Evaluation of different classification methods using electronic nose data to diagnose sarcoidosis

van der Sar, Iris G.; van Jaarsveld, Nynke; Spiekerman, Imme A.; Toxopeus, Floor J.; Langens, Quint L.; Wijsenbeek, Marlies S.; Dauwels, Justin; Moor, Catharina C.

DOI
10.1088/1752-7163/acf1bf

Publication date
2023

Document Version
Final published version

Published in
Journal of Breath Research

Citation (APA)
van der Sar, I. G., van Jaarsveld, N., Spiekerman, I. A., Toxopeus, F. J., Langens, Q. L., Wijsenbeek, M. S., Dauwels, J., & Moor, C. C. (2023). Evaluation of different classification methods using electronic nose data to diagnose sarcoidosis. *Journal of Breath Research*, *17*(4), Article 047104. https://doi.org/10.1088/1752-7163/acf1bf

Journal of Breath Research

PAPER

Evaluation of different classification methods using electronic nose data to diagnose sarcoidosis

To cite this article: Iris G van der Sar *et al* 2023 *J. Breath Res.* **17** 047104

IOP Publishing J. Breath Res. 17 (2023) 047104 https://doi.org/10.1088/1752-7163/acf1bf

Journal of Breath Research

RECEIVED
2 December 2022

REVISED
8 August 2023

ACCEPTED FOR PUBLICATION
18 August 2023

PUBLISHED
29 August 2023

PAPER

Evaluation of different classification methods using electronic nose data to diagnose sarcoidosis

Iris G van der Sar[1], Nynke van Jaarsveld[2,5], Imme A Spiekerman[2,5], Floor J Toxopeus[2,5], Quint L Langens[2,5], Marlies S Wijsenbeek[1], Justin Dauwels[3,4] and Catharina C Moor[1,4,*]

[1] Department of Respiratory Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands
[2] Educational Program Technical Medicine, Leiden University Medical Center, Delft University of Technology & Erasmus University Medical Center, Leiden, Delft & Rotterdam, The Netherlands
[3] Department of Microelectronics, Delft University of Technology, Delft, The Netherlands
[4] These authors share last authorship.
[5] These authors contributed equally.
[*] Author to whom any correspondence should be addressed.

E-mail: c.moor@erasmusmc.nl

Keywords: electronic nose, breath analysis, classification model, diagnostic test, sarcoidosis, interstitial lung disease, volatile organic compounds

Supplementary material for this article is available online

Abstract

Electronic nose (eNose) technology is an emerging diagnostic application, using artificial intelligence to classify human breath patterns. These patterns can be used to diagnose medical conditions. Sarcoidosis is an often difficult to diagnose disease, as no standard procedure or conclusive test exists. An accurate diagnostic model based on eNose data could therefore be helpful in clinical decision-making. The aim of this paper is to evaluate the performance of various dimensionality reduction methods and classifiers in order to design an accurate diagnostic model for sarcoidosis. Various methods of dimensionality reduction and multiple hyperparameter optimised classifiers were tested and cross-validated on a dataset of patients with pulmonary sarcoidosis ($n = 224$) and other interstitial lung disease ($n = 317$). Best performing methods were selected to create a model to diagnose patients with sarcoidosis. Nested cross-validation was applied to calculate the overall diagnostic performance. A classification model with feature selection and random forest (RF) classifier showed the highest accuracy. The overall diagnostic performance resulted in an accuracy of 87.1% and area-under-the-curve of 91.2%. After comparing different dimensionality reduction methods and classifiers, a highly accurate model to diagnose a patient with sarcoidosis using eNose data was created. The RF classifier and feature selection showed the best performance. The presented systematic approach could also be applied to other eNose datasets to compare methods and select the optimal diagnostic model.

1. Introduction

New applications of artificial intelligence (AI) in pulmonary medicine have been increasingly studied and published over the last years. However, no applications have yet been approved for use in clinical practice. Investigated applications range from automatic interpretation of pulmonary function tests and chest computed tomography scans, to predicting disease exacerbations using home monitoring data [1]. AI models sometimes achieve the accuracy level of human experts [2]. Therefore, it is likely that AI will support clinical decision making in the near future.

Electronic nose (eNose) technology is one of the upcoming new technologies for clinical practice that uses AI. An eNose device analyses exhaled breath in real-time, using multiple cross-reactive sensors with different sensitivities. By using classification models to categorize generated sensor data, the eNose device has the potential to be used as non-invasive diagnostic tool. Hence, different eNose devices and clinical applications are currently studied in the field of pulmonary medicine [3].

Interstitial lung diseases (ILDs) comprise a large group of heterogeneous rare individual diseases that affect the interstitium of the lungs. Patients usually present with non-specific symptoms, and disease course and response to therapy widely varies. Sarcoidosis, a form of ILD, is a multisystem granulomatous disease with lung involvement occurring in 89%–99% of patients [4]. In the current guidelines, three main criteria are proposed to diagnose sarcoidosis: a compatible clinical presentation, the finding of nonnecrotizing granulomatous inflammation in tissue samples, and the exclusion of alternative causes of granulomatous disease [5]. However, no objective measures exist to judge whether these criteria are satisfied. Consequently, the established consensus diagnosis always contains a certain margin of uncertainty for each individual, despite multiple diagnostic test, often including invasive tissue biopsy. Therefore, accurate, non-invasive and fast diagnostic modalities are highly needed.

Studies that tested performance of eNose technology as a diagnostic tool for ILD show accuracies varying from 49%–100% [6–12]. The large spread might be explained by differences in study design and eNose devices. Moreover, these studies used different classifiers to analyse the sensor data: neural networks (NNs), canonical discriminant analysis, K-nearest neighbour, linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), random forest (RF), support vector machines (SVMs), and Extreme Gradient Boosting (XGBoost). We previously showed that PLS-DA accurately distinguished sarcoidosis from other forms of ILD, but we did not evaluate the performance of different classifiers or models [11, 12].

In the field of machine learning, various models are usually compared before selecting a final machine learning model [13]. This might also be a good approach for clinical eNose research, as performance might differ per dataset and classification model [14]. Until now, only two eNose studies in ILD evaluated multiple models. They showed fair and comparable model performance on training datasets, but performance in test and validation sets varied [6, 10].

The main aim of this paper is to evaluate the performance of various dimensionality reduction methods and classifiers to design the most accurate diagnostic model for sarcoidosis.

2. Methods

2.1. Dataset and materials

The used dataset includes eNose sensor and clinical data of patients with pulmonary sarcoidosis ($n = 224$) and patients with other ILDs ($n = 317$) from the Erasmus Medical Center (Rotterdam, the Netherlands). Clinical characteristics have been published previously [11]. We collected exhaled breath data using the SpiroNose (Breathomix, Leiden, Netherlands) which is connected to an online secured platform and database called BreathBase. Breath manoeuvres were performed in duplicate. Each manoeuvre included five tidal breaths, followed by a maximal inhalation to vital capacity, 5 s breath hold and slow maximal exhalation leading to a sensor peak value. During the measurements, a mouthpiece with bacterial filter (Pulmosafe 3, Lemon Medical GmbH, Hammelburg, Germany) and a nose clamp were used. The investigator checks the quality of each measurement in real-time during the breath manoeuvre by inspecting the sensor deviation curves that appear in BreathBase. The investigator can provide the patient feedback for the second manoeuvre if necessary. Specifications of the device and manoeuvres have previously been published and are specified in supplementary data A [15]. Sensor characteristics, system verification procedures, and conditions and contraindications for using the device are also described in supplementary data A.

The SpiroNose contains seven different metal oxide semiconductor sensors, present in duplicate on the inside and outside of the device. After data pre-processing (including scaling and correction for ambient air), both the sensor peak value and peak to breath-hold ratio are extracted from each sensor signal, leading to 14 sensor values per patient. The peak value of sensor 2 is set to a constant value and is used for scaling of the other sensor values. The peak value of sensor 2 does not serve as an input variable. The data processing has been described previously [15]. Figure S1 and S2 in supplementary data B shows some examples of sensor diagrams and corresponding input variables.

Clinical characteristics were obtained from medical files and patient questionnaires. The study protocol was approved by the local ethical committee of Erasmus Medical Center (MEC-2019-0230). Analyses were conducted in Matlab (version R2021b), Statistics and Machine Learning Toolbox [16]. The final script to generate the results of this paper was run in June 2022. The full Matlab scripts are freely available on request.

2.2. Model design and testing

Based on previously published eNose studies and compatibilities of Matlab, classifiers k-NN, LDA, NN, RF, and SVM were selected for evaluation of their binary classification performance using eNose sensor data of patients with sarcoidosis and ILD. The overall process of model design and evaluation consisted of several consecutive steps:

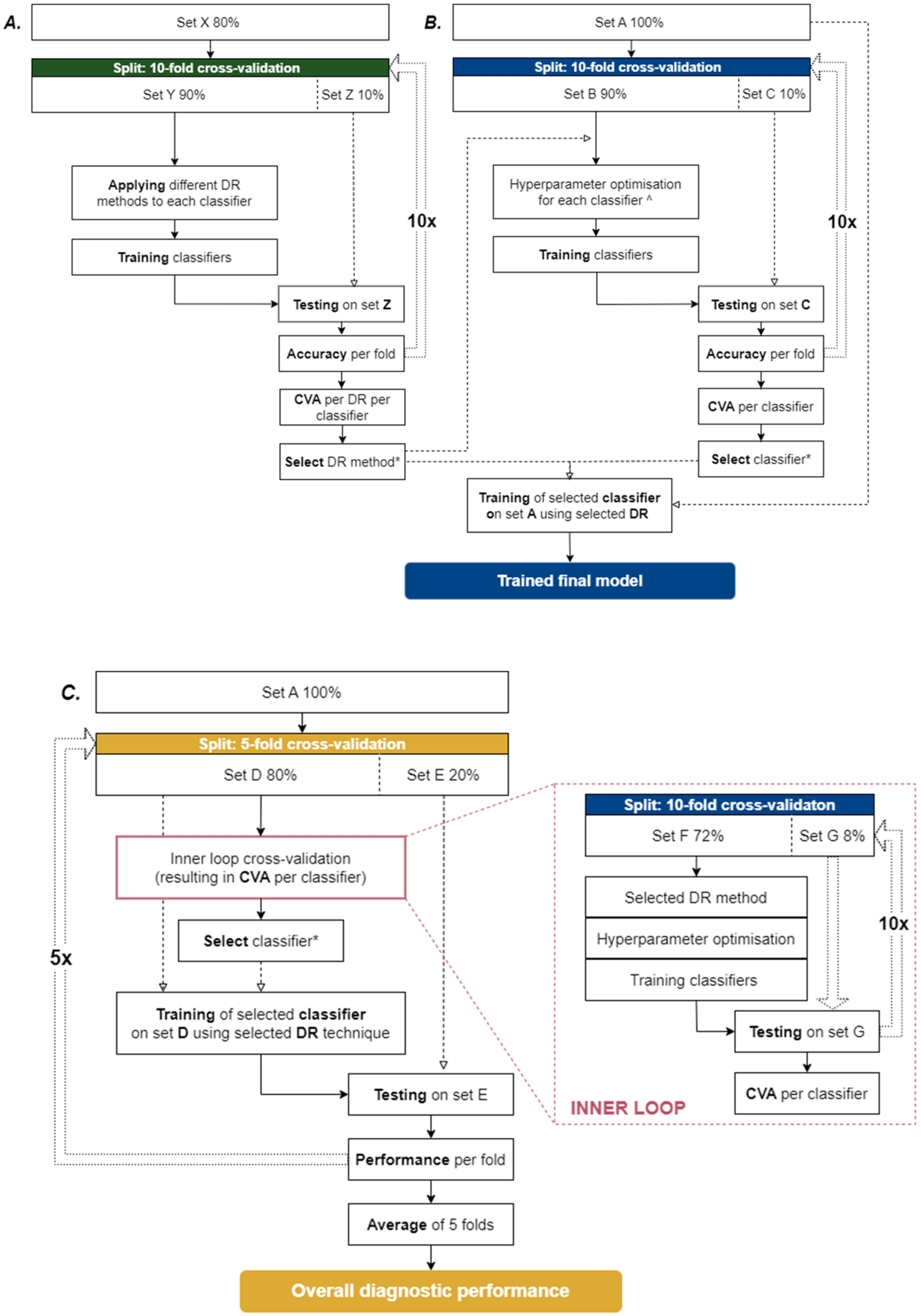
A.
Set X 80%
Split: 10-fold cross-validation
Set Y 90%
Set Z 10%
Applying different DR methods to each classifier
Training classifiers
Testing on set Z
Accuracy per fold
CVA per DR per classifier
Select DR method*
10x
B.
Set A 100%
Split: 10-fold cross-validation
Set B 90%
Set C 10%
Hyperparameter optimisation for each classifier ^
Training classifiers
Testing on set C
Accuracy per fold
CVA per classifier
Select classifier*
10x
Training of selected classifier on set A using selected DR
Trained final model
C.
Set A 100%
Split: 5-fold cross-validation
Set D 80%
Set E 20%
Inner loop cross-validation (resulting in CVA per classifier)
Select classifier*
Training of selected classifier on set D using selected DR technique
Testing on set E
Performance per fold
Average of 5 folds
Overall diagnostic performance
5x
Split: 10-fold cross-validaton
Set F 72%
Set G 8%
Selected DR method
Hyperparameter optimisation
Training classifiers
Testing on set G
CVA per classifier
10x
INNER LOOP

Figure 1. Summary of model design and evaluation. (A) Testing several methods of dimensionality reduction to select the best performing method. (B) Training and testing different classifiers to select the best performing method and train the final model. (C) Validating the overall diagnostic performance of the model using nested cross-validation. Cross-validation in the inner loop is similar to the cross-validation in (B).

*The classifier or DR technique resulting in the best CVA is selected. ^Using Bayesian optimisation. CVA = cross-validated accuracy; DR = dimensionality reduction

1. Testing several methods of dimensionality reduction to select the best performing method to train the model (figure 1(A));
2. Training and testing several hyperparameter optimised classifiers using 10-fold cross-validation to select the most accurate classifier to train the model (figure 1(B));
3. Validating the overall diagnostic performance of the model using nested cross-validation (figure 1(C));
4. Applying the trained final model on random patients to show the individual diagnostic probability;
5. Assessing the sufficiency of dataset size by calculating model accuracies on increasing sample size proportions.

2.2.1. Dimensionality reduction

First, the dimensionality of the dataset was reduced using feature selection or feature extraction, and this was compared to using no dimensionality reduction. The input variables (i.e. features) were the 13 peak sensor and peak to breath-hold values per eNose measurement of a patient. All three methods were tested on 80% of the data using 10-fold cross-validation (figure 1(A)). The method with the highest cross-validated accuracy (CVA) was implemented in training the final model. The dimensionality reduction cross-validation was performed once, as the outcome of dimensionality reduction depends on the dataset itself, not on the classifier [17, 18].

Feature extraction was performed using principal component analysis (PCA) with Matlab's function PCA [19]. PCA results in a set of multivariate components, where each component is a combination of the original 13 sensor values. The first PCA component explains the greatest variance of the data and the last PCA component the least. To determine which components to include, percentage of variability thresholds of $\geqslant$90%, 95% or 99% were compared. The singular value decomposition algorithm was selected within Matlab's PCA function.

Feature selection was performed with Matlab's fscchi2 function [20]. This function was used to calculate the weight of each feature by taking the negative logarithm of the *p*-value resulting from a chi-squared test. This weight represents the extent to which a single feature influences the outcome of the model; a higher score indicates more influence. Following the feature weights calculation, various weight thresholds were tested to select a certain number of contributing features. Three thresholds that resulted in five up to ten contributing features were eventually tested.

2.2.2. Hyperparameter optimisation

In each fold, hyperparameter optimisation was executed while training each classifier. This was done by setting the option OptimizeHyperparameters in each classifier to 'auto'. This led to 2–4 parameters being optimised per classifier. The type of parameters depended on which classifier was being trained. Specifications of the optimizations can be found in supplementary data C. Using Bayesian optimisation, the 5-fold cross-validated loss per set of hyperparameters was calculated over 30 iterations [21]. The set with the minimal cross-validation loss was selected.

The RF method required several other parameters to be defined in the function fitcensemble [22]. In Matlab, the type of learner method was set as 'decision tree' and the aggregation method as 'bag'. Bootstrap aggregation (i.e. bagging) reduces the variance of weak learners such as RF. This specification cannot be combined with the OptimizeHyperparameters option. Thus, hyperparameter optimisation was performed separately using the Bayesopt function in the same manner as for the other classifiers [21].

2.2.3. Model training, testing and selecting

To select the most accurate classifier, 10-fold cross-validation was performed on the full dataset (set A) for each classifier using the selected dimensionality reduction method (section 2.2.1). The data splits for 10-fold cross-validation were made using the function cv-partition [23]. Nine folds formed the training set (set B) and the other fold the test set (set C). The CVA per classifier was calculated as the average of the accuracies of the ten folds and included a range (i.e. minimum and maximum accuracy of the folds). The classifier with the highest CVA was selected and trained on set A. Hyperparameter optimisation was executed anew. For SVM, the selected kernel type was 'linear'. This resulted in the final trained model to classify patients based on eNose data (figure 1(B)).

2.2.4. Diagnostic performance calculation

The overall diagnostic performance of this final model was determined by repeating the initial 10-fold cross-validation within a 5-fold cross-validation (i.e. nested cross-validation) including all five classifiers (*k*-NN, LDA, NN, RF, and SVM) (figure 1(C)). Nested cross-validation leads to less bias than single-loop cross-validation used for training the final model (section 2.2.3) as the results do not depend on a single data split [24].

To execute this validation method, the full dataset (set A) was split into five folds resulting in four folds representing 80% (set D) and one representing 20% of the data (set E), the so-called outer loop. Set D was used for inner loop 10-fold cross-validation and therefore divided into set F and G (figure 1(C)). These sets F and G underwent the exact same process as the initial training and test sets B and C (figure 1(B)).

Each of the five folds resulted in a best performing classifier, and this classifier was subsequently trained on set D and tested on set E to calculate the diagnostic performance (accuracy, specificity, sensitivity

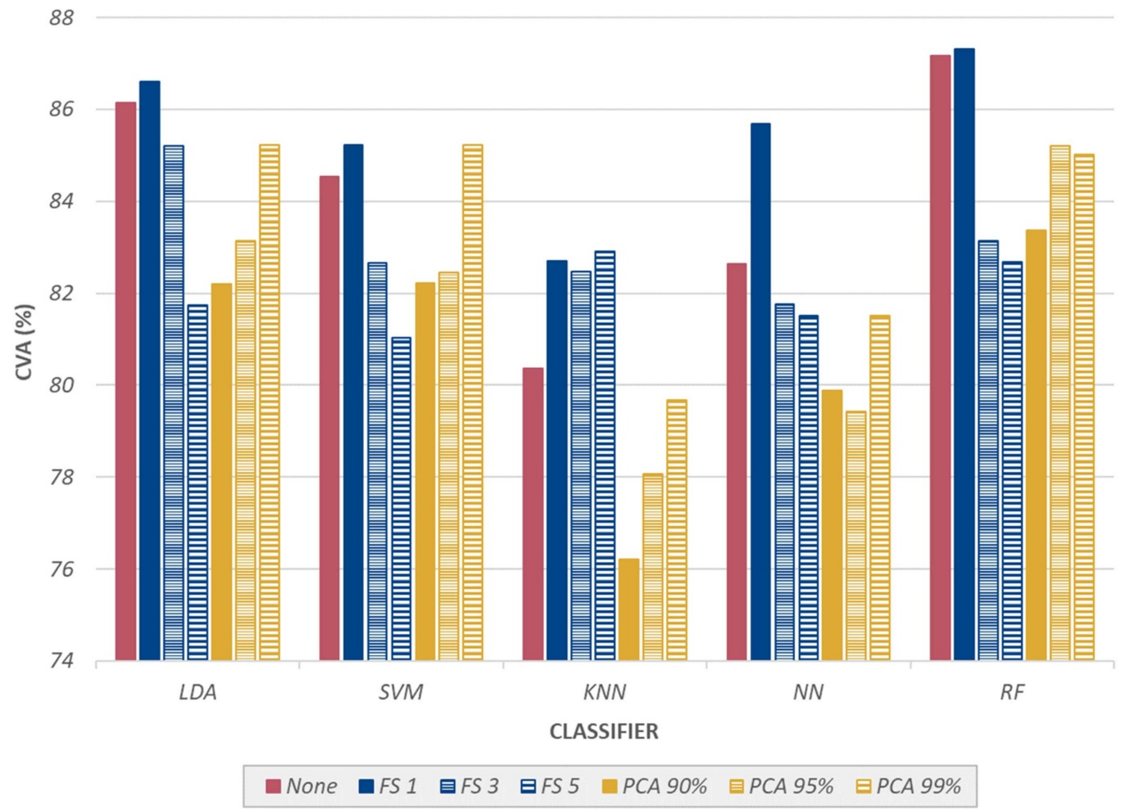

Figure 2. Cross-validated accuracy per classifier when applying different dimensionality reduction method, including three different thresholds for FS (1, 3 and 5) and PCA (90%, 95% and 99%).
CVA = cross-validated accuracy; FS = feature selection; KNN = *K*-nearest neighbour; LDA = linear discriminant analysis; NN = neural networks; PCA = principle component analysis; RF = random forest; SVM = support vector machines.

and AUC values) of that fold using Matlab's function confusionmat. The accuracy was calculated as (true negatives + true positives)/(true negatives + true positives + false negatives + true positives), specificity as true negatives/(true negatives + false positives), and sensitivity as true positives/(true positives + false negatives). Finally, the overall diagnostic performance of the model was the average of these values of the five folds.

2.2.5. Classifying individual patients

To simulate the model's ability in diagnosing a 'new' patient, the trained final model was applied to sensor data from random patients from set A. The model predicted for an individual patient the class it belongs to (sarcoidosis or ILD), including the probability of this prediction and the time it took to complete the prediction. A higher probability means a higher likelihood of the prediction being correct for this individual. The probability was calculated by multiplying the prior probability with multivariate normal density and expressed as an percentage [25].

2.2.6. Evaluation size dataset

In order to evaluate whether the final model would benefit by training on more data or if sample size was sufficient, the model's accuracy was calculated for smaller training dataset sizes. The final trained model was repeatedly trained on an increasing proportion of data to calculate the corresponding accuracy.

The entire dataset was first split into a new training (90%) and test set (10% of data). The model was trained using 1 up to 100% of the training data, each attempt increasing with 1%. The corresponding accuracy was tested using the full test set. Training and testing was repeated 20 times per proportion of training data, resulting in an average accuracy per proportion used.

3. Results

3.1. Dimensionality reduction

CVA values resulting from the five different classifiers after applying 'feature selection', 'feature extraction' (i.e. PCA) and 'no dimensionality reduction' are shown in figure 2. A feature selection weight threshold set on 1 resulted in 10 features, 3 in 6, and 5 in 5 features. A threshold of 1 resulted in the highest CVA in four out of five classifiers. Therefore, this method was chosen for dimensionality reduction to implement in the final model. The weights of each feature are shown in figure 3. The peak value of sensor 3 and 5, and peak to breath-hold ratio of sensor 2 did not reach the optimal threshold of 1, and were excluded for training the final model.

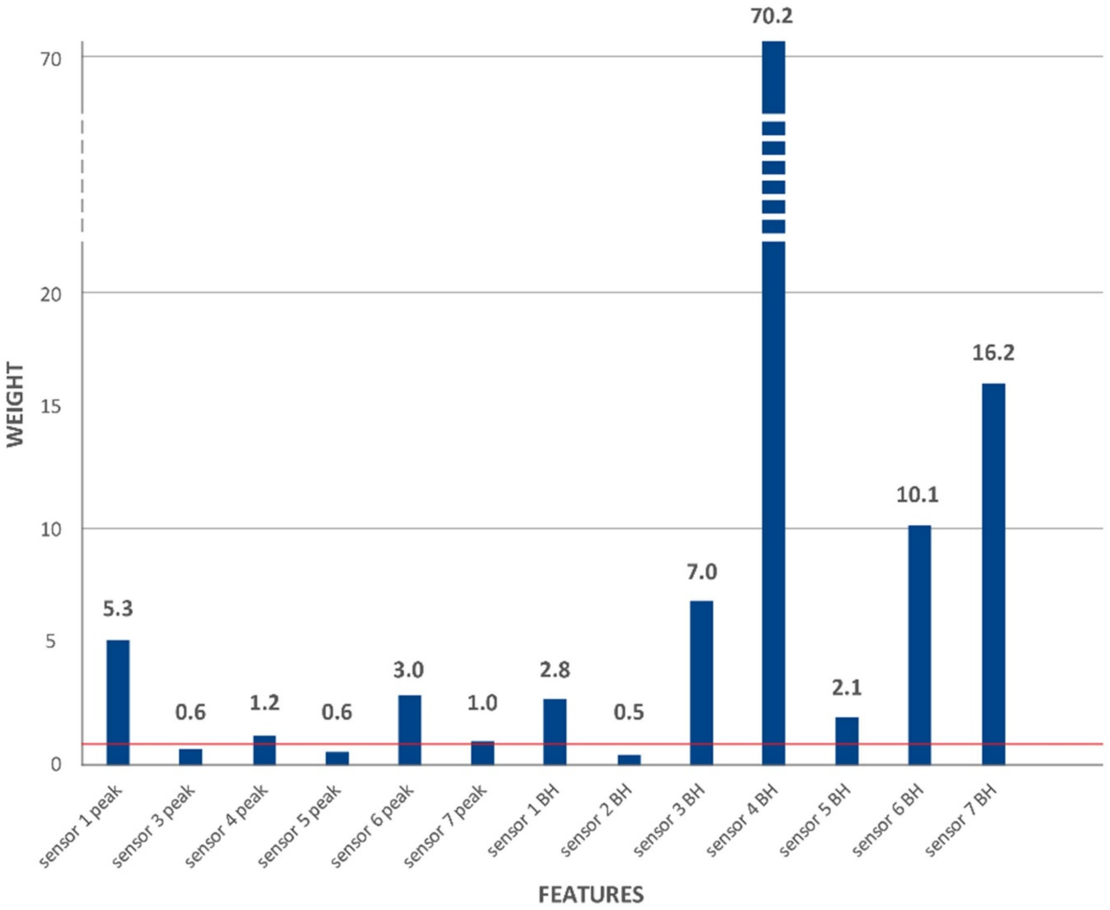

Figure 3. Weight per feature of the trained final model. Features include peak values and peak to breath-hold ratios. Weights represent the extent to which a sensor value influences the response variable of the model.
Weight = −log(p-value per feature). The red line illustrates the weight of 1 used as threshold for feature selection. BH = breath hold.

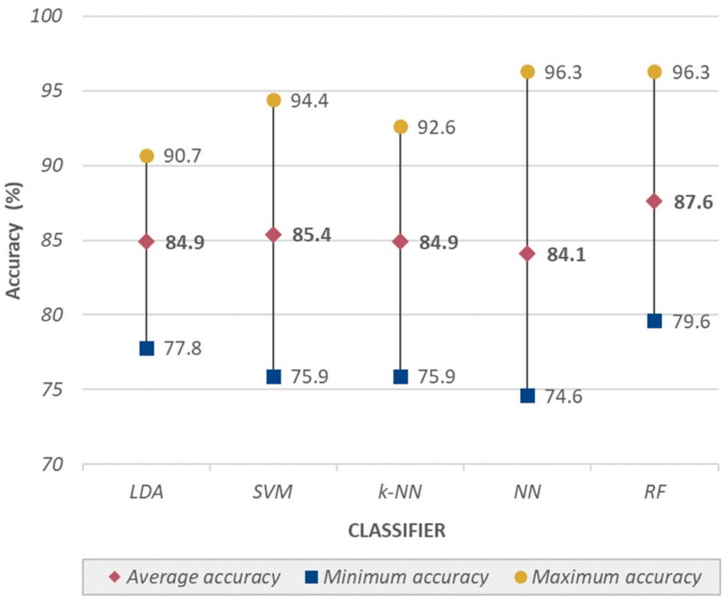

Figure 4. Comparison of the minimum, maximum, and the average accuracy of the ten folds, displayed per classifier calculated by 10-fold cross validation.
K-NN = K-nearest neighbour; LDA = linear discriminant analysis; NN = neural networks; RF = random forest; SVM = support vector machines.

3.2. Model training, testing and selecting

After applying feature selection with threshold of 1, CVAs for all classifiers were calculated separately. RF showed the highest CVA of 87.6% with a range of 79.6%–96.3% (figure 4) and was therefore selected as classifier for the final model.

Hyperparameter optimisation resulted in 100 learning cycles and 23 bins, which were used to train the RF classifier and the final model. All data subsets in the ten folds had approximately the same class distribution of ILD and sarcoidosis as the complete dataset.

3.3. Diagnostic performance calculation

The best performing classifier and the corresponding diagnostic performance values resulting from each fold of the nested cross-validation are shown in table 1. RF performed best in three out of five folds. The CVA resulting from the five folds was 87.1% ranging from 80.7% to 92.6%. The average sensitivity was 91.4% (range 86.4%–96.6%) and specificity 82.2% (range 74.0%—90.5%). The AUC of the receiver operating characteristic curves varied from 83.7%–96.8% with an average of 91.2%. The accuracy for each five classifiers of all five folds and the receiver operating characteristic curve resulting from each fold can be found in supplementary data D (table S3 and figure S3).

3.4. Classifying individual patients

The model's output for each individual patient includes a diagnosis and diagnostic probability based on eNose data. An example of the model's output of ten randomly selected individuals from the full dataset is shown in table 2.

3.5. Evaluation size dataset

Increasing the training dataset from 80%–100% resulted in 0.7% accuracy improvement (87.5–88.2%), indicating that the model is likely trained on sufficient data. The model's accuracy when training with a smaller dataset size is shown in figure S4 in supplementary data E.

4. Discussion

In this paper, we evaluated multiple classification methods to design a highly accurate model using eNose data for diagnosing patients with pulmonary sarcoidosis within a group of patients with ILD. Different dimensionality reduction methods and classifiers were trained, tested and compared systematically. Feature selection and RF resulted in the highest diagnostic performance compared to the other methods assessed and were trained to create a final diagnostic model. Diagnostic performance resulted in a CVA of 87.1%. The presented approach for comparing different dimensionality reduction methods and classifiers to design a diagnostic eNose model has not been described previously. A strength of the designed model is the ability to show a specific diagnostic probability for an individual patient, which will facilitate translation of eNose technology into clinical practice.

When starting to design a diagnostic model for a certain condition using eNose data, the most important factor that determines model performance is whether the selected condition can be detected in exhaled breath accurately. A proof-of-concept study should clarify this first before designing a diagnostic model, like we performed for sarcoidosis previously using the PLS-DA classifier [11]. In the current comparative analysis of classifiers, RF turned out to be the best performing classifier for this dataset. RF has been used previously to classify various medical conditions using eNose data [3, 14]. In general, the majority of eNose papers focus on a single analysis method to classify patients supervised without a clear rationale for the selected method. In this paper, we show a systematic comparative approach to justify the choice for a certain analysis method.

Although RF showed the highest accuracy, differences between classifiers were small and all showed good accuracies. When designing a model for clinical applications, also other factors besides performance have to be considered, such as speed of the model, visualization, and outcome parameters [26]. Our trained final model shows a diagnosis within 1 s for an individual patient including a diagnostic certainty. The latter is important for clinician to interpret the eNose results correctly when using this test in clinical practice.

Before implementing the eNose as a diagnostic tool for sarcoidosis in clinical practice, the current model needs to be trained and tested on an independent heterogeneous multicentre cohort including patients with various related conditions, with respiratory complaints without a diagnosis, and healthy controls matched by possible confounders (e.g. age and sex), to confirm the models robustness and to prevent overfitting [14]. Additionally, analysis of unlabelled patient data need to confirm the hypothesis of this diagnostic tool. This is in particular important for a sarcoidosis cohort due to several reasons. First, patients from different healthcare settings should be included, not only from ILD and sarcoidosis expert centres like the current cohort, as patients' characteristics and diagnostic certainty might differ. Second, given the lack of clear objective diagnostic criteria for sarcoidosis and ILD, the reached consensus diagnosis always includes some uncertainty. It is inevitable that training data of the current dataset are not 100% accurate. Moreover, the time between a patient's diagnosis and eNose

Table 1. Overall diagnostic performance of the final model displayed as the average accuracy (i.e. CVA), sensitivity, specificity and AUC of the five folds. The best performing classifier per fold was selected based on the highest accuracy.

	Classifier	Accuracy (%)	Sensitivity	Specificity	AUC (%)
Fold 1	RF	90.7	90.9	90.5	93.9
Fold 2	SVM	80.7	86.4	74.0	83.7
Fold 3	RF	86.1	88.7	82.6	93.3
Fold 4	SVM	85.2	94.3	76.4	88.1
Fold 5	RF	92.6	96.6	87.8	96.8
Average	—	*87.1*	*91.4*	*82.2*	*91.2*
95% CI		*84.29, 89.91*	*88.99, 93.81*	*78.63, 85.77*	*90.76, 91.64*

AUC = area under the curve; CI = confidence interval; CVA = cross-validated accuracy; RF = random forest; SVM = support vector machines.

Table 2. Example of the diagnostic model's output of ten randomly selected patients including the probability of the assigned class and the time needed to classify. All patients were classified correctly.

	Diagnosis	Probability (%)	Prediction time (s)
Patient 1	ILD	94	0.11
Patient 2	Sarcoidosis	93	0.09
Patient 3	ILD	89	0.09
Patient 4	ILD	88	0.13
Patient 5	Sarcoidosis	97	0.08
Patient 6	Sarcoidosis	97	0.07
Patient 7	ILD	86	0.07
Patient 8	Sarcoidosis	85	0.08
Patient 9	ILD	84	0.06
Patient 10	Sarcoidosis	95	0.06

ILD = interstitial lung disease; s = seconds.

measurement varied. Additionally, class frequencies are assumed a realistic representation of prior probabilities, which might vary in other care settings. Lastly, most patients have received or were receiving therapy, which could have influenced the eNose measurements. Nevertheless, previous analyses of this sarcoidosis cohort suggested that the extent of disease activity and treatment does not significantly affect the accuracy of eNose results [11].

When looking at potential clinical applications of diagnostic AI tools, including eNose technology, it is unlikely that AI will fully replace clinical decision making, as both clinicians and AI systems have unique strengths. It is well recognised that humans outperform machines in detection, perception, improvisation, long-term memory, induction, and judgement, and machines outperform humans in response speed and precision, repetition, short-term memory, deductive reasoning, and handling complex operations [27]. Thus, especially the use of AI combined with clinical decision-making is likely to be of added value. This accounts in particular for diseases without a conclusive diagnostic test, such as sarcoidosis, where pattern recognition is of great importance.

Another prerequisite for a fruitful implementation of eNose technology in clinical practice is trust of clinicians in the capability of the technology [28]. In the current paper, we aim to provide insights to clinicians with regard to data processing, model design and performance. This will build trust in eNose technology and encourage correct interpretation of the model output. Essential for correct model output interpretation and integration in clinical decision-making is the individual diagnostic probability score provided in the current paper. Besides, clinicians should know on what data the model is trained to identify the correct patients for applying the model to.

Several limitations of the developed model and proposed method should be addressed.

The PLS-DA classifier that was used for analyses in the previous proof-of-concept paper on the same sarcoidosis cohort, which led to an accuracy of 83.2% in the validation set, is not evaluated in the current paper. The way PLS-DA reduces and classifies data is substantially different from the other selected classifiers and less commonly used in machine learning [29]. Besides, PLS-DA is not supported by a compatible Matlab package. Moreover, some of the classifiers presented in this paper achieve better accuracy than 83.2%. However, for proof-of-concept studies to explore whether eNose technology is able to distinguish certain patient groups there is no need to compare multiple classifiers and PLS-DA is a reliable method to use [29].

The calculated threshold for feature selection was based on an independent 10-fold cross-validation (figure 1(A)). Preferably, threshold optimisation would have been included in the 10-fold cross-validation when each classifier was trained and tested, to select the most relevant features. This was not executed due to computational limits.

The current results cannot yet be used in clinical practice due to the lack of external validation of the model. Due to the rarity of the disease and the small number of specialized treatment centres, an external patient cohort is difficult to create. To generate robust results and avoid overfitting of the model using the available data, the nested cross-validation was performed as an extra step in testing the model following recommendations from Cawley and Talbot [24]. Another possible source of bias is the absence of data from patients suspected of pulmonary sarcoidosis or ILD.

5. Conclusion

Evaluation of various classification methods resulted in an accurate diagnostic model for sarcoidosis based on exhaled breath eNose data. To design this model, frequently used dimensionality reduction methods and classifiers were assessed and compared systematically by rigorous procedures such as nested cross-validation. For the current eNose dataset, a model based on feature selection followed by RF yield the best results. The proposed strategy to design and evaluate a diagnostic model can serve as an example for other researchers and is applicable to other eNose datasets.

The outcome of the model includes a specific diagnostic probability for an individual patient, which will facilitate translation into clinical practice. After optimising the model with a multicentre training dataset and validating the developed model with eNose data of patients with suspected pulmonary sarcoidosis, eNose models might be integrated in clinical decision making in order to facilitate a fast, accurate and non-invasive diagnosis.

Data availability statements

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

M W reported grants or contracts from Boehringer Ingelheim, Hoffman la Roche, The Netherlands Organisation for Health Research and Development, The Dutch Lung Foundation, The Dutch Pulmonary Fibrosis; consulting fees from Bristol Myers Squibb, Boehringer Ingelheim, Galapagos, Galecto, Hoffman la Roche, Horizon therapeutics, Kinevant Sciences, Molecure, Nerre Therapeutics, Novartis, PureTech Health, Respivant, Thyron, Vicore. All fees are paid to her institution and not related to the current investigation. C M reported grants or contracts from Boehringer Ingelheim, Astra-Zeneca, Daiichi-Sankyo. All fees are paid to her institution and not related to the current investigation. I S reported grants or contracts from Boehringer Ingelheim. All fees are paid to her institution and not related to the current investigation. N J, I S, F T, Q L, J D reported no potential conflicts of interest.

ORCID iDs

Iris G van der Sar https://orcid.org/0000-0001-5566-8129
Marlies S Wijsenbeek https://orcid.org/0000-0002-4527-6962
Justin Dauwels https://orcid.org/0000-0002-4390-1568
Catharina C Moor https://orcid.org/0000-0002-5295-2877

References

[1] Mekov E, Miravitlles M and Petkov R 2020 Artificial intelligence and machine learning in respiratory medicine *Expert Rev. Respir. Med.* **14** 559–64

[2] Topalovic M *et al* 2019 Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests *Eur. Respir. J.* **53** 1801660

[3] van der Sar I G, Wijbenga N, Nakshbandi G, Aerts J G J V, Manintveld O C, Wijsenbeek M S, Hellemons M E and Moor C C 2021 The smell of lung disease: a review of the current status of electronic nose technology *Respir. Res.* **22** 246

[4] Grunewald J, Grutters J C, Arkema E V, Saketkoo L A, Moller D R and Müller-Quernheim J 2019 Sarcoidosis *Nat. Rev. Dis. Primers* **5** 45

[5] Crouser E D *et al* 2020 Diagnosis and detection of sarcoidosis. An official American thoracic society clinical practice guideline *Am. J. Respir. Crit. Care Med.* **201** e26–e51

[6] Yang H-Y, Peng H-Y, Chang C-J and Chen P-C 2017 Diagnostic accuracy of breath tests for pneumoconiosis using an electronic nose *J. Breath Res.* **12** 016001

[7] Dragonieri S, Scioscia G, Quaranta V N, Carratu P, Venuti M P, Falcone M, Carpagnano G E, Foschino Barbaro M P, Resta O and Lacedonia D 2020 Exhaled volatile organic compounds analysis by e-nose can detect idiopathic pulmonary fibrosis *J. Breath Res.* **14** 047101

[8] Krauss E, Haberer J, Maurer O, Barreto G, Drakopanagiotakis F, Degen M, Seeger W and Guenther A 2019 Exploring the ability of electronic nose technology to recognize interstitial lung diseases (ILD) by non-invasive breath screening of exhaled volatile compounds (VOC): a pilot study from the European IPF registry (eurIPFreg) and biobank *J. Clin. Med.* **8** 1698

[9] Dragonieri S, Brinkman P, Mouw E, Zwinderman A H, Carratú P, Resta O, Sterk P J and Jonkers R E 2013 An electronic nose discriminates exhaled breath of patients with untreated pulmonary sarcoidosis from controls *Respir. Med.* **107** 1073–8

[10] Xuan W, Zheng L, Bunes B R, Crane N, Zhou F and Zang L 2022 Engineering solutions to breath tests based on an e-nose system for silicosis screening and early detection in miners *J. Breath Res.* **16** 036001

[11] van der Sar I G, Moor C C, Oppenheimer J C, Luijendijk M L, van Daele P L A, Maitland-van der Zee A H, Brinkman P and Wijsenbeek M S 2022 Diagnostic performance of electronic nose technology in sarcoidosis *Chest* **161** 738–47

[12] Moor C C, Oppenheimer J C, Nakshbandi G, Aerts J G J V, Brinkman P, Maitland-van der Zee A-H and Wijsenbeek M S 2021 Exhaled breath analysis by use of eNose technology: a novel diagnostic tool for interstitial lung disease *Eur. Respir. J.* **57** 2002042

[13] Statnikov A, Aliferis C F, Tsamardinos I, Hardin D and Levy S 2004 A comprehensive evaluation of multicategory

classification methods for microarray gene expression cancer diagnosis *Bioinformatics* **21** 631–43

[14] Leopold J H *et al* 2015 Comparison of classification methods in breath analysis by electronic nose *J. Breath Res.* **9** 046002

[15] de Vries R *et al* 2018 Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label *Eur. Respir. J.* **51** 1701817

[16] Statistics and machine learning toolbox 2021 (The MathWorks Inc.)

[17] Sánchez-Maroño N, Alonso-Betanzos A and Tombilla-Sanromán M 2007 Filter methods for feature selection—a comparative study *IDEAL* pp 178–87

[18] Jolliffe I 2002 *Principal Component Analysis* (Springer)

[19] PCA—Principal component analysis of raw data *MathWorks*

[20] Fscchi2—Univariate feature ranking for classification using chi-square tests (The MathWorks Inc.)

[21] Bayesopt—Select optimal machine learning hyperparameters using Bayesian optimization (The MathWorks Inc.)

[22] Fitcensemble—Fit ensemble of learners for classification *MathWorks*

[23] Cvpartition—Partition data for cross-validation (The MathWorks Inc.)

[24] Cawley G and Talbot N 2010 On over-fitting in model selection and subsequent selection bias in performance evaluation *J. Mach. Learn. Res.* **11** 2079–107

[25] Predict—Predict labels using discriminant analysis classification model (The MathWorks Inc.)

[26] Gromski P S, Correa E, Vaughan A A, Wedge D C, Turner M L and Goodacre R 2014 A comparison of different chemometrics approaches for the robust classification of electronic nose data *Anal. Bioanal. Chem.* **406** 7581–90

[27] Fitts P M 1951 *Human engineering for an effective air-navigation and traffic-control system* (National Research Council)

[28] Asan O, Bayrak A E and Choudhury A 2020 Artificial intelligence and human trust in healthcare: focus on clinicians *J. Med. Internet Res.* **22** e15154

[29] Lee L C, Liong C-Y and Jemain A A 2018 Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps *Analyst* **143** 3526–39