applied sciences

MDPI

Review

Qualitative and Quantitative Analysis of Volatile Molecular Biomarkers in Breath Using THz-IR Spectroscopy and Machine Learning

Akim Tretyakov [1], Denis Vrazhnov [1,2], Alexander Shkurinov [1,3], Viacheslav Zasedatel [1] and Yury Kistenev [1,2,*]

[1] Laboratory of the Molecular Imaging and Machine Learning, National Research Tomsk State University, Tomsk 634050, Russia; dr.akim1998@yandex.ru (A.T.); vda@mail.tsu.ru (D.V.); ashkurinov@physics.msu.ru (A.S.); zevs@ido.tsu.ru (V.Z.)
[2] Institute of Atmospheric Optics, Tomsk 634055, Russia
[3] Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia
* Correspondence: yuk@iao.ru; Tel.: +7-9138286720; Fax: +7-3822529895

Abstract: Exhaled air contains volatile molecular compounds of endogenous origin, being products of current metabolic pathways. It can be used for medical express diagnostics through control of these compounds in the patient's breath using molecular absorption spectroscopy. The fundamental problem in this field is that the composition of exhaled air or other gas mixtures of natural origin is unknown, and content analysis of such spectra by conventional iterative methods is unpredictable. Machine learning methods enable the establishment of latent dependencies in spectral data and the conducting of their qualitative and quantitative analysis. This review is devoted to the most effective machine learning methods of exhaled air sample absorption spectra qualitative and content analysis. The focus is on interpretable machine learning methods, which are important for reliable medical diagnosis. Also, the steps additional to the standard machine learning pipeline and important for medical decision support are discussed.

Keywords: IR spectroscopy; THz spectroscopy; machine learning; gas mixture analysis; qualitative analysis; quantitative analysis; exhaled air; medical diagnostics

Citation: Tretyakov, A.; Vrazhnov, D.; Shkurinov, A.; Zasedatel, V.; Kistenev, Y. Qualitative and Quantitative Analysis of Volatile Molecular Biomarkers in Breath Using THz-IR Spectroscopy and Machine Learning. *Appl. Sci.* **2024**, *14*, 11521. https://doi.org/10.3390/app142411521

Academic Editor: Herbert Schneckenburger

Received: 10 October 2024
Revised: 22 November 2024
Accepted: 28 November 2024
Published: 11 December 2024

1. Introduction

Exhaled air contains a plethora of volatile compounds of endogenous origin, being products of various metabolic pathways. Examination of volatile molecular biomarkers (VMBs) in the patient's breath is a promising way of developing new noninvasive medical screening tests [1,2]. This approach is a kind of "omics" test (genomics, metabolomics, proteomics, etc.), often referred to as "breathomics" [3,4]. Ideally, in the future, exhaled air analysis can become a noninvasive analog of laboratory blood tests with detection of the level of specific VMBs, and medical decisions can be based on analysis of the profile of such VMBs' concentrations. This approach is associated with qualitative and quantitative analysis of gas mixtures. Qualitative analysis means determining the presence or absence of a specific molecular substance in a gas sample or making decisions about the health state of a person from whom this sample was taken. Quantitative analysis includes a volatile substance concentration evaluation.

The qualitative and quantitative analysis of gas mixtures can be implemented experimentally by methods of gas chromatography, absorption spectroscopy, and chemical sensor arrays (e-nose) [5]. Absorption spectroscopy is simple to use and has low-cost operation, high sensitivity, and selectivity. In breathomics, IR and THz absorption spectroscopy are usually used [6–9] because IR spectra represent information about vibrational absorption bands of volatile molecules, while THz spectra describe rotational absorption bands of

polar molecules [10,11]. From the experimental techniques point of view, two options can be highlighted: terahertz (THz) time-domain spectroscopy (THz-TDS) and THz frequency-domain spectroscopy (THz-FDS). These classes can be referred to as time-resolved THz spectroscopy and continuous wave THz spectroscopy, respectively [12,13]. THz-TDS is characterized by a high-speed data acquisition rate (with femtoseconds-long pulses) but with a relatively low spectra resolution. On the other hand, THz-FDS, as a rule, can achieve a high resolution (10^{-3} Hz and higher) but requires more time to measure a wide frequency range spectrum [14,15]. To improve the sensitivity of THz quartz-based spectroscopy, THz quartz-based spectroscopy devices were suggested [12,13,16]. This variant involves the THz quartz tuning fork-based light-induced thermoelastic spectroscopy and quartz-enhanced photoacoustic spectroscopy [8]. These techniques are characterized by an enormously high resolution and sensitivity. But experimental implementation of quartz-enhanced photoacoustic spectroscopy requires achieving several mechanical resonances: a very sharp resonance in a quartz fork and, in addition, tube-shape acoustic resonators [17], which is a nontrivial task due to the influence on them of a studied gas sample pressure, temperature, and composition.

Typical absorption spectral data (spectra) are a plot of the dependence of a studied sample's absorption intensity on the wavelength of light. In the absence of noise and with sufficient resolution of the spectroscopic device, each biomarker has a unique spectrum. However, in reality, the spectra of different substances have overlapping regions with close values that are hard to distinguish, which causes difficulties for IR and THz absorption spectra analysis. From a mathematical point of view, a multicomponent gas mixture absorption spectrum $S(\lambda)$ can be represented as a linear combination of the absorption spectra $s_i(\lambda)$ of the individual molecular substances:

$$S(\lambda) = \sum_i c_i s_i(\lambda), \tag{1}$$

where c_i is the i-th substance concentration, and λ is wavelength. This presentation allows conducting the qualitative ($c_i = 0, 1$) and the quantitative ($c_i \in [0, c_{max}]$) analysis. The problem of finding unknown c_i can be formulated in a matrix form. Let us introduce a matrix S combining absorption spectra of m known substances measured at n wavelengths:

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nm} \end{pmatrix}.$$

Let $c = (c_1, \ldots, c_m)^T$—the vector of unknown concentrations, $b = (b_1, \ldots, b_m)^T$—the measured spectrum of gas composition of known substances with unknown concentrations. This implies that concentrations can be found by solving the system of linear algebraic equations in the matrix form:

$$Sc = b. \tag{2}$$

Iterative algorithms like Multivariate Curve Resolution (MCR) [18–20] and Univariate Calibration [21] combined with least squares [22] and Levenberg–Marquardt methods of extreme search [23–25] were the first methods applied for Equation (2) solution. All these techniques are based on a priori knowledge of all $s_i(\lambda)$ relevant to a studied gas sample. The major problem of such approaches in content analysis of exhaled air or other gas mixtures of natural origin spectra is that their composition is unknown. In this case, $S = \widetilde{S} + S_L$, and $c = \hat{c} + c_L$, where $\widetilde{S}$—the matrix of individual spectra of components, which are definitely present in the studied gas mixture; S_L—the matrix of unaccounted (latent) components, $\hat{c}$—vector of unknown concentrations of accounted components, and c_L—the vector of corresponding concentrations of unaccounted components. Here, $\hat{c} \equiv (c_1, \ldots, c_m)^T$, a symbol with tilde means known data. The experimentally measured matrix b contains

random additive part R associated with noise, measurement error, etc. Therefore, $b = \widetilde{b} + R$. Then Equation (2) takes form:

$$\widetilde{S}\hat{c} = \widetilde{b} + R - S_L(\hat{c} + c_L) - \widetilde{S}c_L \quad (3)$$

Opposite to Equation (2), the system (3) is ill-posed. There are two reasons for the latter: the presence of random noise and unaccounted (latent) components.

Regarding random noise, according to the criterion proposed by J. Adamar, the problem is well-posed if three conditions are satisfied: the solution exists, is unique, and depends continuously on the initial conditions. Otherwise, the problem is ill-posed. In the case of Equation (3), a solution exists, but other conditions are not met; such a task requires using special methods. The random noise R can be reduced by preliminary noise filtration. When continuous dependence of the solution on the initial conditions is absent, regularization methods and iterative procedures of $\hat{c}$ calculation can be used [26].

The problem of the presence of unaccounted (latent) components is related to so-called "grey analytical systems" for which their qualitative chemical composition is incomplete [27]. Regression methods like principal component analysis based on the transformation of spectral responses into orthogonal latent variables (principal components) allow estimating the number of components in a studied spectrum of a gas mixture. It can be completed by estimating the described variance in reduced feature space or using Kaiser's rule [28]. However, there is no strict rule for the described variance threshold selection; Kaiser's rule works well when there are several principal components with eigenvalues being much higher than their average value and the remaining principal components have lesser eigenvalues. Another drawback is that the achieved solution of Equation (1) in principal component space of less dimensionality compared to the initial one is not unique.

The task of qualitative and quantitative analysis in molecular absorption spectroscopy is closely related to the machine learning (ML) field [29]. The qualitative and quantitative absorption spectra analysis based on ML corresponds to classification and regression problems [30]. The difference is that the ML algorithm predicts the value of a logical (presence/absence of specific molecular components in the mixture) [29,31,32] or a real variable (like c_i) [33,34]. The ML algorithm training step is obligatory for both classification and regression tasks. Validation is an estimation of the created data model's efficiency on data, which was not used in the training step.

The useful but often optional step in the ML pipeline is informative feature selection. In the context of absorption spectra analysis, this step means finding more narrow spectral ranges (spectral features) in the experimental data, which keep being relevant to a studied sample. It is equivalent to removing abundant data on some frequencies, thus lowering the dimensionality of the spectral feature space. Here, a spectral feature is any spectral parameter directly or indirectly associated with the presence of a specific molecular component and its concentration, i.e., spectrum intensity value on a definite frequency. The feature selection/extraction depends on a concrete experimental dataset. There are no universal recipes for conducting this step. As a rule, the properties of ML algorithms used for data model creation are known well, and this step comes down to a choice of suitable method(s).

A typical ML pipeline is shown in Figure 1, where two workable solutions for feature extraction are presented. A pattern recognition approach is a mathematical technique based on the analysis of a set of features, which are not related directly to concentrations of specific VMBs. A chemical analytics-based approach implies the usage of spectroscopic information in the form of computed concentration profiles as the basis of a feature set for further ML analysis.

The key characteristic of classification/regression ML data models is the possibility or impossibility of direct interpretability [35]. The latter corresponds to a "black box" ML method [36–39]. Interpretability can be associated with the following properties of an ML algorithm: (a) explicit way of informative features extraction, (b) explicit rules of decision rules, and (c) explicit dependencies between input data parameters and outputs.

Interpretable ML methods are preferable in relation to the "black box" ones, especially in making medical decisions using ML.

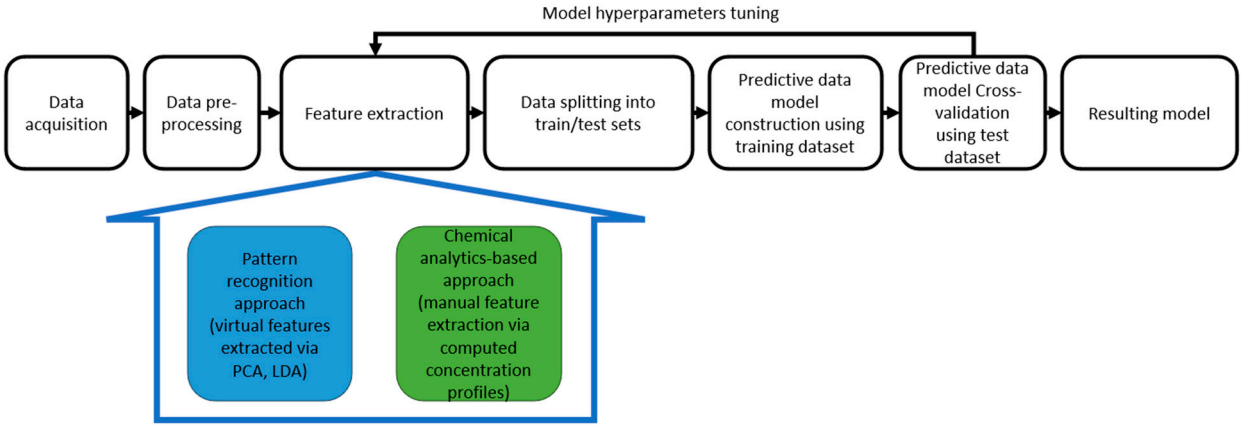

Figure 1. ML pipeline suitable for gas mixtures for absorption spectra analysis.

There are novel and actual review articles for the last four years on THz and infrared (IR) gas phase spectra analysis using ML methods, such as [36,40–44]. With respect to these articles, the presented review includes both THz and IR spectroscopy methods combined with ML methods applications in VMBs spectral data processing and analysis.

The main purpose of the review is to analyze the most suitable approaches in quantitative and qualitative analysis of THz and IR spectra of exhaled air samples using ML methods. At the stage of classification, the focus is on interpretable ML methods. To make a correct medical decision using indirect characteristics of a person's condition, it is very important to have the possibility of clearly understanding reasons, for instance, of classifying a patient as healthy or unhealthy depending on the exhaled air spectra composition. Due to this, not only points for improving a predictive model can be found, but target molecules and their combinations for patients with a certain disease can be identified, or steps for a making decisions process that leads to wrong predictions can be highlighted and fixed [45]. This is a reason to pay less attention in this review to artificial neural networks (ANNs), which lack interpretability, though this field is currently under development [45–47]. Moreover, the study of biological samples involves the complexity of sampling, so ANNs are limited to training datasets of hundreds of examples at best. Such cases require the use of extra sampling techniques like bootstrapping [48]. No doubt, further development of ANNs will provide a powerful tool for such tasks.

For the review, the Google Scholar citation database and the following keywords were used: IR spectroscopy, THz spectroscopy, time-domain spectroscopy, ML methods, gas mixture analysis, qualitative analysis, and quantitative analysis. The period was 2019–2024, and the initial search provided 1080 results. After removing irrelevant and repetitive results, 60 articles remained for analysis.

2. Peculiarities of Implemented ML Pipelines for IR and THz Absorption Spectra Analysis

The main parameters of the IR spectroscopy experimental technique used in breathomics applications and data description, including analyzed volatile molecules, are shown in Table S1 in Supplementary Materials. The ML methods used in the papers are described in Table S1, and the results of their application are shown in Table S2 in Supplementary Materials. The main parameters of the THz spectroscopy experimental technique used in breathomics applications and data description, including analyzed volatile molecules, are shown in Table S3 in Supplementary Materials. The ML methods used for THz spectra analysis and the results of their application are shown in Table S4 in Supplementary Materials.

2.1. Data Pre-Processing

Usually, this step includes noise filtration, normalization, baseline correction, and optimizing the spectral range [49–53]. In general, higher spectral resolution provides better identification of molecular substances in a studied sample. A quite nonstandard absorption spectroscopy step in the ML pipeline associated with spectral resolution enhancement was implemented in [54,55]. In our work [55], the multilayer perceptron artificial network and convolution network models were designed and applied for spectral resolution, improving noisy IR absorption spectra of gas mixtures of eight molecules (C_2H_2, CS, CO, HI, HCl, H_2O, NH_3, O_3). The spectral data were generated using the 2020 HITRAN database [56,57]. As a result, spectral resolution was improved from 5 cm^{-1} to 1 cm^{-1}. The example of decomposition of this gas mixture with concentrations (0.25; 0.00; 0.00; 4.00; 0.00; 0.00; 0.15; 0.20) · 10^5 ppm (the order of concentrations corresponds to the order of molecules presented above) is presented in Figure 2. Here, decomposition was carried out for the high-resolution spectrum, noisy low-resolution spectrum, and the last spectrum with restored spectral resolution.

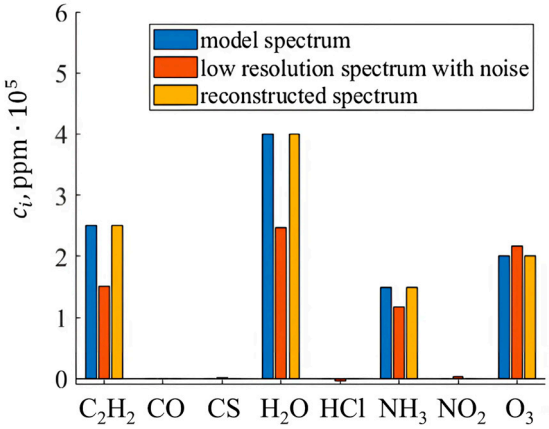

Figure 2. The model spectra decomposition using canonical correlation analysis. Comparison of method's accuracy for original model spectrum, original spectrum with noise and artificially reduced resolution, and the latter, after denoising and high-resolution reconstruction procedures [55]. (Reused under license # 5882601337718).

2.2. Feature Selection and Extraction

According to Table S2, S4 PCA, t-SNE, LDA, and ICA are the most frequently used methods of informative feature extraction. The presented list of methods is supported by other reviews in the field of gas phase THz and IR spectroscopy [40,42–44]. Their characterization is presented in Table 1. It should be pointed out that LDA and ICA have quite strong requirements regarding the statistical properties of analyzed data. Artificial neural networks hold a special place because they, as a rule, simultaneously implement feature extraction and data modeling.

Feature selection is an extremely important step from a practical medicine point of view because it makes the data model more interpretable, but it is used quite rarely. The reason is that there are no formal rules for the informative feature selection, with the exception of an individual feature importance analysis. The latter can be conducted through a model-dependent technique at the stage of data model efficiency estimation by removing an individual feature from the data vector and evaluating the reduced data model efficiency variation. The negligibility of this variation means that this feature is not important. Such a procedure combined with a suitable feature extraction method can increase the efficiency of the latter. In the data-independent technique, latent relations among features are established, for example, by feature correlation analysis. A partial least squares–discriminant analysis (PLS-DA) [65] algorithm can be considered as an alternative

to t-SNE because it can operate with high-dimensional multi-collinear data and reapplied to unseen spectra.

Table 1. Feature selection and extraction methods.

Method	Idea	Advantages	Disadvantages	Ref.
PCA	Data are transformed into a new system of coordinates with the first axis (the first principal component) oriented along a direction of the data maximal dispersion. If necessary, the second and other axes (the second principal component, etc.), are oriented perpendicularly to the previous axis.	PCA reduces effectively the data dimension. The most important original features can be easily established.	PCA is not suitable for the case of data spatial distribution similar to a nonlinear curve (nonlinear manifold). This problem can be overcome using kernel PCA. There are no universal and effective rules for quantity of principal components choice.	[8,58–61]
LDA	Data are transformed into a new system of coordinates of less dimension being linear combinations of original features that should provide maximal between-class and minimal within-class variances of the data. It allows reducing data dimensionality.	This method is justified statistically and addresses effectively an original data multicollinearity problem.	LDA uses strong assumption of multivariate normal distribution of the data, equal covariance matrices for each class, linear separability of the data. This method is not suitable for unlabeled data.	[62]
t-SNE	This method simulates data points in high dimension space by data points in a low dimension space (usually 2D, 3D spaces) in a way that keeps mutual spatial positions of data points.	t-SNE preserves the initial data structure. When the output data presentation corresponds to 2D or 3D space, it allows one to see explicitly the spatial structure of the data.	t-SNE has many tunable parameters that affect performance.	[63]
Independent Component Analysis (ICA)	Effectively uses assumption of statistical independence of the spectral characteristics of the compounds and requires at least one of them was Gaussian.	ICA is effective in removing noise component in a spectrum.	ICA needs to meet very strict conditions for its application.	[64]

2.3. Classification

In the case of qualitative IR gas spectra analysis, the authors of the studied articles report successful ML methods application for a target component presence or gas mixture constituent identification. For instance, there is research devoted to mixture composition estimation relying on 15 functional molecular groups spectra [33]. According to these spectra, it is possible to define whether a molecule belongs to one of the studied functional groups.

It can be highlighted that in the set of articles containing ML methods, the comparative analysis as well as data model architecture variations investigation in order to determine an optimal resulting data classification model is not large [59–61]. The characterization of the most often used ML methods presented in Tables S3 and S5 regarding the interpretability of their results is presented in Table 2. The main outline derived from Table 2 is the fact that, generally, only linear and simple models are interpretable. The complex models require the external tools mentioned above for their prediction results explanation.

Table 2. Classification methods.

Method	Idea	Advantages	Disadvantages	Interpretability	Ref.
ELM-AE	ELM-AE is a variant of single-hidden-layer autoencoder. The ELM-AE is based on construction of the set of three input data representations: (1) decoding–the input feature space is transformed to the low-dimensional feature space; (2) encoding–this low-dimensional feature space is equivalently mapped to the high-dimensional feature space; (3) autoencoding–the input features are equivalently mapped from the original feature space to the equal-dimensional feature space.	Very fast training.	Finding optimal weights and biases for the hidden layer is a nontrivial problem.	"Black box" model.	[66,67]
ANN	Consists of simple computational units (neurons) connected by tunable weights and arranged in layers to learn patterns from input data and provide output responses.	ANNs can distinguish classes, which are characterized by slightly different shapes of spectra.	ANNs are prone to overfitting. It is hard to find the optimal configuration of layers and weights. ANNs require a lot of annotated data for training.	"Black box" model.	[37,68]
CNN	Same as ANN, but 2D convolution masks parameters are optimized instead of weights for neurons outputs.	Excellent performance in image analysis.	Same as for ANN.	"Black box" model.	[58]
SVM	SVM is based on the finding the maximum margin hyperplane, which separates two classes. The support vectors are data points, which are placed on boundaries of this hyperplane. In kernel SVM, data are preliminarily projected into higher dimensional space, where classes are linearly separable.	Good generalization ability.	It is hard to separate slightly different spectra. SVM fails when the number of frequencies is much larger the volume of the dataset.	Linear model can be interpreted. "Black box" model for non-linear kernels.	[69–71]
RF	RF is a combination of decision tree classifiers trained individually. Their joint use increases accuracy of classification. Initial set of spectra can be divided into two subsets, based on computed threshold value for the frequency. A sequence of such splits provides separation of initial data into classes.	The ability to process data with many features, robustness to noise and overfitting. RF can handle missing data.	Many hyperparameters are necessary to optimize. The training and prediction times can be significant.	RF constructs "white model", allowing to estimate the importance of each feature.	[72]

Table 2. *Cont.*

Method	Idea	Advantages	Disadvantages	Interpretability	Ref.
GB	A sequence of decision tree models, where each successor is focused on correcting prediction errors of predecessor.	High model performance, fast learning.	Problems with imbalanced datasets. High number of hyperparameters to tune up. Regularization does not always prevent overfitting.	GB except for some versions can be interpreted.	[73]
K-NN	Formally, the basis of the method is the compactness hypothesis: if the metric of the distance in a feature space is introduced successfully, then neighbor samples are much more often in the same class than in different ones.	Simple model, can be used with different similarity metrics.	The number of nearest neighbors (k) is chosen beforehand and can affect the accuracy of classification. Low performance, dependence on the similarity metrics.	K-NN model can be interpreted.	[74–76]
Soft independent modeling by class analogy (SIMCA)	Based on PCA decomposition of the spectral data. Only data with high explained variance value remained.	Can be used with different similarity metrics, robust to noise.	Same as for PCA.	Like PCA can be interpreted.	[77,78]

2.4. Regression

A summary of regression methods applications in gas sample substances concentration estimation is given in Table 3 [61,79–83]. These methods are mostly linear and do not always show good model performance on complex multivariate data, which include IR and THz spectra. A viable option is multivariate methods with regularization, for example, LASSO (least absolute shrinkage and selection operator), which also provides a selection of variables to remove redundant and noisy data by setting regression coefficients in the model to zero [84,85].

Table 3. Regression methods.

Method	Idea	Advantages	Disadvantages	References
LRA	Output variable approximated as linear combination of input ones (predictors). LRA is an extension of regression analysis for the case of categorical outcome variables.	Fast learning, interpretable model.	Suitable for linear data only. Sensitive to outliers, noise and prone to overfitting.	[79]
PCR	This method uses PCA to transform data into PC space and after that a linear regression is constructed.	Model can be interpreted by loadings matrix analysis. The only one tunable parameter.	It is hard to choose the appropriate value of PCs. The limitations are the same as for PCA.	[61]
PLSR	This method projects predictors and output variables into a new latent space and after that uses LRA.	Simple, interpretable, fast learning.	There is a risk of misinterpretation. Sensitive to the scaling of predictors.	[80]
SVR	Aims to find a hyperplane, which fits the data points while minimizing margin violations, capturing patterns and relationships in the data by passing through as many data points as possible within a specified margin.	Good generalization ability. Linear model can be interpreted.	Separates slightly differing spectra. Provides fair results when number of initial variables is much greater than number of spectra.	[81–83]

The problem of unaccounted components can be solved by methods, which allow to analyze only a target component independently on rest. An example is the Reducing Spectrum Complexity (RSC) method developed by us [8]. RSC uses the fact that a spectrum shape complexity reduces when the target component contribution is removed from a gas mixture absorption spectrum entirely. The same property has Multivariate Curve Resolution with the Addition Method (MCRAD) [86]. MCRAD is based on artificially extending the experimental dataset by numerically adding the target component spectrum with definite concentrations to the experimental spectrum of this component and other unknown components in a mixture.

3. The Optimal ML Pipeline for Classification/Regression Data Model Creation

According to the analysis detailed in the previous section, the optimal ML pipeline for the creation of a classification/regression data model based on IR and THz exhaled air samples absorption spectra can combine (Figure 3):

1. t-SNE for the informative features visual analysis [8,37,58,63,87]. Thus, t-SNE is a proven, reliable tool for exploratory analysis because it preserves the relative distance between samples after projection.
2. PCA or/and PLS for the informative features extraction [8,58,59,61,87,88]. The versatility of the PCA application (including loadings analysis and extra restrictions like non-negativeness of spectra), in addition to the strong mathematical and physical meaning of the results, allows us to conclude that PCA is extremely effective in qualitative and quantitative spectroscopic studies. PLS can be a good replacement for PCA in some cases (and it can produce very similar results), but it is difficult to determine in which case which method is preferable to use.
3. Regression methods for a gas sample substances concentration estimation do not require a priori knowledge of the gas sample composition. This task cannot be solved by direct iterative methods like HAMAND, based on the iteration procedure of a system of linear equations and the additions method [89]. Thus, only the application of an ML-based regression method can give a positive result.
4. Linear SVM, k-NN, RF, or GB are at the stage of creating a prediction data model as methods that allow interpreting the results of classification. These methods are simple, known for good generalization ability even on small training datasets, and interpretable. Some recent ongoing research allows the inclusion of CNN in this list with activation feature maps for interpretability [34].

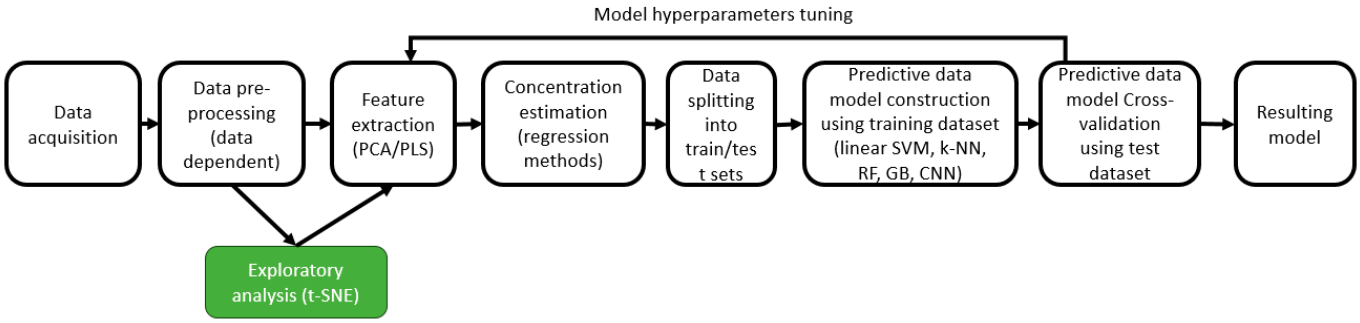

Figure 3. Suggested optimal ML pipeline.

4. Conclusions and Future Steps

The pattern recognition approach requires the extraction of independent informative features, which must provide the best sensitivity and selectivity of a created prediction data model to make accurate medical decisions. Implementation of a pattern recognition approach in breathomics using ML methods does not require solving Equation (3) directly. Instead, an ML algorithm is trained to distinguish spectral patterns of various classes.

In the case of IR and THz absorption spectroscopy, these patterns are absorption spectra profiles of exhaled air samples.

Convolutional neural networks are a golden standard for pattern recognition [34,62]. The advantage of such an approach is that it does not rely on the pure substances' spectra linear superposition model, which is why, in the most interesting experimental spectra processing case, the intermolecular interactions of nonlinear elements of different molecule types are considered. Some recent works show outstanding results in joint analysis using CNN and specially designed interpretability methods like Grad-CAM [34]. Grad-CAM can provide results that are similar to Local Interpretable Model-Agnostic Explanations (LIME) and the Shapley Additive explanations (SHAP) technique in a contrastive metric sense [90] but underperforms when additional layers are placed between the last convolution and the classification layer. Other shortcomings are the drop in localization efficiency with multiple object instances and the poor ability to capture the whole object. These limitations can be overcome by the application of Grad-CAM++ and Integrated Gradients [91,92]. Even so, a full understanding of the extremely complicated deep neural network decision-making process is far from being perfect [93,94].

Often, the application of supervised ML methods requires large-volume training sets, especially deep neural networks. A way to obtain them without using expensive spectrometers is to generate model spectra using spectral databases (NIST, HITRAN, GEISHA, etc.). However, using synthetic spectra for training and validation data sets leads to data model prediction accuracy decreasing when it is tested on new experimental data. Hybrid datasets combine both advantages of the means of data collection: the simplicity of obtaining synthetic spectra and considering the set of physical effects. Therefore, ML methods offer a means of interaction between synthetic and experimental data [61].

The chemical-based approach differs only by using the VMB concentrations profile as an informative features vector (see Figure 1). The step is related to quantitative analysis. The chemical-based approach is based on Equation (3)'s solution [95]. In the case of using conventional iterative methods (like MCR), the objective function is minimized by coefficients c_i prediction variation (by movement along an objective function surface). The exhaled air sample always belongs to "grey analytical systems" [27]. This task can be solved by using supervised ML methods like DNNs [95] and methods like HAMAND and RSC [8,89]. With respect to supervised ML methods, a similar process of optimal c_i values estimation takes place, but with the help of similarity metric(s) relative to the spectra contained in a training dataset. That is, the supervised ML method compares a new spectrum with those obtained earlier and looks for the most similar to the one that was submitted as input. All feasible solutions of Equation (3) (sets of c_i) form a surface with many local extrema. Finding a universal optimal solution is a hard task, and ML methods are specially designed for that, outperforming conventional methods. A comparison of conventional and ML techniques is presented in Figure 4. Therefore, only a chemical-based approach provides interpretability of medical decisions due to the possibility of discovering VMBs of a pathological process.

From the point of view of a typical ML pipeline (see Figure 1), the last steps are the creation of a prediction data model for new data classification and its validation (mostly cross-validation). The step of creating the data model for classification using supervised ML methods is based on pattern recognition and relates to qualitative analysis.

The interpretability of the data model for classification is based on the possibility of analyzing spatial distribution data points in a feature space. When initial spectral data were transformed into new independent variables, for example, using PCA, the biochemical meaning of such new variables was not obvious. It means that for approving a data model designed for a medical decision support system, additional ML pipeline steps could be required. These steps consist of establishing the relationships between informative features and key molecular biomarkers, which are associated with differences in informative features of the target group and the control group. In breathomics studies, such molecular biomarkers are the composition of volatile molecules contained in the

breath, reflecting the shift in metabolism caused by pathology. Therefore, the step of qualitative and quantitative analysis of exhaled air sample composition is obligatory. In the case of the pattern recognition approach, this step should be implemented after cross-validation of the created data model. In the case of a chemical-based approach, a profile of the most specific volatile molecular biomarker concentrations should be established before the data model creation.

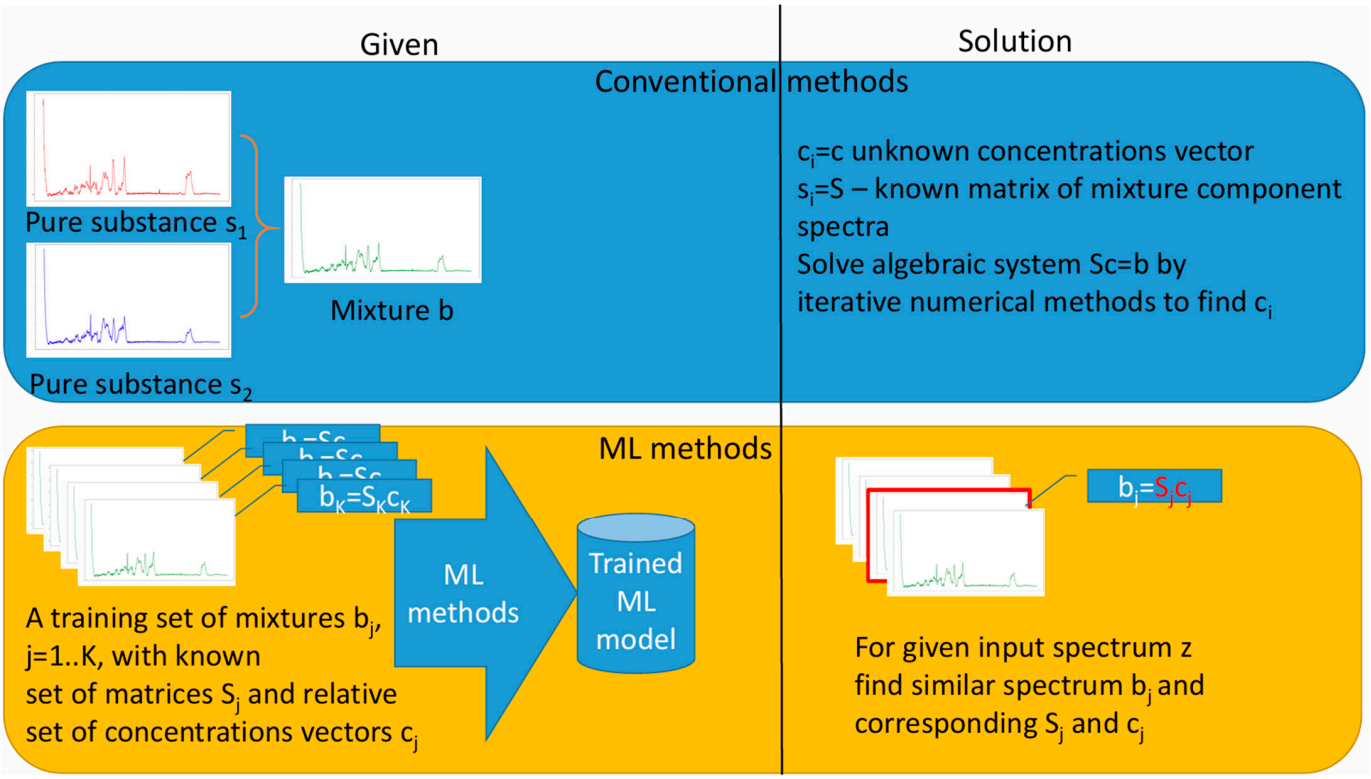

Figure 4. Conventional iterative and ML methods comparison with respect to qualitative/quantitative spectra analysis task.

To our surprise, only a few works related to medical diagnostics using spectral analysis of exhaled air samples using IR and THz absorption spectroscopy combined with ML were found [8,58–60]. The only implementation of pattern recognition and chemical-based approaches on the same date set regarding this task was conducted by us [8]. Two prediction data models were created for acute myocardial infarction diagnosis through exhaled air spectral analysis with IR laser photo-acoustic spectroscopy and ML. The predictive model based on the exhaled air absorption spectrum provided 0.86 of the mean values of both the sensitivity and specificity when linear SVM combined with PCA was used. The created predictive model based on using six VMBs (C_5H_{12}, N_2O, NO_2, C_2H_4, CO, and CO_2) provided 0.82 and 0.93 of the mean values of the sensitivity and specificity, respectively, when linear SVM was used.

Approving this profile of VMB concentrations requires recovery of the most significant metabolic pathways of pathological processes underlying a target disease. Obviously, such knowledge is very limited at the current stage. It means that bringing breathomics into clinical practices is associated with establishing the key volatile molecular biomarkers in the breath and discovering their metabolic origin. Perhaps the crucial step in these studies can be aimed at finding relationships between molecular biomarkers in breath and in the blood. It means that a more suitable ML pipeline should be transformed into something like the one shown in Figure 5.

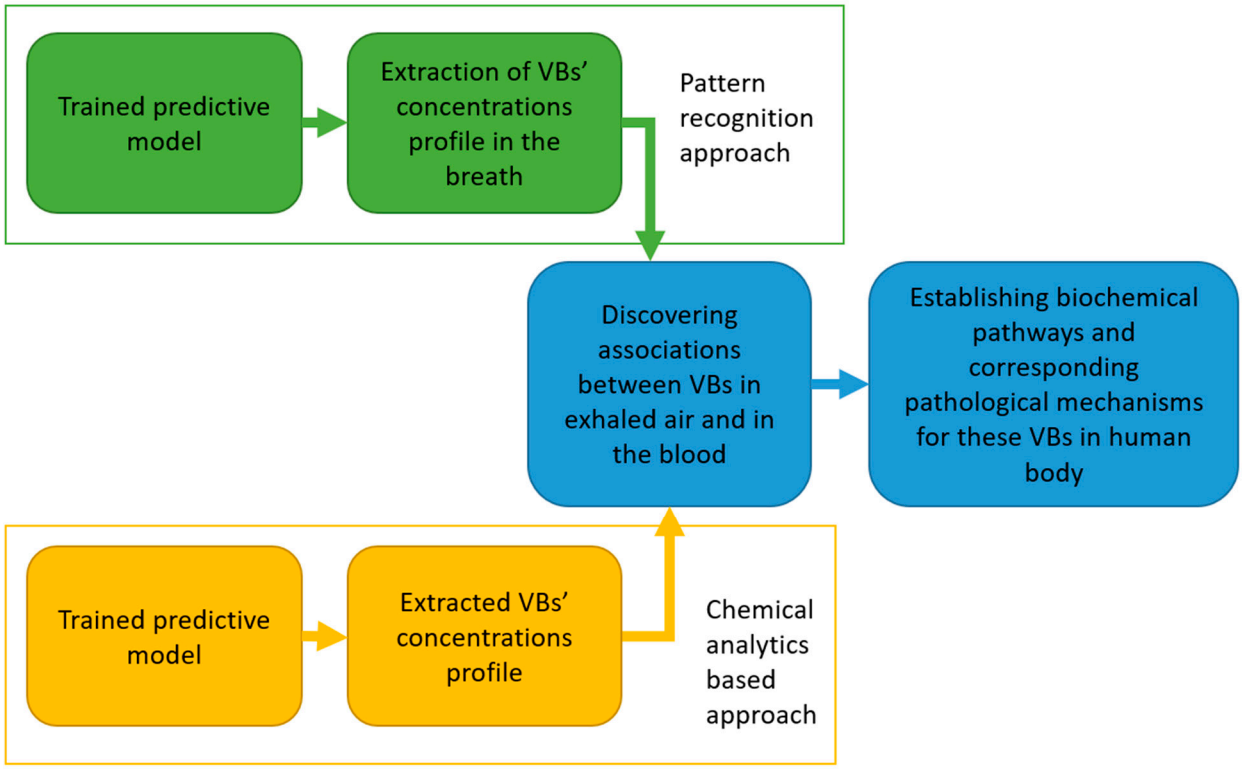

Figure 5. The additional steps in IR-THz spectra processing pipeline for exhaled air samples analysis.

It should be noted that all these additional steps require more complex experimental equipment, time, and expense (see Figure 6).

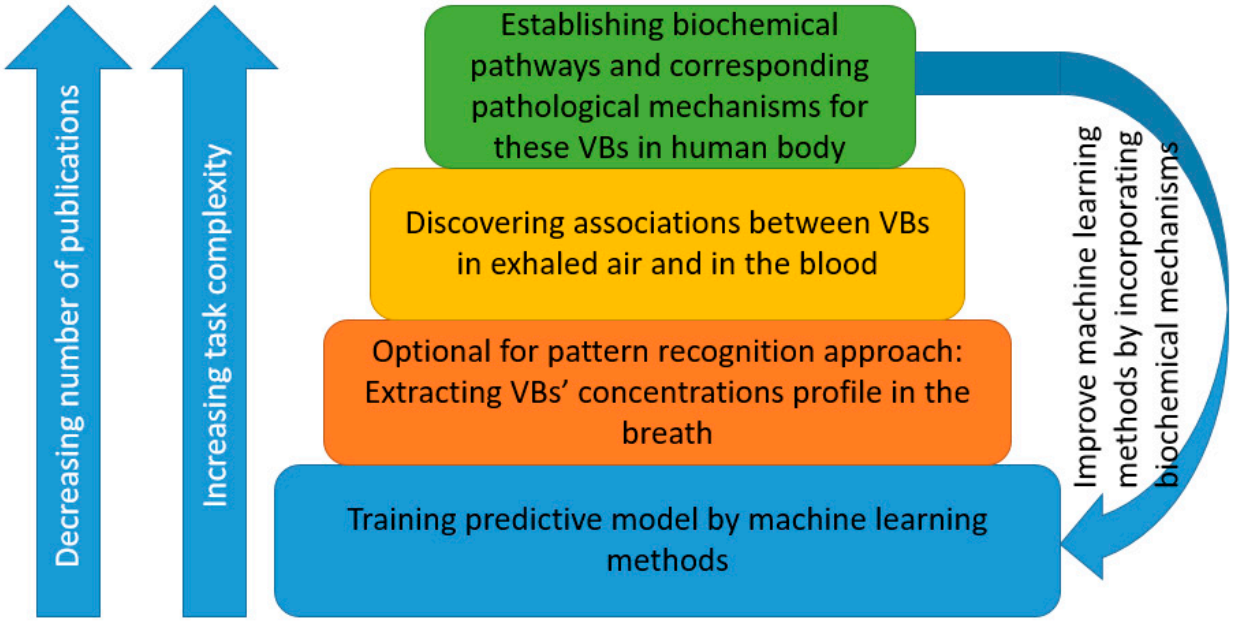

Figure 6. Extended analysis of IR-THz spectra for breathomics applications.

The bottom layer of the pyramid is common processing resulting in predictive model construction without understanding biochemical mechanisms, which are hidden behind. The second layer (orange) necessary for the pattern recognition approach is to extract concentration profiles for specific metabolites for further analysis. The third level (yellow) confirms the relationship between the presence of metabolites in exhaled air and blood and allows screening of metabolites not related to the pathological processes under study. The top-level (green) is associated with establishing a link between the biochemical processes occurring in the body and the detected metabolites. From the bottom to the top of the pyramid, the complexity of the tasks increases, and the number of relevant publications decreases accordingly. The more complicated the step, the less work that will have been completed. In any case, these steps should be completed to bring breathomics methods into routine medical practice.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app142411521/s1, Table S1: IR spectra analysis: input data, materials, and experimental conditions description; Table S2: exhaled air samples IR absorption spectra analysis: ML pipeline description; Table S3: THz spectra analysis: input data, materials, and experimental conditions description; Table S4: THz spectra analysis: ML pipeline description [96].

Author Contributions: Y.K. conceptualization, introduction, discussion. A.T. THz, IR spectroscopy, and machine learning methods review. D.V. THz, IR spectroscopy, and machine learning methods review, draft preparation. A.S. conceptualization, introduction, discussion. V.Z. THz, IR spectroscopy, and machine learning methods review. All authors have read and agreed to the published version of the manuscript.

Funding: The research was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-15-2024-557 dated 25 April 2024).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors thank Peter J. Mitchell, Fellow of the Institute of Linguists of the United Kingdom, for the style review.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kiss, H.; Örlős, Z.; Gellért, Á.; Megyesfalvi, Z.; Mikáczó, A.; Sárközi, A.; Vaskó, A.; Miklós, Z.; Horváth, I. Exhaled biomarkers for point-of-care diagnosis: Recent advances and new challenges in breathomics. *Micromachines* **2023**, *14*, 391. [CrossRef] [PubMed]
2. Sharma, A.; Kumar, R.; Varadwaj, P. Smelling the disease: Diagnostic potential of Breath Analysis. *Mol. Diagn. Ther.* **2023**, *27*, 321–347. [CrossRef] [PubMed]
3. Kistenev, Y.V. Diabetes Noninvasive Diagnostics and Monitoring through Volatile Biomarkers Analysis in the Exhaled Breath Using Optical Absorption Spectroscopy. *J. Biophotonics* **2023**, *16*, e202300198. [CrossRef] [PubMed]
4. Kistenev, Y.V.; Borisov, A.V.; Vrazhnov, D.A. Breathomics for Lung Cancer Diagnosis. *Multimodal Opt. Diagn. Cancer* **2020**, *2020*, 209–243.
5. Kistenev, Y.; Borisov, A.; Nikolaev, V.; Vrazhnov, D.; Kuzmin, D. Laser photoacoustic spectroscopy applications in breathomics. *J. Biomed. Photonics Eng.* **2019**, *5*, 010303. [CrossRef]
6. Lykina, A.A.; Anfertev, V.A.; Domracheva, E.G.; Chernyaeva, M.B.; Kononova, Y.A.; Toropova, Y.G.; Korolev, D.V.; Smolyanskaya, O.A.; Vaks, V.L. Terahertz high-resolution spectroscopy of thermal decomposition gas products of diabetic and non-diabetic blood plasma and kidney tissue pellets. *J. Biomed. Opt.* **2021**, *26*, 043008. [CrossRef]
7. Frater, J.L.; Hurley, M.Y. Complete blood cell count-derived biomarkers and clinical studies: Is it time for new reporting criteria? comment on Anand et al. utility of red cell distribution width (RDW) as a noninvasive biomarker for the diagnosis of acute appendicitis: A systematic review and meta-analysis of 5222 cases. diagnostics 2022, 12, 1011. *Diagnostics* **2022**, *12*, 2329. [CrossRef]
8. Borisov, A.V.; Syrkina, A.G.; Kuzmin, D.A.; Ryabov, V.V.; Boyko, A.A.; Zaharova, O.; Zasedatel, V.S.; Kistenev, Y.V. Application of machine learning and laser optical-acoustic spectroscopy to study the profile of exhaled air volatile markers of acute myocardial infarction. *J. Breath Res.* **2021**, *15*, 027104. [CrossRef]

9. Vaks, V.; Anfertev, V.; Ayzenshtadt, A.; Chernyaeva, M.; Domracheva, E.; Glushkova, K.; Larin, R.; Shakhova, M. Novel approaches in the diagnostics of ear-nose-throat diseases using high-resolution thz spectroscopy. *Appl. Sci.* **2023**, *13*, 1573. [CrossRef]
10. Smith, A.L. Infrared spectrometry. In *Systematic Materials Analysis*; Academic press: Cambridge, MA, USA, 1974; pp. 255–300. [CrossRef]
11. Dexheimer, S.L. *Terahertz Spectroscopy: Principles and Applications*; CRC Press: Boca Raton, FL, USA, 2017.
12. Baxter, J.B.; Guglietta, G.W. Terahertz Spectroscopy. *Anal. Chem.* **2011**, *83*, 4342–4368. [CrossRef]
13. Jepsen, P.; Cooke, D.G.; Koch, M. Terahertz Spectroscopy and Imaging–Modern Techniques and Applications. *Laser Photonics Rev.* **2011**, *5*, 124–166. [CrossRef]
14. Kistenev, Y.V. Potentialities of Small-Size Subterahertz-Wave Spectrometers Based on Cascade Frequency Multiplication for Local Environmental Monitoring of the Atmosphere. *Radiophys. Quantum Electron.* **2023**, *65*, 746–759. [CrossRef]
15. Vogt, D.W.; Erkintalo, M.; Leonhardt, R. Coherent Continuous Wave Terahertz Spectroscopy Using Hilbert Transform. *J. Infrared-Millim. Terahertz Waves* **2019**, *40*, 524–534. [CrossRef]
16. Ma, Y. A Novel Tapered Quartz Tuning Fork-Based Laser Spectroscopy Sensing. *Appl. Phys. Rev.* **2024**, *11*, 041412. [CrossRef]
17. Votintsev, A.P.; Borisov, A.V.; Makashev, D.R.; Stoyanova, M.Y.; Kistenev, Y.V. Quartz-Enhanced Photoacoustic Spectroscopy in the Terahertz Spectral Range. *Photonics* **2023**, *10*, 835. [CrossRef]
18. de Juan, A.; Tauler, R. Multivariate Curve Resolution: 50 Years Addressing the Mixture Analysis Problem—A Review. *Anal. Chim. Acta* **2021**, *1145*, 59–78. [CrossRef]
19. Ishihara, S.; Hattori, Y.; Otsuka, M.; Sasaki, T. Cocrystal formation through solid-state reaction between ibuprofen and nicotinamide revealed using thz and IR spectroscopy with multivariate analysis. *Crystals* **2020**, *10*, 760. [CrossRef]
20. El Haddad, J.; Bousquet, B.; Canioni, L.; Mounaix, P. Review in terahertz spectral analysis. *TrAC Trends Anal. Chem.* **2013**, *44*, 98–105. [CrossRef]
21. Kościelniak, P.; Wieczorek, M. Univariate analytical calibration methods and procedures. A Review. *Anal. Chim. Acta* **2016**, *944*, 14–28. [CrossRef]
22. Merriman, M. *A List of Writings Relating to the Method of Least Squares, with Historical and Critical Notes 1877*; Kessinger Publishing: Whitefish, MT, USA, 2009.
23. Gavin, H.P. *The Levenberg-Marquardt Algorithm for Nonlinear Least Squares Curve-Fitting Problems*; Duke University: Durham, NC, USA, 2019.
24. Madsen, K.; Nielsen, H.B.; Tingleff, O. *Methods for Non-Linear Least Squares Problems*; Informatics and Mathematical Modelling, Technical University of Denmark, DTU: Lyngby, Denmark, 2004.
25. Levenberg, K. A method for the solution of certain non-linear problems in least squares. *Q. Appl. Math.* **1944**, *2*, 164–168. [CrossRef]
26. Bouhamidi, A.; Jbilou, K.; Reichel, L.; Sadok, H. A generalized Global Arnoldi method for ill-posed matrix equations. *J. Comput. Appl. Math.* **2012**, *236*, 2078–2089. [CrossRef]
27. Liang, Y.-Z.; Kvalheim, O.M.; Manne, R. White, grey and black multicomponent systems. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 235–250. [CrossRef]
28. Jolliffe, I.T. *Principal Component Analysis*; Springer: New York, NY, USA, 2002.
29. Ren, T.; Modest, M.F.; Fateev, A.; Sutton, G.; Zhao, W.; Rusu, F. Machine learning applied to retrieval of temperature and concentration distributions from infrared emission measurements. *Appl. Energy* **2019**, *252*, 113448. [CrossRef]
30. Alpaydin, E. *Machine Learning*; The MIT Press: Cambridge, MA, USA, 2021.
31. Kashyap, M.; Bandyopadhyay, A.; Bertling, K.; Sengupta, A.; Rakic, A.D. Quantifying relative moisture content in dielectric models using CW-thz spectroscopy and supervised machine learning regression. *Terahertz Emit. Receiv. Appl. XII* **2021**, *84*, 3. [CrossRef]
32. Zahid, A.; Abbas, H.T.; Ren, A.; Zoha, A.; Heidari, H.; Shah, S.A.; Imran, M.A.; Alomainy, A.; Abbasi, Q.H. Machine learning driven non-invasive approach of water content estimation in living plant leaves using terahertz waves. *Plant Methods* **2019**, *15*, 138. [CrossRef]
33. Enders, A.A.; North, N.M.; Fensore, C.M.; Velez-Alvarez, J.; Allen, H.C. Functional group identification for FTIR spectra using image-based machine learning models. *Anal. Chem.* **2021**, *93*, 9711–9718. [CrossRef]
34. Chowdhury, M.A.; Rice, T.E.; Oehlschlaeger, M.A. VOC-Net: A deep learning model for the automated classification of rotational thz spectra of Volatile Organic Compounds. *Appl. Sci.* **2022**, *12*, 8447. [CrossRef]
35. Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **2019**, *7*, 154096–154113. [CrossRef]
36. Mishra, P.; Passos, D.; Marini, F.; Xu, J.; Amigo, J.M.; Gowen, A.A.; Jansen, J.J.; Biancolillo, A.; Roger, J.M.; Rutledge, D.N.; et al. Deep learning for near-infrared spectral data modelling: Hypes and benefits. *TrAC Trends Anal. Chem.* **2022**, *157*, 116804. [CrossRef]
37. Chowdhury, M.A.; Rice, T.E.; Oehlschlaeger, M.A. Evaluation of machine learning methods for classification of rotational absorption spectra for gases in the 220–330 ghz range. *Appl. Phys. B* **2021**, *127*, 34. [CrossRef]

38. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]
39. Wang, C.-Y.; Ko, T.-S.; Hsu, C.-C. Interpreting convolutional neural network for real-time volatile organic compounds detection and classification using optical emission spectroscopy of plasma. *Anal. Chim. Acta* **2021**, *1179*, 338822. [CrossRef] [PubMed]
40. Mokari, A.; Guo, S.; Bocklitz, T. Exploring the Steps of Infrared (IR) Spectral Analysis: Pre-Processing,(Classical) Data Modelling, and Deep Learning. *Molecules* **2023**, *28*, 6886. [CrossRef] [PubMed]
41. Zhang, W.; Kasun, L.C.; Wang, Q.J.; Zheng, Y.; Lin, Z. A Review of Machine Learning for Near-Infrared Spectroscopy. *Sensors* **2022**, *22*, 9764. [CrossRef] [PubMed]
42. Park, H.; Son, J.-H. Machine Learning Techniques for THz Imaging and Time-Domain Spectroscopy. *Sensors* **2021**, *21*, 1186. [CrossRef]
43. Seo, Y.M.; Goldsmith, P.F.; Tolls, V.; Shipman, R.; Kulesa, C.; Peters, W.; Walker, C.; Melnick, G. Applications of Machine Learning Algorithms in Processing Terahertz Spectroscopic Data. *J. Astron. Instrum.* **2020**, *9*, 2050011. [CrossRef]
44. Helal, S.; Sarieddeen, H.; Dahrouj, H.; Al-Naffouri, T.Y.; Alouini, M.-S. Signal Processing and Machine Learning Techniques for Terahertz Sensing: An Overview. *IEEE Signal Process. Mag.* **2022**, *39*, 42–62. [CrossRef]
45. Haar, L.V.; Elvira, T.; Ochoa, O. An Analysis of Explainability Methods for Convolutional Neural Networks. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105606. [CrossRef]
46. Angelov, P.; Soares, E. Towards Explainable Deep Neural Networks (xDNN). *Neural Netw.* **2020**, *130*, 185–194. [CrossRef]
47. Wu, M.; Wu, H.; Barrett, C. VeriX: Towards Verified eXplainability of Deep Neural Networks. *arXiv* **2022**, arXiv:2212.01051. [CrossRef]
48. Dwivedi, A.K.; Mallawaarachchi, I.; Alvarado, L.A. Analysis of Small Sample Size Studies Using Nonparametric Bootstrap Test with Pooled Resampling Method: Nonparametric Bootstrap Test for Small Sample Size Studies. *Stat. Med.* **2017**, *36*, 2187–2205. [CrossRef] [PubMed]
49. Lee, L.C.; Liong, C.-Y.; Jemain, A.A. A contemporary review on data preprocessing (DP) practice strategy in ATR-Ftir Spectrum. *Chemom. Intell. Lab. Syst.* **2017**, *163*, 64–75. [CrossRef]
50. Guo, S.; Mayerhöfer, T.; Pahlow, S.; Hübner, U.; Popp, J.; Bocklitz, T. Deep learning for 'artefact' removal in Infrared Spectroscopy. *Analyst* **2020**, *145*, 5213–5220. [CrossRef] [PubMed]
51. Helin, R.; Indahl, U.G.; Tomic, O.; Liland, K.H. On the possible benefits of deep learning for spectral preprocessing. *J. Chemom.* **2021**, *36*, e3374. [CrossRef]
52. Kireev, S.V.; Kondrashov, A.A.; Shnyrev, S.L. Application of the Wiener Filtering Algorithm for Processing the Signal Obtained by the TDLAS Method Using the Synchronous Detection Technique for the Measurement Problem of 13CO2 Con-Centration in Exhaled Air. *Laser Phys. Lett.* **2019**, *16*, 085701. [CrossRef]
53. Kistenev, Y.V.; Skiba, V.E.; Prischepa, V.V.; Borisov, A.V.; Vrazhnov, D.A. Gas-Mixture IR Absorption Spectra Denoising Using Deep Learning. *J. Quant. Spectrosc. Radiat. Transf.* **2024**, *313*, 108825. [CrossRef]
54. Elaraby, S.; Sabry, Y.M.; Abuelenin, S.M. Super-resolution infrared spectroscopy for gas analysis using convolutional Neural Networks. *Appl. Mach. Learn.* **2020**, *11511*, 180–187. [CrossRef]
55. Kistenev, Y.V.; Skiba, V.E.; Prischepa, V.V.; Vrazhnov, D.A.; Borisov, A.V. Super-resolution reconstruction of noisy gas-mixture absorption spectra using Deep Learning. *J. Quant. Spectrosc. Radiat. Transf.* **2022**, *289*, 108278. [CrossRef]
56. Kochanov, R.V.; Gordon, I.E.; Rothman, L.S.; Wcisło, P.; Hill, C.; Wilzewski, J.S. HITRAN Application Programming Interface (HAPI): A comprehensive approach to working with spectroscopic data. *J. Quant. Spectrosc. Radiat. Transf.* **2016**, *177*, 15–30. [CrossRef]
57. Hill, C.; Gordon, I.E.; Kochanov, R.V.; Barrett, L.; Wilzewski, J.S.; Rothman, L.S. HITRANonline: An online interface and the flexible representation of spectroscopic data in the HITRAN database. *J. Quant. Spectrosc. Radiat. Transf.* **2016**, *177*, 4–14. [CrossRef]
58. Golyak, I.S.; Kareva, E.R.; Fufurin, I.L.; Anfimov, D.R.; Scherbakova, A.V.; Nebritova, A.O.; Demkin, P.P.; Morozov, A.N. Numerical methods of spectral analysis of multicomponent gas mixtures and human exhaled breath. *Comput. Opt.* **2022**, *46*, 650–658. [CrossRef]
59. Fufurin, I.L.; Golyak, I.S.; Anfimov, D.R.; Tabalina, A.S.; Kareva, E.R.; Morozov, A.N.; Demkin, P.P. Machine learning applications for spectral analysis of human exhaled breath for early diagnosis of diseases. *Opt. Health Care Biomed. Opt. X* **2020**, *68*, 110. [CrossRef]
60. Fufurin, I.L.; Anfimov, D.R.; Kareva, E.R.; Scherbakova, A.V.; Demkin, P.P.; Morozov, A.N.; Golyak, I.S. Numerical techniques for infrared spectra analysis of organic and inorganic volatile compounds for biomedical applications. *Opt. Eng.* **2021**, *60*, 082016. [CrossRef]
61. Ouyang, T.; Wang, C.; Yu, Z.; Stach, R.; Mizaikoff, B.; Liedberg, B.; Huang, G.-B.; Wang, Q.-J. Quantitative analysis of gas phase IR spectra based on Extreme Learning Machine Regression Model. *Sensors* **2019**, *19*, 5535. [CrossRef] [PubMed]
62. Balakrishnama, S.; Ganapathiraju, A. Linear discriminant analysis-a brief tutorial. *Inst. Signal Inf. Process.* **1998**, *18*, 1–8.
63. Barreto, D.F. *An Exploratory Analysis Using t-SNE*; Universidade Federal do Ceará, Centro de Ciências, Curso de Ciência da Computação: Fortaleza, France, 2018. Available online: http://repositorio.ufc.br/handle/riufc/41264 (accessed on 9 October 2024).

64. Nishikawa, T.; Saruwatari, H.; Shikano, K. Comparison of Time-Domain ICA, Frequency-Domain ICA and Multistage ICA for Blind Source Separation. In Proceedings of the European Signal Processing Conference, Toulouse, France, 1 September 2002; IEEE: New York, NY, USA, 2002; pp. 1–4.
65. Ruiz-Perez, D.; Guan, H.; Madhivanan, P.; Mathee, K.; Narasimhan, G. So you think you can pls-da? *BMC Bioinform.* **2020**, *21*, 2. [CrossRef]
66. Zivkovic, M.; Vesic, A.; Bacanin, N.; Strumberger, I.; Antonijevic, M.; Jovanovic, L.; Marjanovic, M. An improved animal migration optimization approach for extreme learning machine tuning. In *Lecture Notes in Networks and Systems, Proceeding of the International Conference on Intelligent and Fuzzy Systems, Izmir, Turkey, 19–21 June 2022*; Springer: Cham, Switzerland, 2022; pp. 3–13. [CrossRef]
67. Wu, C.; Li, Y.; Zhao, Z.; Liu, B. Extreme learning machine with autoencoding receptive fields for image classification. *Neural Comput. Appl.* **2019**, *32*, 8157–8173. [CrossRef]
68. Gurney, K. *An Introduction to Neural Networks*; CRC Press: Boca Raton, FL, USA, 2014.
69. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef] [PubMed]
70. Burges, C.J. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167. [CrossRef]
71. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006; Volume 4.
72. Menze, B.H.; Kelm, B.M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F.A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of Spectral Data. *BMC Bioinform.* **2009**, *10*, 213. [CrossRef]
73. Adler, A.I.; Painsky, A. Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy* **2022**, *24*, 687. [CrossRef] [PubMed]
74. Kramer, O. Dimensionality reduction with unsupervised nearest neighbors. In *Intelligent Systems Reference Library*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 51. [CrossRef]
75. Dudani, S.A. The distance-weighted K-nearest-neighbor rule. *IEEE Trans. Syst. Man Cybern.* **1976**, *SMC-6*, 325–327. [CrossRef]
76. Ripley, B.D. *Pattern Recognition and Neural Networks*; Cambridge university press: Cambridge, UK, 2007.
77. Tominaga, Y. Comparative study of class data analysis with PCA-Lda, Simca, PLS, Anns, and K-NN. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 105–115. [CrossRef]
78. Wheelock, Å.M.; Wheelock, C.E. Trials and tribulations of 'OMICS data analysis: Assessing quality of Simca-based multivariate models using examples from pulmonary medicine. *Mol. BioSyst.* **2013**, *9*, 2589. [CrossRef] [PubMed]
79. Ma, Y.; Wang, Q.; Li, L. PLS model investigation of thiabendazole based on Thz Spectrum. *J. Quant. Spectrosc. Radiat. Transf.* **2013**, *117*, 7–14. [CrossRef]
80. Cramer, R.D. Partial least squares (PLS): Its strengths and Limitations. *Perspect. Drug Discov. Des.* **1993**, *1*, 269–278. [CrossRef]
81. Awad, M.; Khanna, R. Support vector regression. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Apress: Berkeley, CA, USA, 2015; pp. 67–80.
82. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: New York, NY, USA, 2010. [CrossRef]
83. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
84. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
85. Yuan, L.-M.; Yang, X.; Fu, X.; Yang, J.; Chen, X.; Huang, G.; Chen, X.; Li, L.; Shi, W. Consensual regression of lasso-sparse PLS models for near-infrared spectra of food. *Agriculture* **2022**, *12*, 1804. [CrossRef]
86. Kistenev, Y.V.; Borisov, A.V.; Samarinova, A.A.; Colón-Rodríguez, S.; Lednev, I.K. A novel Raman spectroscopic method for detecting traces of blood on an interfering substrate. *Sci. Rep.* **2023**, *13*, 5384. [CrossRef] [PubMed]
87. Lai, W.C.; Zou, Y.; Chakravarty, S.; Zhu, L.; Chen, R.T. *Comparative Sensitivity Analysis of Integrated Optical Waveguides for Near-Infrared Volatile Organic Compounds with 1ppb Detection*; SPIE: Bellingham, WA, USA, 2014; Volume 8990, pp. 202–207.
88. Neumaier, P.; Schmalz, K.; Borngraber, J.; Wylde, R.; Hübers, H.-W.; Hübers, H.-W. Terahertz Gas-Phase Spectroscopy: Chemometrics for Security and Medical Applications. *Analyst* **2015**, *140*, 213–222. [CrossRef]
89. Ando, M.; Lednev, I.K.; Hamaguchi, H.O. Quantitative Spectrometry of Complex Molecular Systems by Hypothetical Addi-Tion Multivariate Analysis with Numerical Differentiation (HAMAND). In *Frontiers and Advances in Molecular Spectroscopy*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 369–378. [CrossRef]
90. Panati, C.; Wagner, S.; Bruggenwirth, S. Feature Relevance Evaluation Using Grad-CAM, LIME and SHAP for Deep Learning SAR Data Classification. In Proceedings of the 2022 23rd International Radar Symposium (IRS), Gdansk, Poland, 12–14 September 2022; IEEE: New York, NY, USA, 2022. [CrossRef]
91. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv* **2017**, arXiv:1703.01365. [CrossRef]
92. Qi, Z.; Khorram, S.; Fuxin, L. Visualizing Deep Networks by Optimizing with Integrated Gradients. *Proc. Conf. AAAI Artif. Intell.* **2020**, *34*, 11890–11898. [CrossRef]
93. Molnar, C. *Interpretable Machine Learning*; Lulu.com: Morrisville, NC, USA, 2020; ISBN 9780244768522.
94. Suara, S.; Jha, A.; Sinha, P.; Sekh, A.A. Is Grad-CAM Explainable in Medical Images? In *Communications in Computer and Information Science*; Springer Nature: Cham, Switzerland, 2024; pp. 124–135, ISBN 9783031581809.

95. Prischepa, V.V.; Skiba, V.; Vrazhnov, D.; Markelov, A. Application of laser absorption spectroscopy and machine learning for component analysis of Multicomponent Gas Media. In Proceedings of the Fourth International Conference on Terahertz and Microwave Radiation: Generation, Detection, and Applications, Tomsk, Russia, 24–26 August 2020; p. 68. [CrossRef]
96. Li, Z.; Rothbart, N.; Deng, X.; Geng, H.; Zheng, X.; Neumaier, P.; Hübers, H.-W. Qualitative and quantitative analysis of terahertz gas-phase spectroscopy using independent component analysis. *Chemom. Intell. Lab. Syst.* **2020**, *206*, 104129. [CrossRef]