Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 6s (2025)
https://www.jneonatalsurg.com

Statistical Techniques of Exhaled Breath Analysis for Disease Diagnosis and Human Health Monitoring

Nilakshi Maruti Mule[1], Dipti Durgesh Patil[2]

[1]Research Scholar, STES' Smt Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune, India.

Assistant Professor, Government College of Engineering, Karad, India.

Email ID: nilmule@gmail.com

[2]Department of Information Technology, MKSSS's Cummins College of Engineering for Women, Pune, India.

Savitribai Phule Pune University, Pune, India.

Email ID: diptivt@gmail.com

Cite this paper as: Nilakshi Maruti Mule, Dipti Durgesh Patil, (2025) Statistical Techniques of Exhaled Breath Analysis for Disease Diagnosis and Human Health Monitoring. *Journal of Neonatal Surgery*, 14 (6s), 332-348

ABSTRACT

Exhaled breath analysis finds patterns of volatile organic compounds in the human body. This paper describes the main types of sensors used for disease diagnosis applications of exhaled breath, data preprocessing, and data analysis techniques of breath data and includes the results of some existing research. First, enlist the sensor mainly used for identifying breath biomarkers. Next, data preprocessing removes the noise from the collected breath data. Then, machine learning algorithms are used to identify discriminators of volatile organic compounds after preprocessing breath data for data analysis. This work comprises data preprocessing and analysis using different breath analysis techniques and machine learning algorithms for real-life situations. The paper includes a brief description of each machine learning algorithm, process, algorithms' condition, and its relevant formula. This work systematically reviews breath data preprocessing and analysis techniques with advantages, disadvantages, applications and some disease results of existing research.

Keywords: *SVM, PCA, VOCs, Exhaled Breath, KNN*

1. INTRODUCTION

This paper comprehensively describes the steps involved in the five data preprocessing techniques used for breath data analysis. It also focuses on the various learning techniques commonly used in this field. There are two main types of statistical techniques used for statistical studies: supervised learning and the second unsupervised. The supervised learning technique is used for classification and regression problems. This technique aims to learn how to predict the model's output. It includes seven different algorithms: Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc. Unsupervised learning is a type of machine learning that uses unlabeled data. Its goal is to find the hidden patterns in the collected data. Unsupervised learning is used for the classification of Clustering and Association problems. It includes three different algorithms as Clustering, KNN, and Apriori.

The human breath samples contain thousands of volatile organic compounds (VOCs). The data generated by the breath is large and complex. Many of the compounds are endogenous, and the others are exogenous [1,2]. The data collected during the exhaled breath contain various sources of variance. These sources include the interest level and the noise level. The objective of the research challenge is to extract useful information from the data by analyzing exhaled breath [3].

Various analytical methods are used to perform the analysis. Mass spectrometry and gas chromatography (GC-MS) are commonly used to analyze breath. This technique is sensitive and reliable. This is used in offline breath analysis. However, it requires sample preconcentration and capillary separation; both are time-consuming [4]. Online real-time breath analysis based on MS, such as SESI-MS, SIFT-MS, and PTR-MS. SIFT-MS is an analytical technique for quantifying real-time trace gases in air or breath samples. This method can detect disease states early and quickly [4,5,39].

PTR-MS is an ionization molecule used for online analysis. It is equipped with a quadrupole mass analyzer with high sensitivity [6]. SESI-MS is an ambient ionization method, a sensitive, robust, and reliable tool. However, calibration is needed for this method, and absolute quantification is impossible.

2. TYPES OF SENSORS FOR BREATH ANALYSIS

For breath analysis, a sensing unit needs breath analysis sensors that can be put into several groups based on how they convert signals. These groups include chemiresistors, electrochemical, optical, piezoelectric (mass-sensitive), and others. A transducer is also essential for accurate measurement because it turns the interaction of the analyte with the sensor into signals that can be measured.

2.1 Chemiresistors Sensing

Sensing is a new non-invasive health care method that is becoming popular because it is sensitive, small, cheap, portable, and easy to make. Changes in how the analyte affects how electricity flows are used in these sensors.[44,40]

2.1.1 Metal oxide semiconductor Sensors (MOS)

These sensors are small in size and have easy operation, low cost, and low maintenance. However, several factors influence the MOS sensors' sensitivity, selectivity, and stability. This multi-factor dependence allows the sensors to be tailored to the application's needs. The most common issues with MOS sensors are power consumption, heat generation, lack of selectivity, and humidity interference. There has been research into these issues and the selection of appropriate materials for fabricating sensors that meet the requirements of a low-cost portable device.[41,49]

2.2 Electrochemical Sensing

Electrochemical sensors are becoming more popular in breath analysis because they are very selective, cheap, small, use little power, and are safe for living organisms. Most of the time, these sensors have been used to find specific gas biomarkers.[42,45]

3. PIEZOELECTRIC SENSORS

Mechanical stress generally causes piezoelectric materials to produce voltage and vice versa. Therefore, piezoelectric sensors, which are mechanically sensitive, are frequently used as mass-sensitive sensors. In addition, piezoelectric sensors use acoustic wave devices, also known as mass, gravimetric, or microbalance sensors. Acoustic waves are generated using an oscillating circuit, which allows the piezoelectric crystal to resonate. Quartz crystal resonators and Surface acoustic wave are the two most common types of piezoelectric gas sensors. [43,46]

3.1 Quartz crystal resonators (QCM)

It is functionalized with various appropriate sensing elements in QCM sensors. The acoustic wave passes through the crystal's bulk parallel or perpendicular to the surface.[47]

3.2 Surface acoustic wave (SAW)

The acoustic wave propagates only parallel to the surface of the piezoelectric crystal in SAW gas sensors, penetrating the crystal to a depth of about one acoustic wavelength. At the surface, motion is both parallel and perpendicular. Therefore, a chemically selective layer is applied to the crystal surface.[48]

4. OPTICAL SENSING

Optical changes occur when an analyte and a biorecognition substance come into contact. It can be measured using colorimetry, fluorescence, chemiluminescence, or scattering mode. For example, exhaled breath coupled with colorimetric sensing has significantly applied to lung cancer diagnosis.[38]

5. FIELD-EFFECT TRANSISTORS (FET) SENSING

Gas detectors' small size, low power consumption, and high FET stability make them famous. In addition, nanomaterials such as carbon nanotubes, nanowires, graphene, and transition metal chalcogenides are used to improve their properties.[50]

6. DATA PREPROCESSING

Data preprocessing is an essential field of breath analysis. Proper preprocessing is essential and may determine whether important information is extracted from the data or not. Figure 1 shows the simplified data preprocessing, highlighting the essential roles of preprocessing.

The raw data used in analytical techniques such as GC-MS, IMS, SIFT-MS, GC-DMS, and GC-tof-MS are often obtained before the data can be analyzed. Data preprocessing is performed on various sub-step procedures, such as denoising, baseline correction, alignment, and peaking. The resulting data is then collected and converted into a statistical data matrix.

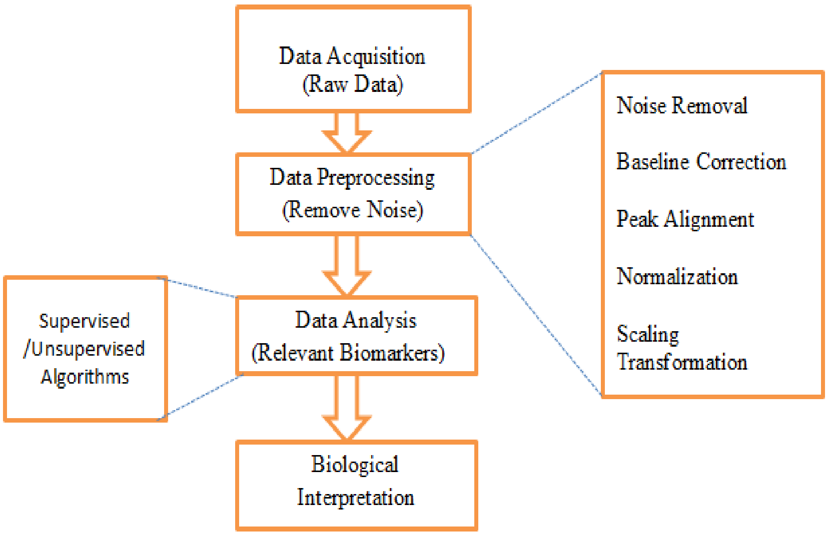

Figure 1. Simplified workflow of Data Preprocessing

A. Data Preprocessing of GC-MS Data

In GC-MS, data preprocessing is essentially involved five sequential steps. These five steps are denoising, baseline correction, alignment of peaks, peak detection, and data matrix. GC-MS data preprocessing involves denoising and second-baseline correction. The main goal of denoising is to reduce the noise introduced by variations in instrumental conditions. The second step is baseline correction done automatically using numerous types of polynomial fitting. After performing denoising and baseline correction, accurate alignment of peaks is conducted. Next, alignment and peak detection is performed on the total ion current [7,8]. Finding all local maxima and minima for each peak is part of this detection activity. In addition, signals to noise ratios are calculated. This step allows each VOC to be represented as a single number across all measured samples in the data matrix. Finally, a data matrix was created with an observation in the row and relative VOCs as values in the columns [7].

B. Data Preprocessing of IMS Data

For detecting VOCs in human breath, ion mobility spectrometers combined with a multi-capillary column (MCC/IMS) are well-known. The first step of IMS data preprocessing is reactant ion peak (RIP) detailing (i.e., each MCC-IMS chromatogram has a characteristic structure). This structure, which appears in a broad vertical line in each chromatogram, can be considered a source of measurement disturbance. The signal descent on the right side of the RIP is known as RIP tailing. Denoising and smoothing are the next steps. The signal-to-noise ratio improves at this stage. Peak picking is the final step in the data preprocessing process. Automated strategies include merged peak cluster localization, growing interval merging, wavelet-based multiscale peak detection, watershed transformation, and peak model estimation [8].

C. Data Preprocessing of SIFT-MS Data

In SIFT-MS, data preprocessing consists of peak detection and peak alignment, each of these steps involves a series of activities. When used for VOCs breath, the SIFT-MS instrument produces anomalous noise signals. Before analyzing the data, these two factors must be filtered. In the first step, normalizing each data point to the sum of the precursor ion peaks allows comparison between samples that receive a different reactant precursor ion [9]. The second step removes significant erroneous peaks more remarkable than the precursor peaks. The next step is to find the tip of the peak before and after the whole mass unit concentration. Next, the process determines whether a suspected peak exists, in reality, the number of non-zero readings within one mass unit. In the last step, the concentration matrix reduces to include only the real mass value for analysis [5].

D. Data preprocessing of GC-DMS

GC-DMS data preprocessing technique includes interpolation, baseline correction, and alignment. This technique is used to adjust for inevitable baseline shifts. The interpolation technique is used when the sample rate is low. Asymmetric least squares are used to adjust the baseline. The alignment is applied to the entire dataset. [10]

E. Data preprocessing of GC-tof-MS

In preprocessing of GC-tof-MS, there is a sequence of five different steps that are performed. These steps are-Denoising, baseline correction, alignment, normalization, and scaling. In this series, the initial step is to choose the retention index of each chromatogram for analysis. It is followed by noise removal using wavelet transformation and baseline correction. In the third phase, optimal correlation warping is used to achieve exact alignment. In the fourth step, the area under the peaks is calculated. The final step is to normalize the data to reduce the impact of random variation.[11,12, 13]

6.1 Normalization, scaling, and transformation

Normalization and scaling are steps between preprocessing raw data and multivariate statistical analysis. Normalization and scaling aim to diminish the impact of undesirable systematic changes between measured samples, such as those resulting from sample collection and analysis methods. The primary transformation goals are to symmetrize data distributions, reduce variance, and bring the distribution closer to normal. In addition, the effect of heteroscedasticity in the data can be removed via transformation. [15] Table 1 shows the overview of typical data preprocessing methods. This mean is estimated by $\bar{x}$ and the standard deviation is estimated by $S_{i.}$ $\tilde{x}$ and $\hat{x}$ represents the data after preprocessing steps.

6.1.1 Normalization

Normalization aims to show each sample in the data appropriately and consistently. Normalization is estimating and applying a scale factor to each measured sample. There are several ways to estimate the scaling factor. One of the ways of estimating scaling factors is total area normalization. This method considers the chromatogram's whole area as a single normalization factor.[15]

Table 1. Overview of Common Preprocessing Methods

	Method	Formula	Goal	Advantages	Disadvantages
Normalization [15]	Integral	$I(i) = \frac{I^{old}(i)}{\sum_k \left(\int_{l_k^l}^{l_k^\mu} (I(x))^n dx\right)^{\frac{1}{n}}}$	Take into account the differences in the samples as a whole.	The spectra are scaled to the same virtual overall concentration.	Robustness, Accuracy
Scaling [14]	Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \tilde{x}_i}{s_i}$	Correlations are used to compare metabolites.	Every metabolite takes on equal importance.	The measurement inaccuracies have inflated.
	Pareto Scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	Reduce the relative importance of tremendous values while maintaining the data structure.	Compared to auto-scaling, this method stays closer to the original measurement.	Sensitive to significant fold changes
	Vast Scaling	$\tilde{x}_{ij} = \frac{(x_{ij} - \bar{x}_i)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	Concentrate on the metabolites with minor changes.	It aims towards robustness and can make use of existing group expertise.	Without group structure, it's not suitable for substantial induced variation.
	Range Scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{max}} - x_{i_{min}})}$	Analyze the metabolites concerning the biological response spectrum.	All metabolites become equally important.	Scaling is related to Biology Inflation of the measurement inaccuracies has inflated. and observant of outliers

	Level Scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	Concentrate on the relative result.	Suited for identification of, e.g., biomarkers	The measurement inaccuracies have inflated.
Transformation [18]	Log transformation	$\tilde{x}_{ij} = log_{10}(x_{ij})$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Pseudo scaling to account for heteroscedasticity. Make additive multiplicative models.	As heteroscedasticity is reduced, multiplying effects become additive.	Values with a significant relative standard deviation and zeros are challenging to work
	Power transformation	$\tilde{x}_{ij} = \sqrt{x_{ij}}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Pseudo-scaling to account for heteroscedasticity	Reduce heteroscedasticity; low values are not a concern.	The square root is chosen at random.

6.1.2 Scaling

When the focus is on statistics, the standard deviation is a measurement of data spread commonly used in scaling. Therefore when it is a significant factor for scaling, then following said three approaches, namely- auto-scaling, Pareto scaling, and extensive scaling, could be more helpful. At the same time, the approach could be more useful when the biological range is significant in scale[14].

The first scaling method involves subtracting the mean value from each measured compound, known as mean centering. This straightforward approach compensates for the discrepancies in abundance levels between high and low abundant chemicals. Mean centering is frequently combined with other scaling methods.[14]

Within scaling, there are two types of subclasses. The first class uses a data dispersion measure as a scaling factor, while the second class uses a size measure. In scaling, the first subclass is related to data dispersion. Different approaches are used for dispersion measures in scaling, and these are four scaling approaches. These approaches are- Autoscaling, Pareto scaling, range scaling, and vast scaling.

The auto-scaling approach is used the standard deviation as the scaling factor. This deviation is commonly applied in all metabolites. The data is acquired in the scaling by using the technique of correlations.[14]

Pareto scaling and Autoscaling are highly similar. The scaling factor in Pareto scaling is the standard deviation squared.

Vast scaling expands auto-scaling and is an acronym for variable stability scaling. This approach uses the coefficient of variation as the scaling factor. This approach focuses on a stable variable that does not show substantial variation. When this scaling approach is applied as a supervised method, the group information about the samples is used to determine the group-specific coefficient of variation for scaling.[16]

Range scaling is used biological range as the scaling factor. Only two values are used to estimate the biological range in this scaling. In some cases, the robustness of range scaling is required. [17]

In scaling, another subclass is related to a size measure. In this type level scaling approach is used. This method utilizes the mean factor as the scaling factor. This mean factor level scaling converts changes from metabolite concentrations into average concentrations. [17]

6.1.3 Transformation

Nonlinear data transformation approaches include log transformation, Box-Cox transformation, rank transformation, and power transformation. Log and power transformation convert extensive value data into small values.[14]

Log transformation is commonly used. A log transformation eliminates heteroscedasticity if the relative standard deviation is constant. However, the log transformation has the disadvantage of handling the value zero. Sometimes the value reaches zero. At that moment, the related log transformation reaches minus infinity. This case creates a few technical challenges and issues.[18]

Like the log transformation, the power transformation has a similar pattern. Therefore, the power transformation results are similar to those achieved after log transformation. Therefore, power transformation with the near-zero value could be utilized in its data transformation process.[18]

7. STATISTICAL TECHNIQUES

Machine learning offers a variety of methods for analyzing and comprehending complex data, resulting in valuable information about biological changes. Data exploration and discovery are usually the first steps in multivariate analysis. This part of the analysis is conducted in a blind and unsupervised manner, providing an unbiased perspective on the data. Typically, multivariate statistical analysis is followed by supervised analysis, which uses a priori data structure knowledge.

7.1 Supervised Analysis

A supervised analysis is one method of statistical analysis. In this method, the data used is labelled. The most common algorithms used for breath analysis are linear and nonlinear.

7.1.1 Linear statistical Techniques

Linear statistical techniques are the starting point of multivariate statistical analysis. Linear techniques are a straightforward interpretation of the outcomes, with fewer parameters to optimize and fast calculation.

A. Linear Discriminant Analysis(LDA)

The LDA technique is one of the most widely used data reduction strategies. The LDA approach projects the original data matrix into a lower-dimensional environment. This technique is used in Biometrics, bioinformatics, and chemistry. In addition, there are two LDA techniques when dealing with classes: class dependent and class independent.

LDA: Class-Independent

- Calculate between class matrix S_B(MxM) as follows:

$$s_B = \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T \qquad (1)$$

- Compute within-class matrix S_W (MxM) as follows:

$$S_W = \sum_{j=1}^{c} \sum_{i=1}^{nj} (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T \qquad (2)$$

LDA: Class Dependent

- Compute the within-class matrix of each class S_{wi} (MxM) as follows:

$$S_{wj} = \sum_{x_i \varepsilon w_j} (x_i - \mu_j)(x_i - \mu_j)^T \qquad (3)$$

- Construct a transformation matrix for each class (W_i) as follows:

$$W_i = S_{w_i}^{-1} S_B \qquad (4)$$

There are two main problems of LDA. The first is a small sample size. This problem occurs because of the limited sample size and insufficient training samples for each class to compare. The second is linearity; this problem occurs because LDA cannot lower-dimensional space.[19,20]

B. Partial Least Squares Discriminant Analysis (PLS-DA)

In dimensionality reduction, PLS-DA is a critical technique. This technique maximizes data's overall variance with only a few components. The PLS-DA components give an excellent graphical representation of the partition. [21]

This combination of conventional and discriminant analyses performs on the principal component of predictor variables. Principal component regression and partial least squares regression (PLS-R) are two multivariate regression approaches. These are used in a variety of industries. PLS-DA is derived from PLS-R with discrete values for the response vector Y. The data matrix X is decomposed into P orthogonal scores T (n *P) and loadings matrix P (J*P) by PLS-R and PLS-DA. The response vector Y is decomposed into P orthogonal scores T (n*P) and loadings matrix Q (1*P) by PLS-R and PLS-DA.In the PLS-DA model, there are two basic equations:

$$X=TP^T+E \qquad (5)$$

$$Y=TQ^T+F \qquad (6)$$

Then, for the data matrix X and the response vector Y, let E and F be the n*J and n*1 error matrices, respectively.[21]

C. Partial least squares regression (PLS-R)

PLS regression is a hybrid of principal component analysis and multiple regressions. PLS-R (partial least squares regression) is the most common form of the PLS technique. It is used in chemistry and technology. PLSR is a technique for estimating parameters in a linear scientific model. This model comprises numerous components, including philosophical, conceptual, technical, numerical, statistical, etc. In addition, the linear PLSR model discovers a few "new" variables, such as LV estimates or rotations. For example, the PLS-regression coefficients b_{mk} can be written as:

$$\boldsymbol{b_{mk}} = \sum_a \boldsymbol{c_{ma} W^*_{ka}} \qquad (\boldsymbol{B} = \boldsymbol{W^* C'}) \qquad (7)$$

Note that until A (the number of PLSR components) equals K, these b's are not independent (the number of X-variables). When new variables are introduced, the variances of regression coefficients grow. [22,23,24]

D. Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA)

The OPLS-DA system was able to distinguish between chemical and physical qualities. OPLS-DA decomposes the X matrix into three different components using information from the categorical response matrix Y:

$$\boldsymbol{X} = \boldsymbol{T_p P_p^T} + \boldsymbol{T_o P_o^T} + \boldsymbol{E} \qquad (8)$$

In the above formula, T_p denotes the predictive score matrix for X. P_p denotes the predictive loading matrix for X. T_o denote the Y-orthogonal score matrix that corresponds. E denotes the residual matrix of X, and P_o denotes the loading matrix of y-orthogonal components. The ability to distinguish between predictive and non-predictive variation is OPLS-main advantage over PLS-DA.[25] Table 2 shows the strength and weaknesses of the supervised linear statistical techniques.

Table 2. Comparison of supervised linear statistical techniques

Sr.No.	Algorithm	Strength	Weakness
1	LDA[53]	It's easy to use, quick, and portable. With assumptions met, it outperforms some algorithms.	That standard distribution assumption is required for features and predictions. For a few types of variables, this isn't always the case.
2	PLS[36]	Because knowledge of the ingredients of interest is required, it can be utilized for exceedingly complicated combinations. It can forecast samples containing elements not present in the calibration mixtures.	For appropriate calibration, a significant number of samples are necessary. It can be challenging to collect calibration samples to avoid collinear ingredient concentrations.
3	PLS-R [37]	Simple to comprehend and compute. Calculations are made in a relatively short amount of time. They are typically used for simple samples, pure compounds, and binary mixes.	It cannot be utilized for complicated mixture samples with overlapping spectral bands between separate elements. Isolated spectral bands entirely relevant to the element of interest are required.
4	PLS-DA [3]	Acceptance of very collinear data.	inability to preserve local data structure
5	OPLS-DA [3,37]	Separate analyses of orthogonal variation are possible.	-------

7.1.2 Nonlinear statistical techniques

Nonlinear techniques are usually more accurate. However, most methods interpret the outcomes and are computationally intensive to calculate. Nonlinear techniques are used when linear methods fail to handle complex biological systems.

Nonlinear statistical learning methods include tree, neural network, and kernel methods. These techniques are usually more precise in their prediction accuracy. However, most methods have a complex understanding of the results and are computationally concentrated to calculate—an approach based on trees, neural networks, and kernel-based nonlinear statistics.

A. K- Nearest Neighbor (KNN)

NN classifiers are well-suited to classification tasks in which the relationships between characteristics and target classes are numerous, convoluted, or otherwise difficult to comprehend.

The KNN algorithm uses the NN technique for classification. This algorithm starts with a training dataset of samples categorized into many categories using a nominal variable.

$$\boldsymbol{dist(p,q)} = \sqrt{(\boldsymbol{p_1}-\boldsymbol{q_1})^2 + (\boldsymbol{p_2}-\boldsymbol{q_2})^2 + \cdots + (\boldsymbol{p_n}-\boldsymbol{q_n})^2} \quad \textbf{(9)}$$

The explanation of the above formula is given here. The distance can be calculated in a variety of ways. For example, KNN uses Euclidean distance. The Euclidean distance formula is as follows:P and q are the cases to be compared, each with n characteristics. The first characteristic value, p, is referred p1, and the first feature value, q, is referred q1.[3,51]

B. Decision Tree Classification

Decision trees are machine learning algorithms. This algorithm uses conversion into smaller segments to reach its particular pattern. The decision tree is a type of flowchart. It is applicable or appropriate for two kinds of situations. One is a situation where the system's classification is needed for legal reasons. The second situation is related to decision-making and is based on clear findings. For example, this algorithm makes medical diagnoses based on lab results, symptoms, or disease progression. The C5.0 algorithm is the most well-known decision tree algorithm. The algorithm calculates the change inhomogeneity caused by a split on each feasible characteristic using entropy. The following is the definition of entropy:

$$\boldsymbol{Entropy(s)} = \sum_{i=1}^{c} -\boldsymbol{p_i log_2(p_i)} \qquad \textbf{(10)}$$

The explanation of the above formula is given here. The first challenge a decision tree must overcome is determining which characteristic to split. The C5.0 measures purity using entropy. The entropy of a data sample reveals how mixed the class values are; a value of 0 indicates that the sample is homogeneous, while a value of 1 indicates that the sample is entirely disordered. [26,34]

The data is separated into many partitions after this split. Therefore, the entropy calculation function (s) must consider the total entropy across all partitions. It accomplishes this by dividing the entropy of each partition by the proportion of records that fall into that division. It can be expressed mathematically as:

$$\boldsymbol{Entropy(s)} = \sum_{i=1}^{n} \boldsymbol{w_i\, Entropy\,(p_i)} \qquad \textbf{(11)}$$

The overall entropy of a split is the sum of entropy. This sum of entropy is derived from each of the n divisions. Then, these divisions are weighed and valued by the proportion of the cases found in that pattern. [3]

C. Decision Tree Regression

The decision tree regression algorithm uses statistics to provide ways for calculating mathematical connections among data pieces. Regression defines the relationship between independent numeric variables and a single numeric dependent variable. This technique can forecast numeric data and measure an outcome's and its predictors' size and intensity. [34]

The decision tree regression is used for three different purposes. These purposes are one- complex modelling relationships among data variables, two- evaluating a treatment's influence on an outcome, and three-extrapolating into the future.

Straight-line regression models are the simplest basic regression models. Simple linear regression is used when there is only one independent variable; otherwise, multiple regressions are used. The dependent variable in both of these models is assumed to be continuous. The formula for simple linear regression is-

$$Y = \alpha + \beta x \qquad \textbf{(12)}$$

The explanation of this formula is given here – It is determined what the best estimates of α and β are. This determination is made using the ordinary least squares (OLS) estimation method. An OLS regression's slope and intercept are chosen to minimize the sum of squared errors or the vertical distance between predicted and actual Y values. [27]

The goal of OLS regression can be expressed mathematically as the task of minimizing the following equation-

$$\sum(\boldsymbol{y_i}-\boldsymbol{\hat{y}_i})^2 = \sum \boldsymbol{e_i^2} \qquad (13)$$

This equation defines the difference between the actual y and projected y values as e (the error). The error values are squared and averaged over all data points. A number that reflects how closely two variables' relationships follow a straight line is called the correlation. Pearson's correlation coefficient is referred to as correlation without a further qualifier. The correlation is in the range of -1 to +1. A correlation close to 0 implies the absence of a linear link, while extreme values suggest a linear relationship. [52]

The following formula defines Pearson's correlation:

$$\boldsymbol{p_{xy}} = \boldsymbol{corr(x,y)} = \frac{\boldsymbol{cov\,(x,y)}}{\boldsymbol{\sigma_x \sigma_y}} \qquad (14)$$

D. Probabilistic Learning- Classification using Naïve Bayes

The Naive Bayes (NB) algorithm provides a basic classification application based on Bayes' theorem. Bayesian classifiers have been employed for: Diagnosing medical diseases. Naive Bayes is a classification method that also employs probability principles. Bayesian classifiers use training data to determine the observed probability of each class based on feature values. Bayesian classifiers are best suited to situations requiring simultaneous consideration of multiple attributes to assess the likelihood of a given result.

$$\boldsymbol{P}\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)p(A)}{P(B)} = \frac{P(A \cap B)}{P(B)} \qquad (15)$$

The explanation of the above formula is given below. The notation P (A/B) denotes the likelihood of occurring if event B has already occurred. Because A's chance is reliant (i.e., conditional) on what happens with event B, this is known as conditional probability. Hence, it is known as the prior probability. Bayes' theorem is used to calculate posterior probability. It is a standard way of estimation. [3, 27,34]

E. Neural Network (NN)

An ANN network is preferred to solve practical problems in the automation of intelligent devices, various tasks based on accurate data, and factual data where input and output are simple. An ANN uses a model developed from our understanding of how the human brain responds to stimuli from sensory inputs to describe the relationship between a set of input signals and an output signal. ANNs are so complex that they are frequently used to automate intelligent devices like office building environmental controls, self-driving automobiles, and self-piloting drones. ANNs are adaptable learners who can perform various tasks such as classification, numerical prediction, and even unsupervised pattern recognition. ANNs work best when the input and output data are well-understood or, at the very least, simple, but the process that connects them is highly complicated.

$$\boldsymbol{y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right)} \qquad \mathbf{(16)}$$

Each n input (x) might contribute significantly or less to the input signals due to the w weights. So the activation function f(x) uses the net total, and the output axon is y(x). [3, 27,34]

F. Support Vector Machines (SVM)

The SVM algorithm aims to find a line that divides the two classes, to create a flat boundary, or hyperplane, with reasonably homogeneous data partitions on both sides. SVMs can be used for classification and numerical prediction. In pattern recognition, SVM is commonly used. Microarray classification and gene expression data are applications to diagnose cancer or other genetic illnesses. Support vectors give a relatively compact approach to storing a classification model, even if the number of features is vast, a significant property of SVMs.

The SVM uses the kernel trick to separate data into higher dimensional space. It is stated as follows-

$$\boldsymbol{k}(\vec{x_i}, \vec{x_j}) = \emptyset(\vec{x_i}) \times \emptyset(\vec{x_j}) \qquad \mathbf{(17)}$$

Where$\emptyset$ (x) is a function that combines the feature vectors x_i and x_j.[52,53]

G. Random Forests (RF)

Random forests (or decision tree forests) are an ensemble-based approach focusing solely on decision tree ensembles. Random forests blend variety and power into a single machine learning approach. Random forests can manage vast datasets because the ensemble uses the fall feature set. Random forest, for example, is easier to use and less prone to overfitting. As a result, this algorithm has potency, versatility, and ease of use. [54] Table 3 shows the strength and weaknesses of the supervised nonlinear statistical techniques.

Table 3:Comparison of supervised nonlinear statistical techniques.

Sr. No.	Algorithm	Strengths	Weaknesses
1	KNN [3]	Easy to use and effective Quick training phase	Slow classification phase Memory intensive Nominal characteristics and insufficient data necessitate extra processing
2	Decision Tree- Classification [27]	It can be used on data with a small or large number of training examples. Creates a model that anyone can understand (for relatively small trees)	Decision tree models frequently favour splits on features with many levels.

		More efficient than other sophisticated models	Large trees can be challenging to understand, and their decisions may appear arbitrary.
3	Decision Tree-Regression [27]	Combines decision tree capabilities with the ability to represent numeric data Performs automatic feature selection, allowing the approach to employ many features. It may be a better match for some types of data than linear regression It does not require statistical understanding to interpret the model	Requires a considerable amount of training data Difficult to establish the total net influence of distinct attributes It may be more difficult to interpret than a regression model
4	Naïve Bayes [53]	Easy to get the estimated probability for a prediction Does well with noisy and missing data Simple to obtain the estimated probability for a prediction	It relies on the frequently incorrect assumption that all features are equally valuable and independent. Predicted probabilities are less trustworthy than estimated probabilities.
5	Neural Network [53]	Able to model more complicated patterns	Multi-collinearity is a risk.
6	Support Vector Machine [3,53]	High accuracy, but not overly sensitive to noisy data or overfitting User-friendly due to the availability of various well-supported SVM algorithms	Finding the optimum model necessitates experimenting with various kernel combinations and model parameters.
7	Random Forest [34]	A general-purpose model that works well in a variety of situations Only the most essential aspects are chosen. It can be applied to data with many features or examples.	It may take some time to fine-tune the model to the data.
8	K-means clustering [34]	It is highly adaptable, and simple changes can be made to fix nearly all of its flaws. It uses simple principles for recognizing clusters that can be stated in non-statistical terms. It's relatively efficient and does an excellent job of grouping data into usable clusters.	Finding the optimal collection of clusters is not guaranteed because it relies on random chance. Requires a fair estimate of how many clusters exist in the data organically.

7.2 Unsupervised Analysis

The unsupervised analysis is the second method of statistical analysis. In this method, the data which is used is unlabeled. The most common algorithms used for breath analysis are Principal component analysis and independent component analysis; these are used for the association rule learner method. In addition, the Hierarchical clustering analysis, K-means, and Fuzzy k-means three are used for the clustering method.

A. Principal Component Analysis (PCA)

The fundamental goal of the Principal component analysis method is to compress data that captures as much variance as possible by using a few latent variables. The goal is to approximate the original data matrix using the fewest possible latent components and loading vectors. It aims to develop new orthogonal features called principal components (PCs) to represent data variance. The PCs explain the data variance and are ideal for visualization and modelling. PCA offers the most efficient data compression of all linear techniques. [28]

PCA is an unsupervised technique used in chemometrics. The PCs are mutually orthogonal linear combinations of the data variables. A few initial PCs are sufficient to accurately represent data variation, allowing display of the data structure.

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} \qquad (18)$$

PCA decomposes the original data into scores, T, and a loadings matrix, p. The PCA is frequently used as the first stage in data analysis, and the PCs are frequently used as input data for additional procedures. [29]

B. Independent Component Analysis (ICA)

ICA is a type of PCA in which the components are believed to be statistically independent of one another rather than just uncorrelated. Independent component analysis uses sparse coding to extract features and compress data. ICA has become a standard method for analyzing multi-variant data. ICA is a probabilistic approach for learning a random vector's linear transform. The goal is to find non-Gaussian, maximally independent (non-normal) components. [30]

The mixing model is stated as follows in this vector-matrix notation:

$$X = AS \qquad (19)$$

Let's consider X the random vector whose elements are the mixtures X_1,... X_n, and S the random vector whose elements are S_1,... S_n. Let us denote by A is a row vector the matrix with elements A_{ij}. [28]

Independent component analysis (ICA) is the statistical model used in the preceding equation. The random vector X is used to estimate both A and S.

[31,32]

Clustering Method

Cluster analysis is the process of grouping individuals together to find patterns in data. Clustering is used to uncover knowledge rather than predict it. It is essentially a collection of data exploration techniques. It is a set of techniques for determining if data contains natural groupings. This algorithm can be employed to produce dissections. The most common algorithms used for clustering are Hierarchical cluster analysis, k-means clustering, and fuzzy k-means clustering. [26]

C. Hierarchical Cluster Analysis (HCA)

Hierarchical clustering is based on sets where different sets are grouped to form different levels of clusters. This algorithm has conceptual simplicity. It is used for summarizing data structure. The procedures used in this algorithm are classified into two types: agglomerative and divisive procedures. Agglomerative (bottom-up, clumping) processes begin with n singleton clusters and merge in sequential order. Divisive (top-down, splitting) procedures begin with all samples in one cluster and build the sequence by splitting clusters one at a time. [33]

D. K-means clustering

The k-means algorithm aims to divide the data into k clusters with the lowest within-group sum of squares. The k-means algorithm includes assigning each n example to one of the k clusters, with k being a predetermined number. The goal is to maximize differences between clusters while minimizing differences within each cluster.

K-means use Euclidean distance; however, Manhattan and Minkowski distances are frequently used.

$$\boldsymbol{dist}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{y}_i)^2} \qquad (20)$$

The formula for Euclidean distance between examples x and y is above, where n is the number of features. The k-means algorithm uses two different alternating grouping object and assignment base calculation methods. [27,34]

E. Fuzzy k-means

Data is separated into different clusters, with each data element belonging to only one cluster. Data are clustered via fuzzy clustering. Each element level has a set of memberships belonging to multiple clusters. These numbers represent the strength of association as a link between that data element and a specific fuzzy cluster. The process of allocating these memberships is known as clustering levels and assigning data pieces to one or several clusters.

$$\boldsymbol{m}_j = \frac{\sum_{i=1}^{n} y_{ji}^{r} x_i}{\sum_{i=1}^{n} y_{ji}^{r}} \qquad (21)$$

The formula is explained here. The fuzzy k-means or c-means method finds a solution for the y_{ji} parameters. The degree of association or membership function of the i^{th} pattern or object with the j^{th} group is represented by the parameter y_{ji}.

The weighting exponent, or r, is a scalar that regulates the resulting 'fuzziness' clusters ($r \geq 1$), and m_j is the 'centroid' of the j^{th} group. [26]

A value of r=1 gives the same problem as the nonlinear optimization scheme. The basic algorithm is iterative. First, select r;

initialize the membership function values $y_{ji,}$ then compute the cluster centers m_j. Then the membership function if d_{il}=0 for some l, y_{li}=1, and y_{ji}=0 for all j≠l; otherwise

$$y_{ji} = \frac{1}{\sum_{k=1}^{g}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{r}-1}} \qquad (22)$$

If not converged, then again computes the cluster centers m_j. [35]

8. RESULTS AND DISCUSSION

The algorithms discussed in this paper have demonstrated the value of feature extraction and selection appropriate for the data being preprocessed and used to train the algorithm. Based on multidimensional data collected from exhaled breath analyzers, all algorithms presented in this paper exhibit very high performance for disease detection tasks. Figure 2 compares selected algorithms used to detect COPD, Diabetes and TB in exhaled breath from patient samples. Machine learning algorithms are used for disease detection. This paper represents a graph of three different diseases. This graph represents the specificity and sensitivity of COPD [55-59], Diabetes [65-68] and TB[60-64] using machine learning algorithms.

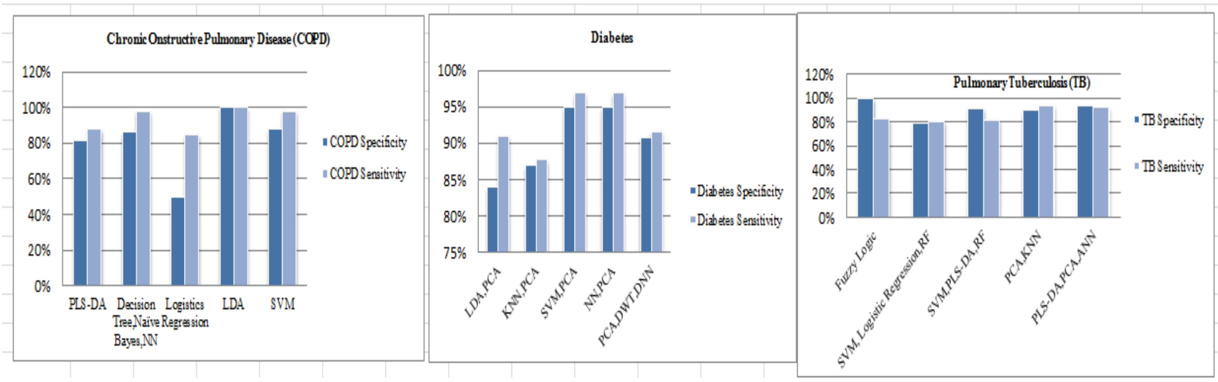

Figure 2. Shows the comparison of selected algorithms used for COPD, Diabetes and TB disease detection

Anand thati et al.[69] proposed a system for estimating the acetone concentration in the exhaled breath. A TGS 822 SnO_2 sensor was used to detect the concentration of acetone. For feature extraction, an artificial neural network algorithm was used. The training and testing range ranges between 80 mg/dl and 180 mg/dl. S.C. van Beek et al.[70] studied 21 pulmonary tuberculosis patients and 50 non-tuberculosis persons to determine the biomarkers of this disease. Support vector machine algorithm was used for analysis and achieved an accuracy of 77%, sensitivity of 62%, and specificity of 84%. E.I.Mohamed et al.[64] developed E-nose with a chemical sensor for TB detection. The breath samples of 260 TB patients and 240 healthy persons were collected. PLS-DA, PCA and ANN were used to recognize the TB patients with 99% accuracy, 92.08% and specificity of 93.5%.

Peter J Mazzone et al.[72] used an E-nose with a colorimetric sensor to analyze the breath samples collected from 92 affected by lung cancer and 137 healthy persons samples. The logistics regression and Pearson's chi-square found 81% of accuracy. In 2007, Peter J Mazzone et al.[71] used the GC-MS technique to identify breath patterns. This study includes 49 lung cancer patients, 18-COPD patients, 20- sarcoidosis patients and 21 healthy persons. The investigation found that VOCs pattern using random forest, colorimetric sensor array and carbon polymer sensor with 82% accuracy.

In 2005, Roberto F. Machado et al.[73] reported their work on lung cancer detection by using an E-nose with 32-polymer composite sensors. They analyzed 76 persons, of which 62 as healthy and 14 had lung cancer. The SVM, PCA and Canonic discriminant analysis were used to classify cancer samples. The result showed 85% of accuracy. Agnieszka smolinska et al.[74] reported a study about the GC-MS to detect asthma. Random forests, PLS-DA was carried out to analyze the signals. The clinical test result showed the distinction between the 252 breath samples, including normal subjects, with a total classification rate of 74.6%.

In 2010, Paolo Montuschi et al.[75] developed GC-MS with E-nose and chemical sensor array for asthma detection. Breath samples of 27 asthma and 24 normal subjects were collected. The PCA and NN were used to recognize the asthma patients. The result showed a classification rate of 87.5%. Using E-nose, N. fens et al.[76] analyzed the exhaled breath samples to discriminate between patients with fixed asthma with COPD, classic asthma with COPD and healthy persons. The breath samples were collected from 60 asthma, 21 positive asthma, 39 magnificent, and 40 COPD. The method could distinguish between fixed asthma with COPD with 88% accuracy, 90% specificity, and 88% sensitivity and classic asthma with COPD with 83% accuracy, 90 % specificity and 91% sensitivity.

9. SUMMARY

This paper summarizes different sensors and data preprocessing methods of GC-MS, IMS, SIFT-MS, GC-DMS, and GC-tof-MS. Normalization, scaling, and transformation are common data preprocessing techniques applied according to the methods used for breath analysis. For data analysis, in the broader context of breath biomarkers and their analysis, this paper focused on different machine learning algorithms used for this analysis. This paper also presents two main aspects of algorithms- the algorithm's goal and its primary technical

formula used in its applications, results and discussion of some existing research. Breath biomarker discovery in combination with machine learning techniques increases medicine's contribution. However, these algorithms' applications may be applied with other advanced techniques to solve real ground issues and assist in the extensive data processing.

REFERENCES

[1] Zhang David, Guo Dongmin, Yan Ke. Breath analysis for medical applications. Springer Nature Singapore Pte Ltd; 2017.

[2] Popov Todor A. Human exhaled breath analysis. CME review, Ann Allergy Asthma Immunol 2011; 106:451–6.

[3] Agnieszka Smolinska, Anne-Christin Hauschild, Jan W Dallinga. Current breathomics - A review on data preprocessing techniques and machine learning in metabolomics breath analysis. Journal of Breath Research 8(2):027105; April 2014.

[4] Kim K-H, Jahan Shamin Ara, Kabir Ehsanul. A review of breath analysis for diagnosis of human health. Trends Anal Chem 2012;33.

[5] Moorhead K T et al 2008 Classifying algorithms for SIFT-MS technology and medical diagnosis Comput. Methods Programs Biomed. 89 226–38.

[6] Yun Sun, Yibing Chen, Chuanqiang Sun, Haipei Liu, Yan Wanga, Xuehui Jiang, Analysis of volatile organic compounds from patients and cell lines for the validation of lung cancer biomarkers by proton-transfer-reaction mass spectrometry, Analytical Methods, The Royal Society of Chemistry 2019.

[7] Rosa Alba Sola Martínez, José María Pastor Hernández, Gema Lozano Terol, Julia Gallego-Jara, Luis García-Marcos, Manuel Cánovas Díaz & Teresa de Diego Puente, Data preprocessing workflow for exhaled breath analysis by GC/MS using open sources, Scientific Reports volume 10, Article number: 22008 (2020).

[8] Hauschild A-C, Baumbach J I and Baumbach J 2012 Integrated statistical learning of metabolic ion mobility - spectrometry profiles for pulmonary disease identification Genet. Mol. Res. 11 2733–44.

[9] Hugues Stefanuto Pierre, Zanella Delphine, Vercammen Joeri, Henket Monique, Florence Schleich, Louis Renaud, Francois Focant Jean. Multimodal combination of GC x GC-HRTOFMS and SIFT-MS for asthma phenotyping using exhaled breath. www.nature.com/scientificreports 2020.

[10] Basanta Maria, Jarvis Roger M, Xu Yun, Blackburn Gavin, Tal-Singer Ruth, Woodcock Ashley, Singh Dave, Goodacre Royston, Paul Thomas CL, Fowler Stephen J. Non-invasive metabolomic analysis of breath using differential mobility spectrometry in patients with chronic obstructive pulmonary disease and healthy smokers. R. Soc. Chem. 2010;135:315–20. Analyst, 2010.

[11] Van Vilet Dillys, Smolinska Agnieszka, Jobsis Quirjin, Rosias Philippe, et al. Can exhaled volatile organic compounds predict asthma exacerbations in children. J. Breath.Res. 2017;11(1). https://doi.org/10.1088/1752-7163/aa5a8b.

[12] Smolinska Agnieszka, Ester M, Klaassen M, Dallinga Jan W, van de Kant Kim DG, Jobsis Quirijn, Edwin J, Moonen C, van Schayck Onno CP, Dompeling Edward, Frederik J, van Schooten. Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children. PloS One 2014;9(4):e95668.

[13] Ronny Schnabel, Rianne Fijten, Agnieszka Smolinska, Jan Dallinga, Marie-Louise Boumans, Ellen Stobberingh, Agnes Boots, Paul Roekaerts, Dennis Bergmans & Frederik Jan van Schooten, Analysis of volatile organic compounds in exhaled breath to diagnose ventilator-associated pneumonia, Scientific Reports volume 5, Article number: 17179 (2015).

[14] Robert A van den Berg, Huub CJ Hoefsloot, Johan A Westerhuis, Age K Smilde and Mariët J van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 2006, 7:142.

[15] Frank Dieterle, Alfred Ross, Go1 tz Schlotterbeck, and Hans Senn, Probabilistic Quotient Normalization as

Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics, Anal. Chem. 2006, 78, 4281-4290.

[16] Smilde A K, van der Werf M J, Bijlsma S, van der Werff-van der Vat B J and Jellema R H 2005 Fusion of mass spectrometry-based metabolomics data Anal. Chem. 77 6729–36.

[17] van den Berg R A, Hoefsloot H C J, Westerhuis J A, Smilde A K and van der Werf M J 2006 Centering, scaling, and transformations: improving the biological information content of metabolomics data BMC Genomics 7 142.

[18] Meloun M and Militky J 1994 Computer-assisted data treatment in analytical chemometrics: 3. Data transformation Chem. Pap-Chem. Zvesti. 48 164–9.

[19] Tharwat, Alaaa, Gaber,Tarekc, Ibrahim,Abdelhameedd, Hassanien,Aboul Ella, Linear discriminant analysis: A detailed tutorial, AI Communications, vol. 30, no. 2, pp. 169-190, 2017.

[20] Mario Fordello, Andrea Bellincontro, Fabio Mencarelli, Partial least squares discriminant analysis: a dimensionality reduction method to classify hyper spectral data, Statistica Applicata - Italian Journal of Applied Statistics Vol. 31 (2).

[21] Richard G. Breretona , Gavin R. Lloyd, Partial least squares discriminant analysis: taking the magic away, Published online in Wiley Online Library: 18 March 2014.

[22] SvanteWolda, Michael Sjöströma,Lennart Eriksson, PLS-regression: a basic tool of chemo metrics, Chemometrics and Intelligent Laboratory Systems,Volume 58, Issue 2, 28 October 2001, Pages 109-130.

[23] Herve Abdi ,Partial least squares (PLS) regression, Program in Cognition and Neurosciences, MS: Gr.4.1.

[24] Agnar Höskuldsson ,PLS Regression Methods, Wiley Analytical Science,Journal of Chemometrics.

[25] Hans Stenlund, András Gorzsás, Per Persson, Björn Sundberg, Johan Trygg, Orthogonal projections to laten structures discriminant analysis modeling on in situ FT-IR spectral imaging of liver tissue for identifying sources of variability, Analytical Chemistry 80(18):6898-906, September 2008.

[26] Webb Andrew 2002 Statistical Pattern Recognition (New York: Wiley).

[27] Chung-Yu Chen, Wei-Chi Lin & Hsiao-Yu Yang, Diagnosis of ventilator-associated pneumonia using electronic nose sensor array signals: solutions to improve the application of machine learning in respiratory research, Respiratory Research volume 21, Article number: 45 (2020).

[28] Michał Daszykowski, From projection pursuit to other unsupervised chemometric techniques, JOURNAL OF CHEMOMETRICS, 2007;21: 270–279.

[29] M.Daszykowski, K.Kaczmarek,Y.Vander Heyden.Walczak, Robust statistics in data analysis- A review: Basic Concepts, Chemometrics and Intelligent Laboratory Systems, Volume 85, Issue 2, 15 February 2007, Pages 203-219.

[30] Aapo Hyvärinen, Independent component analysis: recent advances, Philos Trans A Math Phys Eng Sci. 2013, 371(1984): 20110534.

[31] A.Hyvärinen,E.Oja, Independent component analysis: algorithms and applications, Neural Networks, Volume 13, Issues 4–5, June 2000, Pages 411-430.

[32] Dominic Langlois, Sylvain Chartier, Dominique Gosselin, An Introduction to Independent Component Analysis: InfoMax and FastICA algorithms, March 2010Tutorials in Quantitative Methods for Psychology 6(1).

[33] Angela Serra, Roberto Taliaferro, Unsupervised Learning and clustering, Reference Module in Life Sciences, January 2018.

[34] Lantz B. Machine Learning with R. 2nd ed. Birmingham, UK: Packt Publishing Ltd.; 2015.

[35] Francisco de A.T. de Carvalho, Camilo P. Tenório, Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances, Fuzzy Sets and Systems 161 (2010) 2978–2999.

[36] Frederic Cadet, Miguel de la Guardia, Quantitative Analysis infrared, Encyclopedia of Analytical Chemistry, September 2006.

[37] Muhammad Aminu, Noor Atinah Ahmad, Complex Chemical Data Classification and Discrimination using Locality Preserving PLS-DA, ACS Omega 5(41):26601-26610, October 2020.

[38] Anish Singh Shekhawat, Arnav Jain, Dipti Patil, A Study of ECG Steganography for Securing Patient's Confidential Data based on Wavelet Transformation, International Journal of Computer Applications (0975 – 8887) Volume 105 – No. 12, November 2014.

[39] Shamla Mantri, Vidya Dukare, Smita Yeole, Dipti Patil, V. M. Wadhai, A Survey: Fundamental of EEG,

International Journal of Advance Research in Computer Science and Management Studies ISSN: 2321-7782 (Online), Volume 1, Issue 4, September 2013.

[40] Dipti Patil, Tejashree Chhajed,V. M. Wadhai, Prasad Pomaji, Abhinav Sharma, Bhagyashri Samanta, A Comparative Study of Traditional and Mobile based ECG System Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 41– No.3, March 2012.

[41] Sharwin P. Bobde, Shamla T. Mantri, Dipti D. Patil, Vijay Wadhai, Cognitive Depression Detection Methodology Using EEG Signal Analysis, Intelligent Computing and Information and Communication pp 557-566, Springer link 2018.

[42] Dipti D. Patil, Shamla Mantri, Himangi Pande V.M.Wadhai, M.U.Kharat, Feature Extraction Techniques for Mining ECG Signals in WBAN for Healthcare Applications, International Journal of Advances in Computing and Information Researches, ISSN: 2277-4068, Volume 1– No.1, January 2012.

[43] Shamla Mantri, Dr. Dipti Patil, Dr. Pankaj Agrawal, Dr. Vijay Wadhai, Non Invasive EEG Signal Processing Framework for Real Time Depression Analysis, SAI. Intelligent Systems Conference 2015, November 10-11, 2015 | London, UK.

[44] Maria Kaloumenou,Evangelos Skotadis, Nefeli Lagopati, Efstathios Efstathopoulos, and Dimitris Tsoukalas, Review Breath Analysis: A Promising Tool for Disease Diagnosis—The Role of Sensors, Sensors 2022, 22, 1238. https://doi.org/10.3390/s22031238.

[45] Kaushiki Dixit, Somayeh Fardindoost, Adithya Ravishankara, Nishat Tasnim and Mina Hoorfar, Review Exhaled Breath Analysis for Diabetes Diagnosis and Monitoring: Relevance, Challenges and Possibilities, Biosensors 2021, 11, 476.https://doi.org/10.3390/bios11120476.

[46] Maria Vesna Nikolic, Vladimir Milovanovic, Zorka Z. Vasiljevic, Zoran Stamenkovic, Semiconductor Gas Sensors: Materials, Technology, Design, and Application, Sensors 2020, 20(22), 6694; https://doi.org/10.3390/s20226694.

[47] Andreas T. Güntner, Sebastian Abegg, Karsten Königstein, Philipp A. Gerber, Arno Schmidt-Trucksäss, and Sotiris E. Pratsinis, “Breath Sensors for Health Monitoring”, ACS Sens. 2019, 4, 2, 268–280.

[48] Artur Rydosz, Sensors for Enhanced Detection of Acetone as a Potential Tool for Noninvasive Diabetes Monitoring, Sensors (Basel). 2018 Jul; 18(7): 2298. Published online 2018 Jul 16. doi: 10.3390/s18072298.PMCID:PMC6068483PMID: 30012960.

[49] Milua Masikini, Mahabubur Chowdhury, Ouassini Nemraoui, Review—Metal Oxides: Application in Exhaled Breath Acetone Chemiresistive Sensors, Journal of the Electrochemical Society, 2020 167 037537.

[50] Nidheesh V. R.,Aswini Kumar Mohapatra,Unnikrishnan V. K.,Rajeev Kumar Sinha,Rajesh Nayak,Vasudevan Baskaran Kartha & Santhosh Chidangil, Breath analysis for the screening and diagnosis of diseases, Volume 56, 2021 - Issue 8-10: Special issue of Applied Spectroscopy Reviews on Clinical Applications of Spectroscopy.

[51] Sagnik Das and Mrinal Pal, Review—Non-Invasive Monitoring of Human Health by Exhaled Breath Analysis: A Comprehensive Review, 2020 J. Electrochem. Soc. 167 037562.

[52] Gopi Battineni, Getu Gamo Sagaro, Nalini Chinatalapudi, Francesco Amenta, Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis, Journal of Personalized Medicine, 2020 Mar 31;10(2):21. doi: 10.3390/jpm10020021.

[53] Anna Paleczek, Artur Rydosz, Review of the algorithms used in exhaled breath analysis for the detection of diabetes, Journal of Breath Research, 2022 Jan 26; 16(2). doi: 10.1088/1752-7163/ac4916.

[54] Yuichi Sakumura, Yutaro Koyama, Hiroaki Tokutake, Toyoaki Hida, Kazuo Sato, Toshio Itoh, Takafumi Akamatsu and Woosuck Shin, Diagnosis by Volatile Organic Compounds in Exhaled Breath from Lung Cancer Patients Using Support Vector Machine Algorithm, www.mdpi.com/journal/sensors 2017

[55] Maria Basanta, Roger M. Jarvis,,Yun Xu,Gavin Blackburn,,Ruth Tal-Singer,Ashley Woodcock, Dave Singh, Royston Goodacre, C. L. Paul Thomas ,Stephen J. Fowler," Non-invasive metabolomic analysis of breath using differential mobility spectrometry in patients with chronic obstructive pulmonary disease and healthy smokers", The Royal Society of Chemistry 2010 Analyst, 2010, 135, 315–320.

[56] A.-C. Hauschild1, J.I. Baumbach and J. Baumbach," Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification", Genetics and Molecular Research 11 (3): 2733-2744 (2012).

[57] Maria Basanta, Baharudin Ibrahim, Rachel Dockry, David Douce, Mike Morris, Dave Singh,Ashley Woodcock, Stephen J Fowler," Exhaled volatile organic compounds for phenotyping chronic obstructive pulmonary

disease: a cross-sectional study", Respiratory Research 2012, 13:72.

[58] Akira D.M. Hattesohl, Rudolf A. Jörres, Holger Dressel,Severin Schmid, Claus Vogelmeier, Timm Greulich, Sarah Noeske, Robert Bals ,Andreas Rembert Koczulla," Discrimination between COPD patients with and without alpha 1-antitrypsin deficiency using an electronic nose", Respirology (2011) 16, 1258–1264.

[59] J.J.B.N. Van Berkel, J.W. Dallinga, G.M. Mo¨ller, R.W.L. Godschalk, E.J. Moonen, E.F.M. Wouters, F.J. Van Schooten," A profile of volatile organic compounds in breath discriminates COPD patients from controls", Respiratory Medicine (2010) 104, 557e563.

[60] Michael Phillips, Renee N. Cataneo, Rany Condos, Gerald A. Ring Erickson, Joel Greenberg, Vincent La Bombardi, Muhammad I. Munawar, Olaf Tietje, "Volatile biomarkers of pulmonary tuberculosis in the breath", Tuberculosis (2007) 87, 44–52, Elsevier.

[61] Amandip S. Sahota, Ravi Gowda, Ramesh P. Arasaradnam, Emma Daulton, Richard S. Savage, Jim R. Skinner, Emily Adams, Stephen A. Ward, James A. Covington, "A simple breath test for tuberculosis using ion mobility: A pilot study", Tuberculosis 99 (2016) 143e146, Elsevier.

[62] Marco Beccaria, Carly Bobak, Boitumelo Maitshotlo, Theodore R Mellors, Giorgia Purcaro, Flavio A Franchina, Christiaan A Rees, Mavra Nasir, Andrew Black, and Jane E Hill, "Exhaled human breath analysis in active pulmonary tuberculosis diagnostics by comprehensive gas chromatography-mass spectrometry and chemometric techniques", J Breath Res. ; 13(1): 016005, PMC 2019 May 28.

[63] Nicola M. Zetola, Chawangwa Modongo, Ogopotse Matsiri, Tsaone Tamuhla, Bontle Mbongwe, Keikantse Matlhagela, Enoch Sepako, Alexandro Catini, Giorgio Sirugo, Eugenio Martinelli, Roberto Paolesse, Corrado Di Natale, "Diagnosis of pulmonary tuberculosis and assessment of treatment response through analyses of volatile compound patterns in exhaled breath samples", Journal of Infection,2016, Elsevier.

[64] E. I. Mohamed,M. A. Mohamed, M. H. Moustafa, S. M. Abdel-Mageed, A. M. Moro, A. I. Baess, S. M. El-Kholy, "Qualitative analysis of biological tuberculosis samples by an electronic nose-based artificial neural network", Int J Tuberc Lung Dis 21(7):810–817 Q 2017 The Union.

[65] Siegel A P, Daneshkhah A, Hardin D S, Shrestha S, Varahramyan K and Agarwal M 2017 Analyzing breath samples of hypoglycemic events in type 1 diabetes patients: towards developing an alternative to diabetes alert dogs J. Breath Res. 11 026007.

[66] Guo D, Zhang D, Li N, Zhang L and Yang J 2010 A novel breath analysis system based on electronic olfaction IEEE Trans. Biomed. Eng. 57 2753–63.

[67] Lekha S and Suchetha M 2017 Real-time non-invasive detection and classification of diabetes using modified convolution neural network IEEE J. Biomed. Health Inform. 22 1630–6.

[68] Sarno R, Sabilla S I and Wijaya D R 2020 Electronic nose for detecting multilevel diabetes using optimized deep neural network Eng. Lett. 28 31–42.

[69] Anand Thati , Arunangshu Biswas , Shubhajit Roy Chowdhury and Tapan Kumar Sau, Breath Acetone-Based Non-Invasive Detection Of Blood Glucose Levels, International Journal On Smart Sensing And Intelligent Systems Vol. 8, No. 2, June 2015.

[70] S. C. Van Beek, N. V. Nhung, D. N. Sy, P. J. Sterk, E. W. Tiemersma, F. G. J. Cobelens, "Measurement of exhaled nitric oxide as a potential screening tool for pulmonary tuberculosis", The International Journal of Tuberculosis and Lung Disease, 15(2):185–191© 2011 The Union.

[71] Peter J Mazzone, Jeffrey Hammel, Raed Dweik, Jie Na, Carmen Czich, Daniel Laskowski, Tarek Mekhail, "Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array" Thorax. 2007 Jul; 62(7): 565–568.

[72] Peter J Mazzone, Xiao-Feng Wang, Yaomin Xu, Tarek Mekhail, Mary C Beukemann, Jie Na, Jonathan W Kemling, Kenneth S Suslick, Madhu Sasidhar, "Exhaled Breath Analysis with a Colorimetric Sensor Array for the Identification and Characterization of Lung Cancer" J Thorac Oncol. 2012 January ; 7(1): 137–142.

[73] Roberto F. Machado, Daniel Laskowski, Olivia Deffenderfer, Timothy Burch, Shuo Zheng, Peter J. Mazzone, Tarek Mekhail, Constance Jennings, James K. Stoller, Jacqueline Pyle, Jennifer Duncan, Raed A. Dweik, and Serpil C. Erzurum "Detection of Lung Cancer by Sensor Array Analyses of Exhaled Breath" American Journal of Respiratory and Critical Care Medicine, Volume 171, Issue 11, pp 1286–1291, 2005.

[74] Agnieszka Smolinska, Ester M. M. Klaassen, Jan W. Dallinga, Kim D. G. van de Kant, Quirijn Jobsis, Edwin J. C. Moonen, Onno C. P. van Schayck, Edward Dompeling, Frederik J. van Schooten, "Profiling of Volatile Organic Compounds in Exhaled Breath As a Strategy to Find Early Predictive Signatures of Asthma in

Children" Plos One, Volume 9, Issue 4, e95668, April 2014.

[75] Paolo Montuschi, Marco Santonico, Chiara Mondino, Giorgio Pennazza, Giulia Mantini, Eugenio Martinelli, Rosamaria Capuano, Giovanni Ciabattoni, Roberto Paolesse, Corrado Di Natale, Peter J Barnes, Arnaldo D'Amico," Diagnostic performance of an electronic nose, fractional exhaled nitric oxide, and lung function testing in asthma" Chest, Science Direct, Volume 137, Issue 4, pp 790-796, April 2010.

[76] N. Fens, A. C. Roldaan, M. P. van der Schee, R. J. Boksem, A. H. Zwinderman, E. H. Bel, P. J. Sterk," External validation of exhaled breath profiling using an electronic nose in the discrimination of asthma with fixed airways obstruction and chronic obstructive pulmonary disease", Clinical & Experimental Allergy, 41, 1371–1378.