THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCA Article

Machine Learning Identification of Organic Compounds Using Visible Light

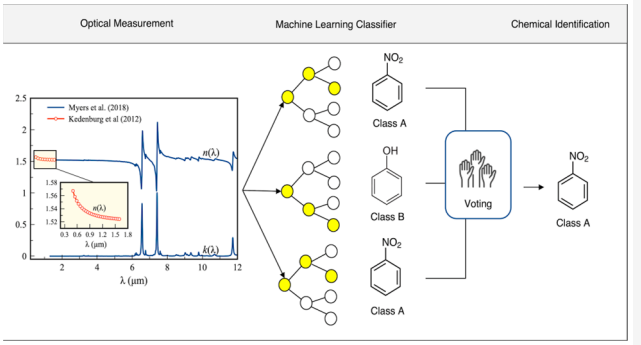
Published as part of The Journal of Physical Chemistry virtual special issue "Early-Career and Emerging Researchers in Physical Chemistry Volume 2".

Thulasi Bikku, Rubén A. Fritz, Yamil J. Colón, and Felipe Herrera*

Cite This: https://doi.org/10.1021/acs.jpca.2c07955 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Identifying chemical compounds is essential in several areas of science and engineering. Laser-based techniques are promising for autonomous compound detection because the optical response of materials encodes enough electronic and vibrational information for remote chemical identification. This has been exploited using the fingerprint region of infrared absorption spectra, which involves a dense set of absorption peaks that are unique to individual molecules, thus facilitating chemical identification. However, optical identification using visible light has not been realized. Using decades of experimental refractive index data in the scientific literature of pure organic compounds and polymers over a broad range of frequencies from the ultraviolet to the far-infrared, we develop a machine learning classifier that can accurately identify organic species based on a single-wavelength dispersive measurement in the visible spectral region, away from absorption resonances. The optical classifier proposed here could be applied to autonomous material identification protocols and applications.



I. INTRODUCTION

Scientific data analysis has been accelerated by machine learning by training models that allow a rapid interpretation of complex data patterns and the automated control of measurement devices.[1,2] The autonomous identification and discovery of chemical compounds for applications in science and industry can therefore benefit from the development of compact, portable, and highly accurate sensors powered by machine learning.[3] Remote molecular sensing based on light exploits the dispersive and absorptive response of a material system to electromagnetic radiation.[4,5] Although chemical methods can be very specific,[6,7] optical sensing techniques can be beneficial because light-matter interaction is nondestructive and can be processed remotely.

The refractive index is an optical property of materials fundamentally related to microscopic physicochemical characteristics such as the dynamic polarizability as well as macroscopic variables such as concentration, temperature, and pressure.[8,9] Refractive indices are commonly used for quantifying the content of target molecules in agricultural,[10,11] pharmaceutical,[8] and manufacturing applications.[12] Therefore, a refractive index database over a broad range of frequencies can serve as useful training data for a machine learning workflow that enables the autonomous identification of chemical compounds using light.

Molecular spectroscopy databases have already been used for training machine learning algorithms in chemical identification problems.[13−15] Efforts have focused on training classifiers with infrared (IR) absorption and Raman scattering databases, in the mid-infrared (mid-IR) spectral range ($\lambda \approx 3-50\ \mu m$).[11−32] In general, the information in the mid-IR is so rich and complex that it would be very unlikely for different molecules to have the same peak structure, particularly in the so-called "fingerprint" region of the IR spectrum.[33−35] Compounds with the same chemical formula but different spatial conformation (isomers) can thus be discriminated by analyzing the position of their Raman peaks, for example.[36] Machine learning classifiers trained on vibrational spectroscopy data can therefore be very accurate, with identification errors of a few percent or less.[13,16,28−32]

Received: November 11, 2022
Revised: February 3, 2023

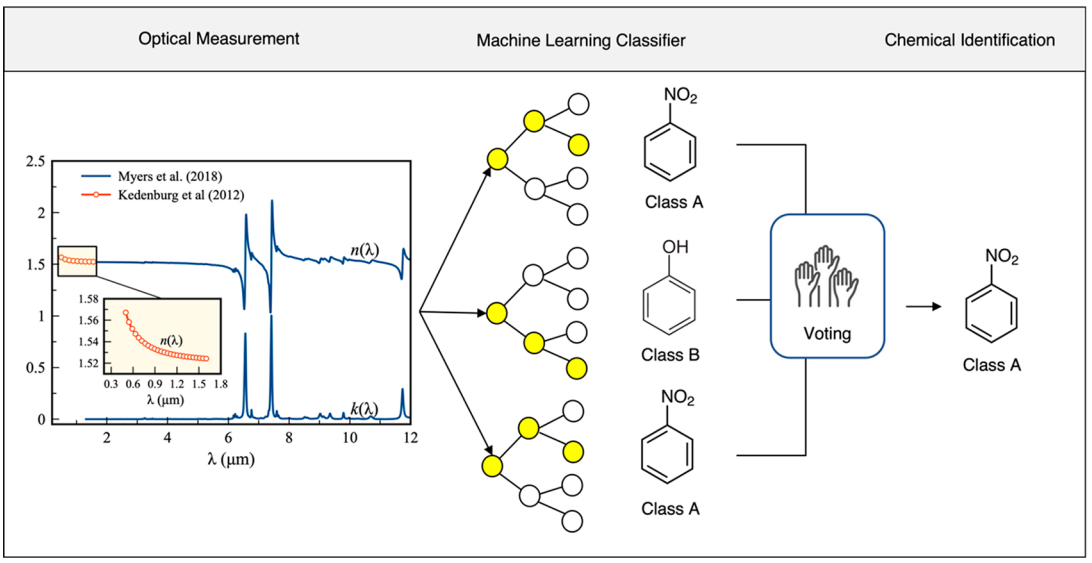


Figure 1. Illustration of the proposed machine learning classification scheme for chemical identification. The left panel shows a sample of the refractive index $n(\lambda)$ and extinction $k(\lambda)$ spectra used for model training. The spectral data is public domain and can be downloaded at https://refractiveindex.info.[52]

Alternative machine learning strategies for molecular classification that are not trained with infrared absorption or scattering spectra have also been reported.[28−32] Away from absorption resonances, molecules and materials experience a dispersive response in the presence of external electromagnetic fields, which is determined by the real part of the dielectric function of the medium $\epsilon(\lambda)$. The imaginary part of $\epsilon(\lambda)$ gives the extinction coefficient $k(\lambda)$[1], which quantifies electromagnetic energy loss due to absorption and scattering. Absorption and dispersion are fundamentally related to each other via the Kramer-Kronig relation.[9] In frequency regions with negligible extinction, the dispersive response is entirely determined by the refractive index $n(\lambda) = \sqrt{\mathrm{Re}[\epsilon]}$, which provides information on the electronic structure of materials ($\lambda \approx 400-800$ nm) or their dielectric properties ($\lambda > 50$ μm). As mentioned above, the refractive index $n(\omega)$ is also an extensive property and thus correlates with the molecular density.

The refractive index has been used for detecting tissue damage in biomedical samples using terahertz transmission images ($\lambda > 100$ μm) and deep learning, reaching recall performances of up to 93% with feature engineering techniques.[15,37] Databases with the static refractive indices ($\lambda \rightarrow \infty$) of small organic compounds, solvents, and polymers have also been developed for training regression models that predict the refractive index given a set of structural and quantum chemical descriptors.[38−43] Predictive models can be useful for the automated discovery of organic optoelectronic materials.[44] Experimental databases with organic refractive indices at the sodium wavelength (589 nm) have also been used for constructing quantitative structure-property relationships (QSPR) that can be used, for example, in the chemical analysis of organic mixtures using visible light.[45]

For molecular classification tasks based on IR spectral data, the target feature of interest for a classifier ("molecule identity") is learned from a high-dimensional feature vector that encodes the spectral peaks of interest.[13] In contrast, the static refractive index of organic compounds is a single-valued feature that does not have the same information content.[46] The same applies to refractive index databases at a single wavelength away from ultraviolet and infrared absorption resonances. New databases and analysis techniques are thus needed for applications that can benefit from accurate organic molecule classification protocols based on the dispersive response of materials over a tunable range of wavelengths in the visible.

Motivated by the growing interest in developing autonomous tools for organic materials discovery,[47−49] here we demonstrate a machine learning classifier of organic compounds based on the measurement of the refractive index $n(\lambda_0)$ at a single optical wavelength λ_0 anywhere in the visible spectral region (400−750 nm), where most organic compounds are fully transparent.[50] The classifier is trained with a publicly available materials science database containing the optical constants of 61 organic molecules and polymers over the spectral range spanning from ultraviolet to far-infrared wavelengths. The classification scheme is illustrated in Figure 1. We envision the proposed classifier for data analysis at the output of a liquid or gas-phase chromatograph that can separate complex chemical mixtures into multiple single-component fractions.[51]

The rest of the article is organized as follows: After describing the structure and information content of the original spectral database in Sec. II.1, we describe the data preprocessing strategies in Sec. II.2 and demonstrate classification errors smaller than 1% in the visible range in Sec. III. Comparisons with recent Raman-based classifiers are given in Sec. III.4. We conclude and suggest future directions and applications in Sec. IV.

II. METHODS

II.1. Experimental Refractive Index Database. We do web scraping on a public domain database of experimental optical constants available at https://refractiveindex.info.[52] The Web site is a repository of published data from the scientific literature since 1940. The site is organized into categories that resemble a virtual bookshelf: The "Shelf" category groups materials into inorganic, organic, glasses, others, and 3D; the "Book" category contains the chemical

https://doi.org/10.1021/acs.jpca.2c07955
J. Phys. Chem. A XXXX, XXX, XXX−XXX

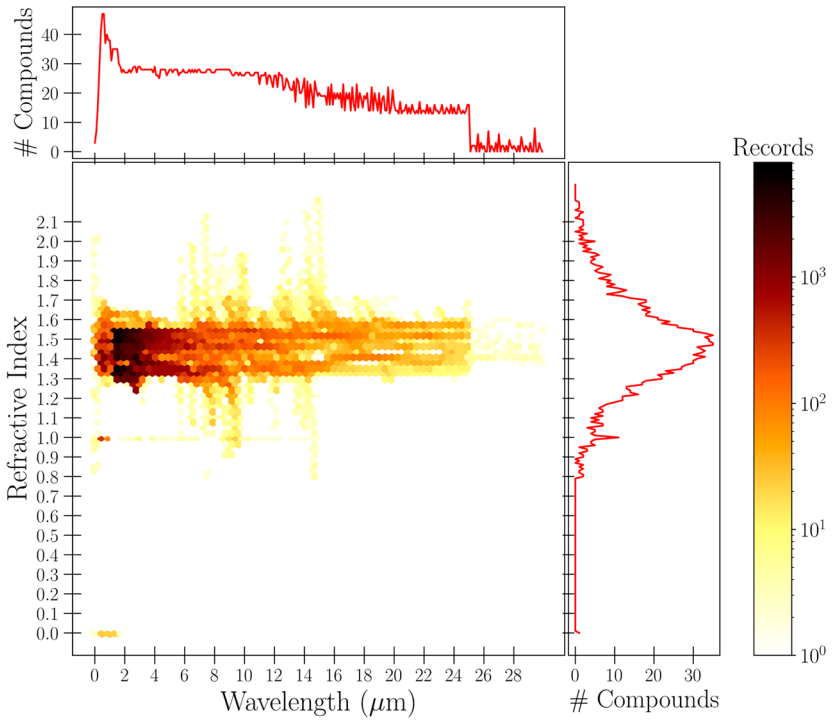


Figure 2. Distribution of the raw data from the https://refractiveindex.info/[52] over a broad range of wavelengths and refractive indices. Side panels show cuts of the data record distribution over wavelengths (top) and refractive indices (right).

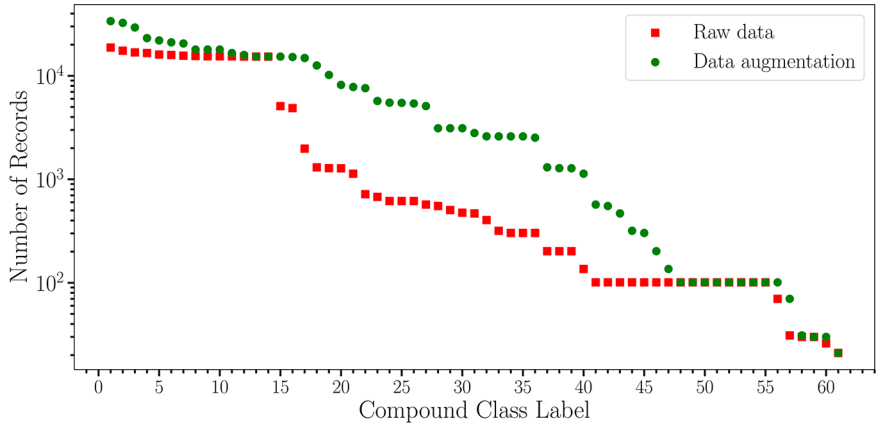


Figure 3. Number of data records per molecular compound (61 compound class labels) available in the raw database from ref 52. The record distribution after data augmentation via Sellmeier fitting in the UV/vis is also shown. The correspondence between the class labels and organic compound names is given Table S6 of the Supporting Information.

compound name, which will be the "class" label predicted by the machine learning classifier on output; the "Page" subsection refers to the source where the optical data was first reported in the literature as well as comments and other information such as the group velocity and group velocity dispersion, the measurement wavelength range, and the state of matter of the sample (gas, liquid, or solid). Each "Page" record has a csv file with the spectral data for n (refractive index) and k (extinction coefficient) over a range of wavelengths (λ). From the general data set, we build a smaller set by selecting "Organic Materials" in the "Shelf" category, which contains 61 organic molecules and polymers. The compiled file has 194 816 data records, sorted in columns with the features "Shelf", "Book", "Page", "λ" (wavelength), "n" (refractive index), and "k" (extinction coefficient). The data has 418 missing n values and 60 944 missing k values. Example data records are shown for the "Acetone" class in Table S1 of the Supporting Information (SM).

In Figure 2, we show a visualization of the organic data set over a grid of wavelengths (λ) and refractive indices (n). The color code indicates the number of data records in each (λ, n) region [The records involve measurements up to 25 μm and refractive indices in the range (0.0−2.3)]. The upper and right-side panels show the number of organic compounds (counted as a label on the "Book" category) with data records in each wavelength and refractive index range.

II.2. Preprocessing the Training Data Set. The original database is naturally heterogeneous and imbalanced, as some compounds and frequency regions have been studied more intensely than others in the literature. This represents a challenge for the implementation of machine learning classification algorithms.[53] Data sets with unequal data records

https://doi.org/10.1021/acs.jpca.2c07955
J. Phys. Chem. A XXXX, XXX, XXX−XXX

per target feature lead to majority and minority classes, thus affecting the overall predictive accuracy of classification models.[54] Figure 3 shows the number of data records per organic molecule class for the raw database (RD) and after preprocessing through data augmentation (DA). The number of records in the original data varies from only a few spectral points (e.g., Polymethyl Pentene, Class Label = 58) to a few thousand (e.g, Ethanol, Class Label = 1), which illustrates the class imbalance problem. In addition, different compounds in the refractive index database can have different types of spectroscopic features over different wavelengths, which is a form of type imbalance.[55]

We tried several preprocessing strategies to overcome class and type imbalance in the raw data set, including oversampling (OS), undersampling (US), and physics-based DA. We also used feature engineering (FE-1 and FE-2) strategies, spectral-based binning (SBB), and spectral-based binning with feature engineering (SBB-FE1 and SBB-FE2), attempting to increase the prediction accuracies with the original imbalanced data set. The details of these preprocessing strategies are given in the Results section.

II.3. Random Forest Classifier. The random forest (RF) algorithm was chosen as the default method for classification. Random forest is a supervised machine learning algorithm based on an ensemble of decision trees.[56] It uses bagging and feature randomness when building each individual decision tree to create an uncorrelated forest of trees whose prediction is more accurate by committee than individually. The data set was divided into 75% for training and 25% for testing. The results below were obtained with a single instance of the split, but due to the intrinsic overfitting evidenced by the relatively large difference between our training and testing accuracies, we carried out additional cross-validation tests over random splits. The cross-validated accuracies do not vary significantly among different splits (see Table S3 in the SM). RF was implemented using Python's Scikit-Learn library with default hyper-parameters.[57] In early stages of this study, we obtained classification accuracies with alternative models such as gradient boosting, support vector classification, and logistic regression on the raw data set, and random forest performed best (see test results Table S5 in the SI). The code and the data sets used in this work are publicly available at https://github.com/fherreralab/organic_optical_classifier.

III. RESULTS AND DISCUSSION

We now discuss the random forest classification performance obtained by training with the original imbalanced data set and after preprocessing the data. We tested several preprocessing techniques to gain insight on the optimal experimental setup necessary for achieving reasonable identification outcomes using minimal optical measurements.

III.1. Classification Accuracies with Imbalanced Class Sets. Incorporating domain knowledge into the features of the training set is known to increase the prediction performance of classification models. Based on this intuition and inspired by the feature structure of Raman-based classifiers[18,24,26] where the target class is associated with a high-dimensional feature vector containing the spectral peaks, we tested whether grouping small fractions of the curves $n(\lambda)$ and $k(\lambda)$ into feature vectors $\boldsymbol{x} = [\lambda_j, n_j, k_j]$ improved the classifier performance. As explained above, the original data set has a three-dimensional feature vector per target class ($j = 1$) representing a single evaluation of the n and k curves at a given value of λ. We then build six-dimensional ($j = 1,2$) and nine-dimensional feature vectors ($j = 1,2,3$) containing the information from two and three consecutive points in the n and k curves. We refer to the six- and nine-dimensional feature vector schemes as feature engineering 1 (FE1) and 2 (FE2), respectively. As a result of the increase in the number of features per target class, the number of records is reduced. For classification problems with large data sets, binning strategies can prove useful for improving the overall class prediction accuracy.[55,56,58] In addition to the FE1 and FE2 strategies, we adopt a spectral-based binning (SBB) strategy based on the division of the wavelength domain into five spectroscopic regions: UV [$\lambda < 0.40\ \mu$m] containing 1773 data records, Visible (VIS) [$0.40 < \lambda < 0.75\ \mu$m] with 5979 records; Near-Infrared (Near-IR) [$0.75 < \lambda < 1.50\ \mu$m] with 35 445 records; Infrared (IR) [$1.50 < \lambda < 4.0\ \mu$m] with 135 407 records; and Far-Infrared (Far-IR) [$\lambda > 4.0\ \mu$m] with 66 678 records. Our splitting of the IR spectrum into subregions is not necessarily standard.[59]

In Figure 4, we compare the training and testing accuracies of the random forest classification training with the raw

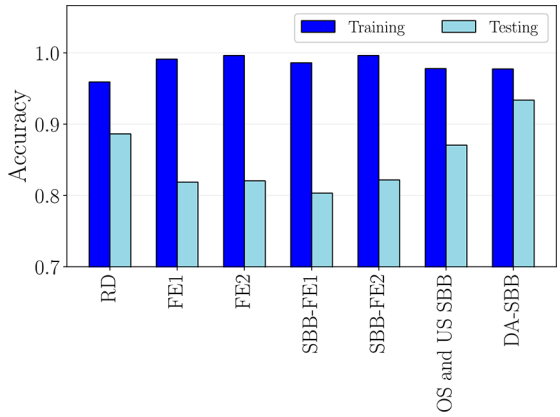


Figure 4. Overall training and testing accuracy for imbalanced data, not separated by wavelength range for the different data preprocessing strategies.

database (RD) as well as FE and SBB preprocessing strategies. Accuracies are also provided in Tables S3 and S4 of the SM. The overall testing accuracies are about 80% with and without preprocessing, although feature engineering and spectral-based binning tend to reduce the prediction accuracy. Combining FE and SBB did not improve performance relative to the raw data. The low performance of the classifier over the entire range of wavelengths (UV to Far-IR) comes roughly speaking from an average of spectral regions of very high accuracies (IR) and regions with very low accuracies (UV/vis). In what follows, we separately studied the performance in different spectral regions.

III.2. Addressing Class Imbalance in the Database. In Figures 3 and 5, we illustrate that the original data set contains a disproportionate number of data records per organic compound in the IR spectral bin (1.50−4.0 μm), relative to the UV and VIS bins. This type of record distribution generates a class imbalance problem[60] that we address using the following strategies: undersampling (US), oversampling (OS), and data augmentation (DA).

Resampling strategies aim to balance classes in the training data by reshaping the data set such that the numbers of records in the different classes becomes comparable.[60] Undersampling

https://doi.org/10.1021/acs.jpca.2c07955
J. Phys. Chem. A XXXX, XXX, XXX−XXX

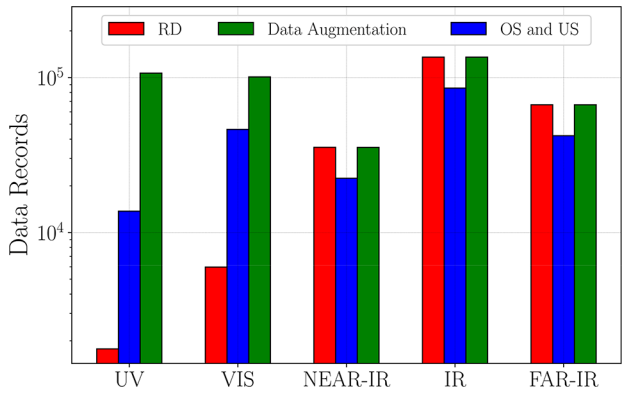


Figure 5. Distribution of data records per spectral bin in the raw database (RD), after oversampling (OS), undersampling (US), and data augmentation (DA).

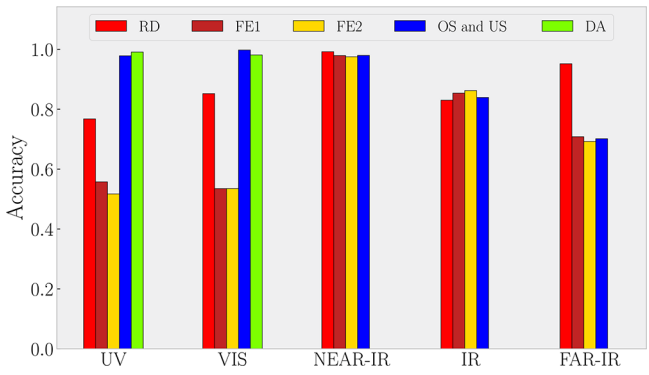


Figure 6. Testing accuracy after spectral-based binning with different data preprocessing strategies: Feature engineering (FE1 and FE2), oversampling (OS), undersampling (US), and data augmentation (DA). The accuracies of the raw database are also shown for comparison.

is a method that randomly reduces the number of data records in the majority class while oversampling duplicates randomly chosen records in the minority class. Both resampling strategies can be effective when used independently or combined. Figure 5 shows the record distribution used for training the random forest model after OS and US are carried out. In the resampled data set, the number of records in the UV/vis is comparable with the IR.

In addition to resampling (OS and US), we augment the data set using physics-based modeling. Specifically, we generate refractive index data using the Sellmeier equation[9]

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

where (A,B_j,C_j) are phenomenological coefficients. For each molecule in the original database that has refractive index measurements in the UV/visible, we fit the experimental curves for n^2 using eq 1 to obtain Sellmeier coefficients. These are then used to interpolate the index data in the region λ = [200,750] nm. Therefore, we exclude the near-infrared wavelengths from the fitting, despite the Sellmeier equation still being valid in this region for most organic materials. This data-augmentation procedure increases the number of records in the UV and VIS bins to about 3000 additional points per chemical compound. Figure 5 shows that the distribution of training records set after DA changed significantly with respect to RD. The augmented data set has about 400 000 records, with comparable numbers of records in all spectral bins. DA has also been used to resolve the imbalance problem in Raman-based classification problems.[28−32]

After addressing class imbalance in the raw data set, we tested the prediction performance of the random forest classifier on the different spectral bins. Figure 6 shows that the accuracies after US and OS improved significantly relative to the imbalanced data set, reaching 97% for UV and 99% for VIS regions (see also Table S4 in the SM). The accuracies after US and OS however did not change significantly in the Near-IR (2% improvement) and IR (0% improvement) bins. There is a decrease of 25% in accuracy for Far-IR, likely due to the reduction of data records on this region during undersampling. On the other hand, data augmentation (DA) significantly improved the performance of the classifier (99% for UV and 98% for VIS) without affecting the accuracies in the infrared bins (see Table S4 in the SM). The accuracies for FE1 and FE2 significantly decreased as expected. The reduced number of data records resulting from the grouping of the features dramatically affects the model performance in the less represented spectral bins.

The results in Figure 6 suggest that measuring the refractive index in the UV/vis region is in principle sufficient for a precise classification of chemical compounds. In comparison with the accuracies reached in the far-infrared spectrum ($\lambda > 5$ μm, fingerprint region), the classification accuracies in the visible region after data augmentation are equally good, without additional steps of hyperparameter optimization for the random forest model.

III.3. Estimating the Required Measurement Precision. The precision with which the refractive index is measured in experiments should affect the classifier performance. Compounds with similar refractive indices at a given wavelength in the visible would not be reliably distinguished if the index measurement does not have enough significant digits. We quantify this intuition by computing the classification accuracy achieved by the random forest model using training data whose precision was manually truncated to a finite number of significant figures. In Figure 7, we show the testing accuracies obtained using different decimal digits for $n(\lambda)$ in the training set, separating the results by preprocessing strategy. While the classification accuracies are in general poor (<82%) for single-wavelength index measurements with

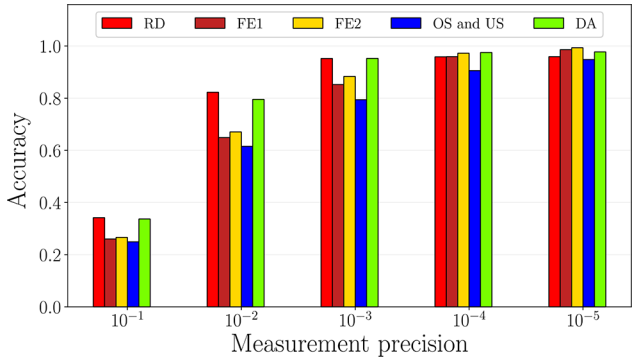


Figure 7. Testing accuracies for refractive index measurements with different precision. Results are shown for different preprocessing strategies: Featured engineering (FE1 and FE2), oversampling (OS), undersampling (US), and data augmentation (DA). The accuracies for the raw database (RD) are also shown for comparison.

https://doi.org/10.1021/acs.jpca.2c07955
J. Phys. Chem. A XXXX, XXX, XXX−XXX

Table 1. Prediction Accuracies Using Raman Spectral Databases[a]

Method	Spectral Database	Testing Accuracies (%)	Data Augmentation	Ref.
1D-CNN	Minerals and organic compounds	100	Standard augmentation transformations in vibrational spectroscopy	31
CRL	72 organic compounds	97.5	Gaussian noise and linear combination	28
DNN	72 organic compounds	92.6* 96.4**	Shifting, Gaussian noise, and interpolation	30
CNN	72 organic compounds	81.9* 86.0**	Shifting, Gaussian noise, and interpolation	30
DRCNN	72 organic compounds	98.1	Shifting and Gaussian noise	29
1D-CNN and KNN	620 mineral and 211 synthetic organic pigments	97.38	Shifting, Gaussian noise, Savitzky-Golay smoothing, spline interpolation, and polynomial reconstruction	32
RF	61 organic compounds	99 (UV) 98.1 (vis) 99.2 (Near-IR) 83.1 (IR) 94.8 (Far-IR)	Sellmeier equation fitting on UV/vis optical ranges	This work

[a]The number of molecules present in the database is specified when the data is available. We also list the data-augmentation methodologies used in these works. Model acronyms: 1D-CNN = 1D convolutional neural network, CRL = contrastive representation learning, DNN = deep neural network, CNN = convolutional neural network, DRCNN = deeply recursive convolutional neural network, KNN = K-nearest neighbor classifier, RF = random forest. *Without transfer learning, **with transfer learning.

two significant figures or less, the raw database already gives better classification accuracies (∼95%) when trained with index data of at least three decimal places. Increasing the number of significant figures beyond four decimals does not significantly improve the accuracy, regardless of the preprocessing strategy used.

These high testing accuracies for single-wavelength measurements are only possible because the trained random forest model has learned the molecular spectra over a broad range of wavelengths spanning hundreds of nanometers. We confirm this by training the same random forest model described above (see Methods Section) but with a reduced training set with only 89 refractive index records at 1000 nm (near IR). Only 42 compounds have measurements reported around this target wavelength. The testing accuracies obtained with such a restricted one-wavelength data set are very poor (30% or less, see Table S3 in the SM). This confirms the limited classification ability expected for conventional single-point index measurements. Our work extends this intuition by exploiting the correlations learned by the decision trees in the refractive index values at different wavelengths. Figure 7 shows that these correlations are more effectively learned when the index measurements have interferometric precision (4 digits or more).

III.4. Comparison with Raman Classifiers. Class imbalance has also been addressed in other molecular classification studies based on Raman spectral databases, which also tend to be heterogeneous with a data record distribution that over-represents compounds of particular interest in chemistry.[28−32] To address class imbalance, researchers have explored data-augmentation strategies such as peak shifting, noise addition, smoothing, spline interpolation, and polynomial reconstruction. These preprocessing strategies were implemented to reconstruct the data set before training deep learning classifier models. In our analysis of the refractiveindex.info database, the Sellmeier fitting procedure is a valid augmentation strategy that can be used to reconstruct the part of the training set corresponding to ultraviolet and visible wavelengths, without introducing data leakage.[9] In Table 1, we compare the refractive index approach with the recent Raman-based classifiers in the literature, showing that the dispersive method can also give high classification performance, using a similar a preprocessing strategy (interpolation) on data sets with a comparable volume of data records and molecular classes.[28−32]

IV. CONCLUSIONS

We have built a machine learning classification scheme to identify organic compounds based on refractive index measurements in the visible spectral region, in which most organic compounds are highly transparent. We trained a random forest classifier using decades of experimental data from the scientific literature. The database contains 194 816 spectral records of refractive index and extinction curves of 61 organic compounds and polymers over a broad range of wavelengths from the UV to the far-infrared. There is a class imbalance problem in the experimental data that restricts the classification accuracy for refractive index inputs at visible wavelengths (400−750 nm) to approximately 80%. This imbalance is primarily due to the disproportionate number of infrared absorption records reported in the mid- and far-infrared regions. Imbalance is a common problem when working with spectroscopic databases as the experimental data is deposited from different sources.[28−32]

We addressed this class imbalance issue by preprocessing the raw data before training the classifier using resampling and physics-based data-augmentation strategies analogous to those employed by other machine learning classifiers trained with Raman spectra.[28−32] By training the random forest model with preprocessed balanced data, we achieve molecular classification testing accuracies in the UV and visible regions better than 98%. Additional improvements can be expected with additional steps of model hyperparameter optimization. Such high accuracies are comparable to those obtained using only Raman spectroscopy databases (see Table 1), thus demonstrating the feasibility of using machine learning tools for enhancing the capabilities of laser-based chemical sensing devices. The high classification precision reached via data augmentation for chemical identification suggests that a similar technique would be applicable with other problems that involve continuous-variable data sets that can be interpolated using physics models. Further research is needed to generalize

https://doi.org/10.1021/acs.jpca.2c07955
J. Phys. Chem. A XXXX, XXX, XXX−XXX

the proposed molecular classifier to identify the structural and other chemical features of the molecules that are present in the Refractive Index Database.[52] Our work thus serves as a starting point for the development of remote chemical sensors based on laser light.

■ ASSOCIATED CONTENT

Data Availability Statement

The data that support the findings of this study are openly available at 10.5281/zenodo.6419970.[61]

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.2c07955.

List of organic compound classes in the training database, additional random forest classification tests, a visualization of the trained trees, and classification metrics using other machine learning models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Felipe Herrera − *Department of Physics, Universidad de Santiago de Chile, 3493 Santiago, Chile; Millennium Institute for Research in Optics*, https://www.miroptics.cl/eng/; orcid.org/0000-0001-8121-1931

Authors

Thulasi Bikku − *Department of Physics, Universidad de Santiago de Chile, 3493 Santiago, Chile; Computer Science and Engineering, Vignan's Nirula Institute of Technology and Science for Women, Guntur, Andhra Pradesh 522009, India*

Rubén A. Fritz − *Department of Physics, Universidad de Santiago de Chile, 3493 Santiago, Chile*

Yamil J. Colón − *Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States*; orcid.org/0000-0001-5316-9692

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.2c07955

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

R.A.F. is supported by DICYT-USACH grant POSTDOC USA1956_DICYT. F.H. and T.B. are supported by ANID through grants FONDECYT Regular No. 1181743 and Millennium Science Initiative Program ICN17_012. Y.J.C. thanks the University of Notre Dame for financial support through start-up funds.

■ ADDITIONAL NOTE

[1]We follow photonics notation and denote the extinction coefficient by k.

■ REFERENCES

(1) Doucet, M.; Archibald, R.; Heller, W. T. Machine Learning for Neutron Reflectometry Data Analysis of Two-Layer Thin Films. *Machine Learning: Science and Technology* **2021**, *2*, 035001.

(2) Ratner, D.; Sumpter, B.; Alexander, F.; Billings, J. J.; Coffee, R.; Cousineau, S.; Denes, P.; Doucet, M.; Foster, I.; Hexemer, A.; et al. *Office of Basic Energy Sciences (BES) Roundtable on Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning*, 2019. https://www.osti.gov/biblio/1630823/ (accessed November 15, 2022).

(3) Ballard, Z.; Brown, C.; Madni, A. M.; Ozcan, A. Machine Learning and Computation-Enabled Intelligent Sensor Design. *Nature Machine Intelligence* **2021**, *3* (7), 556−565.

(4) Wang, X.; Wolfbeis, O. S. Fiber-Optic Chemical Sensors and Biosensors (2015−2019. *Anal. Chem.* **2020**, *92* (1), 397−430.

(5) McDonagh, C.; Burke, C. S.; MacCraith, B. D. Optical Chemical Sensors. *Chem. Rev.* **2008**, *108* (2), 400−422.

(6) Gründler, P.. *Chemical Sensors: An Introduction for Scientists and Engineers*; Springer: Berlin, 2007.

(7) Swager, T. M.; Mirica, K. A. Introduction: Chemical Sensors. *Chem. Rev.* **2019**, *119* (1), 1−2.

(8) Mohan, S.; Kato, E.; Drennen, J. K.; Anderson, C. A. Refractive Index Measurement of Pharmaceutical Solids: A Review of Measurement Methods and Pharmaceutical Applications. *J. Pharm. Sci.* **2019**, *108* (11), 3478−3495.

(9) Boyd, R. W. *Nonlinear Optics*, 4th ed.; Academic Press: Amsterdam, 2020.

(10) Pereira, F. M. V.; de Souza Carvalho, A.; Cabeça, L. F.; Colnago, L. A. Classification of Intact Fresh Plums according to Sweetness Using Time-Domain Nuclear Magnetic Resonance and Chemometrics. *Microchemical Journal* **2013**, *108*, 14−17.

(11) Magwaza, L. S.; Opara, U. L. Analytical Methods for Determination of Sugars and Sweetness of Horticultural Products—a Review. *Scientia Horticulturae* **2015**, *184*, 179−192.

(12) Choudhary, V.; Rönnow, D. A Nondestructive Testing Method for the Determination of the Complex Refractive Index Using Ultra Wideband Radar in Industrial Applications. *Sensors* **2020**, *20*, 3161.

(13) Madden, M. G.; Howley, T. A Machine Learning Application for Classification of Chemical Spectra. *Applications and Innovations in Intelligent Systems XVI* **2009**, 77−90.

(14) Park, H.; Son, J.-H. Machine Learning Techniques for THz Imaging and Time-Domain Spectroscopy. *Sensors* **2021**, *21* (4), 1186.

(15) Cao, Y.; Huang, P.; Chen, J.; Ge, W.; Hou, D.; Zhang, G. Qualitative and Quantitative Detection of Liver Injury with Terahertz Time-Domain Spectroscopy. *Biomedical Optics Express* **2020**, *11* (2), 982.

(16) Madden, M. G.; Ryder, A. G. Machine Learning Methods for Quantitative Analysis of Raman Spectroscopy Data. *Proc. SPIE 4876, Opto-Ireland 2002: Optics and Photonics Technologies and Applications* **2003**.

(17) Zhang, L.; Li, C.; Peng, D.; Yi, X.; He, S.; Liu, F.; Zheng, X.; Huang, W. E.; Zhao, L.; Huang, X. Raman Spectroscopy and Machine Learning for the Classification of Breast Cancers. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2022**, *264*, 120300.

(18) Ryzhikova, E.; Ralbovsky, N. M.; Sikirzhytski, V.; Kazakov, O.; Halamkova, L.; Quinn, J.; Zimmerman, E. A.; Lednev, I. K. Raman Spectroscopy and Machine Learning for Biomedical Applications: Alzheimer's Disease Diagnosis Based on the Analysis of Cerebrospinal Fluid. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2021**, *248*, 119188.

(19) Cui, A.; Jiang, K.; Jiang, M.; Shang, L.; Zhu, L.; Hu, Z.; Xu, G.; Chu, J. Decoding Phases of Matter by Machine-Learning Raman Spectroscopy. *Physical Review Applied* **2019**, *12* (5), 054049.

(20) Guo, S.; Popp, J.; Bocklitz, T. Chemometric Analysis in Raman Spectroscopy from Experimental Design to Machine Learning−Based Modeling. *Nat. Protoc.* **2021**, *16* (12), 5426−5459.

(21) Martinez, J. C.; Guzmán-Sepúlveda, J. R.; Bolaños Evia, G. R.; Córdova, T.; Guzmán-Cabrera, R. Enhanced Quality Control in Pharmaceutical Applications by Combining Raman Spectroscopy and Machine Learning Techniques. *Int. J. Thermophys.* **2018**, *39* (6), 79.

(22) Berghian-Grosan, C.; Magdas, D. A. Application of Raman Spectroscopy and Machine Learning Algorithms for Fruit Distillates Discrimination. *Sci. Rep.* **2020**, *10* (1), 21152.

(23) Zhang, Y.; Cong, Q.; Xie, Y.; JingxiuYang; Zhao, B. Quantitative Analysis of Routine Chemical Constituents in Tobacco by Near-Infrared Spectroscopy and Support Vector Machine.

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **2008**, *71* (4), 1408−1413.
(24) Hu, W.; Ye, S.; Zhang, Y.; Li, T.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. Machine Learning Protocol for Surface-Enhanced Raman Spectroscopy. *J. Phys. Chem. Lett.* **2019**, *10* (20), 6026−6031.
(25) Ralbovsky, N. M.; Lednev, I. K. Towards Development of a Novel Universal Medical Diagnostic Method: Raman Spectroscopy and Machine Learning. *Chem. Soc. Rev.* **2020**, *49* (20), 7428−7453.
(26) Guevara, E.; Torres-Galván, J. C.; Ramírez-Elías, M. G.; Luevano-Contreras, C.; González, F. J. Use of Raman Spectroscopy to Screen Diabetes Mellitus with Machine Learning Tools. *Biomedical Optics Express* **2018**, *9* (10), 4998−5010.
(27) Zou, T.; Dou, Y.; Mi, H.; Zou, J.; Ren, Y. Support Vector Regression for Determination of Component of Compound Oxytetracycline Powder on Near-Infrared Spectroscopy. *Anal. Biochem.* **2006**, *355* (1), 1−7.
(28) Li, B.; Schmidt, M. N.; Alstrøm, T. S. Raman Spectrum Matching with Contrastive Representation Learning. *Analyst* **2022**, *147*, 2238.
(29) Zhou, W.; Tang, Y.; Qian, Z.; Wang, J.; Guo, H. Deeply-Recursive Convolutional Neural Network for Raman Spectra Identification. *RSC Adv.* **2022**, *12* (8), 5053−5061.
(30) Zhang, R.; Xie, H.; Cai, S.; Hu, Y.; Liu, G.; Hong, W.; Tian, Z. Transfer-Learning-Based Raman Spectra Identification. *J. Raman Spectrosc.* **2020**, *51* (1), 176−186.
(31) Hamed Mozaffari, M.; Tay, L.-L. Overfitting One-Dimensional Convolutional Neural Networks for Raman Spectra Identification. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2022**, *272*, 120961.
(32) Shang, L.; Bao, Y.; Tang, J.; Ma, D.; Fu, J.; Zhao, Y.; Wang, X.; Yin, J. A Novel Polynomial Reconstruction Algorithm-Based 1D Convolutional Neural Network Used for Transfer Learning in Raman Spectroscopy Application. *J. Raman Spectrosc.* **2022**, *53* (2), 237−246.
(33) López-López, M.; García-Ruiz, C. Infrared and Raman Spectroscopy Techniques Applied to Identification of Explosives. *TrAC Trends in Analytical Chemistry* **2014**, *54*, 36−44.
(34) Baker, M. J.; Byrne, H. J.; Chalmers, J.; Gardner, P.; Goodacre, R.; Henderson, A.; Kazarian, S. G.; Martin, F. L.; Moger, J.; Stone, N.; et al. Correction: Clinical Applications of Infrared and Raman Spectroscopy: State of Play and Future Challenges. *Analyst* **2018**, *143* (8), 1934−1934.
(35) Tatsch, E.; Schrader, B. Near-Infrared Fourier Transform Raman Spectroscopy of Indigoids. *J. Raman Spectrosc.* **1995**, *26* (6), 467−473.
(36) Liu, C.; Xu, L.; He, H.; Jia, W.; Hua, Z. Discrimination of Phenethylamine Regioisomers and Structural Analogues by Raman Spectroscopy. *Journal of Forensic Sciences* **2021**, *66* (1), 365−374.
(37) Park, H.; Son, J.-H. Machine Learning Techniques for THz Imaging and Time-Domain Spectroscopy. *Sensors* **2021**, *21* (4), 1186.
(38) Katritzky, A. R.; Sild, S.; Karelson, M. Correlation and Prediction of the Refractive Indices of Polymers by QSPR. *J. Chem. Inf. Comput. Sci* **1998**, *38* (6), 1171−1176.
(39) Duchowicz, P. R.; Fioressi, S. E.; Bacelo, D. E.; Saavedra, L. M.; Toropova, A. P.; Toropov, A. A. QSPR Studies on Refractive Indices of Structurally Heterogeneous Polymers. *Chemometrics and Intelligent Laboratory Systems* **2015**, *140*, 86−91.
(40) Bouteloup, R.; Mathieu, D. Improved Model for the Refractive Index: Application to Potential Components of Ambient Aerosol. *Phys. Chem. Chem. Phys.* **2018**, *20* (34), 22017−22026.
(41) Erickson, M. E.; Ngongang, M.; Rasulev, B. A Refractive Index Study of a Diverse Set of Polymeric Materials by QSPR with Quantum-Chemical and Additive Descriptors. *Molecules* **2020**, *25* (17), 3772.
(42) Khan, P. M.; Rasulev, B.; Roy, K. QSPR Modeling of the Refractive Index for Diverse Polymers Using 2D Descriptors. *ACS Omega* **2018**, *3* (10), 13374−13386.
(43) Astray, G.; Cid, A.; Moldes, O.; Ferreiro-Lage, J. A.; Gálvez, J. F.; Mejuto, J. C. Prediction of Refractive Index of Polymers Using Artificial Neural Networks. *Journal of Chemical & Engineering Data* **2010**, *55* (11), 5388−5393.
(44) Liu, J.; Ueda, M. High Refractive Index Polymers: Fundamental Research and Practical Applications. *J. Mater. Chem.* **2009**, *19* (47), 8907.
(45) Redmond, H.; Thompson, J. E. Evaluation of a Quantitative Structure−Property Relationship (QSPR) for Predicting Mid-Visible Refractive Index of Secondary Organic Aerosol (SOA. *Phys. Chem. Chem. Phys.* **2011**, *13* (15), 6872.
(46) Chen, R.-C.; Dewi, C.; Huang, S.-W.; Caraka, R. E. Selecting Critical Features for Data Classification Based on Machine Learning Methods. *Journal of Big Data* **2020**, *7* (1), 52.
(47) Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-Guzik, A. Materials Acceleration Platforms: On the Way to Autonomous Experimentation. *Current Opinion in Green and Sustainable Chemistry* **2020**, *25*, 100370.
(48) Cai, W.; Abudurusuli, A.; Xie, C.; Tikhonov, E.; Li, J.; Pan, S.; Yang, Z. Toward the Rational Design of Mid-Infrared Nonlinear Optical Materials with Targeted Properties via a Multi-Level Data-Driven Approach. *Adv. Funct. Mater.* **2022**, *32*, 2200231.
(49) Stein, H. S.; Gregoire, J. M. Progress and Prospects for Accelerating Materials Science with Automated and Autonomous Workflows. *Chemical Science* **2019**, *10* (42), 9640−9649.
(50) Rocha, F. S.; Gomes, A. J.; Lunardi, C. N.; Kaliaguine, S.; Patience, G. S. Experimental Methods in Chemical Engineering: Ultraviolet Visible Spectroscopy-UV-Vis. *Canadian Journal of Chemical Engineering* **2018**, *96* (12), 2512−2517.
(51) Wade, J. H.; Bailey, R. C. Refractive Index-Based Detection of Gradient Elution Liquid Chromatography Using Chip-Integrated Microring Resonator Arrays. *Anal. Chem.* **2014**, *86* (1), 913−919.
(52) RefractiveIndex.INFO. Refractive index database. https://refractiveindex.info (accessed January 15, 2023).
(53) Fernández, A.; García, S.; Galar, M.; Prati, R. C.; Krawczyk, B.; Herrera, F. *Learning from Imbalanced Data Sets*; Springer Science +Business Media: New York, 2018.
(54) Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature Selection in Machine Learning: A New Perspective. *Neurocomputing* **2018**, *300*, 70−79.
(55) More, A. S.; Rana, D. P. Review of random forest classification techniques to resolve data imbalance. *Proceedings - 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017*, Aurangabad, India, 2017; pp 72−78. .
(56) Parmar, A.; Katariya, R.; Patel, V. A Review on Random Forest: An Ensemble Classifier. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018* **2019**, *26*, 758−763.
(57) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.
(58) Lustgarten, J. L.; Gopalakrishnan, V.; Grover, H.; Visweswaran, S. Improving Classification Performance with Discretization on Biomedical Datasets. *AMIA Annu Symp Proc.* **2008**, *2008*, 445−449.
(59) NASA ipac. *Near, Mid and Far-Infrared*. https://archive.ph/20120529003352/http://www.ipac.caltech.edu/Outreach/Edu/Regions/irregions.html (accessed June 15, 2022).
(60) Ali, A.; Shamsuddin, S. M.; Ralescu, A. L. Classification with Class Imbalance Problem: A Review. *Int. J. Adv. Soft Comput. Its Appl.* **2015**, *7*, 176.
(61) Bikku, T.; Fritz, R. A.; Colón, Y.; Herrera, F. *fherreralab/organic_optical_classifier: v1.0.0*, 2022. DOI: 10.5281/zenodo.6419970 (accessed November 20, 2022).

https://doi.org/10.1021/acs.jpca.2c07955
J. Phys. Chem. A XXXX, XXX, XXX−XXX