Original Article

Check for updates

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

Infolitika Journal of Data Science

Vol. 2, No. 2, 2024

Artificial Neural Network–Particle Swarm Optimization Approach for Predictive Modeling of Kovats Retention Index in Essential Oils

Kurniadinur Kurniadinur [1], Teuku Rizky Noviandy [2], Ghazi Mauer Idroes [3,4], Noor Atinah Ahmad [5], Irvanizam Irvanizam [1], Muhammad Subianto [6] and Rinaldi Idroes [7,*]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; dinur@mhs.unsyiah.ac.id (K.K.); irvanizam.zamanhuri@usk.ac.id (I.I.)
[2] Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia; rizky_si@abulyatama.ac.id (T.R.N.)
[3] Graduate School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; idroesghazi_k3@abulyatama.ac.id (G.M.I.)
[4] Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar 23372, Indonesia
[5] School of Mathematical Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia; nooratinah@usm.my (N.A.A.)
[6] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; subianto@usk.ac.id (M.S.)
[7] 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 30 August 2024
Revised 28 October 2024
Accepted 5 November 2024
Available Online 14 November 2024

Keywords:
ANN-PSO
Gas chromatography
Molecular descriptors
Optimization technique

Abstract

The Kovats retention index is a critical parameter in gas chromatography used for the identification of volatile compounds in essential oils. Traditional methods for determining the Kovats retention index are often labor-intensive, time-consuming, and prone to inaccuracies due to variations in experimental conditions. This study presents a novel approach combining Artificial Neural Networks (ANN) with Particle Swarm Optimization (PSO) to predict the Kovats retention index of essential oil compounds more accurately and efficiently. The ANN-PSO hybrid model leverages the strengths of both techniques: the ANN's capacity to model complex nonlinear relationships and PSO's capability to optimize hyperparameters by finding the global optimum. The model was trained using a dataset of 340 essential oil compounds with molecular descriptors, with the performance evaluated based on Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Results indicate that a simpler ANN configuration with one hidden neuron achieved the lowest RMSE (80.16) and MAPE (5.65%), suggesting that the relationship between the molecular descriptors and the Kovats retention index is not overly complex. This study demonstrates that the ANN-PSO model can serve as an effective tool for predictive modeling of the Kovats retention index, reducing the need for experimental procedures and improving analytical efficiency in essential oil research.

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 Kovats retention index, introduced by Ervin Kovats in 1958, serves as a standardized metric in gas chromatography for identifying and comparing volatile compounds, playing a critical role in essential oil analysis [1, 2]. By consistently calculating retention times across varying chromatographic setups, the Kovats retention

index is widely used to support the qualitative analysis of complex mixtures, such as those found in essential oils [3]. Accurate identification of individual components within essential oils is essential for understanding their aromatic and therapeutic properties [4], and the Kovats retention index facilitates this by offering a reliable reference point [5, 6].

However, traditional methods for determining the Kovats retention index rely on experimental measurements, which are often labor-intensive, time-consuming, and susceptible to variations in experimental conditions [7, 8]. These limitations create challenges in achieving precise and reproducible results, especially when analyzing large datasets or conducting high-throughput studies. As the demand for faster and more accurate analysis grows, these traditional methods increasingly fall short, highlighting the need for efficient alternatives.

In recent years, machine learning has gained significant attention as a powerful tool for solving complex problems across various scientific disciplines [9–12]. Its ability to learn from large datasets and model intricate patterns has led to its widespread adoption in fields such as drug discovery [13–15], finance [16–18], and materials science [19–21]. Among various machine learning methods, Artificial Neural Networks (ANNs) show particular potential due to their ability to model nonlinear relationships [22, 23].

Despite the effectiveness of ANNs in modeling complex nonlinear relationships, they often face challenges such as getting stuck in local minima during the training process and requiring careful tuning of numerous hyperparameters to achieve optimal performance [24, 25]. These limitations can result in suboptimal predictions and reduced generalizability of the model. To address these challenges, hybrid models that integrate ANNs with optimization techniques, such as Particle Swarm Optimization (PSO), offer a compelling solution [26]. PSO, an optimization algorithm inspired by the social behavior of birds, enhances ANN training by efficiently exploring the parameter space to find global optima, leading to improved model accuracy and reliability [27]. While PSO has been applied to optimize machine learning models in various domains, its potential for improving Kovats retention index prediction in essential oils remains underexplored.

This study aims to bridge this gap by developing a hybrid ANN-PSO model to predict the Kovats retention index in essential oils, providing a computational approach that combines the strengths of ANN's nonlinear modeling with PSO's optimization efficiency. This hybrid model is expected to enhance the accuracy and reliability of Kovats retention index predictions, offering an effective alternative to traditional experimental methods. Doing so contributes to essential oil research and proposes a methodology that could streamline workflows in this field.

The structure of this paper is as follows: Section 2 details the materials and methods used in the study, including data acquisition, model architecture, and evaluation criteria. Section 3 presents the results of the study, including a comparison between traditional and hybrid methods, along with a discussion of implications, limitations, and suggestions for future research. Finally, Section 4 concludes the paper, summarizing the findings and outlining potential directions for further investigation in this area.

2. Materials and Methods

2.1. Dataset Collection

This study uses a dataset of 340 essential oil compounds compiled by Babushok et al. [28]. The dataset encompasses volatile compounds commonly present in essential oils, each with distinct structural and chemical properties. These compounds were analyzed using a dimethyl silicone stationary phase, a widely used nonpolar phase in gas chromatography known for delivering consistent retention times across diverse compounds.

We applied a stratified split to partition the data into 80% for training and 20% for testing to facilitate effective model training and evaluation [29]. This stratified approach maintains the original distribution of compound classes or other relevant properties within each subset, ensuring that the test set is representative of the entire dataset. This careful balance prevents skewed model performance and enhances generalizability. The distribution of compounds across the training and testing sets is illustrated in Figure 1, demonstrating that the sampling approach preserves the diversity of compound characteristics within both subsets.

2.2. Molecular Descriptors Calculation

Molecular descriptors are numerical values that represent different properties of a molecule, allowing us to summarize its structure and chemical characteristics in a way that can be used for analysis and prediction. These descriptors capture essential features such as the molecule's size, shape, electronic properties, and the types of atoms or bonds it contains [30]. By translating these complex molecular attributes into simple, quantitative values, molecular descriptors make it easier

Figure 1. Distribution of essential oil compounds: (a) Training set (80% of data) and (b) Testing set (20% of data).

Table 1. Molecular descriptors used in this study and their descriptions.

Descriptor	Description
ATSc1	Centered Broto-Moreau autocorrelation of lag 1 weighted by atomic masses
VCH-7	Valence connectivity index chi-7
SP-1	Simple path count descriptor representing single bonds in a molecule
Kier1	First-order Kier & Hall molecular connectivity index
MlogP	Moriguchi octanol-water partition coefficient (log P)

to analyze and compare different molecules [31]. This is useful in computational modeling, where descriptors allow researchers to predict how a molecule might behave under certain conditions, like how it might interact with other chemicals or how it might perform in a chemical process.

In our study, we selected five key molecular descriptors from a previous study by Idroes et al. [32]. These descriptors were calculated using the Chemical Modelling Environment (OCHEM) platform [33], an online tool that automates calculating various molecular descriptors. From a large set of potential descriptors, the top five were chosen based on their importance to the model, identified using a genetic algorithm approach that helps pick the most relevant factors. The top five descriptors are listed in Table 1.

2.3. Artificial Neural Network-PSO (ANN-PSO)

The ANN-PSO is a hybrid approach that combines an ANN with PSO to improve prediction accuracy. An ANN is a computational model inspired by the neural networks in the human brain. It consists of interconnected nodes or neurons organized into layers: an input layer, hidden layers, and an output layer [34, 35]. This study uses ANN-PSO to model the relationship between molecular descriptors and the Kovats retention index. The ANN component captures complex, nonlinear patterns in the data, while PSO optimizes the ANN's parameters to enhance performance.

The ANN in this study is structured with five input nodes, each representing a molecular descriptor, a single output node for the predicted Kovats retention index, and various hidden layers ranging from 1 to 9 to test different configurations. The forward propagation in the ANN is described by Equation 1:

$$y = f(Wx + b) \tag{1}$$

where y is the output, W represents the weights, x denotes the input features, b is the bias, and f is the activation function.

In PSO, each particle's velocity and position are updated in each iteration based on its own previous experiences (personal best) and the best position found by the entire swarm (global best). The velocity and position updates are governed by Equations 2 and 3, respectively:

$$v_i(t+1) = w * v_i(t) + c1 * r1 * \left(p_i - x_i(t)\right) + c2 * r2 * \left(g - x_i(t)\right) \tag{2}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{3}$$

where $v_i(t)$ is the current velocity of particle i at time t, $x_i(t)$ denotes the current position of particle i , p_i is the particle's personal best position, g is the global best position of the swarm, w is the inertia weight, managing the balance between exploration and exploitation, $c1$ and $c2$ are the cognitive and social coefficients, respectively, $r1$ and $r2$ are random values uniformly distributed between 0 and 1, adding randomness to particle

movements. These equations enable particles to explore the solution space dynamically, iteratively moving toward optimal weight and bias configurations for the ANN.

This study configures PSO with specific parameters that effectively guide the particles toward optimal solutions. Cognitive coefficient $c1$ and social coefficient $c2$ are both set to 2. The cognitive coefficient encourages particles to move toward their own best-known positions (personal learning), while the social coefficient directs them toward the global best position found by the swarm (social learning). The inertia weight w linearly decreases from 0.9 to 0.4 during training, balancing exploration with exploitation to avoid premature convergence on suboptimal solutions.

The swarm size is set to seven particles in this study, calculated based on the structure of the ANN. Given that the ANN has five input nodes and a single output node, the total swarm size is determined by the Equation 4:

$$Swarm\ Size = (num_inputs + 1) \times num_neurons + (num_neurons + 1) \times 1 \quad (4)$$

Each particle represents a possible configuration of weights and biases, and the swarm explores these configurations over 1000 maximum iterations.

During each iteration, the particles adjust their configurations to minimize the root mean square error (RMSE) of predictions on the training set. This iterative process concludes when the global best fitness reaches a pre-defined threshold or the maximum number of iterations is reached, ensuring that the ANN-PSO approach converges on an accurate and generalizable solution for predicting the Kovats retention index.

2.4. Performance Evaluation

To evaluate the proposed ANN-PSO model's performance in predicting the Kovats retention index, four widely used error metrics are employed: coefficient of determination (R^2), RMSE, mean absolute error (MAE), and mean absolute percentage error (MAPE). These metrics comprehensively assess the model's predictive accuracy [36].

The coefficient of determination (R^2) is a statistical measure that explains the proportion of the variance in the actual Kovats retention index values that is predictable from the model's predictions. R^2 values range from 0 to 1, where a value closer to 1 indicates that the model can explain a larger proportion of variance, demonstrating high predictive accuracy. A high R^2 indicates that the model captures most of the variability in the data, signaling good overall fit.

Root Mean Square Error (RMSE) is another important metric, measuring the average magnitude of the errors between predicted and actual values. RMSE is the square root of the mean squared differences between predicted and actual values, which means it penalizes larger errors more heavily. This sensitivity to large errors makes RMSE particularly useful for evaluating models where minimizing substantial deviations is important. A lower RMSE suggests higher accuracy of the model.

Next, Mean Absolute Error (MAE) provides an intuitive measure of prediction error by calculating the average absolute differences between predicted and actual values, without emphasizing larger deviations as in RMSE. MAE offers insight into the model's general performance by highlighting the average error magnitude across all predictions, with lower values indicating better performance.

Lastly, Mean Absolute Percentage Error (MAPE) measures the average error in terms of percentage, making it especially valuable for understanding the model's relative accuracy. MAPE is calculated by taking the average of the absolute percentage errors between predicted and actual values, expressing accuracy in percentage terms. A lower MAPE value suggests that the model provides accurate predictions relative to the true values, which is useful for interpreting accuracy on a relative scale.

The formula for R^2, RMSE, MAE, and MAPE are presented in Equations 5, 6, 7, and 8, respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}|\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}| \quad (7)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (8)$$

where y_i represents the actual Kovats retention index, $\hat{y}_i$ denotes the predicted Kovats retention index, and n is the total number of observations.

In addition to evaluating the proposed ANN-PSO model using these metrics, its performance is also compared with several other machine learning models, including ANN Backpropagation (ANN-BP), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine

Table 2. Performance of ANN-PSO model with different hidden neuron counts.

Hidden Neurons	Training Set				Testing Set			
	R^2	RMSE	MAE	MAPE (%)	R^2	RMSE	MAE	MAPE (%)
1	0.888	105.37	82.61	7.27	0.912	107.01	87.46	7.63%
2	0.913	92.83	72.90	6.15	0.927	97.19	80.57	6.88
3	0.878	109.91	86.28	7.17	0.916	104.19	80.35	6.65
4	0.893	103.06	81.01	6.82	0.925	98.98	79.26	6.71
5	0.858	118.65	91.99	7.68	0.911	107.57	82.18	6.74
6	0.878	110.22	85.86	7.23	0.920	101.93	78.76	6.53
7	0.868	114.46	89.03	7.35	0.913	106.41	80.37	6.46
8	0.857	119.16	93.85	7.72	0.908	109.54	82.49	6.67
9	0.872	112.71	88.06	7.27	0.915	105.06	79.09	6.42

Table 3. Performance comparison of ANN-PSO, ANN-BP, LR, KNN, and SVR-RBF models in predicting the Kovats retention index.

Model	Training Set				Testing Set			
	R^2	RMSE	MAE	MAPE (%)	R^2	RMSE	MAE	MAPE (%)
ANN-PSO	0.913	92.83	72.90	6.15	0.927	97.19	80.57	6.88
ANN-BP	0.888	111.17	86.51	7.38	0.919	105.26	79.84	6.32
LR	0.450	245.97	164.74	14.51	0.431	278.60	226.96	19.01
KNN	0.929	88.46	67.61	5.77	0.887	124.03	93.36	7.16
SVR-RBF	0.888	111.25	85.76	7.26	0.917	106.56	79.98	6.09

with Radial Basis Function kernel (SVM-RBF). This comparison highlights the relative effectiveness of the ANN-PSO model in accurately predicting the Kovats retention index across different algorithms.

3. Results and Discussion

The application of the hybrid ANN-PSO model for predicting the Kovats retention index in essential oils demonstrated that varying the number of neurons in the hidden layer significantly impacted model accuracy and generalizability. Table 2 summarizes the performance metrics for models with hidden neuron counts ranging from 1 to 9, evaluated across both the training and testing datasets.

The results reveal that the choice of hidden neuron count substantially affects the ANN-PSO model's predictive capability for the Kovats retention index. The model with two hidden neurons emerged as the best performer overall, achieving the highest R^2 values (0.913 for training and 0.927 for testing) and the lowest error metrics across RMSE, MAE, and MAPE. This configuration shows that a simpler model architecture can effectively capture key data patterns, avoiding the common pitfalls of overfitting and underfitting.

As the number of hidden neurons increased beyond two, the model's performance showed variability, with a general trend toward reduced accuracy. Models using five or more hidden neurons displayed increased RMSE and MAPE values, suggesting a tendency toward overfitting. In these cases, the models became overly complex and demonstrated reduced generalizability to unseen data. The model maintained a reasonable balance between complexity and accuracy for configurations with three and four hidden neurons, achieving R^2 values close to those observed in the two-neuron configuration, although with slightly elevated error metrics.

Conversely, configurations with three and four hidden neurons presented a more moderate performance, maintaining reasonable R^2 values close to those obtained with two neurons but with slight increases in error metrics. This indicates that these models can capture essential data relationships without the excessive complexity that can impact generalizability. However, they did not match the performance of the simpler two-neuron model, highlighting the importance of limiting the model to the optimal complexity required for the task.

The performance of the ANN-PSO model was compared to other machine learning models, including ANN-BP, LR, KNN, and SVR-RBF. This comparative analysis, summarized in Table 3, highlights the advantages and limitations of each model in predicting the Kovats retention index.

The ANN-PSO model outperformed most of the other models across the majority of evaluation metrics, achieving high R^2 values and low error metrics on the testing set. This indicates that the ANN-PSO model is accurate and generalizes well to new data, showing superior prediction quality compared to the traditional ANN-BP and SVR-RBF models. The optimization introduced by the PSO component significantly contributed to this performance by improving weight and

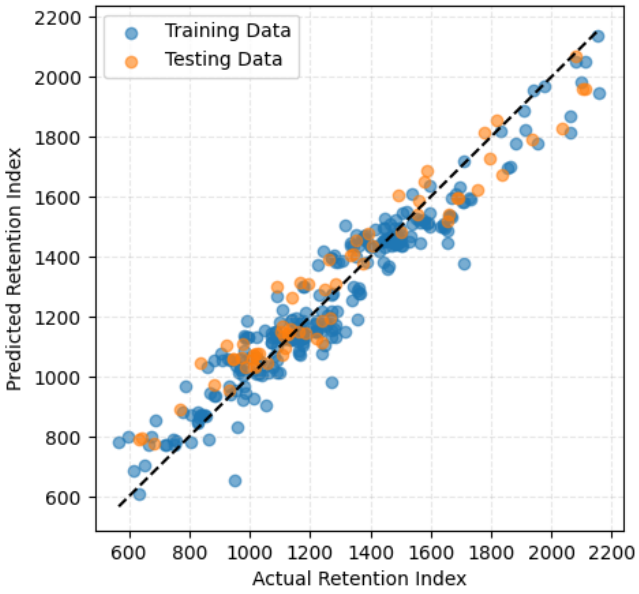

Figure 2. Actual vs. predicted plot of Kovats retention index for ANN-PSO model with two hidden neurons.

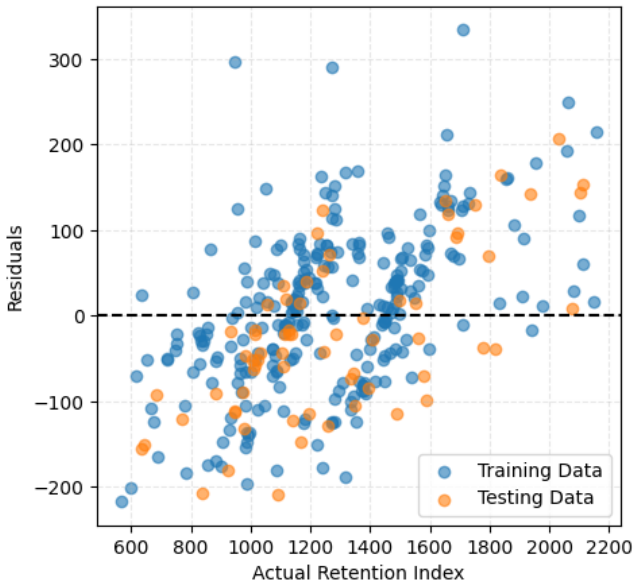

Figure 3. Residual plot Kovats retention index for ANN-PSO model with two hidden neurons.

bias configuration beyond what standard backpropagation methods could achieve.

In comparison, the ANN-BP model performed adequately but with higher RMSE and MAE values, reflecting that while backpropagation is effective, it does not achieve the same degree of optimization and error reduction as the PSO-enhanced version. The KNN model showed strong performance on the training set, with an R^2 of 0.929, the highest among all models tested. However, its performance dropped considerably on the testing set, with an R^2 of 0.887 and increased RMSE and MAPE values, suggesting potential overfitting. The SVR-RBF model demonstrated performance similar to the ANN-BP, with a slight improvement in MAE and MAPE on the testing set but still unable to match the accuracy and generalizability of the ANN-PSO. LR produced the weakest results, with low R^2 values (0.450 for training and 0.431 for testing) and the highest RMSE and MAPE values across both datasets. This outcome reflects the limitations of linear models in capturing the nonlinear relationships present in the dataset, reinforcing the necessity of nonlinear methods for Kovats retention index prediction.

Because the ANN-PSO model with two hidden neurons demonstrated strong predictive accuracy, we conducted further analysis to validate its performance. Figure 2 presents the actual versus predicted Kovats retention index values for both the training and testing datasets. The consistent proximity of both training and testing data points to this line highlights the model's robust accuracy and generalizability across datasets, supporting the efficacy of the ANN-PSO approach in accurately predicting retention indices in essential oil compounds.

Following the model's predictions analysis, we examined the residuals, which represent the differences between the actual and predicted Kovats retention index values. The residual plot, shown in Figure 3, visualizes these residuals across both the training and testing datasets. Ideally, residuals should be randomly distributed around zero, indicating that the model does not systematically overestimate or underestimate the Kovats retention index across the range of actual values. While the residuals for both the training and testing datasets are generally centered around the zero line, there is noticeable scattering in this plot. This dispersion indicates that the ANN-PSO model may struggle slightly with extreme values, leading to occasional under- or overestimations. Although most residuals fall within an acceptable range, several large residuals suggest areas where the model's consistency could be improved, particularly in accurately capturing variability at the upper and lower ends of the retention index range.

This study introduces a hybrid ANN-PSO model that effectively predicts the Kovats retention index for essential oils, underscoring the potential of integrating machine learning with optimization techniques for improved accuracy in chemical analysis. Utilizing molecular descriptors and advanced machine learning methods offers a non-experimental alternative that saves both time and resources, facilitating rapid analysis of complex mixtures found in essential oils. The high predictive accuracy of the ANN-PSO model suggests that similar hybrid techniques could be beneficial for broader applications in analytical chemistry, particularly in fields requiring high-throughput analysis and real-time

predictions. Moreover, the methodology demonstrated here could support advancements in essential oil quality control, formulation, and therapeutic applications, where precise component identification is essential.

While the ANN-PSO model showed promising accuracy and generalizability, several limitations warrant consideration. First, the dataset used was limited to 340 compounds, which may restrict the model's performance when applied to an expanded set of essential oils or volatile compounds with diverse structures. Additionally, while PSO contributed to effective parameter optimization, it requires computational resources that may limit its applicability in low-resource settings. Another limitation is the model's performance on extreme values, as observed in the residual analysis, which suggests that the model may underperform for outliers in the Kovats retention index range. Lastly, PSO optimization can be sensitive to initial parameter choices and may require fine-tuning specific to the dataset to avoid premature convergence, which could limit its scalability and transferability.

Future research should consider expanding the dataset to include a wider range of compounds and molecular structures, which would enhance the model's robustness and generalizability. It is essential to adopt effective data split techniques to ensure proper model evaluation. Utilizing data splits such as an 70/30 training-to-test ratio, or employing k-fold cross-validation, can provide a comprehensive assessment of model performance. Incorporating additional molecular descriptors or using descriptor selection techniques like recursive feature elimination could further refine predictive performance. Investigating ensemble learning techniques, where multiple models contribute to final predictions, may also improve robustness and accuracy, particularly in cases with outliers.

4. Conclusions

this study presents a hybrid ANN-PSO model for predicting the Kovats retention index in essential oils, leveraging the strengths of ANN's nonlinear modeling with PSO's optimization capabilities. The model achieves high predictive accuracy and generalizability, marking a significant advancement over traditional experimental methods. While limitations regarding dataset diversity and parameter sensitivity exist, the ANN-PSO model demonstrates the potential for improving essential oil analysis, offering a fast and reliable alternative. This approach has promising implications for furthering essential oil research and chemical analysis methodologies, laying a foundation for future enhancements in computational prediction models.

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 no conflicts of interest.

References

1. Zenkevich, I. G. (2010). Kovats' Retention Index System, *Encyclopedia of Chromatography*, Vol. 2, 1304–1310.
2. Yabalak, E., Ibrahim, F., and Erdoğan Eliuz, E. A. (2023). Natural Sanitizer Potential of Cuminum cyminum and Applicable Approach for Calculation of Kováts Retention Index of Its Compounds, *International Journal of Environmental Health Research*, Vol. 33, No. 2, 158–169. doi:10.1080/09603123.2021.2011159.
3. 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.
4. Aljaafari, M. N., AlAli, A. O., Baqais, L., Alqubaisy, M., AlAli, M., Molouki, A., Ong-Abdullah, J., Abushelaibi, A., Lai, K.-S., and Lim, S.-H. E. (2021). An Overview of the Potential Therapeutic Applications of Essential Oils, *Molecules*, Vol. 26, No. 3, 628. doi:10.3390/molecules26030628.
5. Sadgrove, N., Padilla-González, G., and Phumthum, M. (2022). Fundamental Chemistry of Essential Oils and Volatile Organic Compounds, Methods of Analysis and Authentication, *Plants*, Vol. 11, No. 6, 789. doi:10.3390/plants11060789.
6. Rojas, C., Cedillo, J. F., Sarmiento, N., Pis Diez, R., and Duchowicz, P. R. (2025). Quantitative Structure-Property Relationship for Gas-Chromatographic Retention Indices of Volatile Organic Compounds in Colored-Quinoa Seeds, *Journal of Food Composition and Analysis*, Vol. 137, 106843. doi:10.1016/j.jfca.2024.106843.
7. Metrani, R., Jayaprakasha, G. K., and Patil, B. S. (2022). Optimization of Experimental Parameters and Chemometrics Approach to Identify Potential Volatile Markers in Seven Cucumis melo Varieties Using HS–SPME–GC–MS, *Food Analytical Methods*, Vol. 15, No. 3, 607–624. doi:10.1007/s12161-021-02119-9.
8. 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.
9. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). An Interpretable Machine Learning Strategy for Antimalarial Drug Discovery with LightGBM and SHAP, *Journal of Future Artificial Intelligence and Technologies*, Vol. 1, No. 2, 84–95. doi:10.62411/faith.2024-16.
10. Qu, C., Schneider, B. I., Kearsley, A. J., Keyrouz, W., and Allison, T. C. (2021). Predicting Kováts Retention Indices Using Graph Neural Networks, *Journal of Chromatography A*, Vol. 1646, 462100. doi:10.1016/j.chroma.2021.462100.
11. Suhendra, R., Husdayanti, N., Suryadi, S., Juliwardi, I., Sanusi, S., Ridho, A., Ardiansyah, M., Murhaban, M., and Ikhsan, I. (2023). Cardiovascular Disease Prediction Using Gradient Boosting

Classifier, *Infolitika Journal of Data Science*, Vol. 1, No. 2, 56–62. doi:10.60084/ijds.v1i2.131.

12. Supriatna, D. J. I., Saputra, H., and Hasan, K. (2023). Enhancing the Red Wine Quality Classification Using Ensemble Voting Classifiers, *Infolitika Journal of Data Science*, Vol. 1, No. 2, 42–47. doi:10.60084/ijds.v1i2.95.

13. Noviandy, T. R., Nisa, K., Idroes, G. M., Hardi, I., and Sasmita, N. R. (2024). Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM, *Journal of Computing Theories and Applications*, Vol. 2, No. 2, 138–147. doi:10.62411/jcta.10129.

14. Staszak, M., Staszak, K., Wieszczycka, K., Bajek, A., Roszkowski, K., and Tylkowski, B. (2022). Machine Learning in Drug Design: Use of Artificial Intelligence to Explore the Chemical Structure–Biological Activity Relationship, *WIREs Computational Molecular Science*, Vol. 12, No. 2. doi:10.1002/wcms.1568.

15. 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.

16. 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.

17. Petropoulos, A., Siakoulis, V., Stavroulakis, E., and Vlachogiannakis, N. E. (2020). Predicting Bank Insolvencies Using Machine Learning Techniques, *International Journal of Forecasting*, Vol. 36, No. 3, 1092–1113. doi:10.1016/j.ijforecast.2019.11.005.

18. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Adam, M., Rusyana, A., and Sofyan, H. (2023). Deep Learning-Based Bitcoin Price Forecasting Using Neural Prophet, *Ekonomikalia Journal of Economics*, Vol. 1, No. 1, 19–25. doi:10.60084/eje.v1i1.51.

19. Wang, Z. H., Liu, Y. F., Wang, T., Wang, J. G., Liu, Y. M., and Huang, Q. X. (2024). Intelligent Prediction Model of Mechanical Properties of Ultrathin Niobium Strips Based on XGBoost Ensemble Learning Algorithm, *Computational Materials Science*, Vol. 231, 112579. doi:10.1016/j.commatsci.2023.112579.

20. Hestroffer, J. M., Charpagne, M.-A., Latypov, M. I., and Beyerlein, I. J. (2023). Graph Neural Networks for Efficient Learning of Mechanical Properties of Polycrystals, *Computational Materials Science*, Vol. 217, 111894. doi:10.1016/j.commatsci.2022.111894.

21. Westermayr, J., Gastegger, M., Schütt, K. T., and Maurer, R. J. (2021). Perspective on Integrating Machine Learning into Computational Chemistry and Materials Science, *The Journal of Chemical Physics*, Vol. 154, No. 23. doi:10.1063/5.0047760.

22. Almeida, J. S. (2002). Predictive Non-linear Modeling of Complex Data by Artificial Neural Networks, *Current Opinion in Biotechnology*, Vol. 13, No. 1, 72–76. doi:10.1016/S0958-1669(02)00288-4.

23. Wu, Y., and Feng, J. (2018). Development and Application of Artificial Neural Network, *Wireless Personal Communications*, Vol. 102, No. 2, 1645–1656. doi:10.1007/s11277-017-5224-x.

24. Safhadi, A. A.-J., Noviandy, T. R., Irvanizam, I., Suhendra, R., Karma, T., and Idroes, R. (2024). Backpropagation Neural Network-Based Prediction of Kovats Retention Index for Essential Oil Compounds, *Infolitika Journal of Data Science*, Vol. 2, No. 1, 28–33. doi:10.60084/ijds.v2i1.197.

25. Lorenzo, P. R., Nalepa, J., Kawulok, M., Ramos, L. S., and Pastor, J. R. (2017). Particle Swarm Optimization for Hyper-Parameter Selection in Deep Neural Networks, *Proceedings of the Genetic and Evolutionary Computation Conference*, ACM, New York, NY, USA, 481–488. doi:10.1145/3071178.3071208.

26. Shariati, M., Mafipour, M. S., Mehrabi, P., Bahadori, A., Zandi, Y., Salih, M. N. A., Nguyen, H., Dou, J., Song, X., and Poi-Ngian, S. (2019). Application of a Hybrid Artificial Neural Network-Particle Swarm Optimization (ANN-PSO) Model in Behavior Prediction of Channel Shear Connectors Embedded in Normal and High-Strength Concrete, *Applied Sciences*, Vol. 9, No. 24, 5534. doi:10.3390/app9245534.

27. Jalal, F. E., Iqbal, M., Khan, W. A., Jamal, A., Onyelowe, K., and Lekhraj. (2024). ANN-Based Swarm Intelligence for Predicting Expansive Soil Swell Pressure and Compression Strength, *Scientific Reports*, Vol. 14, No. 1, 14597. doi:10.1038/s41598-024-65547-7.

28. 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.

29. Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., and Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery, *Malacca Pharmaceutics*, Vol. 2, No. 2, 68–78. doi:10.60084/mp.v2i2.217.

30. Todeschini, R., and Consonni, V. (2000). *Handbook of Molecular Descriptors*, Wiley-VCH Verlag GmbH, Weinheim, Germany. doi:10.1002/9783527613106.

31. Xue, L., and Bajorath, J. (2000). Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening, *Combinatorial Chemistry & High Throughput Screening*, Vol. 3, No. 5, 363–372. doi:10.2174/1386207003331454.

32. 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.

33. 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.

34. Mswahili, M. E., Martin, G. L., Woo, J., Choi, G. J., and Jeong, Y.-S. (2021). Antimalarial Drug Predictions Using Molecular Descriptors and Machine Learning against Plasmodium Falciparum, *Biomolecules*, Vol. 11, No. 12, 1750. doi:10.3390/biom11121750.

35. 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.

36. Agustia, M., Noviandy, T. R., Maulana, A., Suhendra, R., Muslem, M., Sasmita, N. R., Idroes, G. M., Rahimah, S., Afidh, R. P. F., Subianto, M., Irvanizam, I., and Idroes, R. (2022). Application of Fuzzy Support Vector Regression to Predict the Kovats Retention Indices of Flavors and Fragrances, *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*, IEEE, 13–18. doi:10.1109/ICELTICs56128.2022.9932124.