Journal of the Serbian Chemical Society
JSCS@tmf.bg.ac.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. *88* (10) 1013–1023 (2023)
JSCS–5677
Original scientific paper
Published 24 October 2023

Identification of organic compounds using artificial neural networks and refractive index

INNOCENT ABEL KIRIGITI[1], NANIK SITI AMINAH[1]* and SAMSON THOMAS[2]

[1]*Department of Chemistry, Faculty of Science and Technology, Universitas Airlangga, Surabaya 60115, Indonesia and* [2]*Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia*

(Received 1 February, revised 15 February, accepted 4 August 2023)

Abstract: Identification of chemical compounds has many applications in science and technology. However, this process still relies significantly on the knowledge and experience of chemists. Thus, the development of techniques for faster and more accurate chemical compound identification is essential. In this work, we demonstrate the feasibility of using artificial neural networks to accurately identify organic compounds through the measurement of refractive index. The models were developed based on the refractive index measurements in different wavelengths of light, from UV to the far-infrared region. The models were trained with about 250,000 records of experimental optical constants for 60 organic compounds and polymers from published literature. The models performed with accuracies of up to 98 %, with better performance observed for the refractive index measurements across the visible and IR regions. The proposed models could be coupled with other devices for autonomous identification of chemical compounds using a single-wavelength dispersive measurement.

Keywords: machine learning; ANNs; classification, deep learning, materials identification.

INTRODUCTION

Organic material identification is a vital process across various industries such as pharmaceuticals, food, agriculture and environmental science. The ability to identify organic compounds quickly and accurately is imperative for the development of new products, monitoring environmental pollutants and detecting contaminants in food and drugs.[1–3] However, the traditional methods of material identification, such as gas chromatography and mass spectrometry, can be tedious and costly.[4,5] Therefore, it is crucial to explore new ways to facilitate organic

* Corresponding author. E-mail: nanik-s-a@fst.unair.ac.id
https://doi.org/10.2298/ JSC230201049K

materials identification. One such approach that has gained attention recently is the use of machine learning techniques.

Machine learning (ML) is a branch of computer science that focuses on developing algorithms that can learn from and make decisions based on complex data.[6] One recent development in ML is deep learning, a cutting-edge field that uses artificial neural networks (ANNs) to improve the performance of traditional ML models.[7] ANNs are artificial systems that are modelled after biological neural networks and are able to learn and perform tasks without pre-programmed rules by being exposed to various datasets and examples.[8] Deep learning is among the most effective, efficient, and cost-effective approaches to ML.[9] Additionally, ANNs have the advantage of being able to increase their accuracy in production. Unlike traditional ML models like random forests, ANNs do not need to be fully re-trained as more data becomes available; this can lead to significant cost savings in terms of computational resources. Therefore, ANNs are a suitable approach to ML.

ANNs have found applications in various fields such as environmental science, where they are used to predict the percentage of water pollutant removal based on experimental variables such as temperature and treatment time.[10–12] Moreover, Raman spectroscopy imaging has been widely used in combination with machine learning (ML) techniques to identify the properties and structures of organic compounds. Raman spectroscopy is a non-destructive imaging method that provides information about the vibrational modes of a compound, which can be used to determine its chemical structure and composition.

One of the key benefits of using Raman spectroscopy imaging in conjunction with ML algorithms is its ability to accurately identify the chemical structure and composition of various organic compounds.[13,14] Studies have reported ML models based on Raman spectra that were able to classify materials like biomolecules, organics and inorganics.[15–18]

Despite the higher performance of ML models using spectra images as input, the setup and equipment for obtaining associated spectra data is more complex and expensive. Therefore, a simple measurement like the refractive index (n) can offer alternatives. Refractive index of a sample is defined as the ratio of the speed of light in a vacuum to its speed in the sample medium.

The chemical composition of a sample can also affect its refractive index through the presence of certain functional groups or atoms that can interact with light in specific ways.[19] For example, refractive index has been used for detection of components with low chromophoric activities such as sugars, triglycerides, organic acids, pharmaceutical excipients, and polymers.[20] So, the refractive index is an optical property that carries enough information related to chemical composition.

Machine learning models for predicting refractive indices of polymers have been previously reported.[21,22] In another work, refractive index was used as the input to ML models to differentiate between normal and malignant tissues in biomedical.[23]

This work was inspired by a study attempting to apply random forests (RF), a traditional ML algorithm and refractive index to identify organic compounds.[24] In order to train the machine learning models, we use data from a public database of refractive indices for organic compounds and polymers. The database contains data from literature gathered over a long period of time.[25]

To the best of our knowledge, this is the first work to report the use of refractive index for classification of organic compounds with artificial neural networks (ANNs).

EXPERIMENTAL

Database

Version 1.0.0 of a web scrapper built using Python was run on the refractive index website, which is a database for experimental optical constants from published literature since 1940.[25] The scrapper targeted 60 organic compounds and polymers. The scrapped data was stored as comma-separated values (CSV). The file contains four columns: organic compound (book), wavelength (λ), refractive index (n) and extinction coefficient (k). The scrapped database has a total of 248,756 entries and 9645 missing values of k.

The database was split into five categories as follows (Fig. 1): ultraviolet (0–0.4 μm), visible (0.4–0.75μm), near IR (0.75–1.5 μm), IR (1.5–4μm) and far IR (>4 μm).

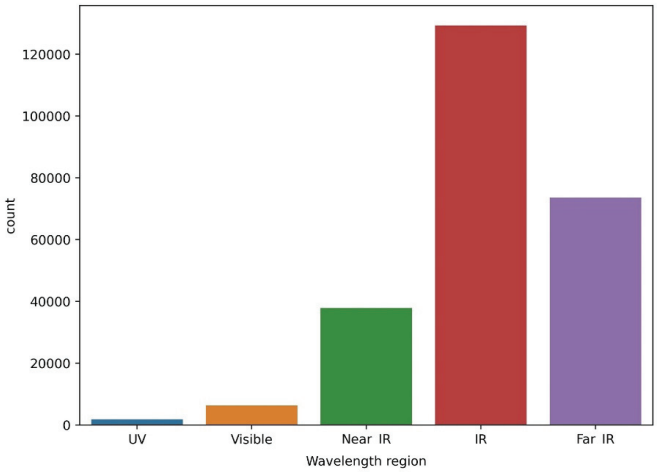

Fig. 1. Raw scrapped data from the refractive index website.

Data pre-processing

All the missing values for extinction coefficients in the raw database were replaced by zero. Since the UV and visible data was not enough for training the models, the data augmentation was performed on these regions using the Sellmeier equation.[26] This is a mathematical

formula that can be used to predict the refractive index of a material as a function of wavelength. The data augmentation was required to artificially synthesize more data using domain knowledge,[27] which is a technique used previously with Raman spectra-based organic classifiers.[16–18]

The Sellmeier equation is given by the following equation:

$$n^2(\lambda) = 1 + \frac{B_1\lambda_2}{\lambda_2 - C_1} + \frac{B_2\lambda_2}{\lambda_2 - C_2} + \ldots \quad (1)$$

Where $n(\lambda)$ is the refractive index at wavelength λ. B_1, B_2,... and C_1, C_2,... are Sellmeier coefficients that are specific to the material. A custom Python script was used to estimate the missing Sellmeier coefficients by curve fitting (see the supporting information).

Artificial neural network classifiers (ANNs) for organic compounds

Scikit-Learn, TensorFlow and Keras Python libraries were used for training and evaluating the accuracy of the ANN classifiers. This was done in a Google Collaboratory environment.[28,29] Seven different models were developed according to available categories (see Fig. 4 later on). Each model contains three main layers: an input layer, hidden layers, and an output layer. The input layer takes in three independent variables: λ, n and the extinction coefficient (k). Hidden layers contain neurons; they extract and represent features from the input data, allowing the network to learn. The output layer performs the final compound classification, which is based on voting among 60 possible compounds. The compound with a high probability is considered the output of the model.[7,28] An overview of the model design is shown in Fig. 2.

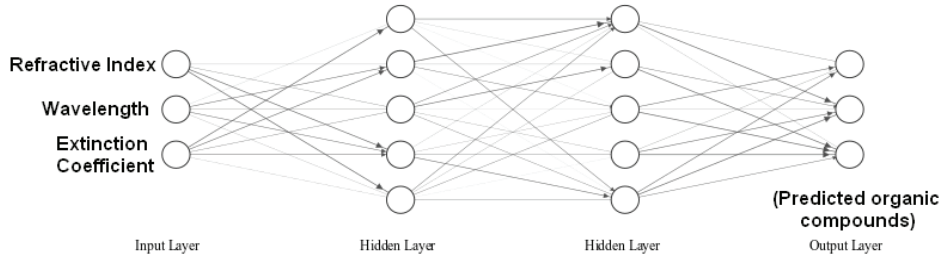

Fig. 2. ANN organic classifier model architecture.

To evaluate the model's performance, monitoring was performed during the training and testing stages. In the training stage, the loss and accuracy of all models were monitored by a validation data set. If the model's prediction is perfect, the loss is zero. This tells how poorly or well a model behaves after each iteration of the optimization.[29,30] On the other hand, the testing stage was performed using a test data set; the test data serves as an estimate of its performance on new, unseen, data.

Accuracy, precision, recall and the $F1$ score were used as evaluation metrics for the classifier. Precision tells us how many of the positive predictions were correct; recall tells us how many of the actual positives were identified while the $F1$ score gives a single metric to evaluate the overall performance of the model by balancing precision and recall.[28,31] These evaluation metrics are mathematically defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1\text{score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{4}$$

where the term true positives (*TP*) refer to the number of samples correctly predicted as positive, while false positives (*FP*) indicate the number of samples incorrectly predicted as positive. Similarly, true negatives (*TN*) refer the number of samples correctly predicted as negative and false negatives (*FN*) represents the number of samples incorrectly predicted as negative.

RESULTS AND DISCUSSION

Data pre-processing

Analysis of the raw data revealed the percent of missing extinction coefficient values in each region as follows (Fig. 3): UV (58.99 %), visible (82.67 %), near IR (7.28 %), IR (0.32 %) and far IR (0.25 %). The final database with data augmentation contains seven categories (Fig. 4). The UV region data was increased from 1807 to 132314, and the visible region data was increased from 6268 to 120455. The amount of data in other regions was left unaltered.

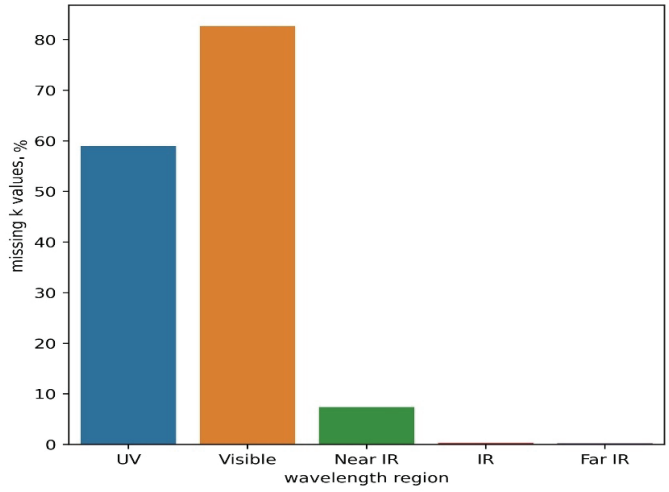

Fig. 3. Missing extinction coefficient values in raw data.

Performance of artificial neural network classifiers

The accuracies of the ANN models are listed in the following order: Near IR (98.44 %) > IR (97.72 %) > visible-augmented (86.60 %) > far IR (84.09 %) > UV-augmented (81.49 %) > visible (69.22 %) > UV (59.00 %). It is observed

that models in the near IR and IR regions outperform other regions (Fig. 5). Moreover, from Figs. 6 and 7, the losses of the near IR and IR regions converge to low and stable values, while the accuracy reaches high and stable values. This indicates that the models in the near IR and IR regions are generalizing well and not overfitting.[28,32] The precision, recall and *F*1 scores of the near IR and IR models shows high performance of above 98 % (Table I).

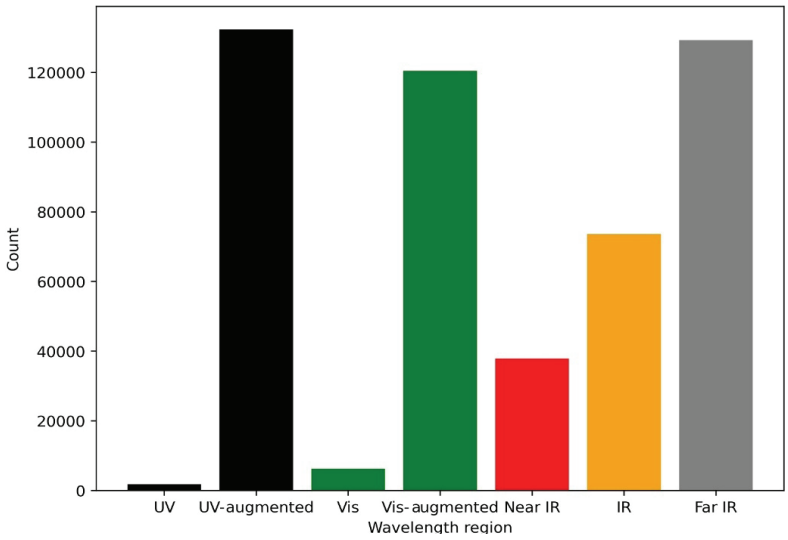

Fig. 4. Augmented refractive index data.

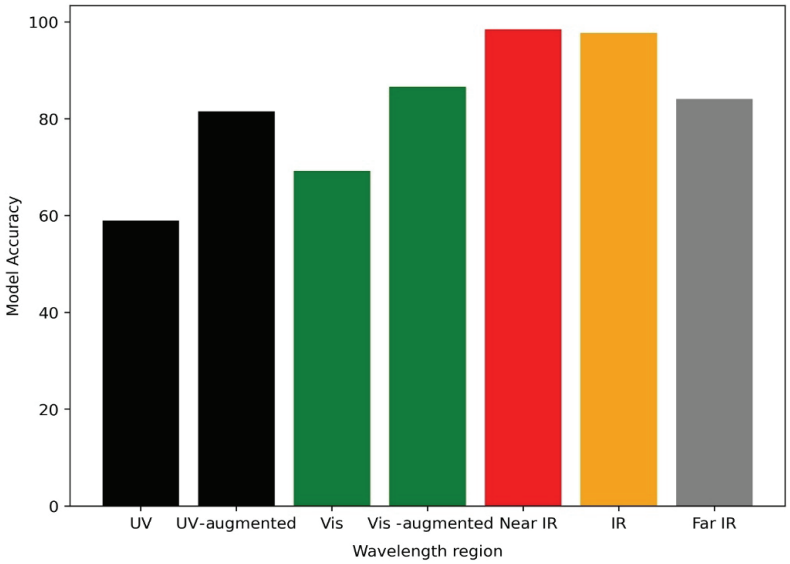

Fig. 5. Testing accuracies for the ANN classifiers.

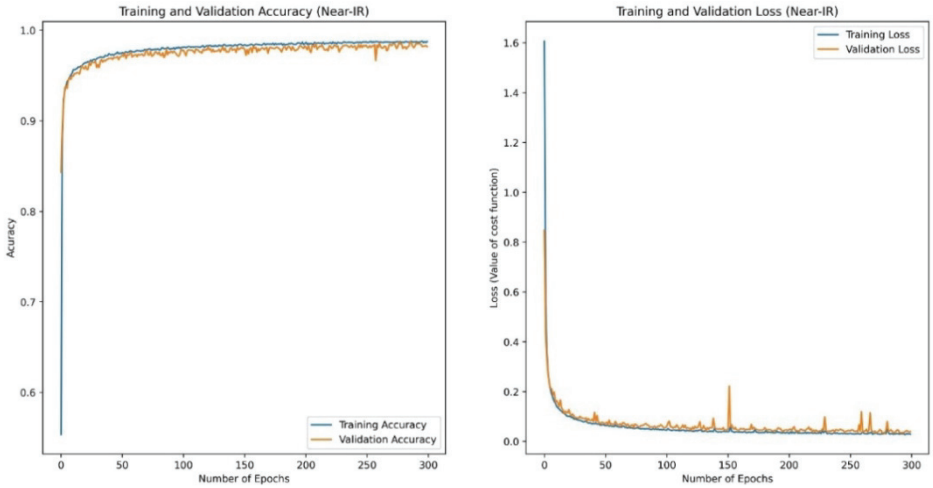

Fig. 6. Training and validation for ANN model in near IR region.

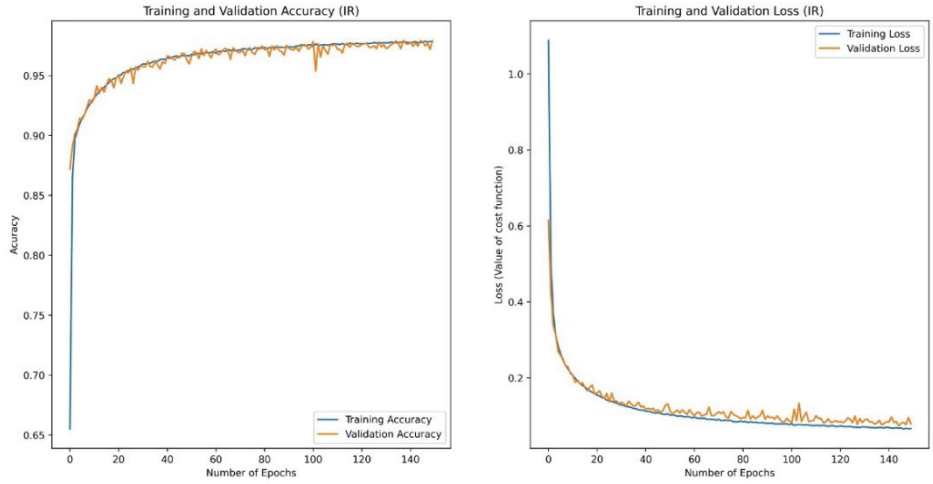

Fig. 7. Training and validation for ANN model in IR region.

TABLE I. Performance evaluation for ANN classifiers

ANN Model	Precision, %	Recall, %	*F*1 score, %	Accuracy, %
UV	79.00	81.00	79.00	59.00
UV – augmented data	82.00	82.00	81.00	81.49
Visible	73.00	69.00	67.00	69.22
Visible – augmented data	86.00	87.00	85.00	86.60
Near IR	99.00	98.00	98.00	98.44
IR	98.00	98.00	98.00	97.72
Far IR	85.00	84.00	84.00	84.09

Meanwhile, the performance of models in the far IR, UV and visible regions is unsatisfactory. The recall, precision and $F1$ scores in these regions are low (Table I). Their loss-accuracy plots show unstable values, and bumpy non-converging lines, indicating their unreliability (Figs. S1–S5 of the Supplementary material tp this paper). For UV and visible regions, more than 58 % of k values were missing in their datasets (Fig. 3), which may provide the model with less information to accurately learn the relationships between input and output variables. As a result, overfitting or poor generalization performance may be observed.[33]

Nevertheless, the accuracies for UV and visible models were observed to increase by more than 17 % with data augmentation (Fig. 5); this suggests the possibility of improving the performance by increasing the amount of training data in these regions.

The accuracy and loss plots obtained from the near IR (Fig. 6) and IR (Fig. 7) models are smoother, and the lines converge well, while plots for models containing augmented data show convergence but with bumpy lines (Fig. S1––S5). This implies that the validation dataset is not a good representation of the training data set; this may be due to artificially synthesized data from the data augmentation process, which introduced noise.[28,32]

Comparison with other machine learning organic classifiers

The developed models are comparable with models from previous studies (Table II), indicating the potential of using the refractive index measurement to facilitate the identification of organic compounds using machine learning.

TABLE II. Comparison of this work with some previous studies using Raman spectra data with machine learning. ResNet = residual neural network, DRCNN = deeply-recursive convolutional neural network, ANN = artificial neural network, CNN = convolutional neural network, *K*NN = *K*-nearest neighbor, ML = machine learning

Method	Dataset	Accuracy, %	Reference
ResNet	Organic biomolecules	100.00	15
CNN	Organic and inorganic compounds	100.00	16
DRCNN	Organic compounds and minerals	98.10	17
KNN	Organic biomolecules	93.90	15
KNN	Edible oils (fatty acids)	88.90	18
ANN classifier	Organic compounds and polymers	81.49 (UV) 86.60 (Vis) 98.44 (Near IR) 97.72 (IR) 84.09 (Far IR)	This work

CONCLUSION

In this study, the artificial neural network classifiers (ANNs) for identifying organic compounds were developed and tested successfully. The models rely on refractive index measurements across the UV and far IR spectral regions. Information related to the refractive index of an organic compound and the wavelength of light used facilitate its accurate identification by artificial neural networks.

ANNs in the near IR and IR regions showed better performance, with the accuracy levels above 97 %, suggesting the potential of refractive index measurements in these regions. The observed performance is comparable to models using Raman spectra as inputs. Although the accuracies for the UV, visible and far IR regions are slightly lower, ranging from 81 to 86 %, the additional data and hyperparameter optimizations showed the possibility of improving performance in the future.

This study demonstrates the feasibility of using artificial neural networks to identify organic compounds using a single wavelength dispersive measurement.

SUPPLEMENTARY MATERIAL

Additional data and information are available electronically at the pages of journal website: https://www.shd-pub.org.rs/index.php/JSCS/article/view/12261, or from the corresponding author on request.

ИЗВОД

ИДЕНТИФИКАЦИЈА ОРГАНСКИХ ЈЕДИЊЕЊА КОРИШЋЕЊЕМ ВЕШТАЧКИХ НЕУРОНСКИХ МРЕЖА И ИНДЕКСА ПРЕЛАМАЊА

INNOCENT ABEL KIRIGITI[1], NANIK SITI AMINAH[1] и SAMSON THOMAS[2]

[1]Department of Chemistry, Faculty of Science and Technology, Universitas Airlangga, Surabaya 60115, Indonesia and [2]Department of Chemistry, Faculty of Mathematics и Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia

Идентификација хемијских једињења има много примена у науци и технологији. Међутим, овај се процес још увек много ослања на знање и искуство хемичара. Тако је од суштинске важности развој техника за брже и поузданије идентификовање хемијских једињења. У овом раду, представићемо изводљивост коришћења неуронских мрежа за поуздано идентификовање органских једињења мерењем индекса преламања. Развијени су модели засновани на мерењима индекса преламања на различитим таласним дужинама светлости, од UV до далеке инфрацрвене области. Модели су тренирани са око 250,000 записа експерименталних оптичких константи за 60 органских једињења и полимера из публиковане литературе. Модели су извођени са поузданошћу до 98 %, са бољим резултатом опаженим за мерења индекса преламања у видљивој и ИЦ области. Предложени модели се могу спрегнути са другим уређајима за аутономну идентификацију хемијских једињења дисперзивним мерењем на једној таласној дужини.

(Примљено 1. фебруара, ревидирано 15. фебруара, прихваћено 4. августа 2023)

REFERENCES

1. W. Shi, W.-E. Zhuang, J. Hur, L. Yang, *Water Res.* **188** (2021) 116406 (https://doi.org/10.1016/j.watres.2020.116406)
2. J. Borrull, A. Colom, J. Fabregas, F. Borrull, E. Pocurull, *J. Chromatogr., A* **1621** (2020) 461090 (https://doi.org/10.1016/j.chroma.2020.461090)
3. L. Díaz de León-Martínez, R. Flores-Ramírez, C. M. López-Mendoza, M. Rodríguez--Aguilar, G. Metha, L. Zúñiga-Martínez, O. Ornelas-Rebolledo, L. E. Alcántara-Quintana, *Clin. Chim. Acta* **522** (2021) 132 (https://doi.org/10.1016/j.cca.2021.08.014)
4. C. Zarfl, *Anal. Bioanal. Chem.* **411** (2019) 3743 (https://doi.org/10.1007/s00216-019-01763-9)
5. B. Nozière, M. Kalberer, M. Claeys, J. Allan, B. D'Anna, S. Decesari, E. Finessi, M. Glasius, I. Grgić, J. F. Hamilton, T. Hoffmann, Y. Iinuma, M. Jaoui, A. Kahnt, C. J. Kampf, I. Kourtchev, W. Maenhaut, N. Marsden, S. Saarikoski, J. Schnelle-Kreis, J. D. Surratt, S. Szidat, R. Szmigielski, A. Wisthaler, *Chem. Rev.* **115** (2015) 3919 (https://doi.org/10.1021/cr5003485)
6. T. F. G. G. Cova, A. A. C. C. Pais, *Front. Chem.* **7** (2019) 809 (https://doi.org/10.3389/fchem.2019.00809)
7. C. Janiesch, P. Zschech, K. Heinrich, *Electron. Mark.* **31** (2021) 685 (https://doi.org/10.1007/s12525-021-00475-2)
8. P. P. Shinde, S. Shah, in *Proceedings of 2018 Fourth Int. Conf. Comput. Commun. Control Autom.,* 2018, pp. 1–6 (https://doi.org/10.1109/ICCUBEA.2018.8697857)
9. S. Dargan, M. Kumar, M. R. Ayyagari, G. Kumar, *Arch. Comput. Methods Eng.* **27** (2020) 1071 (https://doi.org/10.1007/s11831-019-09344-w)
10. E. Yabalak, *J. Environ. Sci. Heal., A* **53** (2018) 975 (https://doi.org/10.1080/10934529.2018.1471023)
11. E. Yabalak, Ö. Yilmaz, *J. Iran. Chem. Soc.* **16** (2019) 117 (https://doi.org/10.1007/s13738-018-1487-8)
12. E. Yabalak, B. Külekçi, A. M. Gizir, *J. Environ. Sci. Heal., A* **54** (2019) 1412 (https://doi.org/10.1080/10934529.2019.1647749)
13. M. H. W. N. Jinadasa, A. C. Kahawalage, M. Halstensen, N.-O. Skeie, K.-J. Jens, in *Recent developments in atomic force microscopy and Raman spectroscopy for materials characterization*. C. S. Pathak, S. Kumar, Eds., InTech Open, Rijeka, 2021 (https://doi.org/10.5772/INTECHOPEN.99770)
14. L. Pan, P. Zhang, C. Daengngam, S. Peng, M. Chongcheawchamnan, *J. Raman Spectrosc.* **53** (2022) 6 (https://doi.org/10.1002/jrs.6225)
15. X. Chen, L. Xie, Y. He, T. Guan, X. Zhou, B. Wang, G. Feng, H. Yu, Y. Ji, *Analyst* **144** (2019) 4312 (https://doi.org/10.1039/C9AN00913B)
16. T. Cooman, T. Trejos, A. H. Romero, L. E. Arroyo, *Chem. Phys. Lett.* **787** (2022) 139283 (https://doi.org/10.1016/J.CPLETT.2021.139283)
17. W. Zhou, Y. Tang, Z. Qian, J. Wang, H. Guo, *RSC Adv.* **12** (2022) 5053 (https://doi.org/10.1039/D1RA08804A)
18. C. Berghian-Grosan, D. A. Magdas, *Talanta* **218** (2020) 121176 (https://doi.org/10.1016/J.TALANTA.2020.121176)
19. J. M. Hollas, *Modern spectroscopy*, 4th ed., Wiley & Sons, Chichester, 2004, ISBN: 978--1-118-68160-2
20. M. W. Dong, *Sep. Sci. Technol.* **6** (2005) 47 (https://doi.org/10.1016/S0149-6395(05)80047-9)

21. J. P. Lightstone, L. Chen, C. Kim, R. Batra, R. Ramprasad, *J. Appl. Phys.* **127** (2020) 215105 (https://doi.org/10.1063/5.0008026)
22. S. A. Schustik, F. Cravero, I. Ponzoni, M. F. Díaz, *Commun. Comput. Inf. Sci.* **1408** (2021) 279 (https://doi.org/10.1007/978-3-030-76310-7)
23. N. Qi, Z. Zhang, Y. Xiang, Y. Yang, X. Liang, P. D. B. Harrington, *Anal. Methods* **7** (2015) 2333 (https://doi.org/10.1039/C4AY02665A)
24. T. Bikku, R. A. Fritz, Y. J. Colón, F. Herrera, *Machine learning identification of organic compounds using visible light,* 2022 (https://doi.org/10.48550/arxiv.2204.11832)
25. M. N. Polyanskiy, *Refractive index database*, https://refractiveindex.info/ (accessed: August 20, 2022)
26. J. W. Gooch, *Encycl. Dict. Polym.* (2011) 653 (https://doi.org/10.1007/978-1-4419-6247-8_10447)
27. J. K. Kim, J. Shao, *Statistical Methods for Handling Incomplete Data*, Chapman and Hall/CRC, Boca Raton, FL, 2021 (https://doi.org/10.1201/9780429321740)
28. A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*, 2nd ed., O'Reilly Media, Inc., Sebastopol, CA, 2019, ISBN: 9781492032649
29. E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Apress, Berkeley, CA, 2019 (https://doi.org/10.1007/978-1-4842-4470-8)
30. *Machine Learning in Chemistry*, H. M. Cartwright, Ed., Royal Society of Chemistry, 2020 (https://doi.org/10.1039/9781839160233)
31. J. Han, J. Pei, H. Tong, *Data mining: concepts and techniques*, Morgan Kaufmann, Burlington, MA, 2011, ISBN 978-0-12-381479-1
32. F. Chollet, *Deep Learning with Python*, Manning Publications Co, Shelter Island, NY, 2017, ISBN 9781617294433
33. M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, A. Fernández-Delgado, *J. Mach. Learn. Res.* **15** (2014) 3133 (http://jmlr.org/papers/v15/delgado14a.html).