Received: 3 July 2022 | Revised: 23 September 2022 | Accepted: 27 September 2022

DOI: 10.1002/qua.27026

RESEARCH ARTICLE

International Journal of QUANTUM CHEMISTRY WILEY

Machine learning approach for the prediction of surface tension of binary mixtures containing ionic liquids using σ-profile descriptors

Widad Benmouloud | **Cherif Si-Moussa** | **Othmane Benkortbi**

Biomaterials and Transport Phenomena Laboratory (LBMPT), Department of Process and Environmental Engineering, University of Yahia Fares Medea, Medea, Algeria

Correspondence
Widad Benmouloud, Biomaterials and Transport Phenomena Laboratory (LBMPT), Department of Process and Environmental Engineering, University of Yahia Fares Medea, Medea 26000, Algeria.
Email: ben.widad36@gmail.com

Funding information
Algerian Ministry of Higher Education and Scientific Research, Grant/Award Number: A16N01UN260120220003; University Yahia Fares of Medea

Abstract

Ionic liquids (IL) are a new class of liquids considered as green solvents; less toxic, less flammable, and less polluting which retain their liquid state over wide temperature ranges and are considered alternatives to volatile organic solvents. The surface tension of IL-organic solvent mixtures plays an important role in the design and development of many industrial processes. This work investigated the capability and feasibility of four ANN model topologies ("trainbr, logsig"; "trainbr, tansig"; "trainlm, logsig"; "trainlm, tansig"), a PSO-SVM model, and an LSSVM model to predict the surface tension of binary systems containing IL. For this purpose, 1623 data points corresponding to the experimental surface tension values of binary mixtures containing IL were collected from the literature. The surface tension values were between 18.9 and 72.7 mN m^{-1}. The temperature, the composition in mole fraction of IL (X_{IL}), descriptors based on the sigma profiles, relating to the H-bond donor and to the H-bond acceptor character, the anion, the cation and the solvent were used as input variables of the model in order to differentiate the different compounds involved in the binary systems. A comparison of the experimental and the predicted values in terms of several statistical metrics showed good agreement, however, the prediction (trainbr, logsig) was better than the other approaches with an overall average absolute relative deviation of .8466% and a mean square error of .4952. These results are very encouraging for future projects modeling other physical and chemical properties of ILs.

KEYWORDS
artificial neural networks, ionic liquids, least-squares support vector machine, support vector machine-particle swarm optimization, surface tension, σ-profile descriptor

1 | INTRODUCTION

In recent years, the situation has evolved considerably in the field of the use of organic molecular solvents in various industrial chemical processes. Due to their harmfulness (toxicity, flammability, and volatile organic compounds (VOC) emission), stringent regulations aim to limit the use of solvents presenting dangers to human health and the environment [1]. Ionic liquids (ILs) have attracted much attention from the scientific community over the past two decades due to their wide variety of applications in many fields of chemistry and chemical industry [2]. They are a new class of liquids, considered as green solvents, that retain their liquid state over wide temperature ranges [1, 3], with high solvation properties, negligible vapor pressure, and high thermal, chemical, and electrochemical stability [4]. Potential applications of ILs require knowledge of physicochemical

 © 2022 Wiley Periodicals LLC.

properties such as density, viscosity, melting point, solvent properties, vapor pressure and surface tension for pure ILs and their mixtures with other solvents [5].

As there are countless combinations of cations and anions that form an IL, the synthesis of all the IL resulting from all these possible combinations is practically impossible. Moreover, measurements of all properties for all synthesized IL and their mixtures are laborious, time consuming and costly. Therefore, it is necessary to develop reliable models that could be considered as a suitable alternative to experimental measurements for predicting the properties of IL mixtures under various conditions [6].A key property in various fields such as oil, gas and chemical industries, is the surface tension of pure and mixed systems [7]. This property plays a particular role in process design by affecting mass and heat transfer at the interface [8, 9].

Several authors have developed predictive models of the surface tension of pure ILs, but few studies have been carried out on their mixtures. Huang et al [6]. used in their studies, semi-empirical models and artificial neural network (ANN) models for the prediction of the surface tension of mixture of ILs. They found that the overall average absolute relative deviation (AARD) of the semi-empirical and ANN models is less than 2%. The study conducted by Atashrouz et al [3]. for the prediction of physico-chemical properties of IL mixtures such as surface tension has allowed the development of flexible computational approaches based on support vector machine (SVM), least square support vector machine (LSSVM) and the group method of data processing type polynomial neural network systems (GMDH-PNN). They found that the LSSVM model is more robust and reliable in predicting physicochemical properties of IL mixtures with AARD of 1.18% for surface tension.

Hashemkhani et al [10]. used three methods namely SVM, CSA-LSSVM and GA-LSSVM to predict the surface tension of binary mixtures containing ILs using 748 data points. They obtained a better precision in the case of CSA-LSSVM model where the average % AARD is 1.3785. A modeling method based on ANN trained by Bayesian regulation back propagation training algorithm (trainbr) has been proposed by Soleimani et al [11]. to predict surface tension of the binary IL mixtures. A comparison with different models such as SVM, GA-SVM, GA-LSSVM, CSA-LSSVM, and GMDH-PNN was carried out. They concluded that the proposed model was better in terms of accuracy with an average AARDof .44%.

In a more recent paper, Cardona and Valderrama [12] proposed a modeling approach based on a cubic equation of state and the concept of geometric similarity for predicting the surface tension of pure substances and mixtures containing organic substances, water and ILs of 90 mixtures binary (2660 data) and 12 ternary mixtures (467 data) considered in a wide range of temperatures from 278.15 to 348.15 K. For primary estimation, they concluded that the model is accurate in design and process simulation.

The use of alternative and complementary methods to experimentation such as quantitative structure–property/activity relationships (QSPR/QSAR) became of great interest and is the most widely adopted methods to augment experimental analytical techniques [13]. The development of QSPR/QSAR mathematical models linking physico-chemical properties and biological activities to a set of molecular descriptors allows to explain the origin of these activities/properties and to predict them for molecules whose experimental data are not available [14].

Klamt et al [15]. developed a quantum chemical approach (COSMO-RS) for the prediction of thermodynamic properties of pure and mixed polarity distribution. From the literature, the distribution area of the σ profile (Sσ profile) has been adopted as a quantitative measure representation of the polar surface screen charge of the molecule on the polarity scale, obtained from the histogram profile function σ given by the COSMO-RS calculation [14, 16]. The modeling of the properties of ILs is mainly based on the use of equations of state and the application of machine learning algorithms. These algorithms show that they have various applications in different fields such as the medical [17], the electrical and electronic engineering [18], the petrochemical [19, 20], chemical engineering [21, 22] and the civil and environmental engineering [23]. Among the alternative methods of computational intelligence, are the ANN, the LSSVM, and the SVM fine tuning with particle swarm optimization (SVM-PSO) are robust and accurate predictive methods that have recently been successfully applied for the prediction of various properties [6].

The aim of this work is to evaluate the ability of different machine learning techniques to model surface tension for correlation and/or prediction purposes. The study intends to estimate the surface tension of binary mixtures (IL-Water and various organic solvents) using three methods: a predictive ANN model, a LSSVM and a SVM fine tuning with PSO (SVM-PSO). The same set of inputs is considered for both types of models including temperature (T), composition of the mixture in mole fraction of IL (X_{IL}), two descriptors based on COSMO-RS sigma profiles of anion, cation and solvent are used as input variables of the model in order to differentiate the different compounds involved in the binary systems.

2 | MATERIALS AND METHODS

2.1 | Database

First of all, it should be mentioned that the experimental data used in this work are not exhaustive of all the data published in the literature. They are limited to binary mixtures containing ILs for which the data of the sigma profiles of anions and cations are available in the published literature. Thus, the experimental data, collected from numerous works, consists of 1623 surface tension (σ) data points relating to 62 binary mixtures of IL and molecular solvent at different temperatures and compositions. The cations involved are: imidazolium (Im), pyridinium (Py), ammonium (N) and

phosphonium (P), whereas the anions are: tetrafluoroborate (BF4), hexafluorophosphate (PF6), bis(trifluoromethylsulfonyl)imide (BTI), bromide (Br), chloride (CL), acetate (AC), alkyl sulfate (RSO4), dimethyl phosphate (DMP), trifluoromethylsulfonate (TfO), nitrate (NO3) and dicyanamide (Dca). Molecular components intended to cover common solvents such as water, alcohols, dimethyl sulfoxide, acetonitrile and tetrahydrofuran. Table 1 shows the source and domains of experimental data of the studied binary mixtures. The global database comprising the experimental data and the calculated sigma profile descriptors of the cation, the anion and the solvent can be found in the supplementary information file.

2.2 | COSMO-RS

Molecular descriptors based on σ profiles derive from the COSMO-RS theory considered as a continuous solvation model that combines quantum chemical theory, dielectric continuum models, surface interactions and statistical thermodynamics [24].

It calculates molecular interactions from the shielding charge (polarization) densities on surface molecular segments. Quantum chemical calculations provide a discrete surface segment around a single molecule embedded in a virtual conductor [25]. The surface of each segment is characterized by its area and the shielding charge density of the segment, which takes into account the electrostatic interaction of the solute molecule by its environment and retro-polarization of the solute molecule [26]. The screening charge density is usually given in the probability diagram of the statistical distribution of the charge density on the surface of a molecule, known as the σ profile which shows the probability of the relative amount of area with the σ polarity at the surface of the molecule considered as the characteristic properties of the molecule [27].

In other words, the σ profile of a molecule includes the main chemical information necessary to predict possible electrostatic, hydrogen bonding and dispersion interactions of the molecule in a fluid. σ-profiles have been shown to be an effective molecular descriptors for establishing QSPR models able to predict the physical, chemical and toxicological properties of ILs [13].

2.3 | Models development

This step consists of the selection of the input variables, which are the independent variables of the model. In this respect, the temperature (T), the mole fraction of IL (X_{IL}), two descriptors based on the sigma profiles, one relating to the donor character of H-bond, the other relating to the acceptor character of H- bond, for the anion, the cation and the solvent are used as input variables of the model in order to differentiate the different compounds involved in the binary systems given as follows:

$$\sigma_m = f(T, X_{IL}, S_{\sigma 1c}, S_{\sigma 2c}, S_{\sigma 1A}, S_{\sigma 2A}, S_{\sigma 1S}, S_{\sigma 2S}) \quad (1)$$

where $S_{\sigma 1c}$ and $S_{\sigma 2c}$ are the sigma profiles descriptors of the cation, $S_{\sigma 1A}$ and $S_{\sigma 2A}$ of the anion and $S_{\sigma 1S}$ and $S_{\sigma 2S}$ of the solvent.

2.3.1 | Artificial neural network models

Artificial neural networks are a computer model used to analyze data. Knowledge is acquired by the network through a learning process, and the connecting strengths of interneurons called synaptic weights are used to store it [28]. Their most important advantages are: their ability to find input traceability and their flexibility to test model interpolation, extrapolation and prediction [29]. ANNs, as parallel distributed systems, are generally composed of an input layer, some hidden layers and an output layer, each neuron is connected to the other neurons of a previous layer thanks to synaptic weights adaptable. Knowledge is usually stored as a set of connection weights [28]. Multilayer perceptron (MLP) networks and radial basis function (RBF) networks are two popular ANNs [30, 31].

The network determines the relationship between the variables and stores the values of the weights and biases that give the lowest error between the calculated and experimental data of the dependent variable through an optimization process using traditional backpropagation algorithms or evolutionary algorithms and genetic algorithms (GAs) [29, 32].

2.3.2 | Least squares support vector machine

The LSSVM which was suggested by Suykens and Vandewalle [33, 34] is considered as an alternative to the supervised SVM learning method proposed by Vapnik in the 1990s used for classification and regression to analyze data and identify patterns [35, 36]. This new version replaces the convex quadratic programming and inequality constraints of the original SVM by solving a linear set of equations and instead using equality

TABLE 1 Source and domains of experimental data of the studied binary mixtures

Solvent	LI	T (K)	X_{LI}	N°	References
Water	[C1MIm][DMP]	298.15	[.0000–1.0000]	16	[51]
	[C1MIm][MeSO4]	[296.8–298.1]	[.0000–.2920]	10	[52]
	[C2MIm][BF4]	[298.15–338.15]	[.0000–1.0000]	53	[7, 53]
	[C2MIm][DEP]	298.15	[.0000–1.0000]	16	[51]
	[C2MIm][DMP]	298.15	[.0000–1.0000]	11	[51]
	[C2MIm][EtSO4]	298.15	[.0062–.5791]	12	[54]
	[C2MIm][MeSO3]	[300.2–303.2]	[.0000–.4850]	27	[52]
	[C4MIm] [AC]	[298.15–338.15]	[.8533–1.0000]	49	[12]
	[C4MIm] [CL]	[298.1–302.1]	[.0000–.4800]	46	[12]
	[C4MIm] [PF6]	303.15	[.9268–.9969]	9	[12]
	[C4MIm][BF4]	[298.15–338.15]	[.0000–1.0000]	59	[7, 53]
	[C4MIm][DBP]	298.15	[.0000–1.0000]	11	[51]
	[C4MIm][DMP]	298.15	[.0000–1.0000]	11	[51]
	[C4Py] [BF4]	[293.15-323.15]	[.0000–1.0000]	72	[12]
	[C4Py][NO3]	298,15	[.0000–.9903]	15	[55]
	[C6MIm] [AC]	[298.15–338.15]	[.8597–1.0000]	49	[12]
	[C6MIm] [CL]	[297.2–298.2]	[.0000–.143]	10	[12]
	[C6MIm][BF4]	[298.15–338.15]	[0, 3–1.0000]	40	[7, 53]
	[C8MIm] [PF6]	[298.05–335.05]	.7827	5	[12]
	[N112(hoe)][Br]	298.15	[.9915–.9981]	3	[56]
	[N311(hoe)][Br]	298.15	[.9900–.9979]	3	[56]
	[P666(14)][BTI]	[298.1–343.3]	.0891	6	[57]
	[P666(14)][Dca]	[298.2–342.8]	.4932	6	[57]
Methanol	[C1MIm][DMP]	298.15	[.0000–1.0000]	11	[51]
	[C1MIm][MeSO4]	298.15	[.0000–1.0000]	9	[56]
	[C2MIm] [AC]	[278.2–318.15]	[.0000–1.0000]	55	[12]
	[C2MIm][DEP]	298.15	[.0000–1.0000]	22	[51]
	[C2MIm][EtSO4]	298.15	[.0000–1.0000]	11	[12]
	[C2MIm][MeSO4]	[293.15–298.15]	[.0000–1.0000]	21	[12, 58]
	[C4MIm][BTI]	[283.15-298.15]	[.1049–.9577]	24	[12]
	[C4MIm][DBP]	298.15	[.0000–1.0000]	11	[51]
	[C4MIm][DMP]	298.15	[.0000–1.0000]	11	[51]
	[C4Py] [BF4]	[293.15–323.15]	[.0000–1.0000]	56	[12]
	[C8MIm] [BTI]	[283.15–298.15]	[.0978–.9368]	20	[12]
Ethanol	[C1MIm][DMP]	298.15	[.0000–1.0000]	11	[51]
	[C2MIm] [AC]	[278.15–338.15]	[.0000–1.0000]	77	[12]
	[C2MIm][C8SO4]	298.15	[.0399–1.0000]	18	[59]
	[C2MIm][DEP]	298.15	[.0000–1.0000]	11	[51]
	[C2MIm][DMP]	298.15	[.0000–1.0000]	11	[51]
	[C2MIm][EtSO4]	298.15	[.0000–1.0000]	11	[12]
	[C2MIm][MeSO4]	293.15	[.0000–1.0000]	14	[12]
	[C4MIm][BF4]	298.15	[.0986–1.0000]	12	[53]
	[C4MIm][BTI]	[283.15–313.15]	[.0999–.9601]	36	[12]
	[C4MIm][DBP]	298,15	[.0000–1.0000]	11	[51]
	[C4MIm][DMP]	298.15	[.0000–1.0000]	11	[51]
	[C6MIm] [EtSO4]	298.15	[.03–1.0000]	13	[12]

TABLE 1 (Continued)

Solvent	LI	T (K)	X_{LI}	N°	References
	[C6MIm][BF4]	298.15	[.0980–1.0000]	10	[53]
	[C8MIm] [BTI]	[283.15–313.15]	[.1042–.9784]	39	[12]
	[C8MIm][BF4]	298.15	[.1012–1.0000]	10	[53]
Tetrahydrofuran	[C2MIm][BTI]	[293.15–308.15]	[.0000–1.0000]	44	[60]
	[C4MIm][BTI]	[293.15–308.15	[.0000–1.0000]	40	[61]
Acetonitrile	[C2MIm][BTI]	[293.15–313.15]	[.0000–1.0000]	45	[60]
	[C4MIm][BTI]	[293.15–313.15]	[.0000–1.0000]	45	[61]
	[PYR-4,1] [BTI]	[288.15–308.15]	[.0000–1.0000]	27	[12]
Dimethyl sulfoxide	[C2MIm][BTI]	[293.15–313.15]	[.0000–1.0000]	45	[61]
	[C4MIm][BTI]	[293.15–313.15]	[.0000–1.0000]	50	[61]
1-Propanol	[C2MIm] [AC]	[288.15–348.15]	[.0000–1.0000]	77	[12]
	[C4MIm][BTI]	[283.15–313.15]	[.0000–1.0000]	31	[12, 62]
	[C8MIm] [BTI]	[283.15–313.15]	[.0974–.9461	33	[12]
1-butanol	[C4MIm][BTI]	[283.15–313.15]	[.0000–1.0000]	29	[12, 62]
	[C8MIm] [BTI]	[298.15–318.15]	[.0000–1.0000]	48	[12]
1-Hexanol	[C8iQuin]BTI]	[298.15–318.15]	[.0000–1.0000]	24	[63]

constraints in the LSSVM method, regression error is applied to the optimization settings. In fact, in SVM algorithms, the regression error is minimized in the learning phase whereas in LSSVM methods, it is mathematically defined and solved [36].

The goal minimization principle in the LSSVM method can be expressed as the following cost function (1) subject to the consequent constraint (2):

$$\text{Minimize}: \text{cost function} = \frac{1}{2} w^T w + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2 \tag{2}$$

$$\text{Subjected to}: y_i = w^T \varphi(X_i) + b + e_i, i = 1, 2, \ldots N \tag{3}$$

where w^T represents the transpose matrix of w, e_i refers to regression errors, b is bias, γ denotes the regularization parameter which controls errors, and the subscript "i" expresses data points for training, and N represents the total number of training points.

Equation (4) is the Lagrangian form, used to solve the LSSVM problem:

$$l(w, b, e, a) = \frac{1}{2} w^T w + \frac{1}{2}\gamma \sum_{i=1}^{N} e_k^2 - \sum_{i=1}^{N} a_k \left(w^T \varphi(x_k) + b + e_k - y_k\right) \tag{4}$$

which a_i are Lagrangian multipliers.

The LSSVM problem is solved by equating the derivatives of the Lagrangian form to zero:

$$\begin{cases} \frac{\partial L}{\partial W} = 0 \Rightarrow w = \sum_{k=1}^{N} a_k \varphi(x_k) \\ \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{k=1}^{N} a_k = 0 \\ \frac{\partial L}{\partial e_k} = 0 \Rightarrow a_k = \gamma e_k, k = 1, 2, \ldots, N \\ \frac{\partial L}{\partial a_k} = 0 \Rightarrow w^T \varphi(x_k) + b + e_k - y_k = 0, k = 1, 2, \ldots, N \end{cases} \tag{5 - 8}$$

The solution of the LSSVM problem is possible by solving the above mentioned set of linear equations instead of the quadratic programming problem. SVM and LSSVM methods are kernel-based approaches. In this study, the RBF kernel function was applied according to Equation (9):

$$k(x, x_i) = exp\left(-\|x_i - x\|^2/\sigma^2\right) \tag{9}$$

where $\|x_i - x\|$ is the Euclidean distance of the i^{th} input from the center of x_c. There are two tuning parameters in LSSVM, that is, γ and σ^2. The parameters are tuned by minimizing the differences between the predicted values and their corresponding experimental values.

2.3.3 | Support vector machine

Support vector machine as an efficient type of supervised machine learning method, which was developed by Vapnik [37], is used for solving classification and nonlinear regression problems [38]. SVM has many attributes including good generalization ability to avoid over-fitting based on regularization, nonlinear classification ability based on kernel trick, and global error minimization based on convex optimization [39]. The regression version of SVM is called SVR with the central goal of finding the best line of fit in the hyperplane [40]. The prediction or approximation function used by a basic SVM is [41]:

$$f(x) = \sum_{i=1}^{l} \alpha_i k(x, x_i) + b \tag{10}$$

where x_i is a feature vector corresponding to a training object, $K(x, x_i)$ is a kernel function and α_i is some real value. The component of vector α and the constant b represent the hypothesis and are optimized during the training. $K(x, x_i)$ is a kernel function, which value is equal to the inner product of two vectors x and x_i in the feature space $\Phi(x)$ and $\Phi(x_i)$, that is,

$$K(x, x_i) = \Phi(x) \cdot \Phi(xi) \tag{11}$$

For a dataset, only the kernel function and the regularity parameter C should be selected to specify an SVM. Any function that satisfies Mercer's condition could be used as a kernel function. The Gaussian kernel:

$$K(u, v) = \exp\left(-\gamma^* |u - v|^2\right) \tag{12}$$

is the most commonly used in support vector regression.

2.3.4 | Particle swarm optimization

In 1995, PSO was introduced by Kennedy and Eberhart based on a social simulation model known as the stochastic optimization algorithm, inspired by the social behavior of flocks of birds or schools of fish. PSO is similar to GAs, in terms of initializing the population with random solutions and finding the optimum by updating the generations. However, unlike GA, PSO does not undergo crossover or mutation, as the particles move through the problem space following the current optimal particles. Since there are only a few parameters to set in PSO, it is easy to implement [42]. The underlying concept is that these particles move over the search area with flexible speed and maintain the best position they have discovered in the search space. Each particle can revise its velocity vector to explore the best position through its flight expertise and the flight expertise of other particles in the search space [19]. Mathematically, a swarm of particles is randomly initialized on the search space and moves through the D-dimensional space to search for new solutions. Let x^i_k and v^i_k respectively be the position and the velocity of the i^{th} particle in the search space at the k^{ith} iteration, then its velocity and the position of this particle at $(k + 1)$ the iteration are updated using the following equations [42]:

$$v^i_{k+1} = w.v^i_k + c_1.r_1.\left(p^i_k - x^i_k\right) + c_2.r_2.\left(p^g_k - x^i_k\right) \tag{13}$$

$$x^i_{k+1} = x^i_k + v^i_{k+1} \tag{14}$$

where r_1 and r_2 represent random numbers between 0 and 1, c_1 and c_2 are constants, p^i_k represents the best ever position of i^{th} particle, and p^g_k corresponds to the global best position in the swarm up to k^{ith} iteration.

2.4 | Statistical metrics

To assess the efficiency and the accuracy of the ANN models, several statistical parameters were used, namely, AARD percentage (AARD%), mean squared error (MSE), mean relative squared error (MRSE), Q2 measurements are briefly presented by three indicators (QF12, QF22, and QF32)

and the concordance correlation coefficient (Q2ccc), determination coefficient (R^2), correlation coefficient (R), accuracy factor (Af), the bias factor (Bf) and Akaike's information criterion (AIC).

Where *N* is the number of data points, Np is the number of parameters in the model and SSE is the sum of the squared error, y_i^{exp} is the value of experimental data sets at the sampling point i, y_i^{cal} is the i^{th} value of the corresponding predicted sampling point i, $\overline{y^{exp}}$, and $\overline{y^{cal}}$ are the average of the experimental and predicted data, k & k' are the slopes of the corresponding regression lines, r_{02} is the square of the correlation coefficient between the observed value and the predicted value of compounds without an intercept, r'_{02} has the same meaning as r_{02}, except that it uses the axes reversed. The mathematical equations of the above mentioned parameters, given by [11, 43–45], are as follows:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{exp} - y_i^{cal}\right)^2 \tag{15}$$

$$\mathrm{AARD}(\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i^{exp} - y_i^{cal}}{y_i^{exp}}\right| \times 100 \tag{16}$$

$$\mathrm{MRSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i^{exp} - y_i^{cal}}{y_i^{exp}}\right)^2 \tag{17}$$

$$Q_{F1}{}^2 = 1 - \frac{\sum_{i=1}^{n_{OUT}}\left(y_i - \widehat{y}_{i/i}\right)^2}{\sum_{i=1}^{n_{OUT}}\left(y_i - \overline{y}_{TR}\right)^2} \tag{18}$$

$$Q_{F2}{}^2 = 1 - \frac{\sum_{i=1}^{n_{OUT}}\left(y_i - \widehat{y}_{i/i}\right)^2}{\sum_{i=1}^{n_{OUT}}\left(y_i - \overline{y}_{OUT}\right)^2} \tag{19}$$

$$Q_{F3}{}^2 = 1 - \frac{\sum_{i=1}^{n_{OUT}}\left(y_i - \widehat{y}_{i/i}\right)^2/n_{OUT}}{\sum_{i=1}^{n_{TR}}\left(y_i - \overline{y}_{TR}\right)^2/n_{TR}} \tag{20}$$

$$Q^2{}_{ccc} = \frac{2.\sum_{i=1}^{n_{OUT}}\left(y_i - \overline{y}_{EXP}\right).\left(\widehat{y}_{i/i} - \overline{y}_{CAL}\right)}{\sum_{i=1}^{n_{OUT}}\left(y_i - \overline{y}_{EXP}\right)^2 + \sum_{i=1}^{n_{OUT}}\left(\widehat{y}_{i/i} - \overline{y}_{CAL}\right)^2 + n_{OUT}.\left(\overline{y}_{EXP} - \overline{y}_{CAL}\right)^2} \tag{21}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i^{exp} - y_i^{cal}\right)^2}{\sum_{i=1}^{N}\left(y_i^{exp} - \overline{y^{exp}}\right)^2} \tag{22}$$

$$R = \frac{\sum_{i=1}^{N}\left(y_i^{exp} - \overline{y^{exp}}\right) \times \left(y_i^{cal} - \overline{y^{cal}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i^{exp} - \overline{y^{exp}}\right)^2} \times \sqrt{\sum_{i=1}^{N}\left(y_i^{cal} - \overline{y^{cal}}\right)^2}} \tag{23}$$

$$k = \frac{\sum y_i^{exp} y_i^{cal}}{\sum \left(y_i^{cal}\right)^2} \tag{24}$$

$$k' = \frac{\sum y_i^{exp} y_i^{cal}}{\sum \left(y_i^{exp}\right)^2} \tag{25}$$

$$R_o{}^2 = 1 - \frac{\sum \left(y_i^{cal} - k y_i^{cal}\right)^2}{\sum \left(y_i^{cak} - \overline{y_i^{cal}}\right)^2} \tag{26}$$

$$R_o'^2 = 1 - \frac{\sum \left(y_i^{exp} - k' y_i^{exp}\right)^2}{\sum \left(y_i^{exp} - \overline{y^{exp}}\right)^2} \tag{27}$$

$$m = \frac{R^2 - {R_0}^2}{R^2} \leq .1 \tag{28}$$

$$n = \frac{R^2 - R_o'^2}{R^2} \leq .1 \tag{29}$$

$${R_m}^2 = R^2 \times \left(1 - \sqrt{\left|R^2 - {R_o}^2\right|}\right) \geq .5 \tag{30}$$

$$\overline{y^{exp}} = \frac{1}{N}\sum_{i=1}^{N} y_i^{exp} \tag{31}$$

$$\overline{y^{cal}} = \frac{1}{N}\sum_{i=1}^{N} y_i^{cal} \tag{32}$$

$$A_f = 10^{\left(\frac{\sum_{i=1}^{N}\left|log\frac{y_i^{cal}}{y_i^{exp}}\right|}{N}\right)} \tag{33}$$

$$B_f = 10^{\left(\frac{\sum_{i=1}^{N} log\frac{y_i^{cal}}{y_i^{exp}}}{N}\right)} \tag{34}$$

$$AIC = N ln\left(\frac{SSE}{N}\right) + 2N_p + \frac{2N_p(N_p+1)}{N-(N_p+1)} \tag{35}$$

$$SSE = \sum_{i=0}^{N}\left(y_i^{obs} - y_i^{pred}\right)^2 \tag{36}$$

The prediction of the model is ideal if AARD%, MSE, MRSE, R^2, R, A*f*, and B*f* are found to be very close to 0, 0, 0, 1, 1, 1, and 1 respectively. A small AIC value indicates the better fitted model.

These metrics were used as acceptability criteria, as shown in Table 2.

3 | RESULTS AND DISCUSSION

Experimental data of surface tension in binary mixtures are published at different temperatures and mixture compositions. The cations and anions constituting the ILs considered in this study are listed in Table 3.

TABLE 2 Acceptability criteria (AC) of a model

Parameters	m, n	Q2	R^2	k, k'	*Bf*, *Af*
AC values	<.1	>.5	>.6	.85 < < 1.15	.9–1.05

3.1 | Data from sigma profile descriptors

The first step in the development of the model consists of defining appropriate molecular descriptors from the σ profiles of the cations, anions and the solvents of the binary mixtures studied. Usually, the σ profile of a solute or solvent is divided into four sections; each section is defined by an interval of σ (e/nm^2). In this work, the sigma profile of IL solvents, anions and cations is divided into two sections. The σ profile descriptor, S_σ, previously mentioned, represents the area under the curve of the σ profile as shown in Figure 1. Thus, $S_{\sigma 1}$, $S_{\sigma 2}$ represent the hydrogen bond donor and acceptor character respectively.

In this work, a MATLAB program was used to calculate the areas $S\sigma_i$ (i = 1,...,6) of the considered cations, anions and solvents after digitizing the σ profile curves obtained from the reference [46] and the digitized σ profiles for the water and organic compounds. The solvents obtained from the bases VT-2005 and VT-2006, respectively [27, 47].

3.2 | Dividing data into training set, test set and validation set

The training dataset consists of 70% randomly chosen data using the divisor function of MATLAB 2018. The remaining 30% was used as a test set for LSSVM and SVM-PSO. In the case of ANN modeling, these 30% were divided into two sets of 15% for testing and 15% for validation. The purpose of this division is to avoid over-fitting during training.

TABLE 3 List of cations and anions

Cation name	Abbreviations	Anion name	Abbreviations
1,3-dimethylimidazolium	[C1MIm]	Dimethylphosphate	[DMP]
1-Ethyl-3-methylimidazolium	[C2MIm]	Methylsulfate	[$MeSO_4$]
1-Butyl-3-methylimidazolium	[C4MIm]	Tetrafluoroborate	[BF4]
1-Hexyl-3-methylimidazolium	[C6MIm]	Bis(trifluoromethylsulfonyl)imide	[BTI]
1-octyl-3-methylimidazolium	[C8MIm]	Octyl sulfate	[C8SO4]
1-butylpyridinium	[C4Py]	Diethyl phosphate	[DEP]
N-octylLsoquinolinium	[C8iQuin]	Ethyl sulfate	[EtSO4]
Ethyl(2-hydroxyethyl)dimethylammonium	[N112(hoe)]	Methane sulfonate	[MeSO3]
Propyl(2-hydroxyethyl)dimethylammonium	[N311(hoe)]	Methylsulfate	[MeSO4]
trihexyltetradecylphosphonium	[P666(14)]	Dibutyl phosphate	[DBP]
		Nitrate	[NO3]
		Bromide	[Br]
		Dicyanamide	[Dca]
1-butyl-1-methylpyrrolidinium	[PYR-4,1]	Acetate	[AC]
		Chloride	[CL]
		Hexafluorophosphate	[PF6]

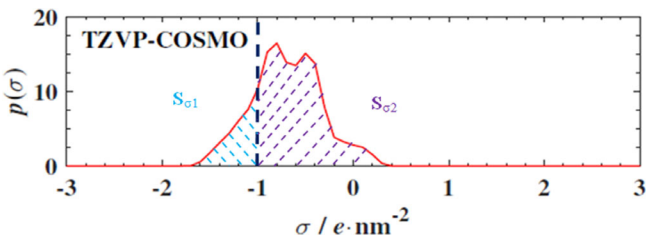


FIGURE 1 Molecular descriptor based on σ profiles.

3.2.1 | ANN modeling

Based on many previous studies, the best algorithms were used in this study such as: the Levenberg-Marqhardt backpropagation learning algorithm (MATLAB function trainlm) and the Bayisian normalization algorithm (MATLAB function trainbr). Concerning the activation function, the logsig and tansig functions for the hidden layers and the purelin function for the output layer were tested.

A program on MATLAB 2018 was developed, in which a modification of the number of neurons in the hidden layer, the learning algorithm and the function of activating the hidden layer was made. The program was performed more than 20 times for the same structure to obtain the model that gives the best MSE. The performance of model was evaluated using statistical metrics mentioned above and graphical tools.

3.2.2 | LSSVM modeling

LSSVM modeling was performed using the LS-SVMlab toolbox. To do this, an appropriate algorithm implemented in LS-SVMlab detects and scales continuous, categorical and binary variables. A step called tuning consists in determining the LSSVM parameters (regularization constant γ and kernel parameter σ^2) by minimizing a selected performance measure. The constants α and b of the LSSVM model were determined from the learning step which uses the regularization constant γ and the kernel parameter σ^2 adjusted in the first step. The last step consisted in testing the generalization capacity of the model on all the data reserved for the test. The optimized parameters were 3548.43 and 1.6510 for γ and σ^2, respectively.

3.2.3 | SVM-PSO modeling

The development of the SVM-PSO model is carried out by the *fitrsvm* function of MATLAB environment. This function has several calculation options relating to the choice of the cross-validation method, the kernel function (gaussian, rbf, and polynomial), the optimal values of the constant C (boxConstraint), epsilon and the parameter of the kernel function (KernelScale) of the model were optimized with PSO algorithm (100 iterations). In this study, we have chosen the Cross-validation method « holdout » which was set at .3 which means 70% of the data is used for learning and 30% for validation.

The optimal values are 200, 1 and .006 for c, k and ep respectively with gaussian as Kernel function.

Table 4 summarizes respectively the calculation of the different statistical metrics for the four topologies of the ANN model ("trainbr, logsig"; "trainbr, tansig"; "trainlm, logsig"; "trainlm, tansig"), for the SVM-PSO model and for the LSSVM model.

Table 5 summarizes respectively the calculation of the different errors for the four topologies of the ANN model, for the SVM-PSO model and for the LSSVM model.

In the light of these results, all the calculated statistical metrics verify the conditions of acceptability. However, the results for the model optimized by using the trainbr training algorithm and the logsig activation function showed the best significant results: a global correlation coefficient R of .9979, the coefficient of determination R^2 of .9958 and the accuracy factor Af of 1.0085. In addition, the error parameters such as: the MSE (.4952) and the AARD% (.8466). This makes it possible to prove the efficiency and robustness of the model developed, which has more precise correlation performance and much better generalization and interpolation capabilities. From Table 6, the smallest value obtained of AIC represents the best fitting model for the six models studied.

Graphical evaluation was performed using scatterplots (for training, testing, validation, and whole data set), % relative error (ER%) plot, and % relative error (ER%) distribution plot. Figure 2 shows the scatter plot of the calculated surface tension as a function of the experimental surface tension for the four topologies of the ANN model, the SVM-PSO model and the LSSVM model for the global ensemble. Figure 3 represents the scatter plot of the calculated surface tension as a function of the experimental surface tension for the optimized model trainbr and logsig. This diagram clearly shows the concordance between the calculated values and the experimental values by a tight dispersion of the points on the first bisector. However, the existence of a number of outliers was noticed.

In order to quantitatively visualize the distribution of relative errors, two types of diagrams were adopted. The first shows the relative error as a function of the surface tension represented in Figure 4 and the second shows the distribution of this error represented in Figure 5. This shows that the majority of the points are located around the zero line. Quantitatively, the relative error for the few outliers did not exceed 20%.

To confirm the generalization capacity of the developed model, we carried out interpolation and extrapolation calculations for the composition and the temperature. Figure 6 shows the isotherms of the variation in surface tension as a function of the composition of the [C2MIm][BTI]-Dimethyl sulfoxide mixture for the established model, the interpolated points follow the shape of the calculated points. In the same way, Figure 7 represent the interpolation and extrapolation calculation of the temperature. The interpolation (305.15 K) and extrapolation (290.15 and 315.15 K) isotherms approximately follow the shape of the calculated points and have the same shapes of the adjacent experimental isotherms.

TABLE 4 Statistical metrics of the four ANN model topologies, the SVM-PSO model, and the LSSVM model

(a): Trainbr, logsig				(b): Trainbr, tansig			
Statistical metric	**Train**	**Test**	**Global**	**Statistical metric**	**Train**	**Test**	**Global**
Q2F1	.9955	.9964	.9958	Q2F1	.9951	.9962	.9955
Q2F2	.9956	.9964	.9958	Q2F2	.9952	.9962	.9955
Q2F3	.9955	.9965	.9958	Q2F3	.9951	.9958	.9953
Q2CCC	.9978	.9982	.9979	Q2CCC	.9976	.9981	.9977
R^2	.9955	.9964	.9958	R^2	.9951	.9962	.9955
R	.9978	.9982	.9979	R	.9976	.9982	.9977
K	1.0000	.9994	.9999	K	1.0000	1.0030	1.0008
k'	.9996	1.0003	.9998	k'	.9996	.9967	.9988
R_0^2	1.0000	1.0000	1.0000	R_0^2	1.0000	.9999	1.0000
$R_0'^2$	1.0000	1.0000	1.0000	$R_0'^2$	1.0000	.9999	1.0000
M	−.0045	−.0036	−.0043	M	−.0049	−.0037	−.0046
N	−.0045	−.0036	−.0043	N	−.0049	−.0037	−.0046
R_m	.9291	.9365	.9309	R_m	.9258	.9358	.9284
A_f	1.0080	1.0098	1.0085	A_f	1.0083	1.0116	1.0091
B_f	1.0001	1.0003	1.0002	B_f	1.0001	.9983	.9996
(c): Trainlm, logsig				**(d): Trainlm, tansig**			
Statistical metric	**Train**	**Test**	**Global**	**Statistical metric**	**Train**	**Test**	**Global**
Q2F1	.9946	.9907	.9934	Q2F1	.9953	.9941	.9949
Q2F2	.9947	.9906	.9934	Q2F2	.9953	.9941	.9949
Q2F3	.9946	.9907	.9935	Q2F3	.9953	.9938	.9949
Q2CCC	.9953	.9973	.9967	Q2CCC	.9977	.9971	.9975
R^2	.9946	.9906	.9934	R^2	.9953	.9941	.9949
R	.9973	.9953	.9967	R	.9977	.9971	.9975
K	.9999	1.0002	1.0000	K	1.0001	1.0013	1.0005
k'	.9997	.9990	.9995	k'	.9995	.9982	.9991
R_0^2	1.0000	1.0000	1.0000	R_0^2	1.0000	1.0000	1.0000
$R_0'^2$	1.0000	1.0000	1.0000	$R_0'^2$	1.0000	1.0000	1.0000
M	−.0054	−.0095	−.0066	M	−.0047	−.0059	−.0051
N	−.0054	−.0095	−.0066	N	−.0047	−.0059	−.0051
R_m	.9217	.8945	.9129	R_m	.9272	.9182	.9242
A_f	1.0101	1.0150	1.0116	A_f	1.0091	1.0128	1.0102
B_f	1.0003	.9990	.9999	B_f	.9999	.9981	.9993
(e): LSSVM modeling				**(f): SVM-PSO modeling**			
Statistical metric	**Train**	**Test**	**Global**	**Statistical metric**	**Train**	**Test**	**Global**
Q2F1	.9906	.9408	.9754	Q2F1	.9721	.9376	.9614
Q2F2	.9907	.9406	.9754	Q2F2	.9722	.9374	.9613
Q2F3	.9394	.9906	.9752	Q2F3	.9721	.9343	.9607
Q2CCC	.9953	.9694	.9875	Q2CCC	.9855	.9677	.9800
R^2	.9906	.9406	.9754	R^2	.9721	.9374	.9613
R	.9953	.9699	.9876	R	.9863	.9687	.9808
K	1.0002	1.0011	1.0005	K	1.0064	1.0081	1.0069
k'	.9991	.9942	.9976	k'	.9915	.9870	.9901
R_0^2	1.0000	1.0000	1.0000	R_0^2	.9995	.9991	.9994
$R_0'^2$	1.0000	.9996	.9999	$R_0'^2$	.9991	.9979	.9988

(Continues)

TABLE 4 (Continued)

(e): LSSVM modeling				(f): SVM-PSO modeling			
Statistical metric	**Train**	**Test**	**Global**	**Statistical metric**	**Train**	**Test**	**Global**
M	−.0095	−.0631	−.0252	M	−.0282	−.0658	−.0396
N	−.0095	−.0627	−.0251	N	−.0278	−.0645	−.0389
R_m	.8947	.7114	.8225	R_m	.8112	.7046	.7739
A_f	1.0082	1.0219	1.0123	A_f	1.0084	1.0217	1.0124
B_f	1.0003	1.0004	1.0003	B_f	.9967	.9928	.9955

TABLE 5 Different errors of the four ANN model topologies, the SVM-PSO model and the LSSVM model

Trainbr, logsig				**Trainbr, tansig**			
Errors	**Train**	**Test**	**Global**	**Errors**	**Train**	**Test**	**Global**
AARD%	.7992	.9812	**.8466**	AARD%	.8256	1.1526	.9108
MRSE	.7235	.6441	.7037	MRSE	.7416	.6917	.7289
MSE	.5235	.4149	**.4952**	MSE	.5499	.4785	.5313
Trainlm, logsig				**Trainlm, tansig**			
Errors	**Train**	**Test**	**Global**	**Errors**	**Train**	**Test**	**Global**
AARD%	1.0111	1.4779	1.1511	AARD%	.9102	1.2568	1.0141
MRSE	.7936	1.0421	.8756	MRSE	.7346	.8414	.7682
MSE	.6298	1.0860	.7667	MSE	.5396	.7079	.5901
LSSVM modeling				**SVM-PSO**			
Errors	**Train**	**Test**	**Global**	**Errors**	**Train**	**Test**	**Global**
AARD%	.8150	2.1653	1.2201	AARD%	.8050	2.0399	1.1755
MRSE	1.0430	2.6514	1.6942	MRSE	1.7923	2.7482	2.1247
MSE	1.0878	7.0301	2.8705	MSE	3.2122	7.5524	4.5143

Note: Bold values signific the best results.

TABLE 6 AIC values of the four ANN model topologies, the SVM-PSO model and the LSSVM model

	AAN model					
Model	**Trainbr, logsig**	**Trainbr, tansig**	**Trainlm, logsig**	**Trainlm, tansig**	**LSSVM model**	**SVM-PSO model**
AIC	**−1178.675**	−1058.967	−435.523	−88.721	1808.694	2578.408

Note: Bold values signific the best results.

3.3 | Methods for quantifying variable importance in ANNs

To determine the relative importance of the input variables, different statistical methods were applied.[[48, 49]]In the present work, Garson's method was used to split the hidden output connection weights into components associated with each input neuron using absolute connection weight values. as well as the relevance factor r, which is in the range of −1 to +1 and is given by the following equation [22]:

$$r_k = \frac{\sum_{i=1}^{n} \left(X_{k,i} - \overline{X}_k\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_{k,i} - \overline{X}_k\right)^2 \sum_{i}^{n} \left(Y_i - \overline{Y}\right)^2}} \tag{37}$$

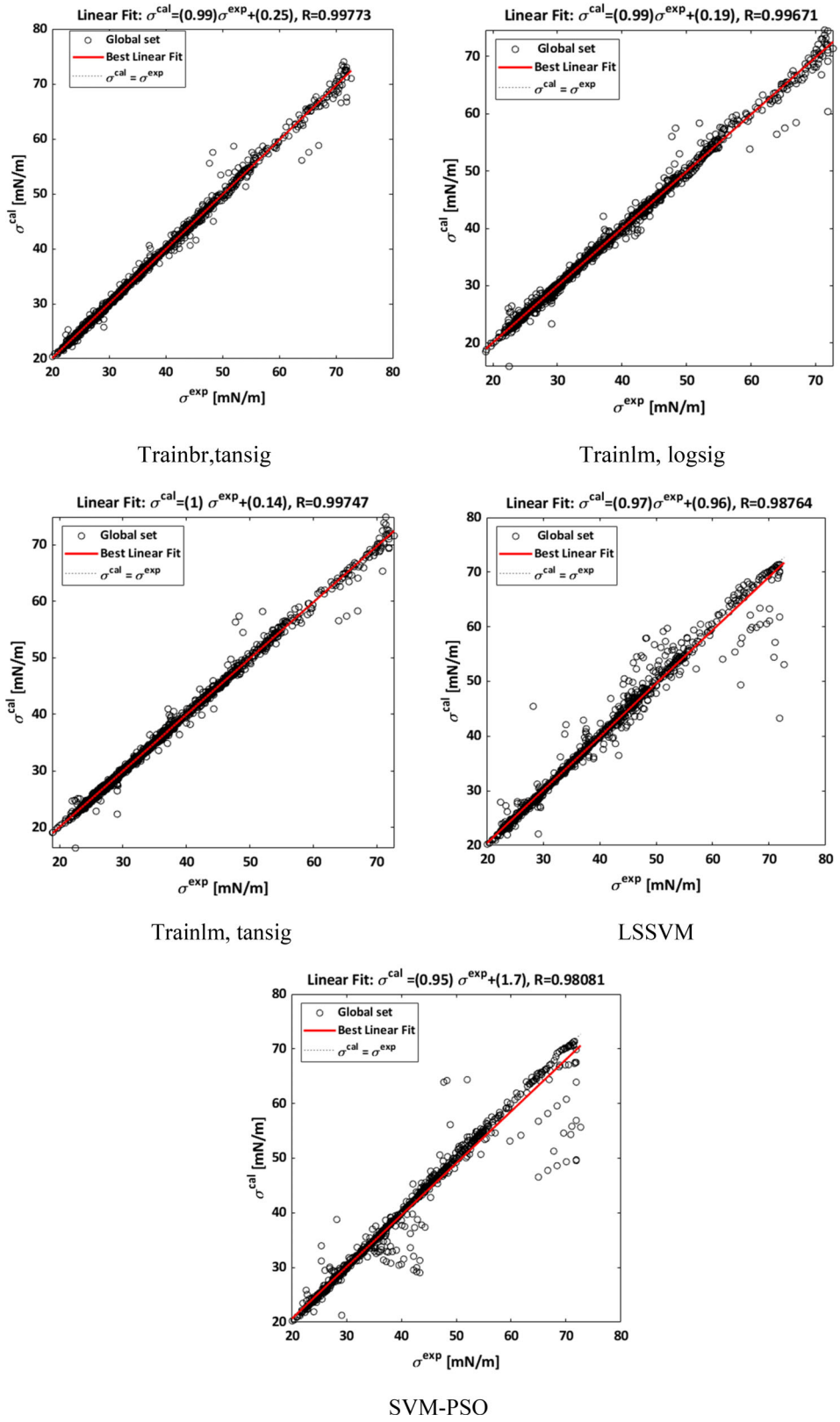


FIGURE 2 Surface tension scatter plot computed as a function of the experimental surface tension for the global basis by the three ANN model topologies, the SVM-PSO model, and the LSSVM model.

where $X_{k,i}$ is the i^{th} importing parameter, Y_i is the i^{th} exporting value, $\overline{X}_k$ is the average value of the k^{th} input, $\overline{Y}$ is the average value of exporting parameter, and n is the number of sets.

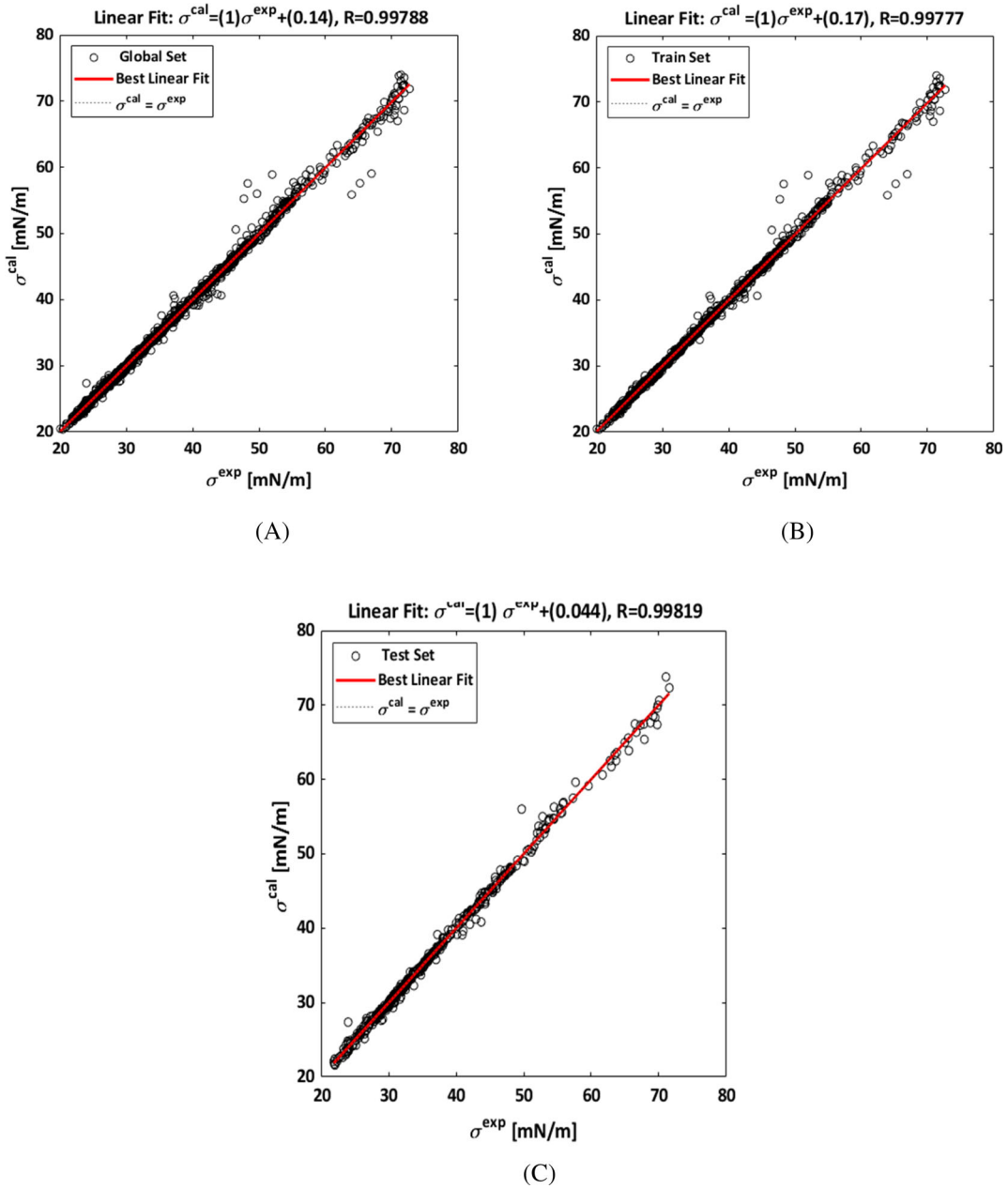


FIGURE 3 Scatter plot of the surface tension calculate as a function of the experimental surface tension: (A) Global data set; (B) Training set; (C) Test set by the best model of the ANN « trainbr, logsig ».

It can be seen from Figure 8 that for the established model, all input variables have almost the same Garson's method importance with the exception of temperature. This can be explained by the low variability of the temperature data, that is, for some binary mixtures the experimental data are available only at one or two temperatures. However, the sigma profile descriptor related to the H-bond acceptor character of the solvent, $S_{\sigma 2S}$, has the highest relative importance (20.26%) in Figure 9, the surface tension showed a straight dependency on the descriptors named: S_σ2A, S_σ1S, and S_σ2S and an opposite dependency to the rest of the descriptors. The most relevant descriptors are the S_σ2S, and temperature with a relevance factor of +.021 and −.0235, respectively. Contrariety, S_σ1A and S_σ2A were the least relevant, with a relevance factor of −.0052 and +.0024, respectively.

3.4 | Outlier diagnostics

The scatter plot showed a number of points far from the first bisector which can be attributed to the probable existence of outliers. This allows to perform a diagnosis of potentially suspicious data according to the method described in the following two references [31, 50].

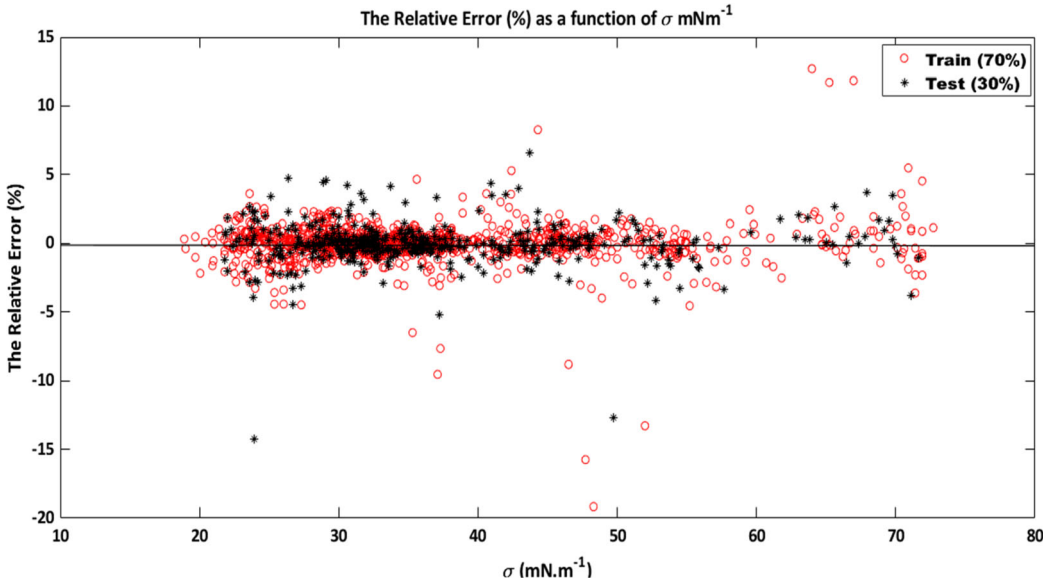


FIGURE 4 The relative error (%) of the surface tension by the ANN model "trainbr, logsig".

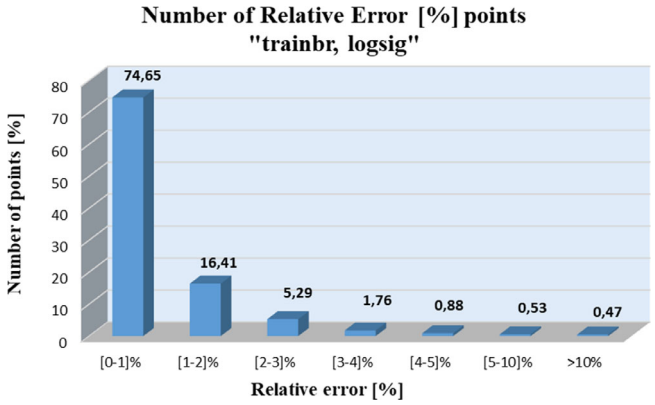


FIGURE 5 Distribution diagrams of RE (%) of the ANN model "trainbr, logsig".

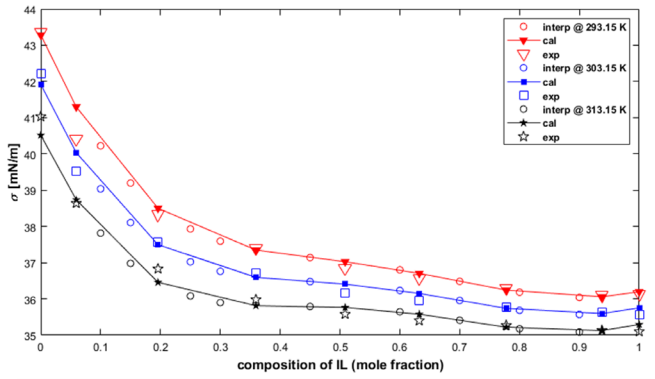


FIGURE 6 The interpolation and extrapolation calculations for the composition for the isotherms of the variation of surface tension as a function of the composition of the mixture [C2MIm][BTI]-Dimethyl sulfoxide for the "trainbr, logsig" ANN model.

Therefore, it is essentially necessary to find rigorous methods to detect outliers in order to remove inaccurate experimental data and improve model accuracy. In this study, an outlier detection method based on the Williams plot was used. This graph explains the relationship between the Hat indices represented by Equation (38) and the residuals defined as differences between the experimental data and the corresponding estimated values (R):

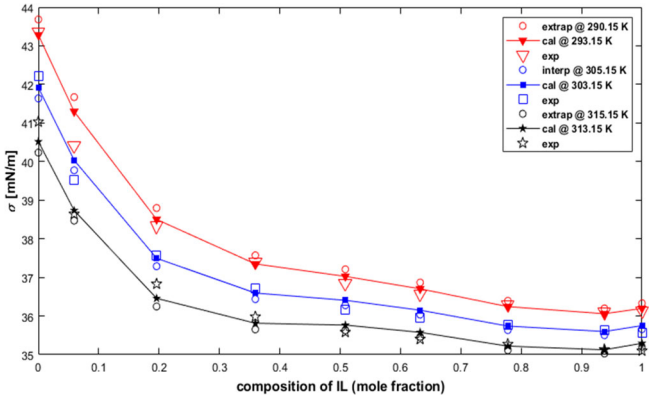


FIGURE 7 The interpolation and extrapolation calculations for the temperature for the isotherms of the variation of surface tension as a function of the composition of the mixture [C2MIm][BTI]-Dimethyl sulfoxide for the "trainbr, logsig" ANN model.

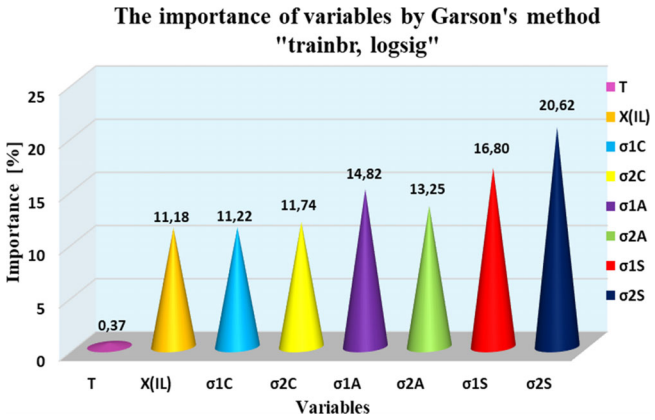


FIGURE 8 The relative importance of the variables by the Garson method of the ANN model "trainbr, logsig".

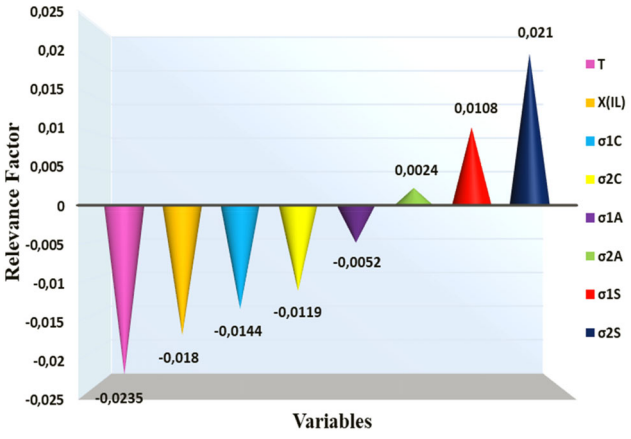


FIGURE 9 The results of the relevance factor performed on the ANN model « trainbr, logsig ».

$$H = X(X^tX)^{-1}X^t \tag{38}$$

where X refers to the $m \times n$ matrix (m and n represent the number of samples and the parameters [input variables] of the model, respectively). The values of Hat are obtained from the main diagonal of the matrix of H.

Figure 7 shows William's plot based on the results of the ANN model. In this chart, the critical leverage (H*) (threshold) is usually set to the value given by the following equation:

$$H^* = \frac{3(n+1)}{m} \tag{39}$$

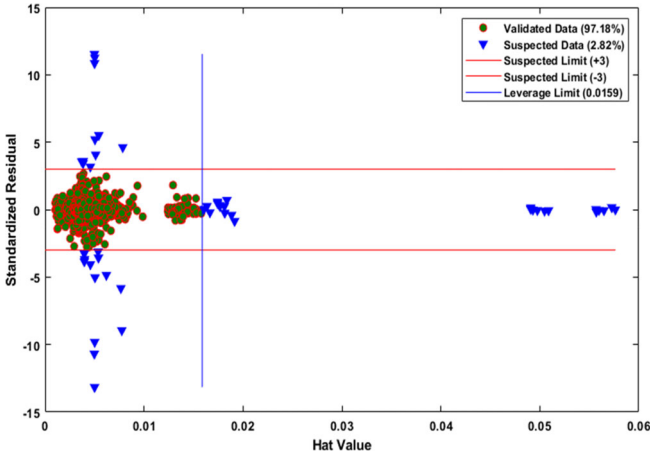


FIGURE 10 Diagnosis of potentially suspicious data and domain of model applicability.

TABLE 7 Comparison between this work and the work of Huang et al [6] in terms of AARD%

		This work		Huang et al [6]. model			
Systems {Ionique liquids + solvents}	Number of points	Average I %Δσ I	I%Δσ I$_{max}$	Average I%Δσ$_{SM}$I	I %Δσ I$_{max}$	Average I%Δσ$_{ANN}$I	I%Δσ I$_{max}$
[C4MIm][BTI] + 1-butanol	13	2.13690	14.2458	2.39918	4.83586	2.99069	5.56928
[C4MIm][BTI] + 1-propanol	11	.94144	2.64873	1.35199	3.04797	1.23155	2.18982
[C2MIm][BTI] + acetonitrile	45	.58221	2.03760	1.10309	3.61209	.35043	1.63301
[C4MIm][BTI] + acetonitrile	45	.28893	1.22930	1.32803	4.15549	.4058	.99813
[C2MIm][BTI] + Dimethyl sulfoxide	45	.48975	2.16922	1.59518	4.79112	.13389	.49756
[C4MIm][BTI] + Dimethyl sulfoxide	50	.90014	19.16957	2.29169	8.29236	.15671	.53247
[C2MIm][C8SO4] + ethanol	18	1.01942	2.10372	1.84492	4.20381	.44156	1.42752
[C4MIm][BF4] + ethanol	12	.86002	2.50953	.70218	1.78356	1.0015	6.5699
[C6MIm][BF4] + ethanol	10	3.20687	2.82941	1.14056	1.50424	.74439	3.40772
[C8MIm][BF4] + ethanol	10	.76441	2.28718	.82575	1.84567	1.17232	2.60078
[C1MIm][MeSO4] + methanol	9	.89797	19.1695	1.04938	9.30710	.56188	3.46873
[C2MIm][MeSO4] + methanol	7	.69485	1.75784	.61541	2.47781	.67376	3.26845
[C2MIm][BTI] + tetrahydrofuran	44	.93363	3.58668	.46162	1.46741	.46749	1.4991
[C4MIm][BTI] + tetrahydrofuran	40	1.22362	4.48224	.58329	2.60047	.63473	11.623
[C1MIm][MeSO4] + water	10	.86816	3.80506	.95716	2.41303	.33636	1.47093
[C2MIm][BF4] + water	9	.63134	1.20231	.18729	1.07727	.19068	.50178
[C2MIm][EtSO4] + water	12	.57436	2.21932	1.59173	3.93446	.60338	1.77915
[C2MIm][MeSO3] + water	27	1.73691	3.70604	3.89171	6.62818	.33508	1.02659
[C4MIm][BF4] + water	15	.73874	1.46092	.55320	1.96164	.4154	1.07962
[C4Py][NO3] + water	15	1.04161	6.56224	3.71712	9.3071	1.6599	11.623
[C6MIm][BF4] + water	8	.33916	.57460	.19374	.38724	.20127	.43478
[P666(14)][BTI] + water	6	.16640	.34975	1.41799	2.47305	.11722	.22418
[P666(14)][Dca] + water	6	.16995	.38829	1.24882	2.36076	.33602	.54937

where *m* is the number of samples, and n is the number of input variables of the model.

The normalized residuals are calculated from the data of the experimental surface tension and that calculated by the model.

$$(R_\mathrm{Norm})_i = \frac{\left(\sigma_i^{exp} - \sigma_i^{cal}\right)}{\sqrt{Var\left(\sigma^{exp} - \sigma^{cal}\right)}} i = 1, \ldots m \tag{40}$$

The normalized residual of three is normally considered a "threshold" value for accepting data points within a range of ±3 *SD*s from the mean (to cover 99% of the normally distributed data). If the majority of the data points fall within the ranges of $0 \le \mathrm{Hat} \le H^*$ and $-3 \le \mathrm{R_Norm} \le 3$, this indicates that the model development and its predictions are performed in the applicability domain, leading to a statistically valid model. Thus, we can affirm that there are "Good High Leverage" points in the domain of $0 \le \mathrm{Hat} \le H^*$ and $-3 \le \mathrm{R_Norm} \le 3$. The points located in the range of R_Norm<−3 or R_Norm>3 (whether larger or smaller than the H^* value) are referred to as model outliers or "Bad High Leverage" points.

Figure 10 shows that the majority of data points are lying inside the applicability domains, except for the existence of 2.82% of suspected data. This confirms the observations made during the interpretations of the scatter plots and the relative error plots.

TABLE 8 Comparison between this work and the work of with the Cardona and Valderrama [12] in terms of AARD%

		This work « trainbr, logsig »		Cardona and Valderrama [12] model	
Systems {ionic liquid + solvent}	***N***	**Average l%Δσl**	**l%Δσl$_{max}$**	**Average l%Δσl**	**l%Δσl$_{max}$**
[C2MIm] [AC] + 1-propanol	77	.727311	3.279583	1.73871	22.48719
[C8MIm] [BTI] + 1-propanol	33	.513259	2.078608	6.64114	14.11421
[PYR-4.1] [BTI] + acetonitrile	27	.805194	2.216206	1.708758	4.497877
[C2MIm] [AC] + ethanol	77	.419793	2.068807	7.379528	20.68375
[C2MIm][DEP] + ethanol	11	1.369486	4.561168	11.7884	22.08567
[C2MIm][EtSO4] + ethanol	11	.994978	2.051358	13.25353	22.58346
[C4MIm][BTI] + ethanol	36	.38637	.784236	6.433625	18.33271
[C4MIm][DMP] + ethanol	11	1.343429	2.666525	14.05437	26.31037
[C6MIm] [EtSO4] + ethanol	13	1.091658	2.951107	3.927588	8.77661
[C8MIm] [BTI] + ethanol	39	.44928	1.518148	5.317242	15.36972
[C1MIm][DMP] + methanol	11	.562193	1.096063	6.937792	11.1255
[C2MIm] [AC] + methanol	55	.728288	3.611441	2.137411	9.412672
[C2MIm][EtSO4] + methanol	11	.955441	2.273106	5.599972	10.49703
[C2MIm][MeSO4] + methanol	14	.519645	2.936068	6.746018	15.85075
[C4MIm][BTI] + methanol	24	.194314	.717207	2.579008	6.910141
[C4MIm][DMP] + methanol	11	.63623	1.737604	2.965154	5.431559
[C4Py] [BF4] + methanol	56	.438397	2.181546	7.621787	18.83593
[C8MIm] [BTI] + methanol	20	1.102813	4.23808	2.704397	6.313109
[C1MIm][DMP] + water	16	1.289812	2.932375	1.380303	71.90
[C2MIm][DEP] + water	16	2.236424	12.66962	4.54	18.54
[C2MIm][DMP] + water	11	1.186484	3.990389	4.665865	9.781574
[C4MIm] [AC] + water	49	.385264	1.20122	3.213681	9.603819
[C4MIm] [CL] + water	46	.882057	2.076755	2.79283	7.822873
[C4MIm] [PF6] + water	9	.416549	1.264713	3.662293	7.550244
[C4MIm][DBP] + water	11	.941997	4.423875	5.830528	71.90
[C4Py] [BF4] + water	72	.694389	3.243486	8.413314	24.03831
[C6MIm] [AC] + water	49	.345018	1.013274	3.203462	10.76396
[C6MIm] [CL] + water	10	2.258253	5.479549	17.48413	26.5232
[C6MIm][BF4] + water	8	.339157	.574605	6.808245	10.50152

3.5 | Comparison with other models

In this section, a comparison study was carried out to evaluate the model's performance compared to previous papers. To the best knowledge of the authors, there is no paper with a similar database to the dataset used in this paper. For that a comparison in terms of common systems was adopted.

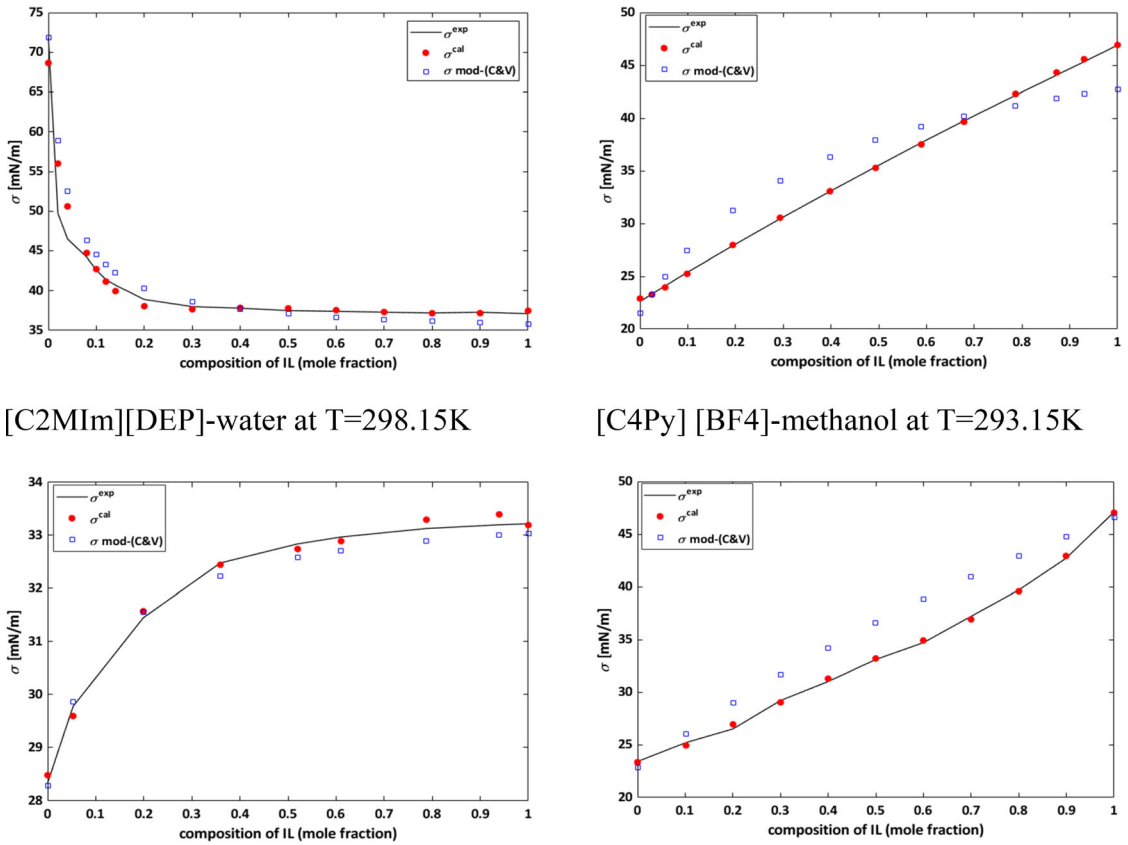


FIGURE 11 Comparison between experimental (lines) and calculated surface tension values by our model (red dots) and the Valderrama model [12].(blue squares) model for some systems.

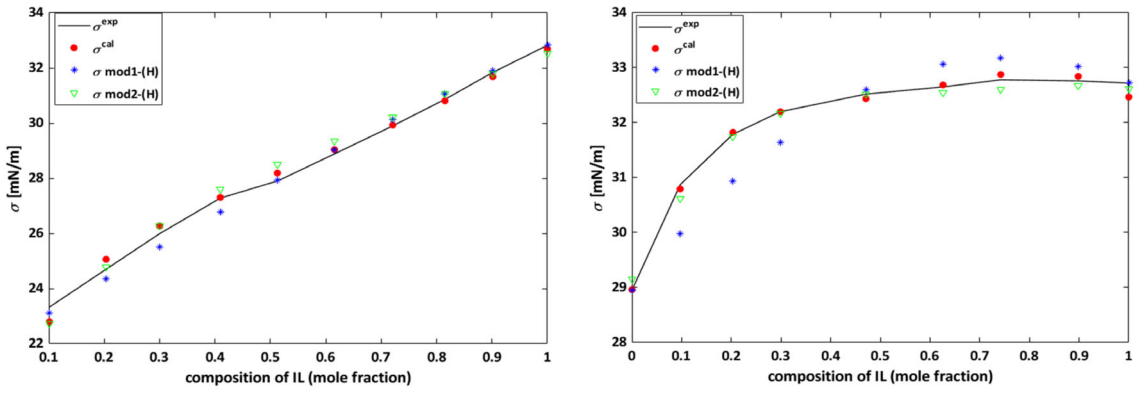


FIGURE 12 Comparison between experimental (lines) and calculated surface tension values by our model (red dots) and the Huang et al [6]. (green triangle) for his ANN model and (blue stars) for his SM model for some systems.

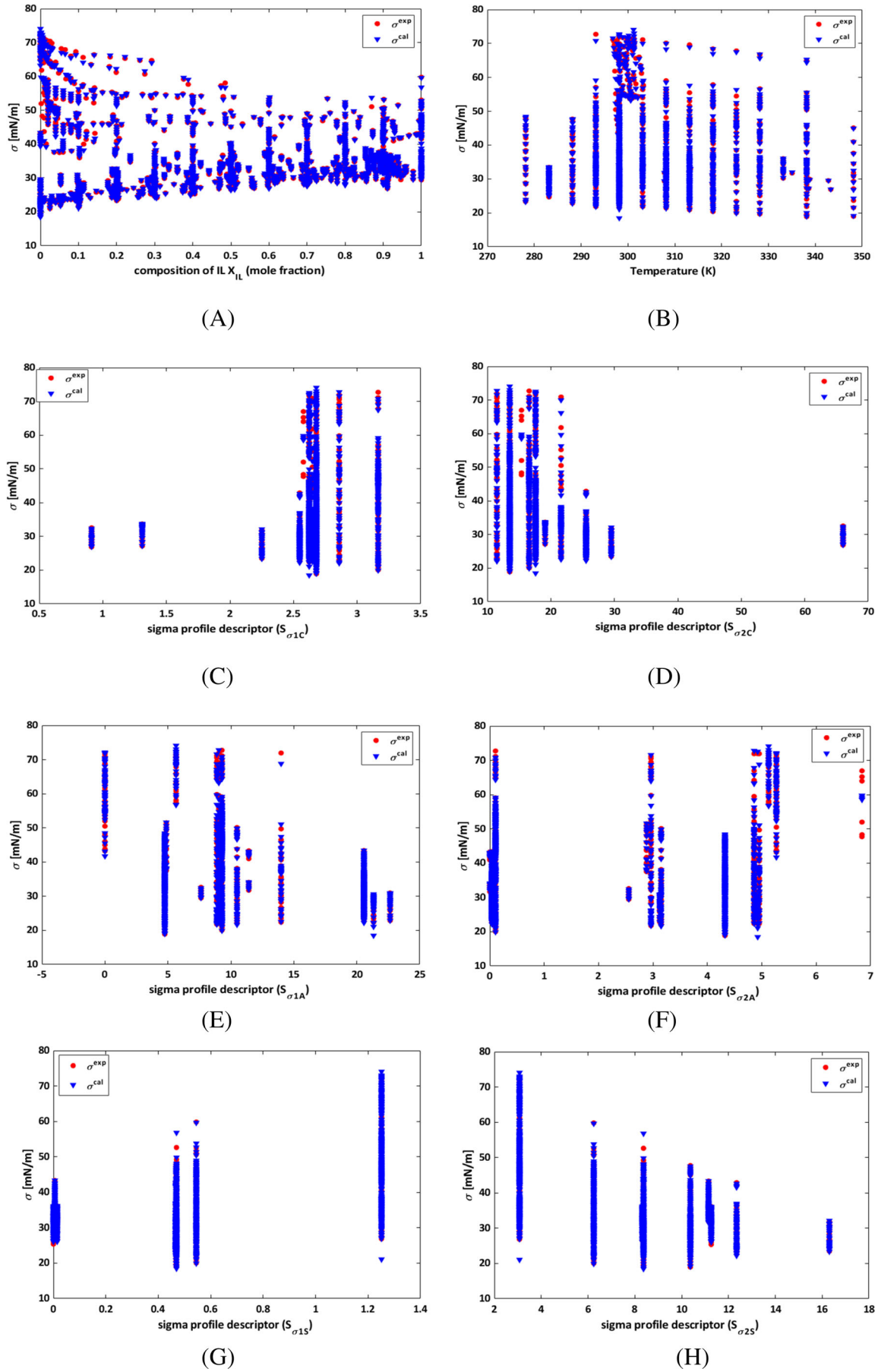


FIGURE 13 Comparison between the predicted based and the experimental data versus the input data: (A) composition of IL, (B) Temperature, (C) S_σ1C, (D) S_σ2C, (E) S_σ1A, (F) S_σ2A, (G) S_σ1S, (H) S_σ2S

A comparison, in term of the error (AARD%), between this work and the work of Huang et al [6]., Cardona and Valderrama.[12]is given in Tables 7 and 8, respectively and the full comparison is included in the supplementary information file. Figures 11 and 12 represent another comparison between the experimental and calculated surface tension of some systems between the main model of this work and the previous mentioned papers.

Figure 13 illustrates a comparison between the calculated versus the experimental values of the ANN model (trainbr, logsig) according to the inputs.

4 | CONCLUSION

A prediction model based on an ANN has been successfully developed to predict the surface tension of binary mixtures. A total of 1623 experimental data points of 62 binary mixtures were collected from various literature resources for use in the ANN model as training, validation and test data points. The best architecture of the feed-forward network, obtained by a constructive approach, consisted of 26 neurones in the hidden layers and was trained by the trainbr algorithm when the Logsigmoid (logsig) activation function in the hidden layer was applied. The following conclusions can be drawn:

1. Significant results were obtained with the proposed ANN method. This fact is supported by the acceptable statistical quality confirmed by various parameters and the low errors of the ANN model results. The overall % AARD obtained and the MSE of .8466% and .4952%, respectively, showed a very good capability and feasibility of using the ANN model for the prediction of surface tension of binary mixtures.
2. The established model is of great practical importance, it allows not only to accurately predict the surface tension of binary mixtures, including ILs, but also promotes this method for other physico-chemical properties of IL mixtures in future studies to overcome limitations in the development of industries and technologies based on the IL.

AUTHOR CONTRIBUTIONS

Widad Benmouloud: Methodology; validation; visualization; writing – original draft; writing – review and editing. **Cherif Si-Moussa:** Supervision; writing – original draft; writing – review and editing. **Othmane Benkortbi:** Supervision; writing – original draft; writing – review and editing.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of the Algerian Ministry of Higher Education and Scientific Research (PRFU Project A16N01UN260120220003) and the University Yahia Fares of Medea.

DATA AVAILABILITY STATEMENT

The supplementary data to this article can be found online at (supporting file.pdf).

ORCID

Widad Benmouloud https://orcid.org/0000-0002-9456-8494
Othmane Benkortbi https://orcid.org/0000-0002-1965-7171

REFERENCES

[1] Q. Zhang, S. Feng, X. Zhang, Y. Wei, *J. Mol. Liq.* **2021**, *328*, 115373.
[2] K. J. Wu, C. X. Zhao, C. H. He, *Fluid Phase Equilib.* **2012**, *328*, 42.
[3] S. Atashrouz, H. Mirshekar, A. Hemmati-Sarapardeh, M. K. Moraveji, B. Nasernejad, *Korean J. Chem. Eng.* **2017**, *34*, 425.
[4] S. Atashrouz, M. Mozaffarian, G. Pazuki, *Ind. Eng. Chem. Res.* **2015**, *54*, 8600.
[5] A. Shojaeian, M. Asadizadeh, *J. Mol. Liq.* **2020**, *298*, 111976.
[6] Y. Huang, X. Zhang, Y. Zhao, S. Zeng, H. Dong, S. Zhang, *Phys. Chem. Chem. Phys.* **2015**, *17*, 26918.
[7] A. Shojaeian, *Thermochim. Acta* **2019**, *673*, 119.
[8] R. Sedev, *Curr. Opin. Colloid Interface Sci.* **2011**, *16*, 310.
[9] M. Tariq, M. G. Freire, B. Saramago, J. A. P. Coutinho, J. N. C. Lopes, L. P. N. Rebelo, *Chem. Soc. Rev.* **2012**, *41*, 829.
[10] M. Hashemkhani, R. Soleimani, H. Fazeli, M. Lee, A. Bahadori, M. Tavalaeian, *J. Mol. Liq.* **2015**, *211*, 534.
[11] R. Soleimani, A. H. Saeedi Dehaghani, N. A. Shoushtari, P. Yaghoubi, A. Bahadori, *Korean J. Chem. Eng.* **2018**, *35*, 1556.
[12] L. F. Cardona, J. O. Valderrama, *Ionics (Kiel)* **2020**, *26*, 6095.
[13] Y. Benguerba, I. M. Alnashef, A. Erto, M. Balsamo, B. Ernst, *J. Mol. Struct.* **2019**, *1184*, 357.
[14] Y. Zhao, Y. Huang, X. Zhang, S. Zhang, *Phys. Chem. Chem. Phys.* **2015**, *17*, 3761.
[15] F. Eckert, A. Klamt, *AIChE J.* **2002**, *48*, 369.
[16] T. Lemaoui, N. E. H. Hammoudi, I. M. Alnashef, M. Balsamo, A. Erto, B. Ernst, Y. Benguerba, *J. Mol. Liq.* **2020**, *309*, 113165.
[17] S. Gambhir, S. Kumar, Y. Kumar, *New Horiz. Transl. Med.* **2017**, *4*, 1.

[18] M. Geethanjali, S. M. Raja Slochanal, R. Bhavani, *Neurocomputing* **2008**, *71*, 904.
[19] M. A. Ahmadi, Z. Chen, *Petroleum* **2019**, *5*, 271.
[20] M. Ahmadi, Z. Chen, *J. Pet. Explor. Prod. Technol.* **2020**, *10*, 2873.
[21] H. Benimam, C. S. Moussa, M. Hentabli, S. Hanini, M. Laidi, *J. Chem. Eng. Data* **2020**, *65*, 3161.
[22] I. Euldji, C. Si-Moussa, M. Hamadache, O. Benkortbi, *Mol. Inform.* **2022**, *2200026*, 1.
[23] L. T. Le, H. Nguyen, J. Dou, J. Zhou, *Appl. Sci.* **2019**, *9*, 2630.
[24] T. Lemaoui, A. S. Darwish, N. E. H. Hammoudi, F. Abu Hatab, A. Attoui, I. M. Alnashef, Y. Benguerba, *Ind. Eng. Chem. Res.* **2020**, *59*, 13343.
[25] J. S. Torrecilla, J. Palomar, J. Lemus, F. Rodríguez, *Green Chem.* **2010**, *12*, 123.
[26] M. Diedenhofen, A. Klamt, *Fluid Phase Equilib.* **2010**, *294*, 31.
[27] E. Mullins, R. Oldland, Y. A. Liu, S. Wang, S. I. Sandler, C. C. Chen, M. Zwolak, K. C. Seavey, *Ind. Eng. Chem. Res.* **2006**, *45*, 4389.
[28] S. A. Kalogirou, *Renew. Sustain. Energy Rev.* **2000**, *5*, 373.
[29] H. Benimam, C. Si-Moussa, M. Laidi, S. Hanini, *Neural Comput. Appl.* **2020**, *32*, 8635.
[30] Z. Wan, Q. De Wang, J. Liang, *Int. J. Quantum Chem.* **2021**, *121*, 1.
[31] A. Baghban, A. H. Mohammadi, M. S. Taleghani, *Int. J. Greenh. Gas Control* **2017**, *58*, 19.
[32] S. Abdel-khalek, A. Alhag, M. Ragab, S. M. Abo-Dahab, A. Algarni, H. Ahmad, *Int. J. Quantum Chem*. **2021**, *121*, e26446.
[33] Y. Sang, H. Zhang, L. Zuo, 2008 IEEE Int. Conf. Cybern. Intell. Syst. CIS 2008, **2008**, 290.
[34] J. A. Suykens, J. Vandewalle, *Neural Process. Lett*. **1999**, *9*, 293.
[35] M. N. Kardani, A. Baghban, J. Sasanipour, A. H. Mohammadi, S. Habibzadeh, *J. Cleaner Prod.* **2018**, *203*, 601.
[36] S. P. Mousavi, S. Atashrouz, M. Nait Amar, F. Hadavimoghaddam, M. R. Mohammadi, A. Hemmati-Sarapardeh, A. Mohaddespour, *J. Mol. Liq.* **2021**, *342*, 116961.
[37] N. H. Farhat, *IEEE Expert. Syst. Appl.* **1992**, *7*, 63.
[38] I. Mehraein, S. Riahi, *J. Mol. Liq.* **2017**, *225*, 521.
[39] Q. Song, G. Yan, G. Tang, F. Ansari, *Mech. Syst. Signal Process.* **2021**, *146*, 107019.
[40] Y. Zhao, X. Zhang, L. Deng, S. Zhang, *Comput. Chem. Eng.* **2016**, *92*, 37.
[41] J. Wang, H. Du, H. Liu, X. Yao, Z. Hu, B. Fan, *Talanta* **2007**, *73*, 147.
[42] H. Garg, *Appl. Math. Comput.* **2016**, *274*, 292.
[43] R. Steele, *Understanding and Measuring the Shelf-Life of Food*, Woodhead Publishing, **2004**.
[44] R. Todeschini, D. Ballabio, F. Grisoni, *J. Chem. Inf. Model.* **1905**, *2016*, 56.
[45] O. Falyouna, O. Eljamal, I. Maamoun, A. Tahara, Y. Sugihara, *J. Colloid Interface Sci.* **2020**, *571*, 66.
[46] K. Paduszyński, *Phys. Chem. Chem. Phys*. **2017**, *19*, 11835.
[47] E. Mullins, Y. A. Liu, A. Ghaderi, S. D. Fast, *Ind. Eng. Chem. Res.* **2008**, *47*, 1707.
[48] J. D. Olden, M. K. Joy, R. G. Death, *Ecol. Modell.* **2004**, *178*, 389.
[49] M. Gevrey, I. Dimopoulos, S. Lek, *Ecol. Modell.* **2003**, *160*, 249.
[50] M. Hosseinzadeh, A. Hemmati-Sarapardeh, *J. Mol. Liq.* **2014**, *200*, 340.
[51] N. N. Ren, Y. H. Gong, Y. Z. Lu, H. Meng, C. X. Li, *J. Chem. Eng. Data* **2014**, *59*, 189.
[52] J. W. Russo, M. M. Hoffmann, *J. Chem. Eng. Data* **2011**, *56*, 3703.
[53] E. Rilo, J. Pico, S. García-Garabal, L. M. Varela, O. Cabeza, *Fluid Phase Equilib.* **2009**, *285*, 83.
[54] J. S. Torrecilla, T. Rafione, J. García, F. Rodrígue, *J. Chem. Eng. Data* **2008**, *53*, 923.
[55] J. Y. Wang, X. J. Zhang, Y. Q. Hu, G. Di Qi, L. Y. Liang, *J. Chem. Thermodyn.* **2012**, *45*, 43.
[56] U. Domańska, A. Pobudkowska, M. Rogalski, *J. Colloid Interface Sci.* **2008**, *322*, 342.
[57] H. F. D. Almeida, J. A. Lopes-Da-Silva, M. G. Freire, J. A. P. Coutinho, *J. Chem. Thermodyn.* **2013**, *57*, 372.
[58] J. Y. Wang, F. Y. Zhao, Y. M. Liu, X. L. Wang, Y. Q. Hu, *Fluid Phase Equilib.* **2011**, *305*, 114.
[59] E. Rilo, M. Domínguez-Pérez, J. Vila, L. M. Varela, O. Cabeza, *J. Chem. Thermodyn.* **2012**, *49*, 165.
[60] M. Geppert-Rybczyńska, J. K. Lehmann, A. Heintz, *J. Chem. Eng. Data* **2011**, *56*, 1443.
[61] M. Geppert-Rybczyńska, J. K. Lehmann, J. Safarov, A. Heintz, *J. Chem. Thermodyn.* **2013**, *62*, 104.
[62] A. Wandschneider, J. K. Lehmann, A. Heintz, *J. Chem. Eng. Data* **2008**, *53*, 596.
[63] U. Domańska, M. Zawadzki, A. Lewandrowska, *J. Chem. Thermodyn.* **2012**, *48*, 101.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: W. Benmouloud, C. Si-Moussa, O. Benkortbi, *Int. J. Quantum Chem.* **2022**, e27026. https://doi.org/10.1002/qua.27026