compared to the accuracy-based strategy in terms of MAE, MSE, ARV, and RMSE by 4.697%, 5.646%, 5.646%, and 4.342%, respectively. The superior generalizability of the proposed MLR model, compared to the conventional version in the chemistry benchmark dataset, underscores the importance and effectiveness of considering the reliability factor in the parameter estimation process to handle uncertainty. Additionally, the statistical significance of the improvement obtained through the reliability-based modeling approach across various chemometrics applications was tested in contrast to the accuracy-based modeling approach. A statistical T-test confirms that the accuracy improvement obtained by the reliability-based method is significant in all of these metrics at the 2%, 2%, 2%, and 3% levels, respectively. This implies that the probability of the Etemadi MLR method's generalizability superiority over the classic method is at least 97%. These findings indicate that both accuracy and reliability play crucial roles in generalization ability, but reliability holds greater importance than accuracy in dealing with uncertain scenarios in chemometric applications. Hence, it is advisable to prefer the Etemadi MLR method over the classic method in unfamiliar situations or when selecting a modeling strategy blindly.

The study aimed to determine the data characteristics that influence the superiority of reliability- or accuracy-based modeling

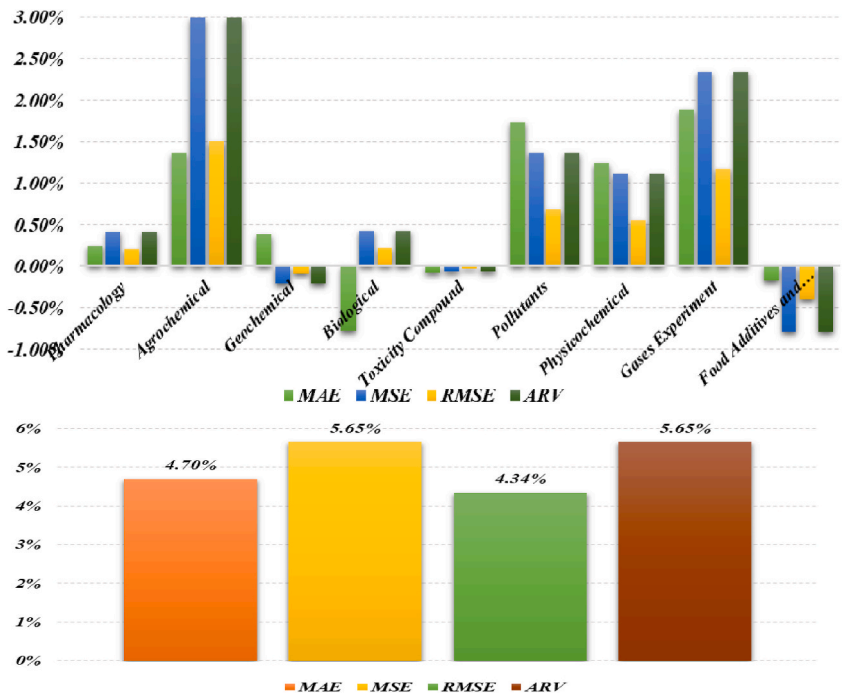


Fig. 1. The achieved improvement by reliability-based modeling strategy than accuracy-based one based on each domain of study and averagely all of them.

strategies. To achieve this, the statistical significance of various factors, including sample size, number of attributes, data type, and application fields, on the generalizability of the reliability-based MLR is examined using the results of 19 case studies. The findings show that only data type significantly influenced the superiority of the Etemadi MLR method, with a significance level of 0. However, other factors, such as sample size, number of attributes, and application fields, do not have a statistically significant impact on superiority, improvement percentage, or ranking of improvement percentage. Therefore, the effect of data characteristics on the proposed method's superiority over the classic one cannot be inferred. Further investigation with additional variables and data is necessary to identify the effective explanatory variables.

Additionally, to offer a more comprehensive evaluation of the proposed reliability-based MLR, its effectiveness is compared to accuracy-based versions using four metrics: MAE, MSE, RMSE, and ARV. This comparison is conducted across ten different fields of chemistry benchmarks, which include Pharmacology, Biochemistry, Agrochemical, Geochemical, Biological, Toxicity Compound, Pollutants, Physicochemical Properties, Gases Experiment, and Chemical Compounds of Food Additives and Ingredients. Results presented in Fig. (1) illustrate that the proposed method can, on average, enhance the generalizability of the classic model in Pharmacology, Biochemistry, Agrochemical, Geochemical, Pollutants, Physicochemical Properties, and Gases Experiment fields by 0.251%, 39.214%, 1.356%, 0.379%, 1.739%, 1.242%, and 1.882%, respectively, based on the MAE metric. Furthermore, the proposed reliability-based MLR model consistently outperforms the conventional MLR model in multiple domains, including Pharmacology, Biochemistry, Agrochemical, Biological, Pollutants, Physicochemical Properties, and Gases Experiment. The improvements are observed in terms of MSE and ARV evaluation measurements, with percentages of 0.419%, 46.781%, 2.991%, 0.430%, 1.372%, 1.108%, and 2.347%, respectively. Additionally, according to the RMSE metric, the improvements are 0.210%, 37.775%, 1.507%, 0.216%, 0.689%, 0.555%, and 1.180%, respectively. However, in the domains of Toxicity Compounds and Chemical Compounds of Food Additives and Ingredients, the proposed model's performance in terms of MAE, MSE, ARV, and RMSE metrics is, on average, lower than the accuracy-based MLR. Additionally, the accuracy-based modeling strategy is, on average, superior to the reliability-based one in the Biological domain based on MAE and in the Geochemical domain based on MSE, ARV, and RMSE metrics. Thus, the empirical outcomes suggest that the Etemadi MLR outperforms the classic MLR in over half of the application domains.

In this analysis, it can be observed that the proposed model shows varying levels of improvement based on the MAE metric across different domains. The category of Biochemistry includes the case with the highest improvement of 78.712%, while Pharmacology shows the lowest improvement at 0.230%. Further investigation reveals that the top five improvements achieved by the proposed model over the classic version are in cases that belong to the fields of Biochemistry, Pollutants, Agrochemical, Gases Experiment, and Geochemical, with MAE improvements of 78.712%, 3.352%, 1.937%, 1.882%, and 1.433%, respectively. To provide a more comprehensive evaluation, the performance of the proposed MLR is also compared to the MSE and ARV of the classic model. Based on the five highest improvements, the proposed model enhances the generalizability of the classic model by 94.166%, 3.731%, 2.347%,

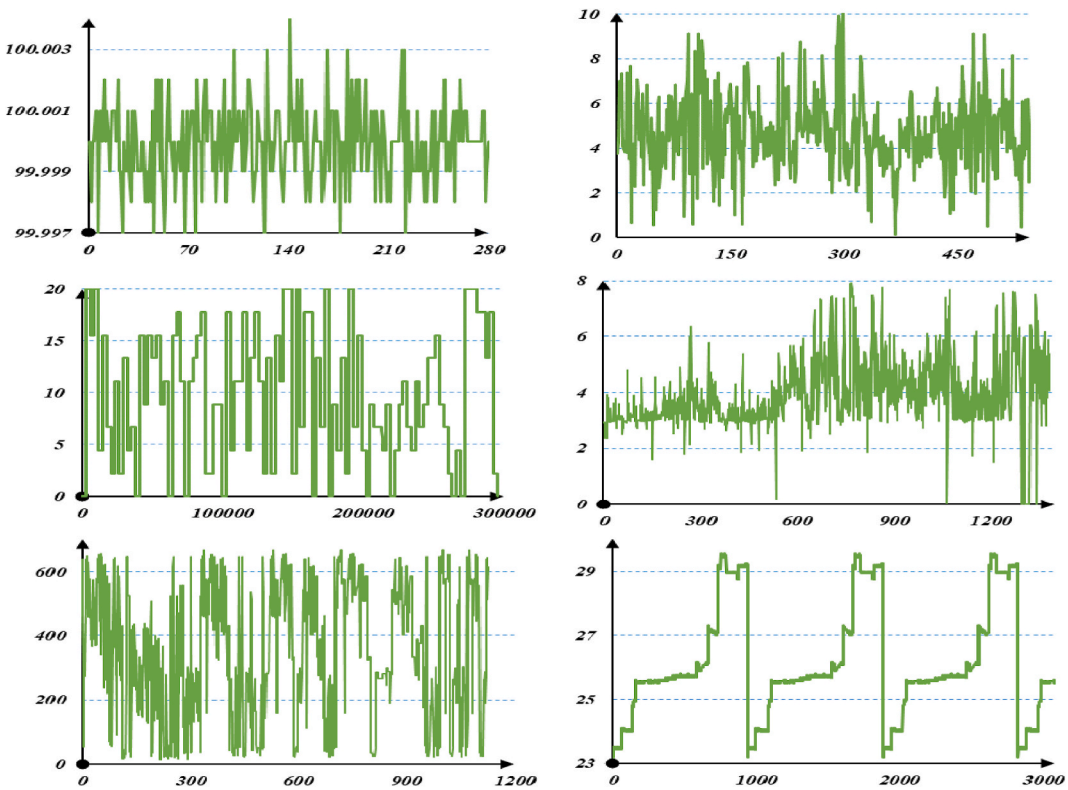


Fig. 2. Pattern of the target variable (Y) of some case studies that the reliability-based modeling strategy is superior to the accuracy-based one.

2.250%, and 1.977% for the cases that fall into the categories of Biochemistry, Agrochemical, Gases Experiment, Agrochemical, and Pollutants, respectively. In general, the proposed MLR model achieves its highest and lowest improvements in cases related to the Biochemistry and Geochemical domains, with improvements of 94.166% and 0.04%, respectively. Similarly, when using the RMSE metric, the highest and lowest improvements are observed in cases dedicated to the Biochemistry and Geochemical domains, with improvements of 75.852% and 0.058%, respectively. Looking more closely, the five domains of cases with the highest improvements over the classic version were Biochemistry, Agrochemical, Gases Experiment, Agrochemical, and Pollutants, with improvements of 75.852%, 1.883%, 1.180%, 1.131%, and 0.994%, respectively. While it may be expected that the proposed MLR approach is generally superior to the classic MLR in various fields of chemometric applications, it highlights the importance of the proposed parameter estimation process in minimizing uncertainty in the machine learning procedure.

Looking from a different angle, Fig. (2) displays charts of the target variable (Y) from several case studies, demonstrating the superiority of the reliability-based modeling strategy over the accuracy-based one. The charts reveal that the Y-variable data exhibit highly complex patterns characterized by irregular fluctuations and periods, resulting in significant uncertainty. As anticipated, in chemometric applications that involve highly sensitive and variable chemical processes and experiments, result uncertainty is notably increased. The Etemadi modeling strategy, which incorporates the reliability concept in parameter estimation and learning procedures, has emerged as the leading approach to minimizing result uncertainty in the modeling paradigm. Consequently, it is generally expected that reliability-based modeling strategies will achieve better results on average than accuracy-based approaches.

6. Discussion

The primary focus of this section is to provide a summary of significant research efforts. Subsequently, it offers an analysis and discussion of the research findings.

1) This paper presents pioneering research that introduces the Etemadi reliability-based modeling methodology for the first time in the field of chemometrics. The study investigates ten significant chemistry application domains, encompassing various fields such as Pharmacology, Biochemistry, Agrochemical, Geochemical, Biological, Toxicity Compound, Pollutants, Physicochemical Properties, Gases Experiment, and Chemical Compounds of Food Additives and Ingredients.
2) The study examined the contribution of each reliability and accuracy factor to achieving a high level of generalization, particularly in the field of chemometrics. To ensure the study's accuracy, several evaluation metrics, including MAE, MSE, RMSE, and ARV, were taken into account, ensuring that the outcomes of this study are not affected by the type of measurement.
3) The varying levels of generalizability among models, resulting from differences in their structure, complexity, and learning processes, present a challenge, regardless of the modeling strategy used. To overcome this issue, a solution was proposed: the utilization of multiple linear regression, a widely-used chemometrics model belonging to the category of simple linear statistical models that employ a direct optimal learning procedure. The reliability-based modeling strategy was specifically implemented for multiple linear regression to enhance its efficacy.
4) This paper aimed to assess and compare the generalizability of the two most prominent methodologies utilized in the modeling paradigm that employs reliability or accuracy concepts when designing parameter estimation processes. To achieve this, the study eliminated the impact of factors such as data characteristics and model complexity on the generalizability of these two strategies. This was accomplished by considering several benchmark datasets and applying the MLR model.
5) The outcome of the proposed reliability-based chemometric regression was compared to that of the accuracy-based regression model, with no consideration given to other types of machine learning models. This decision was made because the generalizability of accuracy-based models, which do not belong to the class of statistical linear models, is influenced not only by the modeling approach but also by complexity, structure, and learning algorithms. Hence, comparing the performances of Etemadi regression to other categories of models, such as non-linear, fuzzy, or intelligent models, would not align with the purpose of this study. Nevertheless, the reliability-based modeling strategy can be applied to a range of other models, including non-linear, fuzzy, intelligent, and shallow/deep models, and their generalization ability can be compared with their accuracy-based counterparts.
6) The empirical evidence supports the effectiveness of reliability in enhancing generalizability. Additionally, it was statistically proven that reliability has a greater impact on improving the generalizability of chemometric models than accuracy, with a significance level of at least 3%. Consequently, adopting a modeling approach based on reliability can improve generalizability by an average of 5.1% compared to an accuracy-based approach, according to various metrics such as MAE, MSE, RMSE, and ARV.
7) From another standpoint, the study conducted an analysis of the effectiveness of a reliability-based strategy versus an accuracy-based approach across ten different chemometric fields. The results revealed that the proposed Etemadi MLR outperformed the classic MLR in eight domains, namely Pharmacology, Biochemistry, Agrochemical, Geochemical, Biological, Pollutants, Physicochemical Properties, and Gases Experiment. However, in only two domains, namely Toxicity Compound and Chemical Compounds of Food Additives and Ingredients, the Etemadi MLR averagely exhibited slightly lower performance than the classic MLR, with differences of 0.1% and 0.5%, respectively.

7. Conclusion

This paper introduces a novel formulation of the reliability concept into the parameter estimation process of chemometric models, aiming to design models that are both more accurate and more generalizable. The proposed model, employing a reliability-based learning process for training and parameter estimation, has demonstrated its success in achieving accurate and generalizable

results. The study confirms the effectiveness and superior performance of the reliability-based model when compared to the accuracy-based model across a range of chemistry benchmark datasets. The key findings can be summarized as follows.

(1) Incorporating the concept of reliability into the design framework of chemometric models to address the challenges posed by the variable and highly sensitive nature of chemistry experiments and processes is a promising approach that significantly contributes to achieving high generalizability.
(2) The proposed reliability-based model was evaluated using several assessment metrics and a variety of chemometrics applications with diverse characteristics, including Pharmacology, Biochemistry, Agrochemical, Geochemical, Biological, Toxicity Compound, Pollutants, Physicochemical Properties, Gases Experiment, and Chemical Compounds of Food Additives and Ingredients. Based on empirical results, the reliability-based model exhibited an average performance improvement of 5.1% in terms of generalization ability compared to the accuracy-based model.
(3) The statistically significant superiority of the proposed model emphasizes the pivotal role of the reliability factor within the learning algorithm of causal models, contributing to their generalizability. In this context, incorporating reliability-based learning and parameter estimation processes into causal models stands out as a remarkable approach for addressing high-risk and sensitive scenarios, as well as for modeling uncertainty. This approach has resulted in the superior performance of the proposed model when compared to the accuracy-based model.
(4) The proposed reliability-based method represents a logical, viable, and efficient approach that can be implemented in all categories of statistical and intelligent chemometric models, making it a promising area for future research. Furthermore, future studies can investigate the influence of data features and characteristics on the superior generalizability of the proposed method compared to traditional models.

Funding

No specific financial support was received to carry out this study.

Conflicts of interest/Competing interests

The authors declare that they have no competing interests.

Data availability

Data used in the study is freely available at the UCI web site [54], or can be downloaded from the corresponding author home page [55], and also is available from the corresponding author on reasonable request. Data will be made available on request.

Code and Software availability

Code and Software used in the study is available from the corresponding author on reasonable request.

CRediT authorship contribution statement

Sepideh Etemadi: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Visualization. **Mehdi Khashei:** Supervision, Validation.

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] H. Moussa, F. Dahmoune, M. Hentabli, H. Remini, L. Mouni, Optimization of ultrasound-assisted extraction of phenolic-saponin content from Carthamus caeruleus L. rhizome and predictive model based on support vector regression optimized by dragonfly algorithm, Chemometr. Intell. Lab. Syst. 222 (2022) 104493, https://doi.org/10.1016/j.chemolab.2022.104493.
[2] R. Min, Z. Wang, Y. Zhuang, X. Yi, Application of semi-supervised convolutional neural network regression model based on data augmentation and process spectral labeling in Raman predictive modeling of cell culture processes, Biochem. Eng. J. 191 (2023) 108774, https://doi.org/10.1016/j.bej.2022.108774.
[3] X. Sun, D. She, Y. Fei, X. Han, L. Gao, An improved pore-solid fractal model for predicting coastal saline soil hydraulic properties based on changepoints determined by genetic algorithm-support vector regression, Soil Tillage Res. 224 (2022) 105502, https://doi.org/10.1016/j.still.2022.105502.
[4] H. Haghnazar, K.H. Johannesson, R. González-Pinzón, M. Pourakbar, E. Aghayani, A. Rajabi, A.A. Hashemi, Groundwater geochemistry, quality, and pollution of the largest lake basin in the Middle East: comparison of PMF and PCA-MLR receptor models and application of the source-oriented HHRA approach, Chemosphere 288 (2022) 132489, https://doi.org/10.1016/j.chemosphere.2021.132489.
[5] A. Das, N. Bar, S.K. Das, Pb (II) adsorption from aqueous solution by nutshells, green adsorbent: adsorption studies, regeneration studies, scale-up design, its effect on biological indicator and MLR modeling, J. Colloid Interface Sci. 580 (2020) 245–255, https://doi.org/10.1016/j.jcis.2020.07.017.
[6] A.A. Toropov, M.R. Di Nicola, A.P. Toropova, A. Roncaglioni, E. Carnesecchi, N.I. Kramer, A.J. Williams, M.E. Ortiz-Santaliestra, E. Benfenati, J.L.C. Dorne, A regression-based QSAR-model to predict acute toxicity of aromatic chemicals in tadpoles of the Japanese brown frog (Rana japonica): calibration, validation,

and future developments to support risk assessment of chemicals in amphibians, Sci. Total Environ. 830 (2022) 154795, https://doi.org/10.1016/j.scitotenv.2022.154795.
[7] L. Yu, T. Zheng, R. Yuan, X. Zheng, APCS-MLR model: a convenient and fast method for quantitative identification of nitrate pollution sources in groundwater, J. Environ. Manag. 314 (2022) 115101, https://doi.org/10.1016/j.jenvman.2022.115101.
[8] A. Shahi, H.V. Molamahmood, N. Faraji, M. Long, Quantitative structure-activity relationship for the oxidation of organic contaminants by peracetic acid using GA-MLR method, J. Environ. Manag. 310 (2022) 114747, https://doi.org/10.1016/j.jenvman.2022.114747.
[9] L. Jin, H. Ye, Y. Shi, L. Li, R. Liu, Y. Cai, J. Li, F. Li, Z. Jin, Using PCA-APCS-MLR model and SIAR model combined with multiple isotopes to quantify the nitrate sources in groundwater of Zhuji, East China, Appl. Geochem. 143 (2022) 105354, https://doi.org/10.1016/j.apgeochem.2022.105354.
[10] M. Varol, G. Karakaya, K. Alpaslan, Water quality assessment of the Karasu River (Turkey) using various indices, multivariate statistics and APCS-MLR model, Chemosphere 308 (2022) 136415, https://doi.org/10.1016/j.chemosphere.2022.136415.
[11] B. Souyei, S. Meneceur, A. Khechekhouche, QSPR study on thermal energy of aliphatic Aldehydes using molecular descriptors and MLR technique, Mater. Today: Proc. 51 (2022) 2157–2162, https://doi.org/10.1016/j.matpr.2022.01.302.
[12] J. Niimi, O. Tomic, T. Næs, D.W. Jeffery, S.E. Bastian, P.K. Boss, Application of sequential and orthogonalised-partial least squares (SO-PLS) regression to predict sensory properties of Cabernet Sauvignon wines from grape chemical composition, Food Chem. 256 (2018) 195–202, https://doi.org/10.1016/j.foodchem.2018.02.120.
[13] A. Courand, M. Metz, D. Héran, C. Feilhes, F. Prezman, E. Serrano, R. Bendoula, M. Ryckewaert, Evaluation of a robust regression method (RoBoost-PLSR) to predict biochemical variables for agronomic applications: case study of grape berry maturity monitoring, Chemometr. Intell. Lab. Syst. 221 (2022) 104485, https://doi.org/10.1016/j.chemolab.2021.104485.
[14] Y. Yang, X. Shang, Z. Chen, K. Mei, Z. Wang, R.A. Dahlgren, M. Zhang, X. Ji, A support vector regression model to predict nitrate-nitrogen isotopic composition using hydro-chemical variables, J. Environ. Manag. 290 (2021) 112674, https://doi.org/10.1016/j.jenvman.2021.112674.
[15] K. Chen, Q. Liu, Q. Jiang, X. Hou, W. Gao, Source apportionment of surface water pollution in North Anhui Plain, Eastern China, using APCS-MLR model combined with GIS approach and socioeconomic parameters, Ecol. Indicat. 143 (2022) 109324, https://doi.org/10.1016/j.ecolind.2022.109324.
[16] D.E. Arthur, A. Uzairu, P. Mamza, S.E. Abechi, G. Shallangwa, Activity and toxicity modelling of some NCI selected compounds against leukemia P388ADR cell line using genetic algorithm-multiple linear regressions, J. King Saud Univ. Sci. 32 (1) (2020) 324–331, https://doi.org/10.1016/j.jksus.2018.05.023.
[17] Y. Yang, X. Wang, X. Zhao, M. Huang, Q. Zhu, M3GPSpectra: a novel approach integrating variable selection/construction and MLR modeling for quantitative spectral analysis, Anal. Chim. Acta 1160 (2021) 338453, https://doi.org/10.1016/j.aca.2021.338453.
[18] H. Kaneko, K. Funatsu, Fast optimization of hyperparameters for support vector regression models with highly predictive ability, Chemometr. Intell. Lab. Syst. 142 (2015) 64–69, https://doi.org/10.1016/j.chemolab.2015.01.001.
[19] B.C. Lago, C.A. Silva, L.C.A. Melo, E.G. de Morais, Predicting biochar cation exchange capacity using Fourier transform infrared spectroscopy combined with partial least square regression, Sci. Total Environ. 794 (2021) 148762, https://doi.org/10.1016/j.scitotenv.2021.148762.
[20] T.N. Brown, Empirical regressions between system parameters and solute descriptors of polyparameter linear free energy relationships (PPLFERs) for predicting solvent-air partitioning, Fluid Phase Equil. 540 (2021) 113035, https://doi.org/10.1016/j.fluid.2021.113035.
[21] F. Raposo, R. Borja, C. Ibelli-Bianco, Predictive regression models for biochemical methane potential tests of biomass samples: pitfalls and challenges of laboratory measurements, Renew. Sustain. Energy Rev. 127 (2020) 109890, https://doi.org/10.1016/j.rser.2020.109890.
[22] I. Barra, M. Kharbach, E.M. Qannari, M. Hanafi, Y. Cherrah, A. Bouklouze, Predicting cetane number in diesel fuels using FTIR spectroscopy and PLS regression, Vib. Spectrosc. 111 (2020) 103157, https://doi.org/10.1016/j.vibspec.2020.103157.
[23] I.A. Naguib, F.F. Abdallah, Ultraviolet cutoff area and predictive ability of partial least squares regression method: a pharmaceutical case study, Spectrochim. Acta Mol. Biomol. Spectrosc. 231 (2020) 118116, https://doi.org/10.1016/j.saa.2020.118116.
[24] M.M. Ávila, M.L. Durán, D. Caballero, T. Antequera, T. Palacios-Pérez, E. Cernadas, M. Fernández-Delgado, Magnetic Resonance Imaging, Texture Analysis and Regression Techniques to Non-destructively Predict the Quality Characteristics of Meat Pieces, vol. 82, Engineering Applications of Artificial Intelligence, 2019, pp. 110–125, https://doi.org/10.1016/j.engappai.2019.03.026.
[25] D.R. Michanowicz, J.L. Shmool, L. Cambal, B.J. Tunno, S. Gillooly, M.J.O. Hunt, S. Tripathy, K.N. Shields, J.E. Clougherty, A hybrid land use regression/line-source dispersion model for predicting intra-urban NO2, Transport. Res. Transport Environ. 43 (2016) 181–191, https://doi.org/10.1016/j.trd.2015.12.007.
[26] S.E. Adeniji, S. Uba, A. Uzairu, Theoretical modeling for predicting the activities of some active compounds as potent inhibitors against Mycobacterium tuberculosis using GFA-MLR approach, J. King Saud Univ. Sci. 32 (1) (2020) 575–586, https://doi.org/10.1016/j.jksus.2018.08.010.
[27] N. Xu, F. Meng, G. Zhou, Y. Li, B. Wang, H. Lu, Assessing the suitable cultivation areas for Scutellaria baicalensis in China using the Maxent model and multiple linear regression, Biochem. Systemat. Ecol. 90 (2020) 104052, https://doi.org/10.1016/j.bse.2020.104052.
[28] K. Nakamura, T. Yasutaka, T. Kuwatani, T. Komai, Development of a predictive model for lead, cadmium and fluorine soil–water partition coefficients using sparse multiple linear regression analysis, Chemosphere 186 (2017) 501–509, https://doi.org/10.1016/j.chemosphere.2017.07.131.
[29] C. Bertelkamp, A.R.D. Verliefde, J. Reynisson, N. Singhal, A.J. Cabo, M. De Jonge, J.P. van der Hoek, A predictive multi-linear regression model for organic micropollutants, based on a laboratory-scale column study simulating the river bank filtration process, J. Hazard Mater. 304 (2016) 502–511, https://doi.org/10.1016/j.jhazmat.2015.11.003.
[30] S.H. Ewaid, S.A. Abed, S.A. Kadhum, Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis, Environ. Technol. Innovat. 11 (2018) 390–398, https://doi.org/10.1016/j.eti.2018.06.013.
[31] A. Olaya-Abril, L. Parras-Alcántara, B. Lozano-García, R. Obregón-Romero, Soil organic carbon distribution in Mediterranean areas under a climate change scenario via multiple linear regression analysis, Sci. Total Environ. 592 (2017) 134–143, https://doi.org/10.1016/j.scitotenv.2017.03.021.
[32] D. Palmer, J.O. Pou, L. Gonzalez-Sabaté, J. Díaz-Ferrero, Multiple linear regression based congener profile correlation to estimate the toxicity (TEQ) and dioxin concentration in atmospheric emissions, Sci. Total Environ. 622 (2018) 510–516, https://doi.org/10.1016/j.scitotenv.2017.11.344.
[33] K. Dieguez-Santana, H. Pham-The, P.J. Villegas-Aguilar, H. Le-Thi-Thu, J.A. Castillo-Garit, G.M. Casañola-Martin, Prediction of acute toxicity of phenol derivatives using multiple linear regression approach for Tetrahymena pyriformis contaminant identification in a median-size database, Chemosphere 165 (2016) 434–441, https://doi.org/10.1016/j.chemosphere.2016.09.041.
[34] A. Mandal, N. Bar, S.K. Das, Phenol removal from wastewater using low-cost natural bioadsorbent neem (Azadirachta indica) leaves: adsorption study and MLR modeling, Sustainable Chemistry and Pharmacy 17 (2020) 100308, https://doi.org/10.1016/j.scp.2020.100308.
[35] A.G. More, S.K. Gupta, Predictive modelling of chromium removal using multiple linear and nonlinear regression with special emphasis on operating parameters of bioelectrochemical reactor, J. Biosci. Bioeng. 126 (2) (2018) 205–212, https://doi.org/10.1016/j.jbiosc.2018.02.013.
[36] K. Abrougui, K. Gabsi, B. Mercatoris, C. Khemis, R. Amami, S. Chehaibi, Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR), Soil Tillage Res. 190 (2019) 202–208, https://doi.org/10.1016/j.still.2019.01.011.
[37] Z. Du, Y. Hu, N.A. Buttar, Analysis of mechanical properties for tea stem using grey relational analysis coupled with multiple linear regression, Sci. Hortic. 260 (2020) 108886, https://doi.org/10.1016/j.scienta.2019.108886.
[38] B.M. Robert, G.R. Brindha, B. Santhi, G. Kanimozhi, N.R. Prasad, Computational models for predicting anticancer drug efficacy: a multi linear regression analysis based on molecular, cellular and clinical data of oral squamous cell carcinoma cohort, Comput. Methods Progr. Biomed. 178 (2019) 105–112, https://doi.org/10.1016/j.cmpb.2019.06.011.
[39] A. Hosseinzadeh, M. Baziar, H. Alidadi, J.L. Zhou, A. Altaee, A.A. Najafpoor, S. Jafarpour, Application of artificial neural network and multiple linear regression in modeling nutrient recovery in vermicompost under different conditions, Bioresour. Technol. 303 (2020) 122926, https://doi.org/10.1016/j.biortech.2020.122926.
[40] W. Tang, Y. Li, Y. Yu, Z. Wang, T. Xu, J. Chen, J. Lin, X. Li, Development of models predicting biodegradation rate rating with multiple linear regression and support vector machine algorithms, Chemosphere 253 (2020) 126666, https://doi.org/10.1016/j.chemosphere.2020.126666.

[41] R. Rendall, A.C. Pereira, M.S. Reis, Advanced predictive methods for wine age prediction: Part I–A comparison study of single-block regression approaches based on variable selection, penalized regression, latent variables and tree-based ensemble methods, Talanta 171 (2017) 341–350, https://doi.org/10.1016/j.talanta.2016.10.062.
[42] S. Etemadi, M. Khashei, Etemadi multiple linear regression, Measurement 186 (2021) 110080, https://doi.org/10.1016/j.measurement.2021.110080.
[43] S. Etemadi, M. Khashei, Accuracy versus reliability-based modelling approaches for medical decision making, Comput. Biol. Med. 141 (2022) 105138, https://doi.org/10.1016/j.compbiomed.2021.105138.
[44] S. Etemadi, M. Khashei, S. Tamizi, Etemadi reliability-based multi-layer perceptrons for classification and forecasting, Inf. Sci. 651 (2023) 119716, https://doi.org/10.1016/j.ins.2023.119716.
[45] M. Khashei, N. Bakhtiarvand, S. Etemadi, A novel reliability-based regression model for medical modeling and forecasting, Diabetes Metabol. Syndr.: Clin. Res. Rev. 15 (6) (2021) 102331, https://doi.org/10.1016/j.dsx.2021.102331.
[46] Z. Hajirahimi, M. Khashei, S. Etemadi, A novel class of reliability-based parallel hybridization (RPH) models for time series forecasting, Chaos, Solit. Fractals 156 (2022) 111880, https://doi.org/10.1016/j.chaos.2022.111880.
[47] X.Q. Li, L.K. Song, Y.S. Choy, G.C. Bai, Multivariate ensembles-based hierarchical linkage strategy for system reliability evaluation of aeroengine cooling blades, Aero. Sci. Technol. 138 (2023) 108325, https://doi.org/10.1016/j.ast.2023.108325.
[48] X.Q. Li, L.K. Song, G.C. Bai, D.G. Li, Physics-informed distributed modeling for CCF reliability evaluation of aeroengine rotor systems, Int. J. Fatig. 167 (2023) 107342, https://doi.org/10.1016/j.ijfatigue.2022.107342.
[49] H. Zhang, L.K. Song, G.C. Bai, X.Q. Li, Active extremum Kriging-based multi-level linkage reliability analysis and its application in aeroengine mechanism systems, Aero. Sci. Technol. 131 (2022) 107968, https://doi.org/10.1016/j.ast.2022.107968.
[50] X.Q. Li, L.K. Song, G.C. Bai, Deep learning regression-based stratified probabilistic combined cycle fatigue damage evaluation for turbine bladed disks, Int. J. Fatig. 159 (2022) 106812, https://doi.org/10.1016/j.ijfatigue.2022.106812.
[51] A. Roy, S. Chakraborty, Support Vector Machine in Structural Reliability Analysis: A Review, Reliability Engineering & System Safety, 2023 109126, https://doi.org/10.1016/j.ress.2023.109126.
[52] T. Mazhar, R.N. Asif, M.A. Malik, M.A. Nadeem, I. Haq, M. Iqbal, M. Kamran, S. Ashraf, Electric vehicle charging system in the smart Grid using different machine learning methods, Sustainability 15 (3) (2023) 2603, https://doi.org/10.3390/su15032603.
[53] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2019. http://archive.ics.uci.edu/ml.
[54] http://archive.ics.uci.edu/ml/datasets.php.
[55] http://iutbox.iut.ac.ir/index.php/apps/files/Econometrics data sets.