



Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 2, No. 1, 2024



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 Saintifika, Aceh Besar 23771, Indonesia; trizkynoviandy@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

Article History

Received 12 March 2024
Revised 3 May 2024
Accepted 14 May 2024
Available Online 23 May 2024

Keywords:

ANN
Machine learning
Supervised learning
Gas chromatography

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 R^2 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.



Copyright: © 2024 by the authors. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License. (<https://creativecommons.org/licenses/by-nc/4.0/>)

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 adjust 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 Noviandy 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.

Molecular Descriptor	Definition
ATSc1	A topological descriptor indicating the presence and arrangement of certain chemical groups within the molecule.
VCH-7	A descriptor related to electronic properties, particularly electron distribution and charge
SP-1	A spatial descriptor that captures the three-dimensional arrangement of atoms.
Kier1	Molecular flexibility based on the kappa shape index.
MLogP	A descriptor of lipophilicity, estimating how well the compound mixes with fats and oils.

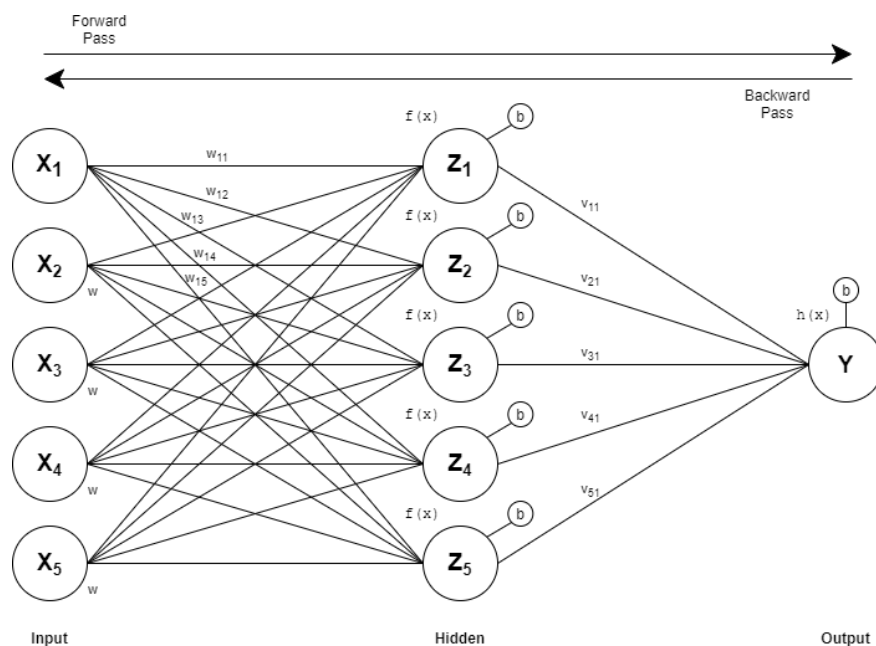


Figure 1. ANN architecture used in this study.

Table 2. Range of hyperparameter used in this study.

Hyperparameter	Values	Steps
Epoch	50, 100, 150, 200, 250	50
Learning rate	0.01 - 0,09	0.01
Hidden layer size	1 - 10	1
Activation function	Linear, Sigmoid, ReLU, tanH	-

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

Table 3. Optimal hyperparameters for the BPNN model.

Hyperparameter	Values
Epoch	100
Learning rate	0.1
Hidden layer size	10
Activation function	ReLU

Table 4. Performance metrics of the BPNN model.

Metric	Value
R ²	0.934
RMSE	76.98

is assessed using the root mean square error (RMSE) and the coefficient of determination (R²) [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 R² 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 shown in Equation 1 and 2, respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (1)$$

$$R^2 = \frac{\sum_{i=1}^n (y_i^{fit} - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

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 R² value of 0.934 and a Root Mean Squared Error (RMSE) of 76.98. The R² 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 R² 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 R² 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

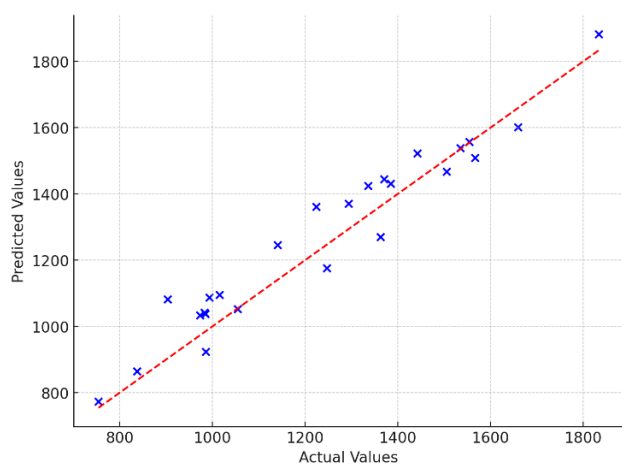


Figure 2. Actual vs. predicted plot.

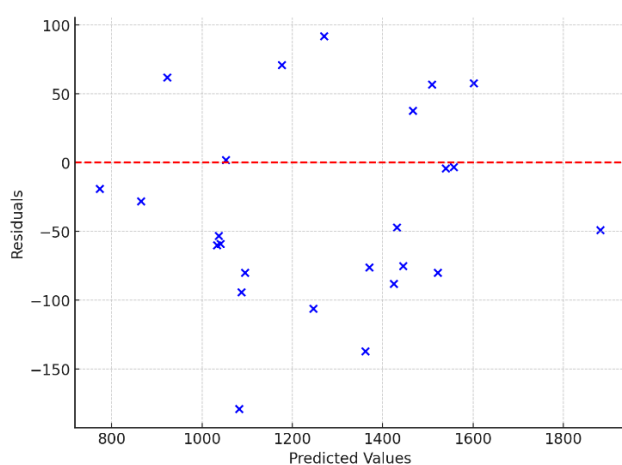


Figure 3. Residual plot.

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.

Author Contributions: Conceptualization, A.A.J.S., T.R.N., I.I. and R.I.; methodology, I.I. and R.I.; software, A.A.J.S. and R.S.; validation, I.I., T.K. and R.I.; formal analysis, A.A.J.S., T.R.N. and T.K.; investigation, A.A.J.S. and R.S.; resources, R.I.; data curation, R.I.; writing—original draft preparation, A.A.J.S., T.R.N. and R.S.; writing—review and editing, I.I., T.K. and R.I.; visualization, R.S.; supervision, I.I. and R.I.; project administration, R.I.; funding acquisition, R.I.. All authors have read and agreed to the published version of the manuscript.

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

1. Babushok, V. I. (2015). Chromatographic Retention Indices in Identification of Chemical Compounds, *TrAC - Trends in Analytical Chemistry*, Vol. 69, 98–104. doi:10.1016/j.trac.2015.04.001.

2. Zenkevich, I. G. (2010). Kovats' Retention Index System, *Encyclopedia of Chromatography*, Vol. 2, 1304–1310.
3. Goodner, K. L. (2008). Practical Retention Index Models of OV-101, DB-1, DB-5, and DB-Wax for Flavor and Fragrance Compounds, *LWT - Food Science and Technology*, Vol. 41, No. 6, 951–958. doi:10.1016/j.lwt.2007.07.007.
4. Strehmel, N., Hummel, J., Erban, A., Strassburg, K., and Kopka, J. (2008). Retention Index Thresholds for Compound Matching in GC–MS Metabolite Profiling, *Journal of Chromatography B*, Vol. 871, No. 2, 182–190. doi:10.1016/j.jchromb.2008.04.042.
5. von Mühlen, C., and Marriott, P. J. (2011). Retention Indices in Comprehensive Two-Dimensional Gas Chromatography, *Analytical and Bioanalytical Chemistry*, Vol. 401, No. 8, 2351–2360. doi:10.1007/s00216-011-5247-1.
6. Qu, C., Schneider, B. I., Kearsley, A. J., Keyrouz, W., and Allison, T. C. (2021). Predicting Kovats Retention Indices Using Graph Neural Networks, *Journal of Chromatography A*, Vol. 1646, 462100. doi:10.1016/j.chroma.2021.462100.
7. Goel, P., Bapat, S., Vyas, R., Tambe, A., and Tambe, S. S. (2015). Genetic Programming Based Quantitative Structure–Retention Relationships for the Prediction of Kovats Retention Indices, *Journal of Chromatography A*, Vol. 1420, 98–109. doi:10.1016/j.chroma.2015.09.086.
8. Lyu, Z., Yu, Y., Samali, B., Rashidi, M., Mohammadi, M., Nguyen, T. N., and Nguyen, A. (2022). Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam, *Materials*, Vol. 15, No. 4, 1477. doi:10.3390/ma15041477.
9. Nguyen, T.-A., Ly, H.-B., and Pham, B. T. (2020). Backpropagation Neural Network-Based Machine Learning Model for Prediction of Soil Friction Angle, *Mathematical Problems in Engineering*, Vol. 2020, 1–11. doi:10.1155/2020/8845768.
10. Ayyappa, Y., Neelakanteswara, P., Bekkanti, A., Tondeti, Y., and Basha, C. Z. (2021). Automatic Face Mask Recognition System With FCM AND BPNN, *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, 1134–1137. doi:10.1109/ICCMC51019.2021.9418243.
11. Samantaray, S., and Sahoo, A. (2020). Appraisal of Runoff Through BPNN, RNN, and RBFN in Tentulikhunti Watershed: A Case Study, 258–267. doi:10.1007/978-981-13-9920-6_26.
12. Mailinda, I., Ruldeviyani, Y., Tanjung, F., Mikoriza T, R., Putra, R., and Fauziah A, T. (2021). Stock Price Prediction During the Pandemic Period with the SVM, BPNN, and LSTM Algorithm, *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 189–194. doi:10.1109/ISRITI54043.2021.9702865.
13. Sadeghi, M., Mohammadinasab, E., and Momeni Isfahani, T. (2023). QSPR models for predicting the Kovats retention indices of synthetic ester derivatives based on pyrethrin essential oil, *Journal of Essential Oil Research*, Vol. 35, No. 6, 542–562. doi:10.1080/10412905.2023.2265376.
14. Maulana, A., Noviandy, T. R., Idroes, R., Sasmita, N. R., Suhendra, R., and Irvanizam, I. (2020). Prediction of Kovats Retention Indices for Fragrance and Flavor using Artificial Neural Network, *2020 International Conference on Electrical Engineering and Informatics (ICELTICS)*, IEEE, 1–5. doi:10.1109/ICELTICS50595.2020.9315391.
15. Babushok, V. I., Linstrom, P. J., and Zenkevich, I. G. (2011). Retention Indices for Frequently Reported Compounds of Plant Essential Oils, *Journal of Physical and Chemical Reference Data*, Vol. 40, No. 4, 043101. doi:10.1063/1.3653552.
16. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.
17. Fernández-Torras, A., Comajuncosa-Creus, A., Duran-Frigola, M., and Aloy, P. (2022). Connecting chemistry and biology through molecular descriptors, *Current Opinion in Chemical Biology*, Vol. 66, 102090. doi:10.1016/j.cbpa.2021.09.001.
18. Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V. V., Tanchuk, V. Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I. I., Palyulin, V. A., Radchenko, E. V., Welsh, W. J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-De-Sousa, J., Zhang, Q. Y., Bender, A., Nigsch, F., Patiny, L., Williams, A., Tkachenko, V., and Tetko, I. V. (2011). Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information, *Journal of Computer-Aided Molecular Design*, Vol. 25, No. 6, 533–554. doi:10.1007/s10822-011-9440-2.
19. Noviandy, T. R., Maulana, A., Sasmita, N. R., Suhendra, R., Irvanizam, I., Muslem, M., Idroes, G. M., Yusuf, M., Sofyan, H., Abidin, T. F., and Idroes, R. (2022). The Prediction of Kovats Retention Indices of Essential Oils at Gas Chromatography Using Genetic Algorithm-Multiple Linear Regression and Support Vector Regression, *Journal of Engineering Science and Technology*, Vol. 17, No. 1, 306–326.
20. Maulana, A., Faisal, F. R., Noviandy, T. R., Rizkia, T., Idroes, G. M., Tallei, T. E., El-Shazly, M., and Idroes, R. (2023). Machine Learning Approach for Diabetes Detection Using Fine-Tuned XGBoost Algorithm, *Infolitika Journal of Data Science*, Vol. 1, No. 1, 1–7. doi:10.60084/ijds.v1i1.72.
21. Idroes, R., Noviandy, T. R., Maulana, A., Suhendra, R., and Sasmita, N. R. (2023). ANFIS-Based QSRR Modelling for Kovats Retention Index Prediction in Gas Chromatography, *Infolitika Journal of Data Science*, Vol. 1, No. 1, 1–14. doi:10.60084/ijds.v1i1.73.
22. Noviandy, T. R., Maulana, A., Idroes, G. M., Irvanizam, I., Subianto, M., and Idroes, R. (2023). QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction, *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, IEEE, 220–225. doi:10.1109/COSITE60233.2023.10250039.
23. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). Enhancing Loan Approval Decision-Making: An Interpretable Machine Learning Approach Using LightGBM for Digital Economy Development, *Malaysian Journal of Computing (MJOC)*, Vol. 9, No. 1, 1734–1745. doi:10.24191/mjoc.v9i1.25691.
24. Roy, K., and Kar, S. (2015). How to Judge Predictive Quality of Classification and Regression Based QSAR Models?, Z. Ul-Haq; J. D. B. T.-F. in C. C. Madura (Eds.), , Bentham Science Publishers, 71–120. doi:10.1016/B978-1-60805-979-9.50003-2.