A Neural Networks Model for Accurate Prediction of the Flash Point of Chemical Compounds

Mirshahvalad, HamidReza; Ghasemiasl, Ramin*
Department of Mechanical Engineering, West Tehran Branch, Islamic Azad University, Tehran, I.R. IRAN

Raufi, Nahid
Department of Chemical Engineering, South Tehran Branch, Islamic Azad University, Tehran, I.R. IRAN

Malekzadeh dirin, Mehrdad
Department of Mechanical Engineering, West Tehran Branch, Islamic Azad University, Tehran, I.R. IRAN

ABSTRACT: Flashpoint is one of the most important flammability characteristics of chemical compounds. In the present study, we developed a neural network model for accurate prediction of the flashpoint of chemical compounds, using the number of hydrogen and carbon atoms, critical temperature, normal boiling point, acentric factor, and enthalpy of formation as model inputs. Using a robust strategy to efficiently assign neural network parameters and evaluate the authentic performance of the neural networks, we could achieve an accurate model that yielded average absolute relative errors of 0.97, 0.96, 0.99 and 1.0% and correlation coefficients of 0.9984, 0.9985, 0.9981 and 0.9979 for the overall, training, validation and test sets, respectively. These results are among the most accurate ever reported ones, to date.

KEYWORDS: Flashpoint; Predictive models; Neural Networks; QSPR; Group contribution method.

INTRODUCTION
Flash Point (FP) is one of the most important flammability properties of flammable liquids in evaluating their quality in logistic chains including storage and handling [1,2]. Compounds with lower FP can be flamed at lower temperatures and after ignition, the rate of flame spread over fuels with lower FP is faster which necessitates more fire safety considerations for them [3-6].

The aim of the present work is to develop a model to predict the closed cup flash point [7] of pure chemical compounds using neural networks. Neural networks are one of the most robust modeling tools which works based on machine learning and has already been used in modeling various properties of chemical compounds [8-11]. FP is defined as the lowest temperature at which the vapors of a compound can be flamed by an ignition source. Developing accurate predictive models has always been an active area in science as it is highly required for many scientific and industrial applications. The number of compounds with undetermined properties are uncountable. Sometimes, experimental determination of those properties is not always possible, e.g. when the experimental measurement is dangerous, expensive or not

*To whom correspondence should be addressed. + E-mail: ramin2@ghasemiasl.ir

1021-9986/2020/4/297-304 8/8/5.08
operationally feasible. For such cases, using reliable predictive models result in a considerable saving in time, effort and operational costs.

The predicted FP of pure compounds is also used to evaluate the FP of mixture of chemical compounds [12, 13].

The models which predict FP of pure compounds can be generally divided into three main categories. The first category includes the Group Contribution Method (GCM) based models which relate a property to the constituting functional groups of a compound. The most accurate GCM based models use artificial neural networks to model the nonlinear relationship between the functional groups and FP [14-16]. A drawback of the classic GCM models is that they cannot distinguish between the properties of isomers which are characterized by the same functional groups. To overcome this limitation, a combination of Normal Boiling point (NBP) and functional groups as input variables was suggested in some studies [3, 17-19].

The second category of FP predictive models are Quantitative Structure Property Relationship (QSRP) models which use a more extensive set of structure related parameters, known as molecular descriptors, to predict a property [20]. Similar to the GCM based models, the most successful QSPR models also exploit ANNs to map the nonlinear relationship between the molecular descriptors and FP [21-24].

The third category is empirical models which predict a property using other more readily accessible or more convenient to measure properties. The main advantage of the correlations compared to the GCM and QSPR models is that their higher accuracy and more straightforward application. The most successful empirical correlations typically exploit vapor-liquid equilibrium related properties e.g. Normal Boiling Point (NBP), critical temperature ($T_c$), vapor pressure and enthalpy of vaporization (Δ$H_v$) as they represent the fuel volatility which is inversely proportional to the FP [25-28].

For example, Patil [29] and Hshieh [30] proposed a quadratic correlation between the NBP (K) and FP (K) as follows:

$$FP = a + b \times NBP + c \times NBP^2 \quad (1)$$

where a, b and c are constants. Riaizi and Daubert used NPB to predict FP by a nonlinear correlation in the [31]:

$$\frac{1}{FP} = -1.4568 \times 10^{-2} + 2.84947 \times NBP - 1.903 \times 10^7 \times \ln(NBP) \quad (2)$$

Catoire and Naudet proposed a correlation which predicts $FP$ as a function of the NBP, number of carbons (n) and enthalpy of vaporization (Δ$H_v$) as follows [32]:

$$FP = 1.477 \times NBP^{0.79686} \times \Delta H_v^{0.16443} \times n^{0.05948} \quad (3)$$

Gharagheizi et. al. proposed the following correlation to predict the FP using NBP, acentric factor ($\omega$), critical temperature ($T_c$) and critical pressure ($P_c$) [33]:

$$FP = 0.7327 \times NBP + 0.53 \times \frac{\omega T_c}{NBP} \times A + 5.4226 \quad (4)$$

where

$$A = P_c + \frac{M_w}{\omega P_c} + \frac{4.1848}{\omega} + \frac{2.7283 \times NBP}{\omega P_c} + \frac{1.5132 \times NBP}{P_c} \quad (5)$$

In the present study, we introduce an empirical model for prediction of the $FP$ using the number of carbon ($n_c$) and hydrogen ($n_H$) atoms, $T_c$, NBP, enthalpy of formation of ideal gas state (Δ$H_f$), and $\omega$ as model inputs. The relationships between the model inputs and FP is studied using feed forward neural networks, based on the strategies proposed by Alibakhshi [34]. Except for the ideal gas enthalpy of formation and number of hydrogen atoms, other parameters have already been used in other FP predictive models, as discussed before. The data of Δ$H_f$ can be measured either experimentally or by available highly accurate theoretical approaches [35].

**EXPERIMENTAL SECTION**

**Dataset**

The reliable data of $T_c$, NBP, Δ$H_f$ and $\omega$ for a large number of compounds from diverse families was supplied by the DIPPR 801 database [36]. Considering the accuracy of the data indicated by DIPPR, only the experimentally determined data with the uncertainty of less than 3% were selected which resulted in a dataset of 393 compounds. The full list of the studied compounds is reported in the supporting materials.
Developing ANN models

The correlation between the input variables and FP was studied via feedforward neural networks with one hidden layer. The reliable performance of ANNs as well as the efficient number of neurons, assigned transfer functions and training algorithm and ANN parameters were evaluated using the strategies proposed by Alibakhshi [34]. Based on that, the dataset was first randomly divided into three subsets, namely training, validation and test sets. The training data set containing randomly selected 75% of the compounds was used for training the ANNs and finding the optimum weight and bias constants of the networks. 13% of the compounds were randomly selected for cross validation and the increase in their root mean square error for 6 successive iterations was used as a condition to stop the training. The rest 12% of the compounds were used for testing the performance of the ANN after training. Considering the size of the training dataset, we studies ANNs with 1-5 neurons in the hidden layer, where the upper bond was selected to fulfill the existence of roughly 10 compounds per ANN parameter, as suggested by Alibakhshi [34]. The randomly division of dataset to training, validation and test sets were repeated 20 times and for each one, 20 different set of ANN parameters were randomly assigned and for each one, Levenberg-Marquardt backpropagation (trainlm) and Gradient descent backpropagation (traind) training algorithms and tangent sigmoid (tansig) and log-sigmoid (logsig) transfer functions were studied, using a Matlab code [53].

The performance of the model was evaluated using Average Absolute Deviation (AAD), average absolute relative error (AARE%), and correlation coefficient (R) stated as follows:

$$\text{AAD} = \frac{1}{N} \sum \left( \left| y_i^{\text{exp}} - y_i^{\text{pred}} \right| \right)$$

$$\text{AARE}\% = \frac{1}{N} \sum \left( \left| \frac{y_i^{\text{exp}} - y_i^{\text{pred}}}{y_i^{\text{exp}}} \right| \times 100 \right)$$

$$R = \frac{N \sum y_i^{\text{exp}} y_i^{\text{pred}} - (\sum y_i^{\text{exp}})(\sum y_i^{\text{pred}})}{\sqrt{N \sum (y_i^{\text{exp}})^2 - (\sum y_i^{\text{exp}})^2} \sqrt{N \sum (y_i^{\text{pred}})^2 - (\sum y_i^{\text{pred}})^2}}$$

Where $y_i^{\text{exp}}$ and $y_i^{\text{pred}}$ are the experimentally determined and predicted data of FP for the compound $i$ respectively.

In the next step, the ANNs for which the overall AARE% was lower than 1.5% were selected for further refinement and evaluation of their reliability and authentic performance, based on the method suggested by Alibakhshi [34]. Accordingly, the ANNs selected in the next step were retrained again for the same initially assigned parameters but 20 different randomly division of the dataset and for each repeat, the t-test statistical method was used to compare the errors of the test and training sets. The ANNs for which in all of the 20 repeats the AARE% of the test and training dataset were not significantly different with 0.95 of significance level, were selected as the efficiently trained models with low risk of overfitting and the average of 20 repeats also were considered as the authentic performance of that model.

RESULTS AND DISCUSSION

Considering 20 different dataset division and for each one studying 20 different initialization, 2 different training algorithm, 2 different transfer function and 5 different assigned number of neurons for the hidden layer, a total number of 8000 ANNs were initially studied. Among initially studied ANNs, 3947 of them yielded an overall AARE% of lower than 1.5% and were selected for retraining using 20 different randomly assigned training, validation and test sets.

After retraining in the next step, for only 495 out of 3947 models, the AARE% of the training and test sets were found to be not significantly different in all 20 repeats as evaluated by t-test statistical test. Those models were selected as the efficiently trained ANNs. Fig. 1 depicts the number of efficiently trained models based on the assigned number of hidden layer neurons.

Among the efficiently trained models, the best result was observed for an ANN with 4 neurons in the hidden layer, trainlm training algorithm and logsig transfer function for the hidden layer. This model yielded AARE% of 0.97, 0.96, 0.99 and 1.0% and correlation coefficients of 0.9984, 0.9985, 0.9981 and 0.9979 for the overall, training, validation and test sets, respectively.

For the selected ANN, the initially assigned and final optimized values for the weight and bias constant of each neuron are reported in Tables 1.

According to the results, the new model for 67.9% of compounds could predict the FP with accurate relative...
Table 1: Initially assigned and final optimized weight and bias constants of the network.

<table>
<thead>
<tr>
<th>Initial weights (inputs to neuron 1)</th>
<th>Initial weights (inputs to neuron 2)</th>
<th>Initial weights (inputs to neuron 3)</th>
<th>Initial weights (inputs to neuron 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.54358</td>
<td>-0.65686</td>
<td>0.625666</td>
<td>0.326525</td>
</tr>
<tr>
<td>0.097088</td>
<td>1.377202</td>
<td>0.635659</td>
<td>0.333176</td>
</tr>
<tr>
<td>0.732389</td>
<td>0.13732</td>
<td>-1.00035</td>
<td>-0.09939</td>
</tr>
<tr>
<td>0.880638</td>
<td>-0.27203</td>
<td>-0.62159</td>
<td>0.160549</td>
</tr>
<tr>
<td>-0.3301</td>
<td>0.725605</td>
<td>0.635374</td>
<td>-0.59474</td>
</tr>
<tr>
<td>1.177079</td>
<td>-0.40471</td>
<td>-0.72457</td>
<td>-1.5825</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optimized weights (inputs to neuron 1)</th>
<th>Optimized weights (inputs to neuron 2)</th>
<th>Optimized weights (inputs to neuron 3)</th>
<th>Optimized weights (inputs to neuron 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.2879</td>
<td>1.777513</td>
<td>0.941787</td>
<td>-1.39325</td>
</tr>
<tr>
<td>2.707046</td>
<td>0.154273</td>
<td>-0.27656</td>
<td>0.910771</td>
</tr>
<tr>
<td>2.39006</td>
<td>-0.38348</td>
<td>0.137078</td>
<td>-0.42175</td>
</tr>
<tr>
<td>0.498074</td>
<td>0.196007</td>
<td>0.755638</td>
<td>2.528354</td>
</tr>
<tr>
<td>0.059078</td>
<td>-0.1885</td>
<td>0.399716</td>
<td>-1.14864</td>
</tr>
<tr>
<td>0.617524</td>
<td>-0.037</td>
<td>0.084474</td>
<td>0.028263</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial weights (hidden to output)</th>
<th>Optimized weights (hidden to output)</th>
<th>Initial bias (hidden layer)</th>
<th>Optimized bias (hidden layer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.37353</td>
<td>-1.49238</td>
<td>1.763889</td>
<td>6.933026</td>
</tr>
<tr>
<td>0.916688</td>
<td>2.271763</td>
<td>0.587965</td>
<td>-1.04681</td>
</tr>
<tr>
<td>0.276929</td>
<td>1.542926</td>
<td>0.587965</td>
<td>1.866295</td>
</tr>
<tr>
<td>-0.58372</td>
<td>-0.17034</td>
<td>1.763889</td>
<td>2.149752</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial bias (output layer)</th>
<th>Optimized bias (output layer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.33376</td>
<td>-0.31287</td>
</tr>
</tbody>
</table>

Fig. 1: The number of efficiently trained ANNs for each assigned number of hidden layer neurons.

Fig. 2: Distribution of compounds versus AARE% range.

The results obtained via the proposed model are compared with the reported results for other most successful models in Table 2. As can be seen in Table 2, the new model provides lowest error compared to other available models. However, a more realistic comparison can be made only if different models are compared for a same dataset.

error less than 1%. The maximum AARE% with the value of 6.15% was observed for methane amine. The distribution of the compounds versus the relative error of the predicted FPs is depicted in Fig. 2 and Fig. 3 depicts the comparison of the predicted and experimentally determined data for the training, validation and test datasets.
Table 2: Comparison of the results of the developed model with other accurate models reported in the literature.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>No. data</th>
<th>AAD (k)</th>
<th>AARE (%)</th>
<th>Max. AARE (%)</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>The new model</td>
<td>ANN</td>
<td>393</td>
<td>2.68</td>
<td>0.97</td>
<td>6.15</td>
<td>0.9984</td>
</tr>
<tr>
<td>Albahri (2015)</td>
<td>GCM+ANN</td>
<td>375</td>
<td>3.55</td>
<td>1.1</td>
<td>6.62</td>
<td>0.9961</td>
</tr>
<tr>
<td>Mathieu (2010)</td>
<td>Semi-empirical</td>
<td>92</td>
<td>3.75</td>
<td>1.37</td>
<td>5.4</td>
<td>0.9922</td>
</tr>
<tr>
<td>Pan et al. (2007)</td>
<td>QSPR</td>
<td>92</td>
<td>3.75</td>
<td>1.38</td>
<td>10.18</td>
<td>0.9907</td>
</tr>
<tr>
<td>Alibakhshi et.al. (2017)</td>
<td>Correlation</td>
<td>740</td>
<td>4.066</td>
<td>1.225</td>
<td>9.81</td>
<td>0.9934</td>
</tr>
<tr>
<td>Alibakhshi et. al. (2015)</td>
<td>Semi-empirical</td>
<td>740</td>
<td>4.11</td>
<td>1.23</td>
<td>9.49</td>
<td>0.9935</td>
</tr>
<tr>
<td>Rowley et al. (2011)</td>
<td>Correlation ($\Delta H_v$+NBP)</td>
<td>1062</td>
<td>4.65</td>
<td>1.32</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Meshahvalad et. al. (2019)</td>
<td>QSPR+ANN</td>
<td>87</td>
<td>4.66</td>
<td>1.42</td>
<td>13.5</td>
<td>–</td>
</tr>
<tr>
<td>Lazsés (2010)</td>
<td>GCM+ANN+PSO</td>
<td>505</td>
<td>6.2</td>
<td>1.8</td>
<td>8.6</td>
<td>–</td>
</tr>
<tr>
<td>Keshavarz and Ghanbarzadeh (2011)</td>
<td>Correlation</td>
<td>173</td>
<td>6.35</td>
<td>2.21</td>
<td>12.8</td>
<td>0.9899</td>
</tr>
<tr>
<td>Catoire &amp; Naudet (2004)</td>
<td>Correlation ($\Delta H_v$+NBP)</td>
<td>600</td>
<td>6.36</td>
<td>1.84</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gharagheizi et al. (2012)</td>
<td>Correlation (NBP, $P_c$, $T_c$, $\omega$, $M_w$)</td>
<td>1471</td>
<td>–</td>
<td>1.94</td>
<td>7.5</td>
<td>0.9935</td>
</tr>
<tr>
<td>Mathieu and Alaime (2014)</td>
<td>GCM</td>
<td>488</td>
<td>8.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tetteh et al. (1999)</td>
<td>QSPR+ANN</td>
<td>400</td>
<td>9.59</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rowley et al. (2010)</td>
<td>Correlation ($\Delta H_v$+NBP)</td>
<td>1062</td>
<td>9.68</td>
<td>2.84</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Hukkerikar et al. (2012)</td>
<td>GC</td>
<td>512</td>
<td>10.66</td>
<td>3.27</td>
<td>–</td>
<td>0.89</td>
</tr>
<tr>
<td>Khaje and Modarres (2010)</td>
<td>ANFIS</td>
<td>95</td>
<td>11.5</td>
<td>31.1</td>
<td>1500</td>
<td>0.986</td>
</tr>
</tbody>
</table>

**Fig. 3:** Comparison of experimentally determined and predicted data for training, validation and test sets.
CONCLUSIONS

In the present study, we developed an ANN model to predict the FP using the number of carbon (n_C) and hydrogen (n_H) atoms, critical temperature (T_c), Normal Boiling Point (NBP), ideal gas enthalpy of formation (ΔH_f), and acentric factor (ω). We used a robust scheme to train the ANNs and also to evaluate their authentic performance. The results show that the new model, compared to other available models, produces the lowest error for FP prediction and the applied efficient validation strategies allow using the predicted results confidently for practical applications.

REFERENCES


