Backpropagation Neural Network-Based Prediction of Kovats Retention Index for Essential Oil Compounds

Aulia Al-Jihad Safhadi 1, Teuku Rizky Noviandy 2, Irvanizam Irvanizam 1, Rivansyah Suhendra 3, Taufiq Karma 4 and Rinaldi Idroes 5.*

1 Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; jiadaulia@gmail.com (A.A.J.S.); irvanizam.zamanhuri@usk.ac.id (I.I.)
2 Interdisciplinary Innovation Research Unit, Graha Primera Saintifikasi, Aceh Besar 23771, Indonesia; trizkynovianty@gmail.com (T.R.N.)
3 Department of Information Technology, Faculty of Engineering, Universitas Teuku Umar, Aceh Barat 23681, Indonesia; rivansyahsuhendra@utu.ac.id (R.S.)
4 Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar 23372, Indonesia; taufiqkarma94@gmail.com (T.K.)
5 Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; rinaldi.idroes@usk.ac.id (R.I.)

* Correspondence: rinaldi.idroes@usk.ac.id

Abstract

The identification of chemical compounds in essential oils is crucial in industries such as pharmaceuticals, perfumery, and food. Kovats Retention Index (RI) values are essential for compound identification using gas chromatography-mass spectrometry (GC-MS). Traditional RI determination methods are time-consuming, labor-intensive, and susceptible to experimental variability. Recent advancements in data science suggest that artificial intelligence (AI) can enhance RI prediction accuracy and efficiency. However, the full potential of AI, particularly artificial neural networks (ANN), in predicting RI values remains underexplored. This study develops a backpropagation neural network (BPNN) model to predict the Kovats RI values of essential oil compounds using five molecular descriptors: ATSc1, VCH-7, SP-1, Kier1, and MLogP. We trained the BPNN on a dataset of 340 essential oil compounds and optimized it through hyperparameter tuning. We show that the optimized BPNN model, with an epoch count of 100, a learning rate of 0.1, a hidden layer size of 10 neurons, and the ReLU activation function, achieves an R² value of 0.934 and a Root Mean Squared Error (RMSE) of 76.98. These results indicate a high correlation between predicted and actual RI values and a low average prediction error. Our findings demonstrate that BPNNs can significantly improve the efficiency and accuracy of compound identification, reducing reliance on traditional experimental methods.

1. Introduction

The identification and characterization of chemical compounds, particularly those found in essential oils, play a crucial role in various fields such as pharmaceuticals, perfumery, and food industries [1]. One significant aspect of characterizing these compounds is determining their Kovats Retention Index values, a measure widely utilized in gas chromatography-mass spectrometry (GC-MS) analysis for compound identification and comparison [2].
Traditionally, the determination of Kovats retention index values has been reliant on experimental procedures conducted through gas chromatography, involving the measurement of retention times for compounds relative to a series of standard reference compounds [3, 4]. However, this conventional approach can prove to be time-consuming, labor-intensive, and costly. Additionally, variability in experimental conditions can introduce uncertainties in retention index measurements, potentially impacting the reliability and reproducibility of results [5]. Consequently, researchers and analysts have sought alternative methods to expedite and streamline the determination of retention index values.

Recent advancements in analytical chemistry and data science have shed light on the intrinsic relationship between a compound’s chemical structure and its Kovats retention index, offering promising avenues for predictive modeling using artificial intelligence (AI) techniques [6]. Through comprehensive analyses of molecular descriptors, researchers have identified robust correlations between structural features and RI values across diverse chemical compounds [7].

With the advancements in computational techniques, particularly artificial neural networks (ANN), there arises an opportunity to predict Kovats retention index values accurately and efficiently. Among ANN architectures, the Backpropagation Neural Network (BPNN) stands out as a powerful tool for such predictions due to its ability to learn complex patterns and relationships within datasets [8]. BPNN iteratively adjusts the weights and biases of individual neurons through backpropagation, minimizing the error between predicted and actual retention index values during the training phase [9]. Through this iterative optimization process, BPNNs can effectively capture non-linear dependencies and interactions between molecular descriptors and RI values, thereby yielding highly accurate predictive models.

BPNN has been widely utilized across various research domains, including facial recognition [10], rainfall-runoff modelling [11], and stock price prediction [12]. Notably, studies have investigated the use of BPNN to model the relationship between compound structures and Kovats retention indices [13, 14]. These efforts have demonstrated promising results, indicating the potential of machine learning in this domain. However, further investigation and refinement are warranted to enhance prediction accuracy and robustness across diverse compound datasets.

This study aims to harness the capabilities of BPNN for predicting Kovats retention indices of essential oil compounds. By leveraging a comprehensive dataset sourced from previous research, comprising 340 essential oil compounds, the primary objectives include refining predictive models to accurately estimate Kovats retention indices. Additionally, this research endeavors to contribute insights into the efficacy of machine learning techniques for enhancing the understanding and prediction of essential oil compound properties, thereby facilitating advancements in compound analysis and application.

The article is structured as follows: Section 2 explains the dataset and methods used, providing a detailed overview of the research approach. In Section 3, the results are presented and discussed thoroughly, exploring their implications. Lastly, Section 4 wraps up the study with conclusions, summarizing key findings.

2. Materials and Methods

2.1. Dataset

The dataset used in this study comprises 340 essential oil compounds, as sourced from the Babushok et al. [15]. To predict the Kovats retention index for each compound, it is necessary to calculate molecular descriptors. Molecular descriptors are quantifiable properties of molecules that can be used to predict chemical behavior, including structural, electronic, and hydrophobic characteristics [16, 17]. The calculation of these descriptors is conducted through the Online Chemical Modelling Environment [18]. In this study, we specifically chose five molecular descriptors from the essential oils dataset: ATSc1, VCH-7, SP-1, Kier1, and MLogP. These descriptors were selected based on their efficacy in representing molecular structure and properties that influence retention indices in gas chromatography. This selection criteria was supported by Noviandj et al. [19], who employed a genetic algorithm-multiple linear regression approach to identify the five most optimal descriptors.

Table 1. Molecular descriptors used in this study.

<table>
<thead>
<tr>
<th>Molecular Descriptor</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATSc1</td>
<td>A topological descriptor indicating the presence and arrangement of certain chemical groups within the molecule.</td>
</tr>
<tr>
<td>VCH-7</td>
<td>A descriptor related to electronic properties, particularly electron distribution and charge</td>
</tr>
<tr>
<td>SP-1</td>
<td>A spatial descriptor that captures the three-dimensional arrangement of atoms.</td>
</tr>
<tr>
<td>Kier1</td>
<td>Molecular flexibility based on the kappa shape index.</td>
</tr>
<tr>
<td>MLogP</td>
<td>A descriptor of lipophilicity, estimating how well the compound mixes with fats and oils.</td>
</tr>
</tbody>
</table>
Table 2. Range of hyperparameter used in this study.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Values</th>
<th>Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoch</td>
<td>50, 100, 150, 200, 250</td>
<td>50</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.01 - 0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>Hidden layer size</td>
<td>1 - 10</td>
<td>1</td>
</tr>
<tr>
<td>Activation function</td>
<td>Linear, Sigmoid, ReLU, tanH</td>
<td>-</td>
</tr>
</tbody>
</table>

2.2. Data Preprocessing

The dataset is scaled using a standard scaler to ensure that all features have similar magnitudes. This step is crucial for improving the performance and convergence of certain machine learning algorithms [20]. After scaling, the dataset is split into two parts: 80% for the training set and 20% for the testing set [21]. The training set is used to train the model, enabling it to learn patterns and relationships within the data. The testing set is used to independently evaluate the model's performance on unseen data, ensuring it can generalize to new examples.

2.3. Backpropagation Neural Network (BPNN)

The Kovats indices prediction model is built on a BPNN, which undergoes a training process involving two key phases: forward pass and backward pass. During the forward pass, input data is processed through the network’s layers, with weights and biases adjusted accordingly. This process follows the architecture of the BPNN employed in this study, comprising an input layer, hidden layers, and an output layer (Figure 1). Specifically, the input layer features five neurons corresponding to selected descriptors: ATSc1, VCH-7, SP-1, Kier1, and MLogP. The optimal neuron count for the hidden layer is determined through experimentation. The output layer, consisting of one neuron, focuses on predicting essential oil compound properties. The forward pass generates an error value essential for the subsequent backward pass. In the backward pass, weights and biases are updated based on this error and the derivative of the activation function, employing Stochastic Gradient Descent (SGD). SGD facilitates weight updates based on error from a single data point per epoch.

2.4. Hyperparameter Tuning

Hyperparameter tuning is a critical step in optimizing the performance of the BPNN model. This process involves adjusting various parameters that control the learning process and structure of the network [22]. In this study, we explored a range of values for several key hyperparameters to identify the optimal configuration for predicting Kovats retention indices of essential oil compounds. We employed a grid search approach to systematically evaluate the combination of hyperparameters. This method involves exhaustively searching through a manually specified subset of the hyperparameter space [23]. The hyperparameters tuned and their respective ranges are summarized in Table 2.

2.5. Evaluation of BPNN Model

The final phase involves evaluating the predicted Kovats indices of the essential oil compounds using the ANN backpropagation model. The model’s prediction accuracy
In this plot, the $y$-e at zero). This pattern indicates that the
ons, and the
sociated with chemical compound data.
eterm, leading to highly accurate predictions.
linear dependencies between molecular structure and
hyperparameters, efficiently maps the complex and non
relationships between the molecular descriptors and the
model to effectively learn and generalize the intricate
ReLU activation function.
rate of 0.1, a hidden layer size of 10 neur
configuration includes an epoch count of 100, a learning
hyperparameters for our BPNN model. The optimal
presented in Table 3, which identifies the best set of
The results of the hyperparameter tuning process are
observed values.
value of the data point
value of the observed data point
is assessed using the root mean square error (RMSE) and
the coefficient of determination ($R^2$) [24].
The root mean square error (RMSE) represents the
square root of the average of the squared differences
between predicted and actual values. A lower RMSE
indicates better agreement between predicted and
actual values, signifying higher model accuracy. The $R^2$
assesses the proportion of the variance in the dependent
variable that is predictable from the independent
variable(s). It ranges from 0 to 1, where 1 indicates a
perfect fit. Mathematically, these metrics are calculated
as shows in Equation 1 and 2, respectively.
\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (1)
\]
\[
R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i^2 - \bar{y}^2)}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)
\]
Where $n$ is the number of observations, $y_i$ is the actual
value of the observed data point $i$, $\hat{y}_i$ is the predicted
value of the data point $i$, and $\bar{y}$ is the mean of the
observed values.

3. Results and Discussion

The results of the hyperparameter tuning process are
presented in Table 3, which identifies the best set of
hyperparameters for our BPNN model. The optimal
configuration includes an epoch count of 100, a learning
rate of 0.1, a hidden layer size of 10 neurons, and the
ReLU activation function. This configuration allows the
model to effectively learn and generalize the intricate
relationships between the molecular descriptors and the
Kovats retention indices. By integrating these
descriptors, the BPNN model, with its optimal
hyperparameters, efficiently maps the complex and non
linear dependencies between molecular structure and
retention indices, leading to highly accurate predictions.

With these optimal hyperparameters, the BPNN model
was trained and evaluated on the testing set, yielding the
results presented in Table 4. The model achieved an $R^2$
value of 0.934 and a Root Mean Squared Error (RMSE) of
76.98. The $R^2$ value of 0.934 indicates a high degree of
correlation between the predicted and actual Kovats
retention indices, signifying that the model can explain
approximately 93.4% of the variance in the data. This
high $R^2$ value demonstrates the effectiveness of the
BPNN in capturing the complex relationships between
molecular descriptors and retention indices. The RMSE of
76.98 suggests that the predictions made by the model
are reasonably close to the actual values, with an average
error of approximately 77 units. This level of accuracy is
particularly notable given the inherent variability and
complexity associated with chemical compound data.
These results underscore the capability of BPNN models
in predicting Kovats retention indices with high accuracy,
highlighting the potential of machine learning techniques
in streamlining and enhancing the analysis of essential oil
compounds.

indices, shown in Figure 2. The plot displays a strong
linear relationship between the actual and predicted
values, indicated by the clustering of points around the
red dashed line, which represents the ideal scenario
where predicted values perfectly match the actual values.
This visual confirmation of high correlation is consistent
with the model’s $R^2$ value of 0.934, underscoring its ability
to accurately predict retention indices. The slight
deviations of some points from the red line suggest
minor prediction errors, which are quantitatively
reflected in the RMSE of 76.98. Overall, the plot illustrates
the model’s robustness and reliability in capturing the
complex relationships between molecular descriptors
and retention indices, validating the effectiveness of the
chosen hyperparameters and the BPNN approach.

The residuals plot for the Kovats retention indices
predicted by the BPNN model, shown in Figure 3,
provides insight into the model’s prediction errors. The
residuals, which are the differences between the actual
and predicted values, are plotted against the predicted
values. Ideally, these residuals should be randomly
scattered around the horizontal axis (represented by a
red dashed line at zero). This pattern indicates that
the model’s errors are distributed randomly and are not
affected by the predicted values. In this plot, the
residuals appear to be scattered without a clear pattern,
suggesting that the BPNN model does not exhibit
systematic bias in its predictions. Some larger residuals
indicate outliers where the prediction errors

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoch</td>
<td>100</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Hidden layer size</td>
<td>10</td>
</tr>
<tr>
<td>Activation function</td>
<td>ReLU</td>
</tr>
</tbody>
</table>

Table 3. Optimal hyperparameters for the BPNN model.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.934</td>
</tr>
<tr>
<td>RMSE</td>
<td>76.98</td>
</tr>
</tbody>
</table>

Table 4. Performance metrics of the BPNN model.
are more significant, but overall, the majority of residuals fall within a reasonably small range. This pattern supports the effectiveness of the BPNN model and its hyperparameters in predicting Kovats retention indices, highlighting its potential utility in accurately modeling the relationship between molecular descriptors and retention indices.

The implications of these results are significant for the field of analytical chemistry and essential oil compound analysis. The high accuracy of the BPNN model in predicting Kovats retention indices demonstrates the potential of machine learning techniques to streamline and enhance traditional methods of compound identification and characterization. By reducing reliance on labor-intensive and time-consuming experimental procedures, this approach can expedite the analysis process, making it more efficient and cost-effective. However, there are limitations to this study that warrant consideration. The model's performance is highly dependent on the quality and representativeness of the dataset used for training. Any biases or gaps in the data could affect the model's generalizability to new compounds. Additionally, while the BPNN model demonstrates strong predictive capabilities, it requires significant computational resources for training, which may limit its accessibility for some applications.

Future studies should focus on expanding the dataset to include a broader range of essential oil compounds, thereby enhancing the model's robustness and applicability. Investigating the integration of additional molecular descriptors could also improve predictive accuracy by capturing more nuanced chemical properties. Moreover, exploring other machine learning techniques and hybrid models might yield further improvements in performance.

4. Conclusions

In this study, we successfully developed a BPNN model to predict the Kovats retention indices of essential oil compounds using molecular descriptors. The BPNN model demonstrated high accuracy in capturing the complex and non-linear relationships between molecular descriptors and retention indices, achieving an $R^2$ value of 0.934 and an RMSE of 76.98. These results underscore the potential of machine learning techniques, specifically BPNNs, in streamlining and enhancing the traditional methods of compound identification and characterization in analytical chemistry.


Funding: This research was funded by Universitas Syiah Kuala, Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi through the "Penelitian Profesor" scheme (grant number: 141/UN11/SPK/PNBP/2022).

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available from the corresponding author upon reasonable request.

Conflicts of Interest: All the authors declare that there are no conflicts of interest.

References